Collider Effect in the Analysis of Epidemiological Observational Data: Statistical Insights and Interactive Visualization

Paradoxical effects due to conditioning on a third variable correlated with both exposure and outcome in the analysis of observational data is common but correlation is not causation

Miguel Angel Luque-Fernandez*[1],BSc, MA, MSc, PhD
Daniel Redondo-Sanchez[1], BSc
Maria Jose Sanchez Perez[1], MD, PhD,
Anand Vaidya [2] MD, MSc
Mireille Schnitzer [3] BSc, MSc, PhD
Michael Schomaker[4], BSc, MSc, PhD

1 Introduction

Classical epidemiology during the last 30 years has focussed on the control of confounding [1]. However, it is just recently that epidemiologists have started to describe the bias produced by other source of structures such as colliders and mediators [2, 3]. In the epidemiological literature different explanations have been proposed to describe the paradoxical protective effect of established risk factors on an outcome such as the birth weight and the pre-eclampsia smoking paradoxes [4, 5]. The use of direct acyclic graphs (DAGs) help to visualize these new structures and distinguishes between biases resulting from lack of conditioning on common causes of exposure and outcome (confounding) (Figure 1A), conditioning on intermediates variables (mediation) (Figure 1B) or on common effects (collider bias) (Figure 1C) [6, 7].

A collider for a certain pair of variables such as an exposure and an outcome is a third variable that is causally influenced by both of them. Controlling for, or conditioning analysis on, such a variable can introduce a spurious association between its causes (exposure and outcome) explaining why medical literature is plenty of paradoxical findings [8]. In DAG terminology, a collider is the variable in the middle of an inverted fork (i.e., variable Z in A → Z ← Y) [9] (Figure 1C). While this methodological note will not close the vexing gap between correlation and causation, but it will contribute to increase awareness and the general understanding of colliders among applied epidemiologists and medical researchers.

The rest of the methodological note is structured as follows: i) we unveil the statistical structure of the collider bias through a simulated data generation using a linear system of structural equations, ii) we illustrate the effect of conditioning on a collider based on a non-communicable disease epidemiology example, iii) we provide R-code in easy-to-read boxes throughout the manuscript and in a GitHub repository: https://github.com/migariane/ColliderApp for replicability and iv) we provide readers with a Shiny application allowing real-time interaction to visualize the paradoxical effect of conditioning on a collider http://watzilei.com/shiny/collider/.

[1] Biomedical Research Institute. Non-Communicable and Cancer Epidemiology Group (ibs.Granada), Andalusian School of Public Health, University of Granada, Granada, Spain, miguel.luque.easp@juntadeandalucia.es

[2] Brigham and Women's Hospital. Harvard Medical School, Harvard University, Boston, MA, USA.

[3] Faculty of Pharmacy, University of Montral, Canada

[4] University of Cape Town, Centre for Infectious Disease Epidemiology and Research, Observatory, 7925; Cape Town, South Africa

2 Statistical structure of a collider

2.1 Reviewing confounding

Regression is the tool that classically epidemiologist use to address causal questions in observational studies. Let's begin reviewing what we mean with controlling for a confounder (W) when assessing the association between a treatment or exposure (A) and an outcome (Y) via the following causal graph (Figure 1A). In the above diagram, W act as a confounder perturbing the perceived causal relationship between A and Y if unaccounted for. We now review adjustment for confounding via linear regression model. In the Box-one we show how to generate data consistent with the direct acyclic graph from Figure 1A and we run regression models. Both, the coefficients from the regression models and the visualization of the linear fit for the adjusted model help us to illustrate and interpret the ”confounding” effect.

Box-one:

```
N <- 1000
library(visreg)
W <- rnorm(N)
A <- 0.5 * W + rnorm(N)
Y <- 0.3 * A + 0.4 * W + rnorm(N)
fit1 <- lm(Y ˜ A)
fit2 <- lm(Y ˜ A + W)
visreg(fit1, "A", gg = TRUE, line = list(col = "blue"),
       points=list(size = 2, pch = 1, col = "black")) + theme_classic()
```

Note that our confounder W is the only variable that is exogenous i.e. it is not dependent or generated from the outcome Y or the exposure A and is the only variable that we can set in any way we want. However, both A and Y depend on W, and they are endogenous to the system of linear equations. This is too the mechanism of randomization in an experimental design where by design W is exogenous. Thus it does not depends of A or Y. The first regression ignores W and incurs on a upward bias in As coefficient due to the confounders positive effects on both W and Y. The second regression including the confounder W recovers As true coefficient while increasing the R-squared by a few percentage points thus statistically speaking, adjusting for W improves the goodness of fit of the model (Figure 2A, Table 1: columns 1, 2).

2.2 Collider deconstruction

Unlike in Figure 1A where the causal arrows start from W, in Figure 1C they now point towards W from A and Y. If we condition on W, we will have created a collider effect i.e. a spurious association between A and Y (i.e. the association between A and Y collides). First, we generate the data, again using a simple linear data generating mechanism. Notice that A is now exogenous while W is ”endogenous” (i.e. it depends or is generated from the outcome Y or the exposure A).

Box-two:

```
N <- 1000
library(visreg)
W <- rnorm(N)
A <- 0.5 * W + rnorm(N)
Y <- 0.3 * A + 0.4 * W + rnorm(N)
W <- 1.2 * A + 0.9 * Y + rnorm(N)
fit3 <- lm(Y ˜ A + W)
visreg(fit3, "A", gg = TRUE, line = list(col="red"),
       points=list(size = 2, pch = 1, col = "black")) + theme_classic()
```

Unlike in previous section, the simpler regression without W recovers the true coefficient of A, while the regression adjusting for W biased importantly it. The collider model after a statistical point of view is not unequivocally inferior. For instance, it has an R-squared that is approximately 20 percentage points higher than the first model. However, conditioning on the collider W has paradoxically change the direction of the association between A and Y (Figure 2B, Table 1: column 3).

Note that even if the collider change the sense of the association it still helps to predict Y given that it is part of the conditional expectation function E(Y|A,W). It is important to highlight that if you do not care about prediction and your main interest is understanding how the world works (i.e. explicative or causal) you probably will not condition on W. Subject-matter knowledge (i.e. plausible biological associations in clinical epidemiological settings) is mandatory in explicative or causal modeling approaches. Thus, the use of DAGs for causal modeling help enormously to explicitly display the casual association between variables and elucidate the presence of colliders [10, 9].

3 Motivating Example

3.1 Data generation

Based on a non-communicable chronic disease epidemiology motivation, we generated a dataset with 1,000 observations to contextualize the effect of conditioning on a collider. We simulated the effect of age in years on systolic blood pressure (SBP) in mmhg assuming that SPB increases with age. Then, we simulated the effect of diary sodium intake in grams as a confounder (i.e. increasing sodium intake is positively associated with both SBP and age). Finally, we also simulated 24 hours excretion of proteinuria as a function of age, systolic blood pressure and sodium intake. Note that for the simulation we used exactly the same principles as the above highlighted when explaining the statistical foundations of the collider effect. Supplementary Table 1 shows the descriptive statistics (minimum, maximum, mean, median, first and third quartiles) for the generated data. We assured that the range of values of the simulated data was biologically plausible and as close to reality as possible [11, 12].

Box-three shows the code used for the data generation based on the structural relationship of the variables depicted on a DAG (Figure 3). The simulation implies a linear relationship between the variables. Thus, the interpretation of the beta coefficients in the formulae of the code on box-three is straightforward (i.e. Systolic blood presure = β_1 × sodium + β_2 × age; where $\beta_1 = 10$ and $\beta_2 = 1.25$). Supplementary Figure 1 shows the functional form for each variable and their multivariable correlation matrix.

Box-three:

```
generateData <- function(n){
    age <- rnorm(n, 65, 5)
    sodium  <- age/15 + rnorm(n)
    sbp <- 10*sodium + 1.25*age + rnorm(n)
    proteinuria  <- 1.5*age + 1.8*sbp - 0.9*sodium + rnorm(n)
    data.frame(sbp, age, sodium, proteinurie)
}
set.seed(777)
ObsData <- generateData(n = 1000)
```

We fit three different models to evaluate the effect of age on SBP: i) univariate unadjusted model; ii) bivariate adjusted model for sodium intake (sodium acting as a confounder); iii) multivariate adjusted model for sodium intake acting as a confounder and proteinuria acting as a collider. Models fit specification is shown here below, in box-four we show the model fit in R.

Models fit specification:

$$\text{Systolic Blood Pressure} = \beta_0 + \beta_1 \times \text{age}$$

$$\text{Systolic Blood Pressure} = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{sodium}$$
$$\text{Systolic Blood Pressure} = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{sodium} + \beta_3 \times \text{proteinuria}$$

Box-four:

```
## Models Fit
fit0 <- lm(sbp ~ age, data = ObsData);tidy(fit0)
fit1 <- lm(sbp ~ age + sodium , data = ObsData);tidy(fit1)
fit2 <- lm(sbp ~ age + proteinuria + sodium, data = ObsData);tidy(fit2)

## Models visualization
par(mfrow=c(1,3))
visreg(fit0, ylab = "SBP in mmhg", line=list(col = "blue"),
points = list(cex = 1.5, pch = 1), jitter = 10, bty = "n")
visreg(fit1, ylab = "SBP in mmhg", line=list(col = "blue"),
points = list(cex = 1.5, pch = 1), jitter = 10, bty = "n")
visreg(fit2, ylab = "SBP in mmhg", line=list(col = "red"),
points = list(cex = 1.5, pch = 1), jitter = 10, bty = "n")

## Comparing models and models goodness of fit
stargazer(fit0, fit1, fit2, type = "latex", multicolumn = FALSE)
```

3.2 Collider effect

Table 3 shows the model coefficients and goodness of fit and Figure 4 shows the linear fit for the association between age and SBP for each of the three different models both illustrating the effect of conditioning on a collider. As you can see the coefficient of age paradoxically becomes a protective factor with a SBP decrease of 0.32 mmhg for each increase of one year of age. Likewise, the linear fit showed in Figure 4C (collider model) in comparison with the fits from the unadjusted and bivariate adjusted model (Figures 4A and 4B) changes dramatically the interpretation of the main effect of age. We also provide the link to a shiny app (http://watzilei.com/shiny/collider/) where users can dynamically modify the input values of the beta coefficients of the collider model for the data generation process. The range of values of the slider input allows users to dynamically visualize the collider effect and to revert the collider effect. Very small values of the input for the coefficients of the model for the generation of the collider effect revert it while the effect is augmented when the strength of the association between proteinuria (collider) with the outcome (SBP) and the exposure (age) increases.

4 Conclusion

We investigated a situation where adding a certain type of variable to a linear regression model, called a collider, biased coefficients while still increasing predictive power. DAGs based on a great subject matter knowledge are vital to identify collider. Determining if a variable is a collider involves thinking critically about the data generation process and the relationship of the variables in a real world scenario. Then, the decision whether to include or exclude it in a regression model using observational data in epidemiology lies on the main interest of your modeling strategy. Whether it is prediction or explanation (causation) is the the key. If the main objective of your study is to obtain a predictive model that makes accurate predictions of Y (response) you probably will condition on the collider to increase the precision of your prediction. However, if your main objective is to create a model of reality that is useful in making decisions based on a structural causal framework, then conditioning on the collider would introduce selection bias and probably paradoxical effects. To conclude, making causal inferences on the basis of correlational data is very hard but if you have an explanatory variable W that could actually be caused by the response Y as well as the exposure or treatment A, then you probably are in front of a collider. Most research in epidemiology tries to explain how the world works (i.e. it is causal), thus to prevent paradoxical associations between your outcome and exposure you probably would not like to condition on a collider.

Contributors and sources

5 Key message

- Paradoxical associations between and outcome and exposure are common in epidemiological studies from observational data.
- A collider is a variable that is causally influenced by an exposure and an outcome.
- Controlling in multivariable analyses for a collier can introduce a spurious association between its causes (exposure and outcome).
- Using Directed Acyclic Graphs based on substantial clinical knowledge helps identifying colliders.
- Whether or not to adjust for a collider will depend on the main analytical interest i.e. predictive modeling approaches will condition on it to increase predictive accuracy while explicative modeling approaches will not condition on it to prevent bias.

Acknowledgements

Miguel Angel Luque Fernandez is supported by the Spanish National Institute of Health, Carlos III Miguel Servet I Investigator Award (CP17/00206).

6 Figures and Tables

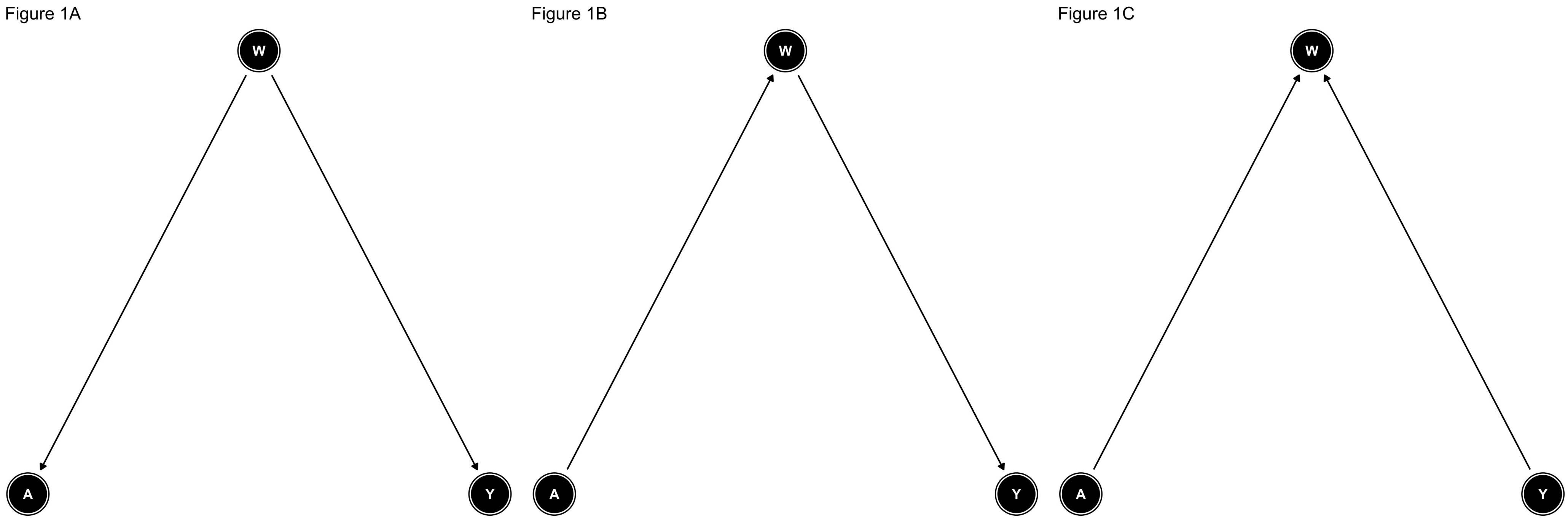

Figure 1: Basic structural associations between exposure and outcome

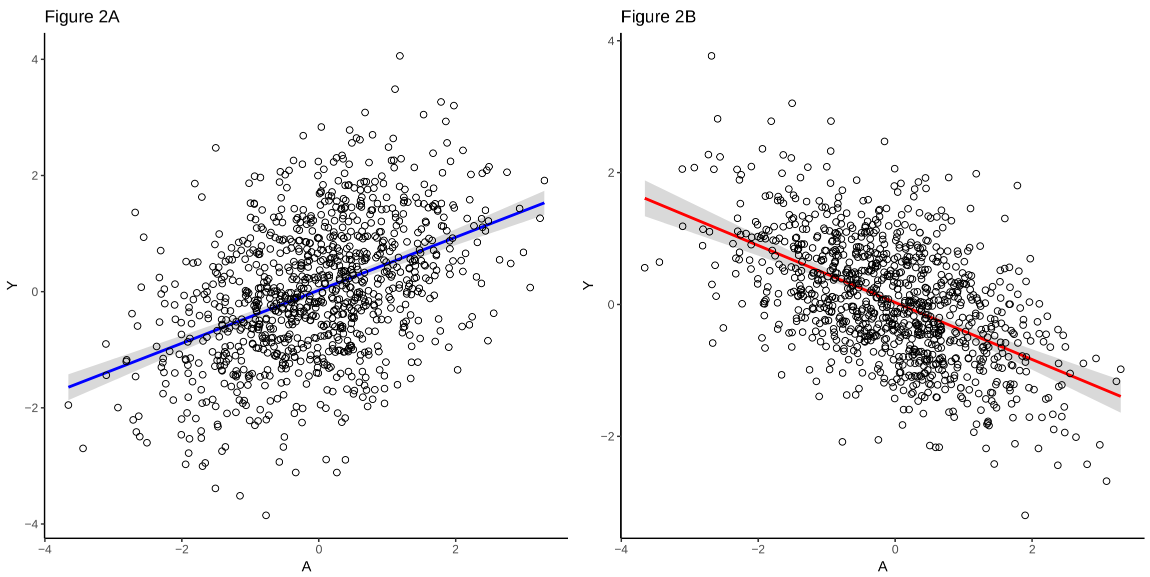

Figure 2: Visualization of the collider effect

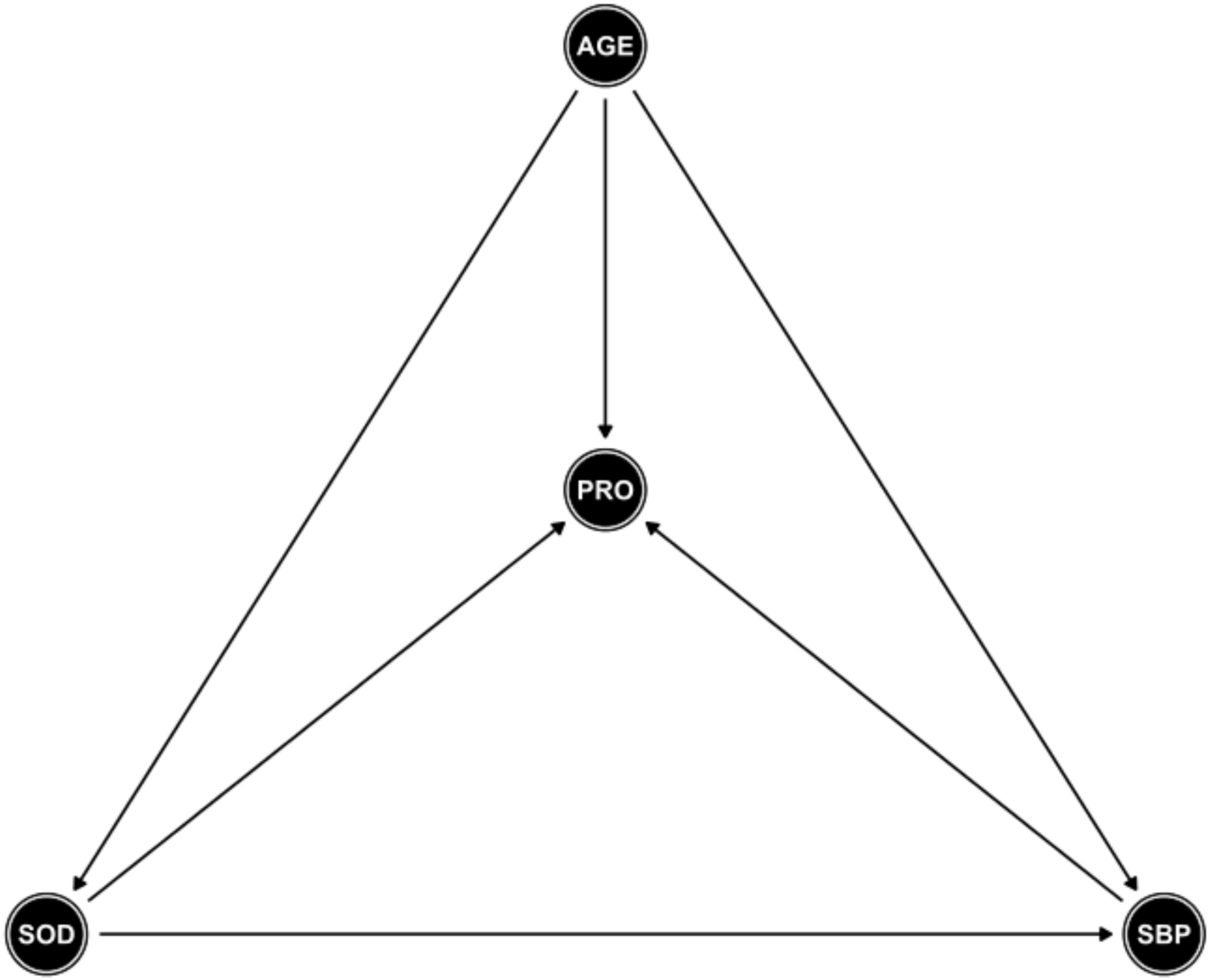

Figure 3: DAG for the collider

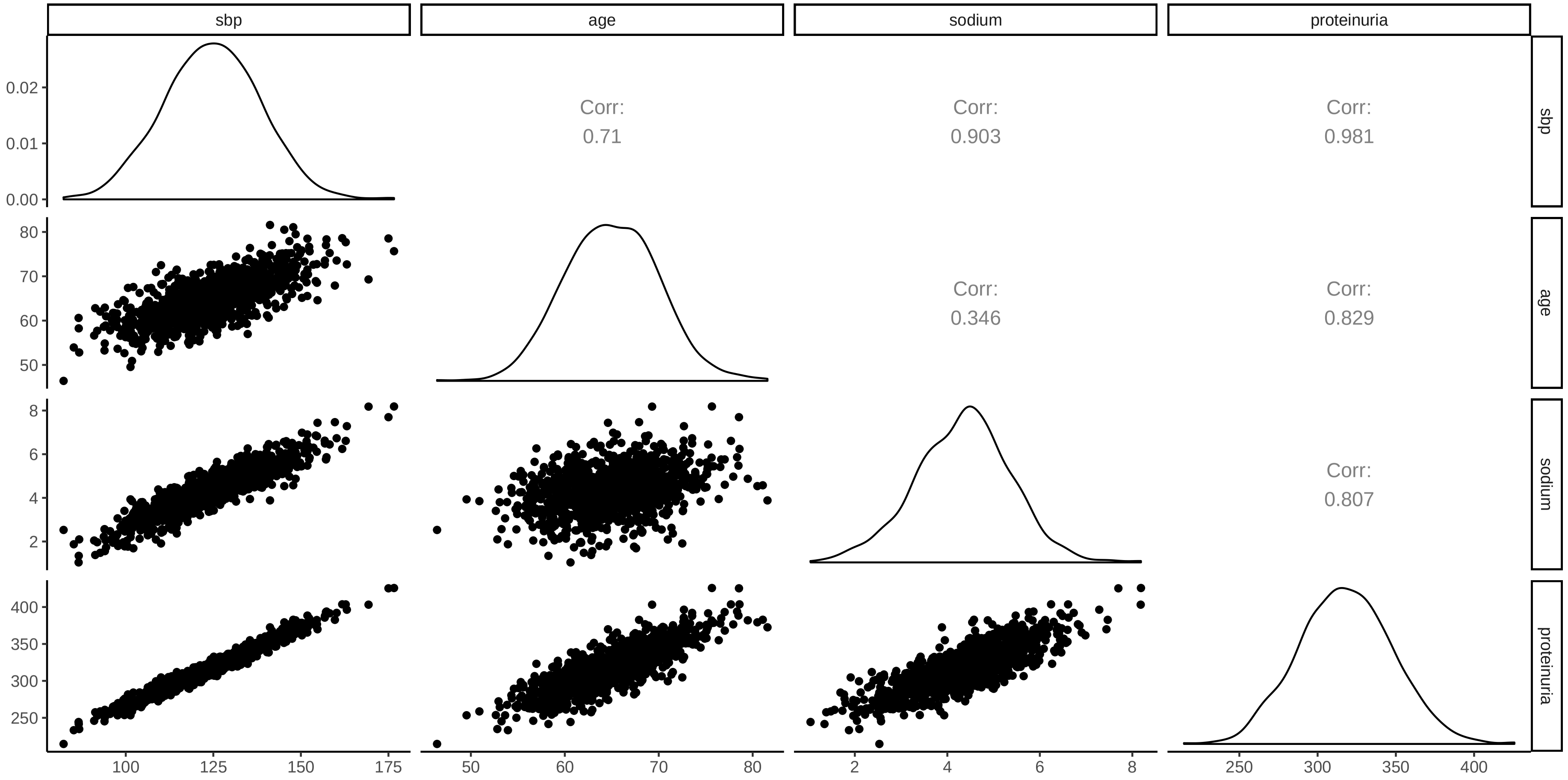

Figure 4: Visualization of the multivariate structure of the data generation

Table 1: Illustration of confounding and collider effects when adjusting for W the association between and outcome (Y) and an exposure (A), n = 1,000

	Dependent variable (Y)		
	Crude association	W as confounder	W as collider
	(1)	(2)	(3)
A	0.458***	0.285***	−0.384***
	(0.031)	(0.031)	(0.036)
W		0.441***	0.515***
		(0.035)	(0.017)
Constant	−0.009	0.014	0.015
	(0.034)	(0.031)	(0.024)
Observations	1,000	1,000	1,000
R^2	0.181	0.308	0.573
Adjusted R^2	0.180	0.307	0.572
Residual Std. Error	1.071 (df = 998)	0.975 (df = 997)	0.766 (df = 997)
F Statistic	219.864*** (df = 1; 998)	221.972*** (df = 2; 997)	669.137*** (df = 2; 997)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 2: Univariate, bivariate and multivariate association between systolic blood pressure and age adjusted for sodium intake acting as a confounder and proteinuria acting as a collider, n = 1,000

	Dependent variable (Y)		
		Systolic Blood Pressure in mmhg	
	(Univariate)	(Bivariate)	(Multivariate collider)
Age in years	1.972^{***}	1.255^{***}	-0.319^{***}
	(0.062)	(0.007)	(0.030)
Proteinuria in mg			0.419^{***}
			(0.008)
Sodium intake in g		9.989^{***}	2.825^{***}
		(0.032)	(0.136)
Constant	−3.570	−0.311	−0.105
	(4.039)	(0.407)	(0.208)
Observations	1,000	1,000	1,000
R^2	0.504	0.995	0.999
Adjusted R^2	0.503	0.995	0.999
Residual Std. Error	9.757 (df = 998)	0.983 (df = 997)	0.502 (df = 996)
F Statistic	$1,013.503^{***}$ (df = 1; 998)	$98,648.120^{***}$ (df = 2; 997)	$252,906.500^{***}$ (df = 3; 996)
Note:			*p<0.1; **p<0.05; ***p<0.01

Supplementary Table 1: Descriptive distribution of the simulated data, n = 1,000

Systolic blood pressure in mmhg	Age in years	Sodium in gr	Proteinuria mg in 24h
Min. : 82.25	Min. :46.40	Min. :1.042	Min. :214.6
1st Qu.:115.14	1st Qu.:61.57	1st Qu.:3.661	1st Qu.:297.1
Median :124.61	Median :64.91	Median :4.393	Median :317.8
Mean :124.65	Mean :65.01	Mean :4.344	Mean :318.0
3rd Qu.:133.96	3rd Qu.:68.35	3rd Qu.:4.998	3rd Qu.:337.9
Max. :176.59	Max. :81.58	Max. :8.186	Max. :425.8

7 Annex

```
library(visreg)
library(broom)
library(stargazer)

## Data Generation based on Noncommunicable Disease Epidemiology
generateData <- function(n){
    age <- rnorm(n, 65, 5)
    sodium  <- age/15 + rnorm(n)
    sbp <- 10*sodium + 1.25*age + rnorm(n)
    proteinuria  <- 1.5*age + 1.8*sbp - 0.9*sodium + rnorm(n)
    data.frame(sbp,age,sodium,proteinuria)
}

set.seed(777)
ObsData <- generateData(n=1000)

## Models Fit
fit0 <- lm(sbp ~ age, data = ObsData);tidy(fit0)
fit1 <- lm(sbp ~ age + sodium , data = ObsData);tidy(fit1)
fit2 <- lm(sbp ~ age + proteinuria + sodium, data = ObsData);tidy(fit2)

## Models visualization
par(mfrow=c(1,3))
visreg(fit0, ylab = "SBP in mmhg", line=list(col="blue"),
 points = list(cex = 1.5, pch = 1), jitter = 10, bty = "n")
visreg(fit1, ylab = "SBP in mmhg", line=list(col="blue"),
 points = list(cex = 1.5, pch = 1), jitter = 10, bty = "n")
visreg(fit2, ylab = "SBP in mmhg", line=list(col="red"),
 points = list(cex = 1.5, pch = 1), jitter = 10, bty = "n")

## Comparing Models
stargazer(fit0, fit1, fit2, type = "latex", multicolumn = FALSE)
```

References

[1] Sander Greenland and Hal Morgenstern. Confounding in health research. *Annual Review of Public Health*, 22(1):189–212, May 2001.

[2] Stephen R Cole, Robert W Platt, Enrique F Schisterman, Haitao Chu, Daniel Westreich, David Richardson, and Charles Poole. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, 39(2):417–420, Nov 2009.

[3] Tyler J. Vanderweele and Stijn Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2(4):457–468, 2009.

[4] Miguel Angel Luque-Fernandez, Helga Zoega, Unnur Valdimarsdottir, and Michelle A. Williams. Deconstructing the smoking-preeclampsia paradox through a counterfactual framework. *European Journal of Epidemiology*, 31(6):613–623, Jun 2016.

[5] S. Hernandez-Diaz, E. F. Schisterman, and M. A. Hernan. The birth weight "paradox" uncovered? *American Journal of Epidemiology*, 164(11):1115–1120, Sep 2006.

[6] Miguel A. Hernán, Sonia Hernández-Díaz, and James M. Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, Sep 2004.

[7] James M. Robins, Miguel Ángel Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, Sep 2000.

[8] Julia M Rohrer. Thinking clearly about correlations and causation: Graphical causal models for observational data. 2017.

[9] JUDEA PEARL. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[10] Sander Greenland, Judea Pearl, and James M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, Jan 1999.

[11] Linda Van Horn, Jo Ann S Carson, Lawrence J Appel, Lora E Burke, Christina Economos, Wahida Karmally, Kristie Lancaster, Alice H Lichtenstein, Rachel K Johnson, Randal J Thomas, Miriam Vos, Judith Wylie-Rosett, Penny Kris-Etherton, and American Heart Association Nutrition Committee of the Council on Lifestyle and Cardiometabolic Health; Council on Cardiovascular Disease in the Young; Council on Cardiovascular and Stroke Nursing; Council on Clinical Cardiology; and Stroke Council. Recommended dietary pattern to achieve adherence to the american heart association/american college of cardiology (aha/acc) guidelines: A scientific statement from the american heart association. *Circulation*, 134(22):e505–e529, Nov 2016.

[12] Michael F Carroll. Proteinuria in adults: A diagnositc approach. *American family physician*, 62(6), 2000.