

Educational Note: Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data

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

Michael Schomaker², BSc, MSc, PhD

Daniel Redondo-Sanchez¹, BSc

Maria Jose Sanchez Perez¹, MD, PhD,

Anand Vaidya³ MD, MSc

Mireille E. Schnitzer⁴ BSc, MSc, PhD

1 Summary

In the epidemiological literature different explanations have been proposed to describe the paradoxical protective effect of established risk factors. A collider for a certain pair of variables (exposure and outcome) is a third variable that is influenced by both of them. Controlling for, or conditioning the analysis on (i.e., stratification or regression) a collider, can introduce a spurious association between its causes (exposure and outcome). Based on a motivating example in non-communicable disease epidemiology, we generated a dataset with 1,000 observations to contextualize the effect of conditioning on a collider. We estimate the effect of 24-hour dietary sodium intake in grams on systolic blood pressure in mmHg accounting for the effect of age in years to illustrate the paradoxical effect of 24-hour dietary sodium intake on systolic blood pressure after conditioning on a collider (24-hour urinary protein excretion in mg). If the main objective of an study is to obtain a predictive model that makes accurate predictions of the outcome researchers probably will condition on the collider to increase the precision of the prediction. However, if the 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 collider bias and probably paradoxical effects.

2 Key messages box

- Paradoxical associations between an 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 collider can introduce a spurious association between its causes (exposure and outcome).
- Using Directed Acyclic Graphs based on substantial clinical knowledge helps to identify colliders.

¹Biomedical Research Institute. Non-Communicable and Cancer Epidemiology Group (ibs.Granada), Andalusian School of Public Health, University of Granada, Granada, Spain. Department of NonCommunicable Disease Epidemiology. London School of Hygiene and Tropical Medicine. Centre de Recherche en Epidemiologie, Biostatistique et Recherche Clinique Ecole de Sante Publique, Universite Libre de Bruxelles, Belgium. Department of Epidemiology. Harvard School of Public Health. Harvard University. miguel.luque.easp@juntadeandalucia.es

²University of Cape Town, Centre for Infectious Disease Epidemiology and Research, Observatory, 7925; Cape Town, South Africa

³Brigham and Women's Hospital. Harvard Medical School, Harvard University, Boston, MA, USA.

⁴Faculty of Pharmacy, University of Montreal, Montreal, Canada

- 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.

3 Introduction

During the last 30 years, classical epidemiology has focussed on the control of confounding [1]. However, it is only recently that epidemiologists have started to focus on the bias produced by colliders and mediators in addition to confounders [2, 3]. In the epidemiological literature different explanations have been proposed to describe the paradoxical protective effect of established risk factors. Such as, for example, the protective effect of maternal smoking on infant mortality and the incidence of pre-eclampsia, namely the birthweight and the smoking pre-eclampsia paradoxes [4, 5]. Directed acyclic graphs (DAGs) help to visualize the assumed structural relationship of the variables under analysis. Using DAGs we can distinguish between i) biases resulting from lack of conditioning on common causes of exposure and outcome (confounding) (Figure 1A); ii) conditioning on intermediates variables (mediation) (Figure 1B) or iii) on common effects (collider bias) (Figure 1C) [6, 7].

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

The reminder of the methodological note is structured as follows: i) we demonstrate the statistical structure of the collider bias through a simulated data generation using a linear system of structural equations; ii) we then illustrate the effect of conditioning on a collider based on a realistic non-communicable disease epidemiology example (hypertension and dietary sodium intake); iii) we provide R-code in easy-to-read boxes throughout the manuscript and a GitHub repository: <https://github.com/migariane/ColliderApp> for reproducibility and iv) we provide readers with an educational web application allowing real-time interaction to visualize the paradoxical effect of conditioning on a collider <http://watzilei.com/shiny/collider/>.

4 Statistical structure of a collider

4.1 Reviewing confounding

Regression is the tool that epidemiologist generally use to answer questions in observational studies, many of which are causal in principle. Let's begin by reviewing what we mean by controlling for a confounder (W) when assessing the association between treatment or exposure (A) and an outcome (Y) via the following causal graph (Figure 1A). In the above diagram, W acts as a confounder distorting the perceived causal relationship between A and Y if unaccounted for. Thus, confounding introduces bias, but it can be controlled conditioning on the confounder by regression or stratification. We now review adjustment for confounding via a linear regression model. In Box 1 we show how to generate data consistent with the DAG from Figure 1A after which we run two different regression models. W is generated as a standard normally randomly distributed variable with i.e. mean 0 ($\mu = 0$) and variance 1 ($\sigma^2 = 1$). The generation of A was forced to depend on W plus a standard normal distributed random noise and Y is generated depending on both the A and W plus a standard normal randomly distributed noise. Note that the true simulated effect of A is 0.3. Then, we run unadjusted (fit1) and adjusted (fit2: adjusted for W) regression models of the effect of A on Y and visualize the linear fit of both models using the package visreg

implemented in the statistical software R (R Foundation for Statistical Computing, Vienna, Austria).

Box 1

```
library(visreg)
N <- 1000
set.seed(777)
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()
```

Table 1 (columns 1, 2) shows the coefficients from the fitted regression models helping us to illustrate and interpret the confounding effect. 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 can be specified independently of the other variables in the model. However, both A and Y depend on W and they are thus endogenous to the system of linear equations, i.e., depends or is generated from the outcome Y or the exposure A . Thus, it does not depend on A or Y . The first regression ignores W and incurs an upward bias in the coefficient of A (0.471). The second regression including the confounder W recovers the true value of the coefficient of A (0.3) with a beta coefficient for A of 0.289 while increasing the R-squared from 0.197 to 0.298. Thus, statistically speaking, adjusting for W improves the goodness of fit of the model (Figure 2A, Table 1: columns 1, 2).

4.2 Collider deconstruction

Unlike in Figure 1A where the causal arrow starts from W , in Figure 1C they now point towards W from A and Y . If we condition on W (i.e., regression or stratification), we will have created a collider bias, i.e. a spurious association between A and Y . Classically, to describe the effect of the collider bias we use the expression *association is not causation*. It means that the association in a multiplicative scale $P(Y = 1 | A = 1) / P(Y = 1 | A = 0)$ or similarly that $P(Y = 1 | A = 1) - P(Y = 1 | A = 0)$ in an additive scale is not causation. Conditioning on the collider by regression or stratification introduces bias while ignoring the collider does not add bias. To emulate this scenario, in box 2 we generate the data, again using a simple linear data generating mechanism. First, we simulate A as a standard normal randomly distributed variable. Y is forced to depend on A plus a standard normal noise but W is generated depending on both A and Y plus a standard normal random noise. Notice that A is now exogenous while W is endogenous. Then we fit the unadjusted (fit3) excluding the collider and the collider model (fit4: collider model). Note that the true simulated coefficient of A is 0.3.

Box 2

```
library(visreg)
N <- 1000
set.seed(777)
A <- rnorm(N)
Y <- 0.3 * A + rnorm(N)
W <- 1.2 * A + 0.9 * Y + rnorm(N)
fit3 <- lm(Y ~ A)
fit4 <- lm(Y ~ A + W)
visreg(fit3, "A", gg = TRUE, line = list(col = "red"),
       points = list(size = 2, pch = 1, col = "black")) + theme_classic()
```

Table 1 (columns 3, 4) shows the coefficients for unadjusted (fit3) and the collider models (fit4). Unlike the previous section, the simpler regression without W approximately recovers the true coefficient of

A (0.3) with a value of the beta coefficient for A of 0.326, while the regression adjusting for W is substantially biased (-0.416). The collider model (fit4) after a statistical point of view is not unequivocally inferior. For instance, the goodness of fit of the collider model increases significantly with an R-squared that is approximately 80 percentage points higher than the unadjusted model. However, conditioning on the collider W has paradoxically changed the direction of the association between A and Y (Figure 2B, Table 1: column 3).

Note that even if the collider changed 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 should not condition on W. Subject-matter knowledge (i.e., plausible biological associations in clinical epidemiological settings) is mandatory in explicative or causal modeling approaches [10]. Thus, the use of DAGs as a knowledge translating tool helps enormously to explicitly display the believed or known casual relationships between variables and elucidate the presence of colliders [9, 11].

5 Motivating Example

5.1 Data generation

Based on a motivating example in non-communicable disease epidemiology, we generated a dataset with 1,000 observations to contextualize the effect of conditioning on a collider. Comorbidities are defined as the coexistence of disorders, in addition to a primary disease of interest, which are causally unrelated to the primary disease (e.g., cancer) [12]. High blood pressure is one of the most common comorbidities. Nearly 1 in 3 Americans suffer from high blood pressure and more than half do not have it under control [13]. Sustained levels of systolic blood pressure over time are associated with increased cardio-vascular morbidity and mortality [14]. Summative evidence shows that exceeding the recommendations for 24-hour dietary sodium intake in grams is associated with increased levels of systolic blood pressure (SBP) in mmHg [15]. Patients affected by high blood pressure are advised to decrease 24-hour dietary sodium intake but high SBP is associated with increasing age in years [15]. Thus, age is a confounder for the association between sodium intake and SBP. However, high levels of 24-hour excretion of urinary protein (proteinuria) are associated with sustained high SBP, advanced age and increased 24-hour dietary sodium intake. Therefore, as depicted in Figure 3, proteinuria acts as a collider (Figure 3).

We estimate the effect of 24-hour dietary sodium intake in grams on systolic blood pressure (SBP) accounting for the effect of age. The objective of the illustration is to show the paradoxical effect of 24-hour dietary sodium intake on SBP after conditioning on a collider (proteinuria).

Box 3 shows the data generation for the simulation based on the structural relationship between the variables depicted on a DAG (Figure 3). We assumed that SBP increases with increasing age and dietary sodium intake. We also simulated 24-hour excretion of urinary protein as a function of age, SBP, and sodium intake. We assured that the range of values of the simulated data was biologically plausible and as close to reality as possible [16, 17]. Supplementary Table 1 shows the descriptive statistics (minimum, maximum, mean, median, first and third quartiles) for the generated data.

The simulation implies a linear relationship between the variables. Thus, the interpretation of the beta coefficients in the formulae of the code on box 3 is straightforward (i.e., Systolic blood pressure = $\beta_1 \times \text{sodium} + \beta_2 \times \text{age} + \varepsilon$; where $\beta_1 = 2.25$, $\beta_2 = 2.00$, and ε is standard normal distributed error). Supplementary Figure 1 shows the functional form for each variable and the multivariable correlation matrix.

Box 3

```
generateData <- function(n, seed){
  set.seed(seed)
  Sodium_gr <- rnorm(n, 3.50, 0.50)
  Age_years <- Sodium_gr * 18 + rnorm(n)
  sbp_in_mmHg <- 2.25 * Sodium_gr + 2.00 * Age_years + rnorm(n)
  Proteinuria_in_mg <- 0.90 * Age_years + 1.80 * sbp_in_mmHg + 3.50 * Sodium_gr + rnorm(n)
  data.frame(sbp_in_mmHg, Sodium_gr, Age_years, Proteinuria_in_mg)
}

ObsData <- generateData(n = 1000, seed = 777)
```

We fit three different models to evaluate the effect of age on SBP: i) univariate unadjusted model; ii) bivariate adjusted model for age acting as a confounder; iii) multivariate adjusted model for age and proteinuria. The model specifications are shown here below; in Box 4 we show how to fit and visualize the corresponding models in R.

Models fit specification:

Systolic Blood Pressure = $\beta_0 + \beta_1 \times \text{Sodium in gr} + \varepsilon$

Systolic Blood Pressure = $\beta_0 + \beta_1 \times \text{Sodium in gr} + \beta_2 \times \text{Age in years} + \varepsilon$

Systolic Blood Pressure = $\beta_0 + \beta_1 \times \text{Sodium in gr} + \beta_2 \times \text{Age in years} + \beta_3 \times \text{Proteinuria in mg} + \varepsilon$

Box 4

```
## Models Fit
fit0 <- lm(sbp_in_mmHg ~ Sodium_gr, data = ObsData);tidy(fit0)
fit1 <- lm(sbp_in_mmHg ~ Sodium_gr + Age_years , data = ObsData);tidy(fit1)
fit2 <- lm(sbp_in_mmHg ~ Sodium_gr + Age_years + Proteinuria_in_mg, 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")
```

5.2 Effect of conditioning on a collider

Table 2 shows the model coefficients and goodness of fit and Figure 4 shows the linear fit for the association between sodium and SBP for each of the three different models illustrating the effect of conditioning on a collider. The coefficient for the effect of sodium on SBP paradoxically becomes a protective risk factor (i.e., for one unit increase in sodium intake SBP decreases 0.7 mmHg). Likewise, the linear fit shown in Figure 4C (collider model) in comparison with the fits from the unadjusted and bivariate adjusted model (Figures 4A and 4B) changes the interpretation of the main effect of sodium dramatically. We also provide the link to a shiny app (<http://watzilei.com/shiny/collider/>) (Supplementary Figure 2) 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 visualize the collider effect and revert it interactively. Very small values of the input slider for the coefficients of the model for the generation of the collider effect revert it while higher values create the collider effect.

6 Conclusion

We investigated a situation where adding a certain type of variable to a linear regression model, called a collider, biased coefficient estimates while still improving model fit (i.e., increasing predictive power). DAGs based on subject matter knowledge are vital for identifying colliders. Determining if a variable is a collider involves thinking critically about the true unobserved data generation process and the relationship between the variables in a real-world scenario. Then, the decision whether to include or exclude the variable in a regression model using observational data in epidemiology lies in the main interest of your modeling strategy (i.e., whether it is prediction or explanation/causation). If the main objective of an study is to obtain a predictive model that makes accurate predictions of Y (response) researchers probably will condition on the collider to increase the precision of the prediction. However, if the 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 collider bias and probably paradoxical effects. To conclude, making causal inferences on the basis of observational 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 faced with a collider. Most research in epidemiology tries to explain how the world works (i.e., it is causal), thus to prevent paradoxical associations, epidemiologist probably should not condition on a collider.

Contributors and sources

7 Contributors and sources

The article and shiny application arise from the motivation to disseminate the principles of modern epidemiology among clinicians and applied researchers. MALF developed the concept, designed the study, carried out the simulation, analysed the data, and wrote the article. DRS and MALF developed the shiny application. All authors interpreted the data and drafted and revised the manuscript, code for the manuscript, and code for the shiny application. All authors read and approved the final version of the manuscript. MALF is the guarantor of the article.

Acknowledgements

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

Anand Vaidya was supported by the National Institutes of Health (grants DK107407 and DK115392) and by the Doris Duke Charitable Foundation (award 2015085).

Mireille E. Schnitzer is supported by a New Investigator Salary Award from the Canadian Institutes of Health Research.

8 Figures and Tables

Figure 1A

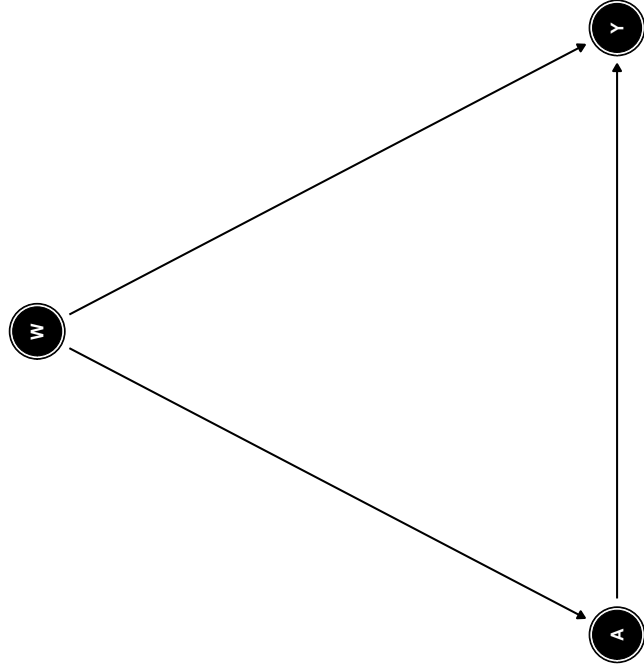


Figure 1B

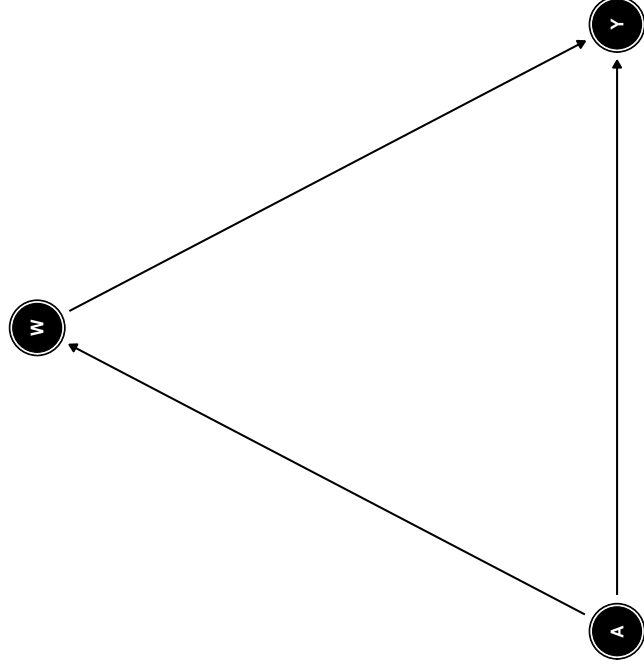


Figure 1C

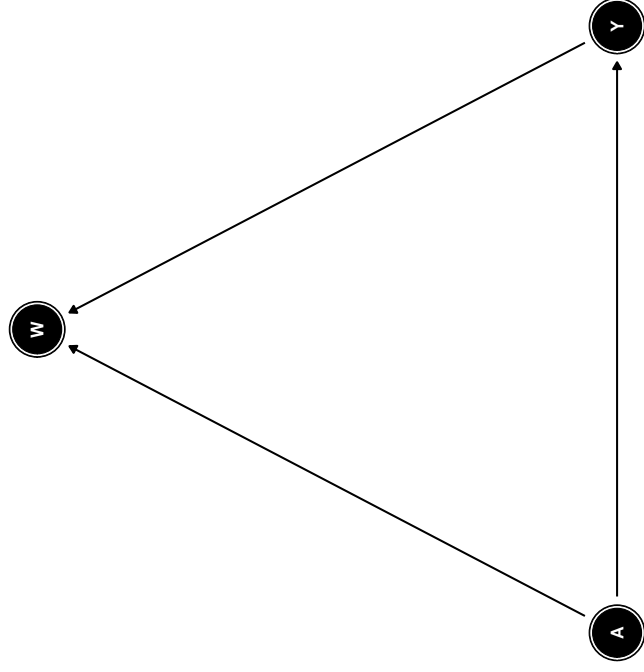


Figure 1: Basic structural associations between exposure and outcome

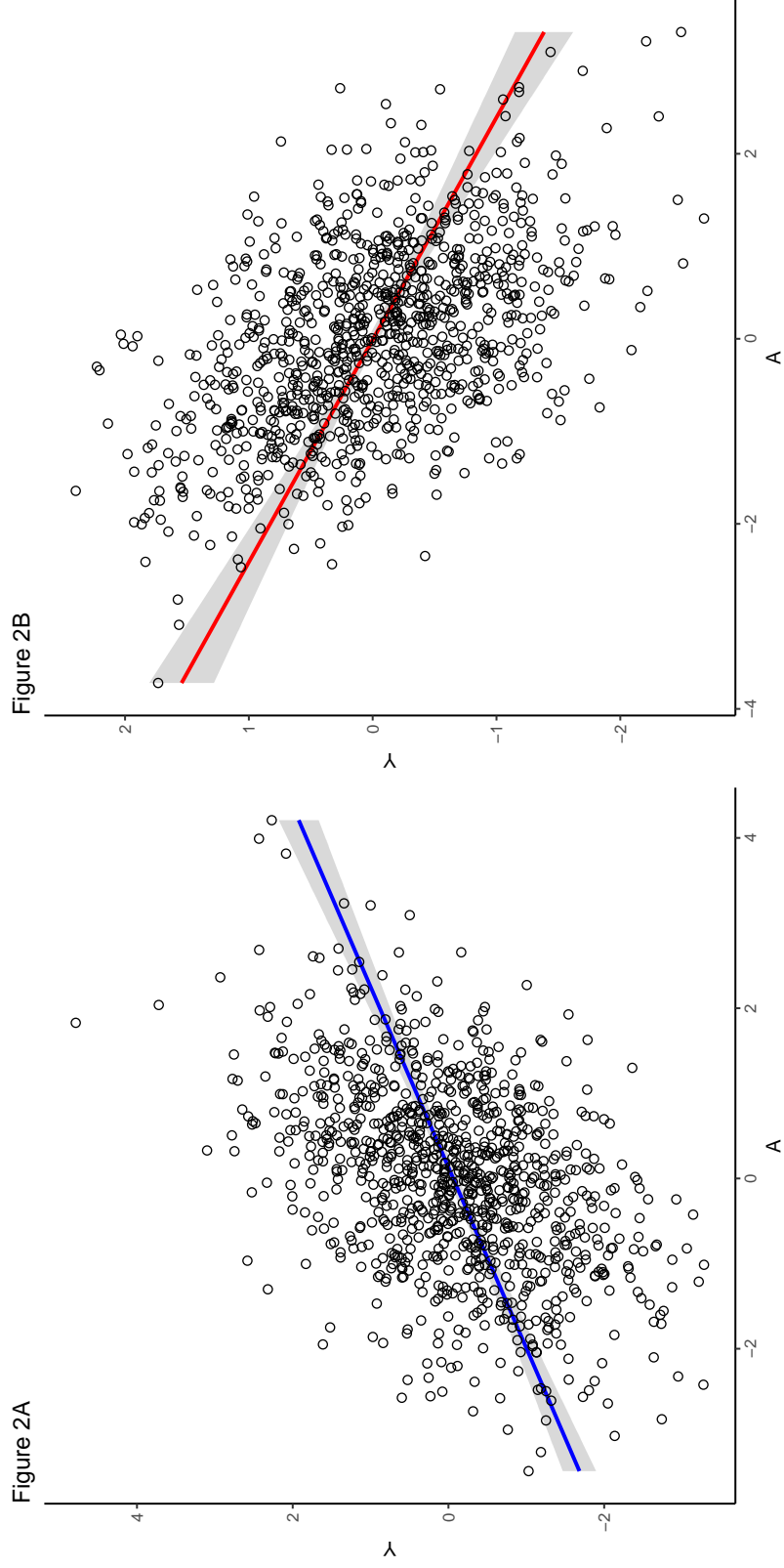


Figure 2: Visualization of the collider effect

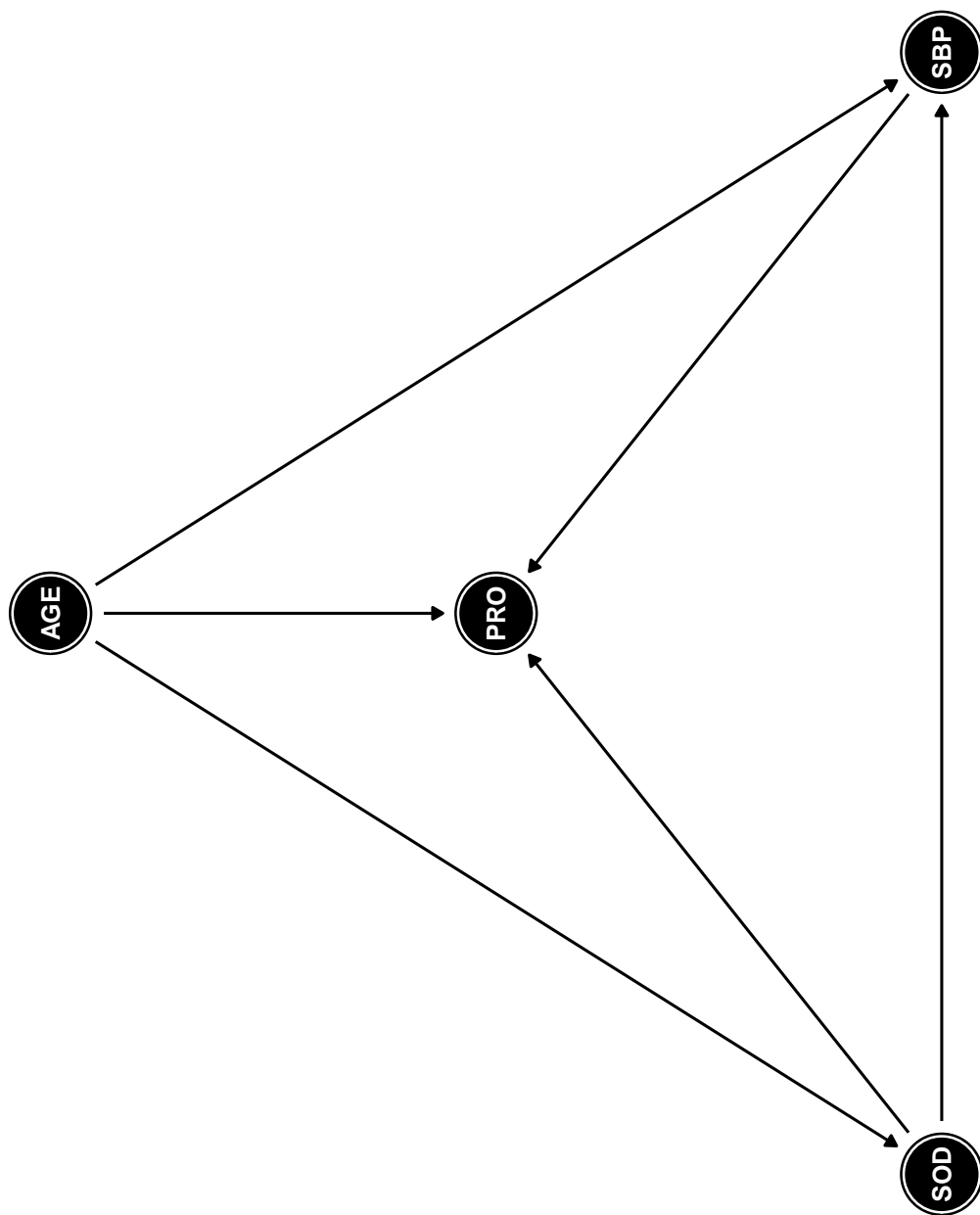


Figure 3: DAG for the collider

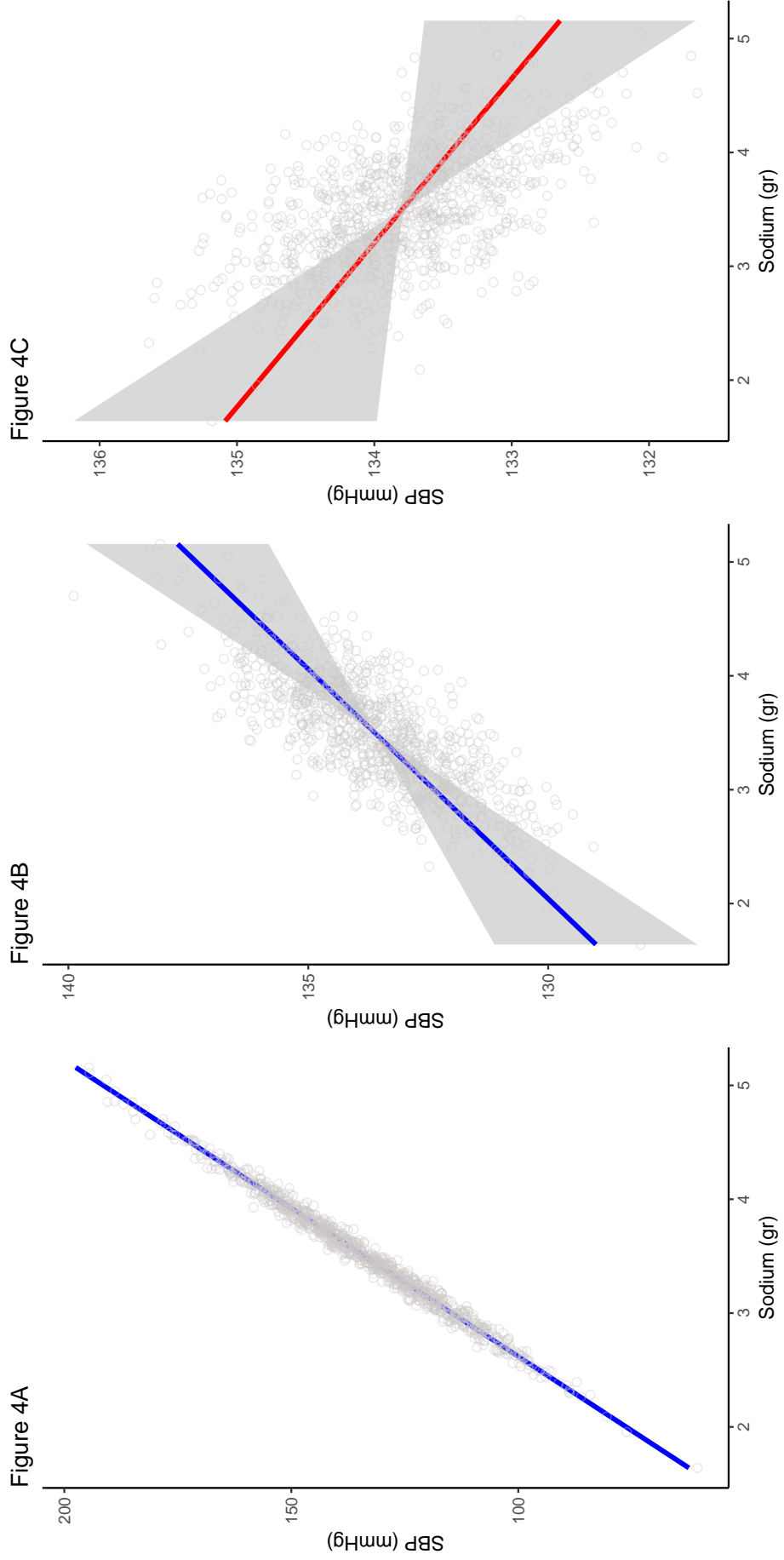


Figure 4: Collider effect for the illustration

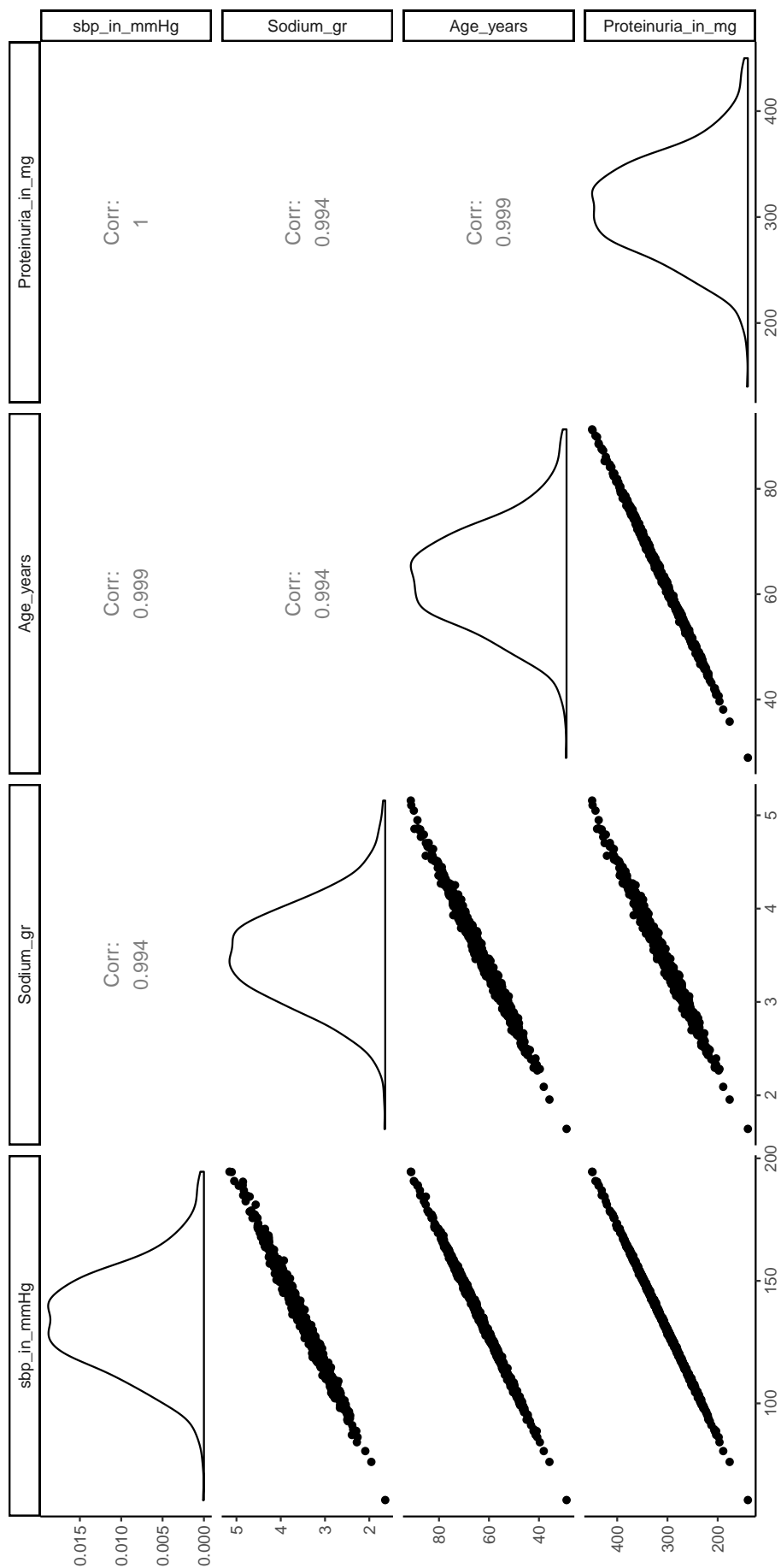


Figure 5: Visualization of the multivariate structure of the data generation

Colliders in Epidemiology: an educational interactive web application

Motivation Collider Visualization Data generation Article Credits & Acknowledgment

Effect of dietary sodium intake on systolic blood pressure for different models' specifications.

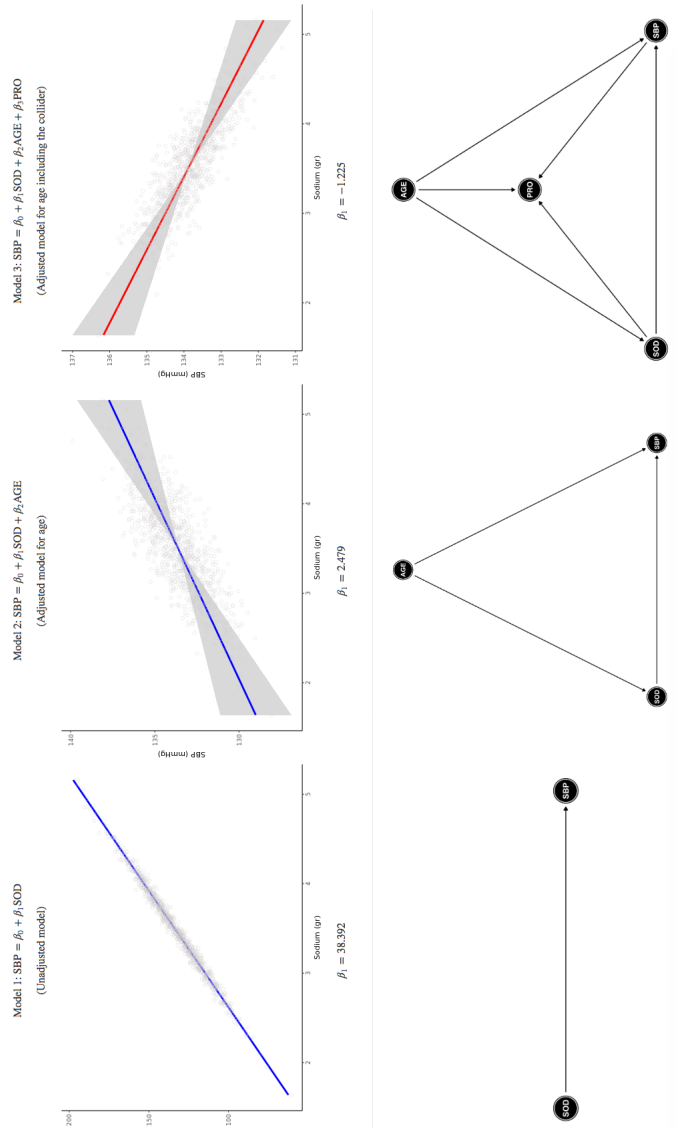


Figure 6: Collider shiny application screenshot

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

		Dependent variable (Y):			
		W as confounder		W as collider	
		Unadjusted (Fit1)	Adjusted (Fit2)	Unadjusted (Fit3)	Adjusted (Fit4)
A		0.471*** (0.030)	0.289*** (0.032)	0.326*** (0.031)	-0.416*** (0.035)
W			0.425*** (0.035)		0.491*** (0.018)
Constant		-0.061* (0.033)	-0.060* (0.031)	0.010 (0.031)	0.035 (0.023)
Observations		1,000	1,000	1,000	1,000
R ²		0.197	0.298	0.101	0.493
Adjusted R ²		0.196	0.297	0.100	0.492
Residual Std. Error		1.050 (df = 998)	0.983 (df = 997)	0.972 (df = 998)	0.730 (df = 997)
F Statistic		245.138*** (df = 1; 998)	211.778*** (df = 2; 997)	111.585*** (df = 1; 998)	484.929*** (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 24-hour sodium dietary intake adjusted for age acting as a confounder and proteinuria acting as a collider, n = 1,000

	Dependent variable (Y)		
	(Univariate)	Systolic Blood Pressure in mmHg (Bivariate)	(Multivariate collider)
Sodium in gr	38.392*** (0.138)	2.479*** (0.581)	-0.693** (0.303)
Age in years		1.989*** (0.032)	0.103*** (0.039)
Proteinuria in mg			0.419*** (0.008)
Constant	-0.535 (0.487)	-0.195 (0.221)	-0.072 (0.113)
Observations	1,000	1,000	1,000
R ²	0.987	0.997	0.999
Adjusted R ²	0.987	0.997	0.999
Residual Std. Error	2.168 (df = 998)	0.983 (df = 997)	0.502 (df = 996)
F Statistic	77,749.270*** (df = 1; 998)	191,223.900*** (df = 2; 997)	489,361.800*** (df = 3; 996)
<i>Note:</i>			
*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	Sodium in gr	Age in years	Proteinuria mg in 24h
Min. : 60.55	Min. :1.640	Min. :28.96	Min. :140.0
1st Qu.:120.38	1st Qu.:3.157	1st Qu.:56.70	1st Qu.:278.5
Median :133.93	Median :3.491	Median :62.90	Median :309.8
Mean :133.87	Mean :3.501	Mean :63.03	Mean :310.0
3rd Qu.:147.16	3rd Qu.:3.835	3rd Qu.:69.24	3rd Qu.:341.1
Max. :194.55	Max. :5.158	Max. :91.28	Max. :449.9

9 Annex

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

## Data Generation based on Noncommunicable Disease Epidemiology
generateData <- function(n, seed){
  set.seed(seed)
  Sodium_gr <- rnorm(n, 3.50, 0.50)
  Age_years <- Sodium_gr * 18 + rnorm(n)
  sbp_in_mmHg <- 2.25 * Sodium_gr + 2.00 * Age_years + rnorm(n)
  Proteinuria_in_mg <- 0.90 * Age_years + 1.80 * sbp_in_mmHg + 3.50 * Sodium_gr + rnorm(n)
  data.frame(sbp_in_mmHg, Sodium_gr, Age_years, Proteinuria_in_mg)
}

ObsData <- generateData(n = 1000, seed = 777)

## Models Fit
fit0 <- lm(sbp_in_mmHg ~ Sodium_gr, data = ObsData);tidy(fit0)
fit1 <- lm(sbp_in_mmHg ~ Sodium_gr + Age_years, data = ObsData);tidy(fit1)
fit2 <- lm(sbp_in_mmHg ~ Sodium_gr + Age_years + Proteinuria_in_mg, 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] M. A. Hernan. Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology*, 155(2):176–184, Jan 2002.
- [11] Sander Greenland, Judea Pearl, and James M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, Jan 1999.
- [12] Miquel Porta. *A dictionary of epidemiology*. Oxford university press, 2008.
- [13] Emelia J Benjamin, Michael J Blaha, Stephanie E Chiuve, Mary Cushman, Sandeep R Das, Rajat Deo, J Floyd, M Fornage, C Gillespie, CR Isasi, et al. Heart disease and stroke statistics-2017 update: a report from the american heart association. *Circulation*, 135(10):e146–e603, 2017.
- [14] Qiuping Gu, Vicki L Burt, Ryne Paulose-Ram, Sarah Yoon, and Richard F Gillum. High blood pressure and cardiovascular disease mortality risk among us adults: the third national health and nutrition examination survey mortality follow-up study. *Annals of epidemiology*, 18(4):302–309, 2008.
- [15] Frank M Sacks, Laura P Svetkey, William M Vollmer, Lawrence J Appel, George A Bray, David Harsha, Eva Obarzanek, Paul R Conlin, Edgar R Miller, Denise G Simons-Morton, et al. Effects on blood pressure of reduced dietary sodium and the dietary approaches to stop hypertension (dash) diet. *New England journal of medicine*, 344(1):3–10, 2001.
- [16] 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.
- [17] Michael F Carroll. Proteinuria in adults: A diagnostic approach. *American family physician*, 62(6), 2000.