Collider Effects and Paradoxical Results in the Analysis of Observational Studies: A Reproducible Illustration and Educational Shiny Application

Miguel Ángel Luque Fernández, Michael Schomaker, Daniel Redondo Sánchez, María José Sánchez Pérez, Anand Vaidya, Mireille E. Schnitzer

XXXVII SEE 2019 (Oviedo)

https://maluque.netlify.com/

http://watzilei.com/shiny/collider/

Background

International Journal of
Epidemiology

Issues | Advance articles | Submit ▾ | Purchase | Alerts | About ▾ | All International Jou ▾

Volume 48, Issue 2
April 2019

Article Contents

Abstract

Educational Note: Paradoxical collider effect in the analysis of non-communicable disease epidemiological data: a reproducible illustration and web application FREE

Miguel Angel Luque-Fernandez ✉, Michael Schomaker, Daniel Redondo-Sanchez, Maria Jose Sanchez Perez, Anand Vaidya, Mireille E Schnitzer

International Journal of Epidemiology, Volume 48, Issue 2, April 2019, Pages 640–653, https://doi.org/10.1093/ije/dyy275
Published: 14 December 2018 **Article history** ▾

41 Vie

A

International Journal of Epidemiology

Background

Colliders

- Classical epidemiology has focused on explicative modelling **Causal Inference** but it is only recently that epidemiologists have started to integrate predictive modelling **Machine Learning** in their causal models ("Two worlds").

Background

Colliders

- Classical epidemiology has focused on explicative modelling **Causal Inference** but it is only recently that epidemiologists have started to integrate predictive modelling **Machine Learning** in their causal models ("Two worlds").

- Therefore, classical epidemiology has focused on the control of **confounding** but it is only recently that epidemiologists have started to focus on the bias produced by other structures such as **colliders**.

Background

Colliders

- Classical epidemiology has focused on explicative modelling **Causal Inference** but it is only recently that epidemiologists have started to integrate predictive modelling **Machine Learning** in their causal models ("Two worlds").

- Therefore, classical epidemiology has focused on the control of **confounding** but it is only recently that epidemiologists have started to focus on the bias produced by other structures such as **colliders**.

Background: Colliders

Colliders

- A **collider** for a certain pair of variables (e.g., an outcome Y and an exposure A) is a third variable (C) that is caused by both.

Background: Colliders

Colliders

- A **collider** for a certain pair of variables (e.g., an outcome Y and an exposure A) is a third variable (C) that is caused by both.

- In a **directed acyclic graph** (DAG), a collider is the variable in the middle of an inverted fork (i.e., the variable C in A $\rightarrow$ **C** $\leftarrow$ Y).

Background: Colliders

Colliders

- A **collider** for a certain pair of variables (e.g., an outcome Y and an exposure A) is a third variable (C) that is caused by both.

- In a **directed acyclic graph** (DAG), a collider is the variable in the middle of an inverted fork (i.e., the variable C in A $\rightarrow$ **C** $\leftarrow$ Y).

Background

Figure 1A

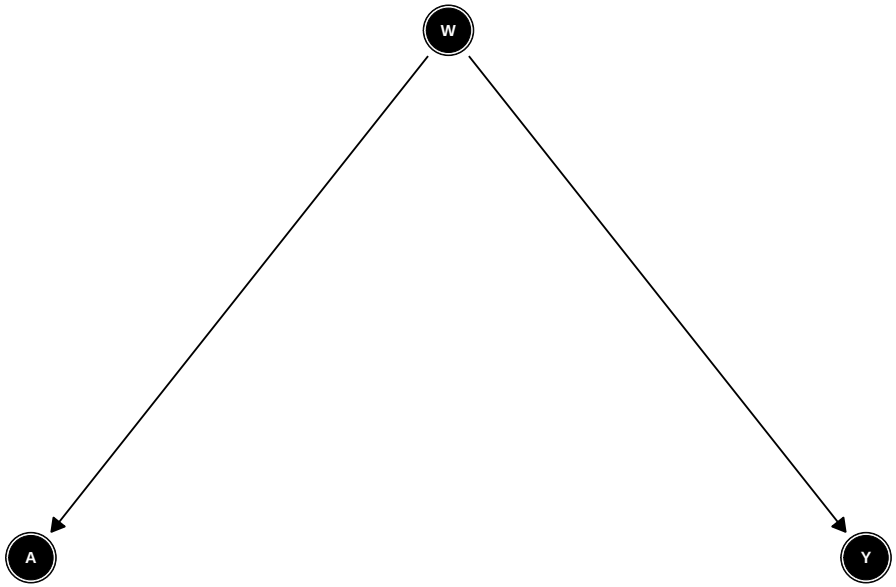
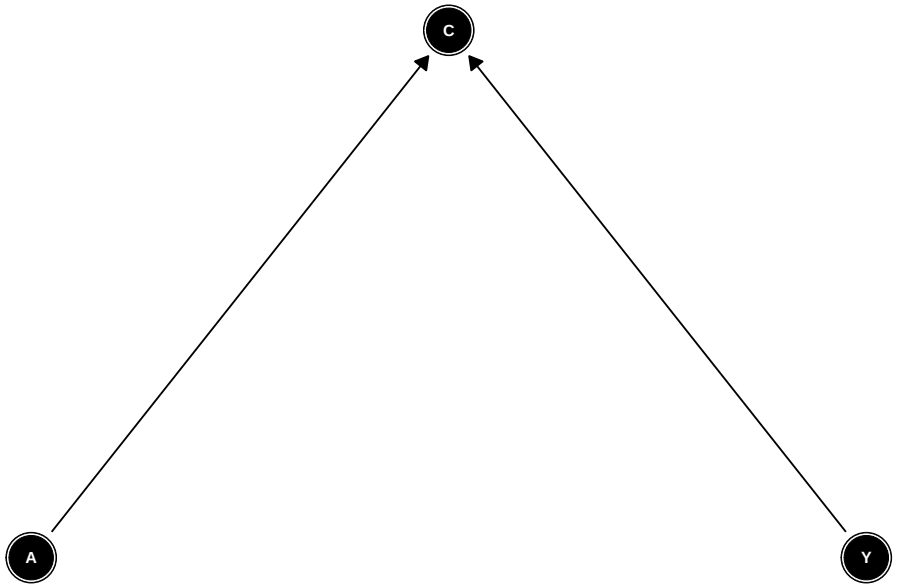


Figure 1B



Directed Acyclic Graphs

Background: Colliders

Colliders

- Controlling for, or conditioning an analysis on a collider (i.e., through stratification or regression) can introduce a **spurious association** between its causes.

Background: Colliders

Colliders

- Controlling for, or conditioning an analysis on a collider (i.e., through stratification or regression) can introduce a **spurious association** between its causes.

- This potentially explains many **paradoxical findings** in the medical literature, where established risk factors for a particular outcome appear protective.

Background: Colliders

Colliders

- Controlling for, or conditioning an analysis on a collider (i.e., through stratification or regression) can introduce a **spurious association** between its causes.

- This potentially explains many **paradoxical findings** in the medical literature, where established risk factors for a particular outcome appear protective.

- Desconstructing paradoxical effects in medical litterature: Luque-Fernandez MA et al. Deconstructing the **smoking-preeclampsia paradox** through a counterfactual framework. Eur J Epidemiol. 2016;31:613-623 (https://www.ncbi.nlm.nih.gov/pubmed/26975379).

Simple linear simulation

Confounder structure

```
N <- 1000                                  # sample size
set.seed(777)
W <- rnorm(N)                              # confounder
A <- 0.5 * W + rnorm(N)                    # exposure
Y <- 0.3 * A + 0.4 * W + rnorm(N)          # outcome
fit1 <- lm(Y ~ A)                          # crude model
fit2 <- lm(Y ~ A + W)                      # adjusted model
```

Collider structure

```
N <- 1000                                  # sample size
set.seed(777)
A <- rnorm(N)                              # exposure
Y <- 0.3 * A + rnorm(N)                    # outcome
C <- 1.2 * A + 0.9 * Y + rnorm(N)          # collider
fit3 <- lm(Y ~ A)                          # crude model
fit4 <- lm(Y ~ A + C)                      # adjusted model
```

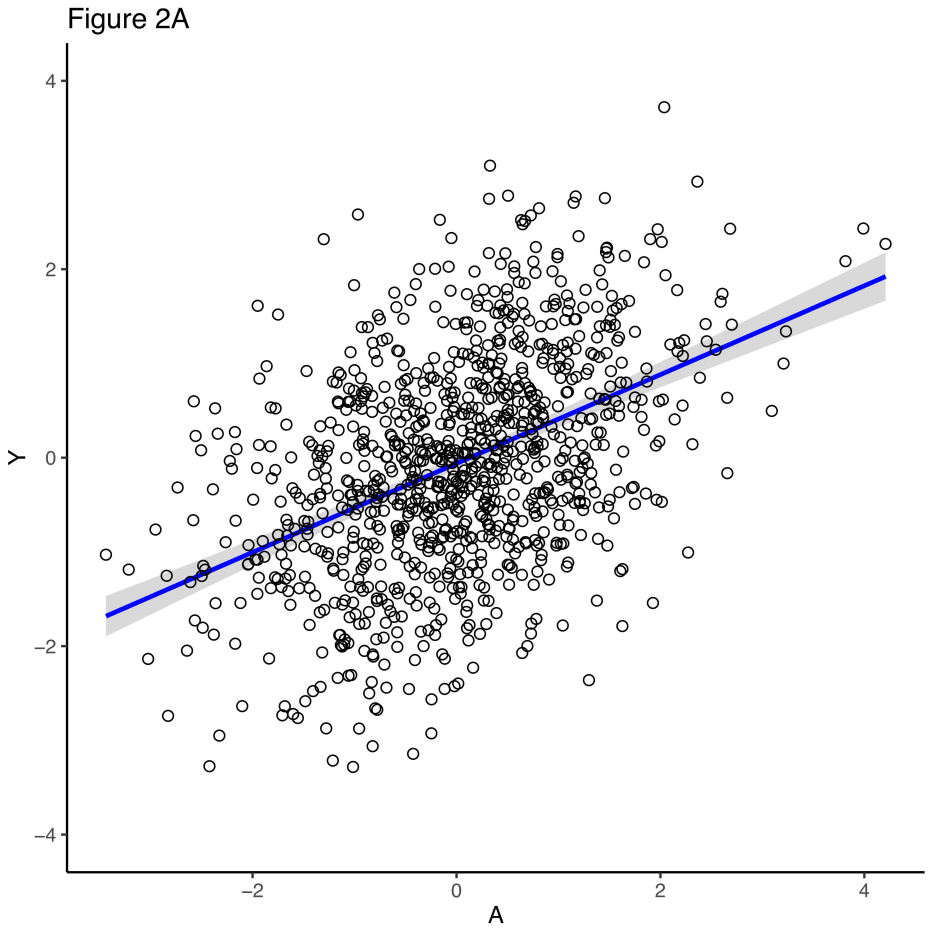
Collider and confounding effects

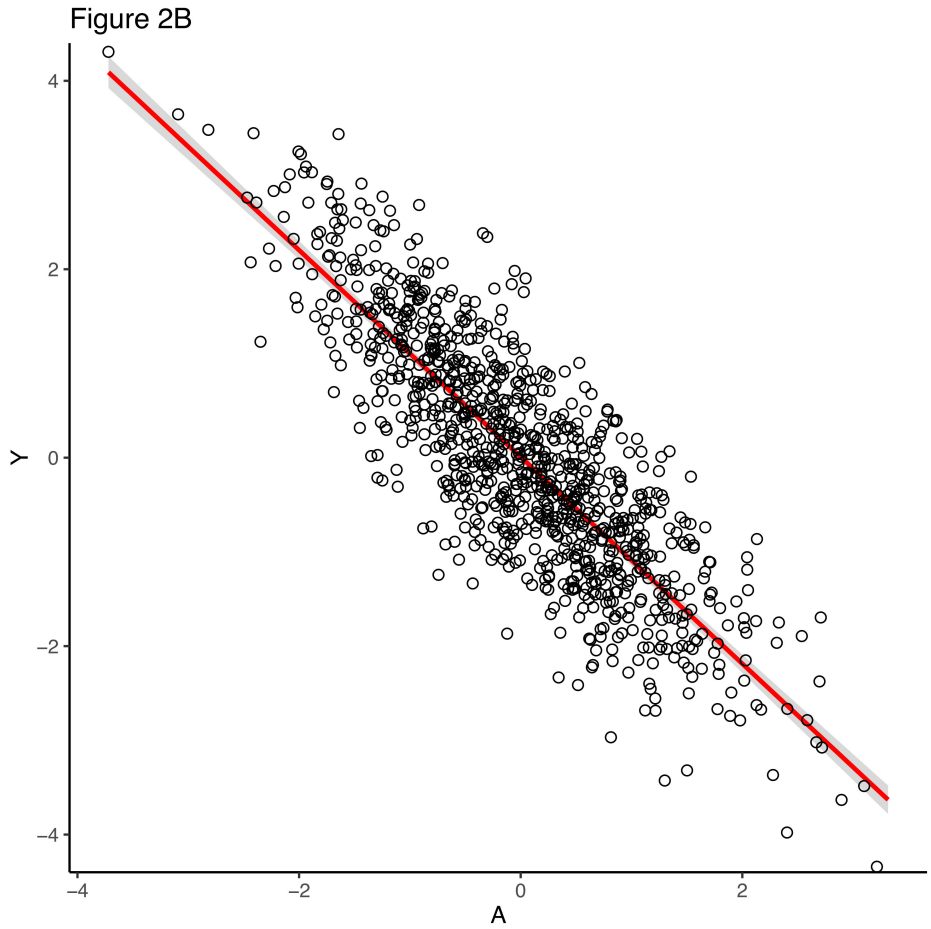
	Dependent variable (Y)			
	W (confounder)		C (collider)	
	Unadjusted β (SE)	Adjusted β (SE)	Unadjusted β (SE)	Adjusted β (SE)
	(Fit 1)	(Fit 2)	(Fit 3)	(Fit 4)
A	**0.471**	**0.289**	**0.326**	**-0.416**
	(-0.030)	(-0.032)	(-0.031)	(-0.035)
W		0.425		
		(-0.035)		
C				0.491
				(-0.018)
Intercept	-0.061	-0.06	0.01	0.035
	(-0.033)	(-0.031)	(-0.031)	(-0.023)
AIC	100.42	-31.992	-55.369	-626.824

Note: Lower AIC is better

Luque-Fernandez et al. Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application. International Journal of Epidemiology, Volume 48, Issue 2, April 2019. https://doi.org/10.1093/ije/dyy275

Display Linear Fit: models (fit2) and (fit4)

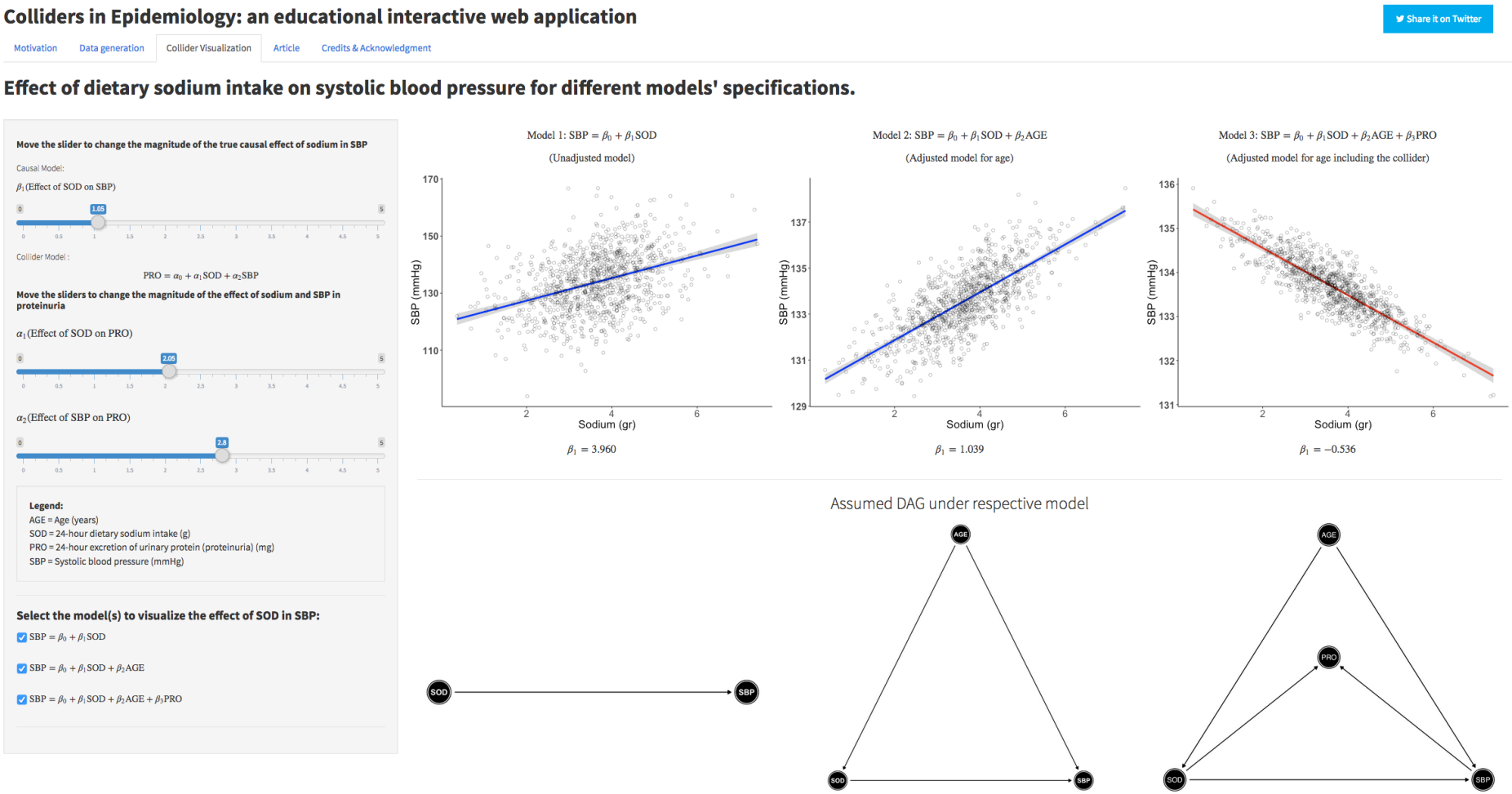




Collider Effect

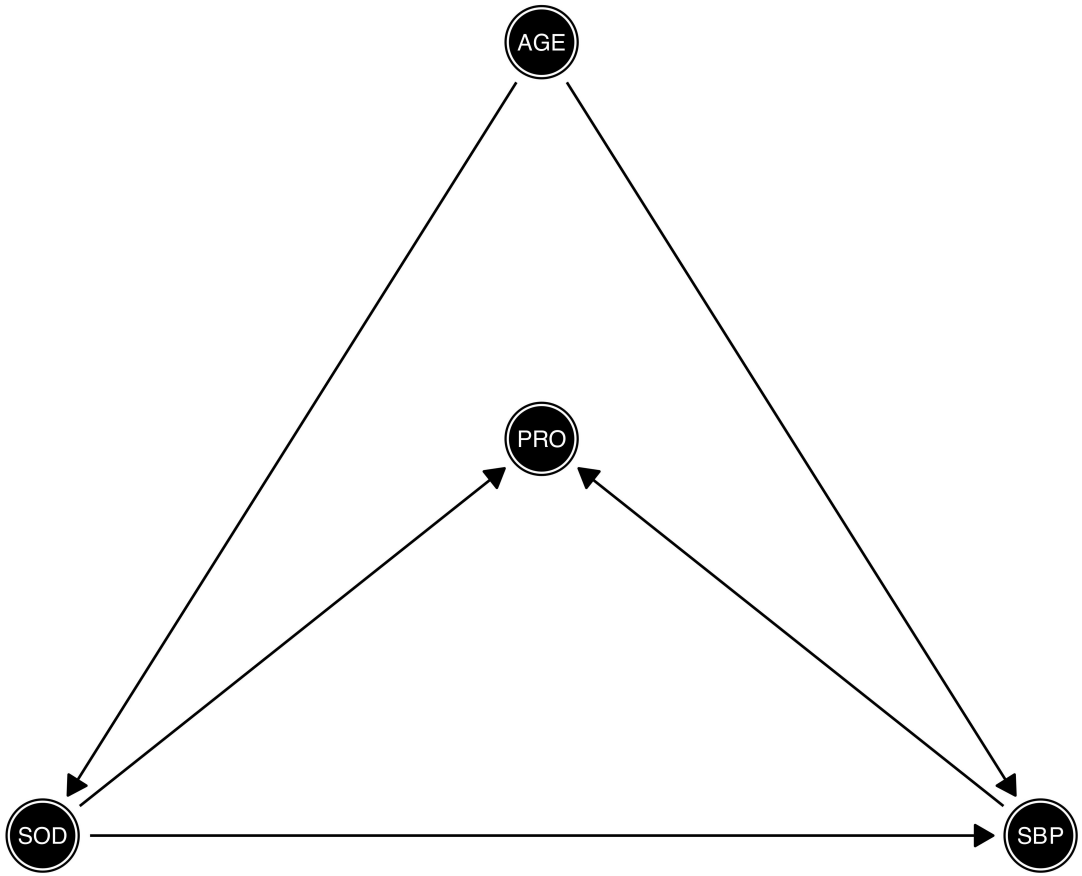
Luque-Fernandez et al. Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application. International Journal of Epidemiology, Volume 48, Issue 2, April 2019. https://doi.org/10.1093/ije/dyy275

Shiny web application



Colliders in Epidemiology: an educational interactive Shiny web application

Directed Acyclic Graph



Directed acyclic graph depicting the structural causal relationship of the exposure and outcome, confounding and collider effects. Exposure: 24-hour sodium dietary intake in gr (SOD), outcome: systolic blood pressure in mmHg (SBP), confounder: age in years (AGE), collider: 24-hour urinary protein excretion, proteinuria (PRO).

Luque-Fernandez et al. Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application. International Journal of Epidemiology, Volume 48, Issue 2, April 2019. https://doi.org/10.1093/ije/dyy275

Seeting Monte Carlo simulations

Data Generation

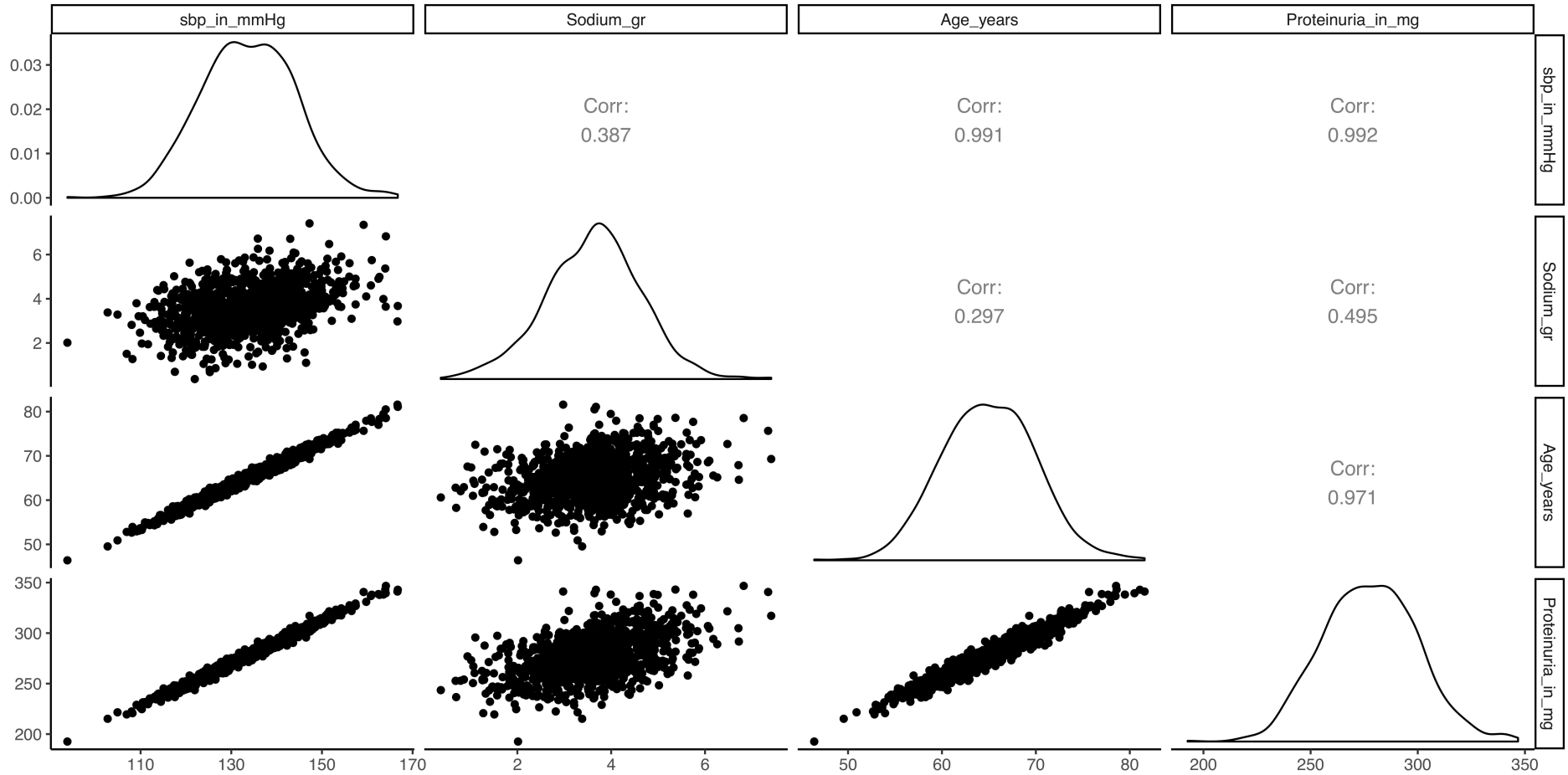
```
generateData <- function(n, seed){
set.seed(seed)
Age_years <- rnorm(n, 65, 5)
Sodium_gr <- Age_years / 18 + rnorm(n)
sbp_in_mmHg <- 1.05 * Sodium_gr + 2.00 * Age_years + rnorm(n)
hypertension <- ifelse(sbp_in_mmHg>140,1,0)
Proteinuria_in_mg <- 2.00*sbp_in_mmHg + 2.80*Sodium_gr + rnorm(n)
data.frame(sbp_in_mmHg, hypertension, Sodium_gr, Age_years,
Proteinuria_in_mg)
}
ObsData <- generateData(n = 1000, seed = 777)
```

Monte Carlo simulations

MC simulations

```
R<-1000
true <- rep(NA, R)
collider <- rep(NA,R)
se <- rep(NA,R)
set.seed(050472)
for(r in 1:R) {
if (r%%10 == 0) cat(paste("This is simulation run number", r, "\n"))
ObsData <- generateData(n=10000)
# True effect
true[r] <- summary(lm(sbp_in_mmHg ~ Sodium_gr + Age_years, data = ObsData))$coef[2,1]
# Collider effect
collider[r] <- summary(lm(sbp_in_mmHg ~ Sodium_gr + Age_years + Proteinuria_in_mg,
               data = ObsData))$coef[2,1]
se[r] <- summary(lm(sbp_in_mmHg ~ Sodium_gr + Age_years + Proteinuria_in_mg, data = ObsData))$coef[2,2]
}
# Estimate of sodium true effect
mean(true)
# Estimate of sodium biased effect in the model including the collider
mean(collider)
# simulated standard error/confidence interval of outcome regression
lci <- (mean(collider) - 1.96*mean(se)); mean(lci)
uci <- (mean(collider) + 1.96*mean(se)); mean(uci)
# Bias
Bias <- (true - abs(collider));mean(Bias)
# % Bias
relBias <- ((true - abs(collider)) / true); mean(relBias) * 100
# Plot bias
plot(relBias)
```

One sample MC simulations



Visualization of the multivariate structure of the data generation, n = 1,000.

Luque-Fernandez et al. Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application. International Journal of Epidemiology, Volume 48, Issue 2, April 2019. https://doi.org/10.1093/ije/dyy275

Models specifications

Unadjusted model

$$\text{SBP in mmHg} = \beta_0 + \beta_1 \times \text{Sodium in gr} + \varepsilon$$

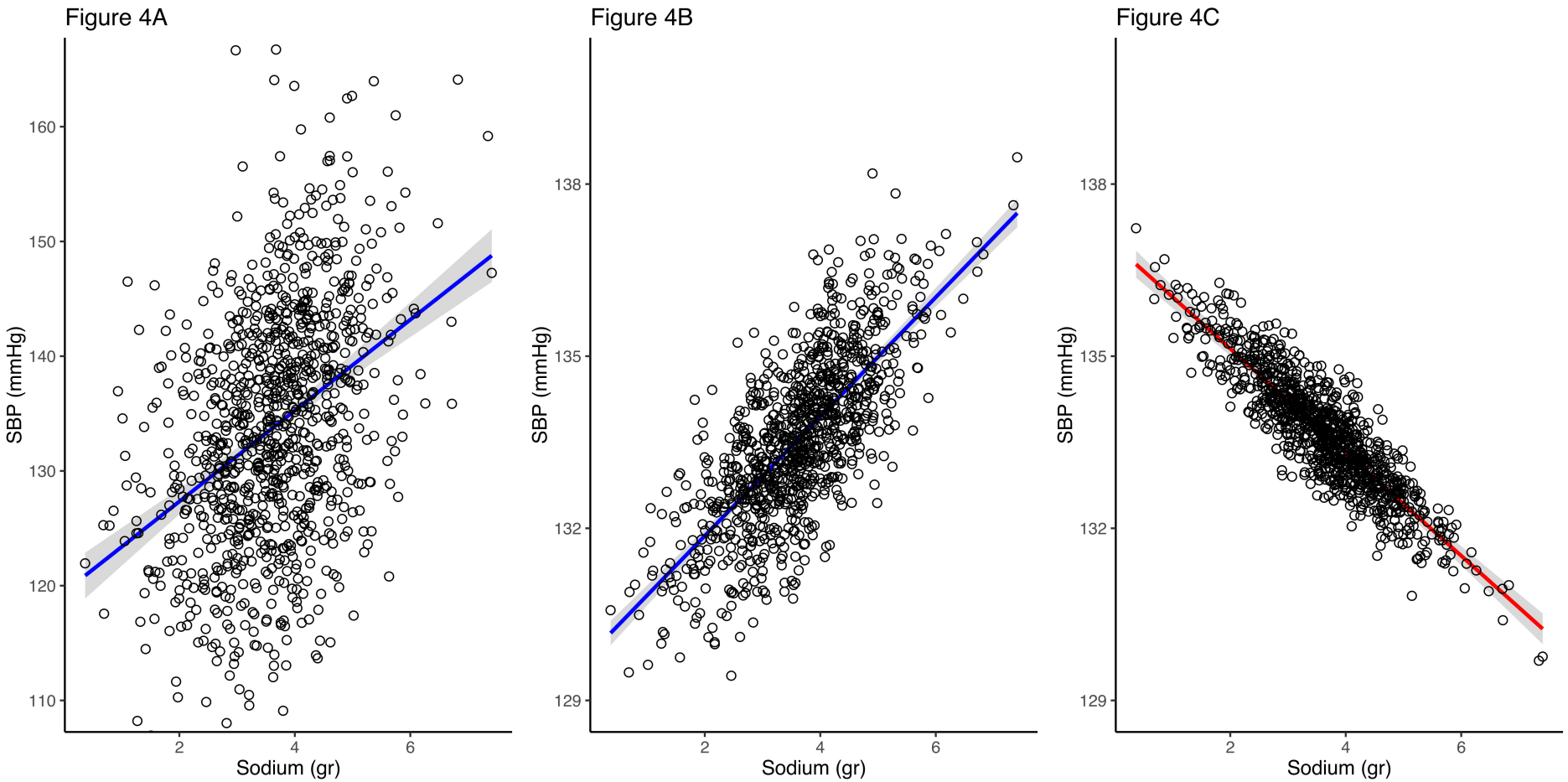
Adjusted model (confounder)

$$\text{SBP in mmHg} = \beta_0 + \beta_1 \times \text{Sodium in gr} + \beta_2 \times \text{Age in years} + \varepsilon$$

Adjusted model (confounder and collider)

$$\text{SBP} = \beta_0 + \beta_1 \times \text{Sodium} + \beta_2 \times \text{Age} + \beta_3 \times \text{Proteinuria} + \varepsilon$$

Models fit visualization



Luque-Fernandez et al. Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application. International Journal of Epidemiology, Volume 48, Issue 2, April 2019. https://doi.org/10.1093/ije/dyy275

Collider and confounding effects

	Dependent variable: SBP in mmHg		
	Univariate (SE)	Bivariate (SE)	Multivariate (SE)
True effect of Sodium in gr: 1.05			
Sodium in gr	**3.960**	**1.039**	**-0.902**
	(0.298)	(0.032)	(0.036)
Age in years		2.004	0.416
		(0.007)	(0.027)
Proteinuria in mg			0.396
			(0.007)
Intercept	119.420	-0.311	-0.091
	(1.122)	(0.407)	(0.192)
AIC	7363.45	2807.89	1302.66

Note: Lower AIC is better

Luque-Fernandez et al. Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application. International Journal of Epidemiology, Volume 48, Issue 2, April 2019. https://doi.org/10.1093/ije/dyy275

Tutorial Causal Inference

Introduction to Causal Inference (short course)

https://ccci.netlify.com/

Collider Shiny App

http://watzilei.com/shiny/collider/

GitHub Open source Collider files

https://github.com/migariane/ColliderApp

Causal Inference tutorial: TMLE

https://www.ncbi.nlm.nih.gov/pubmed/29687470

¡Gracias por vuestra atención!

Miguel Ángel Luque-Fernández

miguel.luque.easp@juntadeandalucia.es @watzilei

Carlos III Institute of Health, Grant/Award Number: **CP17/00206**
Andalusian Department of Health, Grant Number: **PI-0152/2017**

Background Causal Modelling: Potential Outcomes

Rubin and Heckman

- This framework was developed first by statisticians (Rubin, 1983) and econometricians (Heckman, 1978) as a new approach for the estimation of **causal effects** from observational data.

- We will keep separate the **causal framework** (a conceptual issue briefly introduce here) and the **"how to estimate causal effects"** (an statistical issue also introduced here)

Notation and definitions

Observed Data

- Treatment **A**.

  Often, A = 1 for treated and A = 0 for control.

- Confounders **W**.
- Outcome **Y**.

Potential Outcomes

- For patient i $\mathbf{Y_i(1)}$ and $\mathbf{Y_i(0)}$ set to A = a $Y^{(a)}$, namely A = 1 and A = 0.

Causal Effects

- Average Treatment Effect: **E[Y(1) - Y(0)]**.

Background: Causal effects with observational data

Potential Outcomes

Treatment (A) effect on outcome (Y) in real world:

$$Y_i(1) = Y_i(A = 1) \text{ and } Y_i(0) = Y_i(A = 0)$$

However we would like to know what would have happened if:

Treated $\mathbf{Y_i(1)}$ would have been non-treated $\mathbf{Y_i(A = 0) = Y_i(0)}$.

Controls $\mathbf{Y_i(0)}$ would have been treated $\mathbf{Y_i(A = 1) = Y_i(1)}$.

Identifiability

- How we can identify the effect of the potential outcomes $Y^{\mathbf{a}}$ if they are not observed?
- How we can estimate the expected difference between the potential outcomes **E[Y(1) - Y(0)]**, namely the **ATE**.

Background: Causal Inference Assumptions

IGNORABILITY

$$(\mathbf{Y_i(1)}, \mathbf{Y_i(0)}) \perp \mathbf{A_i} \mid \mathbf{W_i}$$

POSITIVITY

POSITIVITY: $\mathbf{P(A = a \mid W) > 0}$ for all a, W

SUTVA

- We have assumed that there is only on version of the treatment (**consistency**) $\mathbf{Y(1)}$ if $\mathbf{A = 1}$ and $\mathbf{Y(0)}$ if $\mathbf{A = 0}$.
- The assignment to the treatment to one unit doesn't affect the outcome of another unit (**no interference**) or **IID** random variables.
- The model used to estimate the assignment probability has to be **Correctly Specified**.

G-Formula, (Robins, 1986)

G-Formula for the **identification** of the ATE with observational data

$$
\begin{aligned}
E(Y^a) &= \sum_y E(Y^a \mid W = w) P(W = w) \\
&= \sum_y E(Y^a \mid A = a, W = w) P(W = w) \text{ by consistency} \\
&= \sum_y E(Y = y \mid A = a, W = w) P(W = w) \text{ by ignorability}
\end{aligned}
$$

The **ATE**=

$$
\sum_{\mathbf{w}} \left[\sum_{\mathbf{y}} \mathbf{P(Y = y \mid A = 1, W = w)} - \sum_{\mathbf{y}} \mathbf{P(Y = y \mid A = 0, W = w)} \right] \mathbf{P(W = w)}
$$

$$
P(W = w) = \sum_{y,a} P(W = w, A = a, Y = y)
$$

G-Formula, (Robins, 1986)

G-Formula for the identification of the ATE with observational data

The **ATE**=

$$\sum_{\mathbf{w}} \left[\sum_{\mathbf{y}} \mathbf{P(Y = y \mid A = 1, W = w)} - \sum_{\mathbf{y}} \mathbf{P(Y = y \mid A = 0, W = w)} \right] \mathbf{P(W = w)}$$

$$P(W = w) = \sum_{y,a} P(W = w, A = a, Y = y)$$

G-Formula

- The sums is generic notation. In reality, likely involves sums and integrals (we are just integrating out the W's).
- The **g-formula** is a **generalization of standardization** and allow to estimate unbiased treatment effect estimates.

Regression-adjustment

$$\widehat{ATE}_{RA} = N^{-1} \sum_{i=1}^{N} [E(Y_i \mid A = 1, W_i) - E(Y_i \mid A = 0, W_i)]$$

$$m_A(w_i) = E(Y_i \mid A_i = A, W_i)$$

$$\widehat{ATE}_{RA} = N^{-1} \sum_{i=1}^{N} [\hat{m}_1(w_i) - \hat{m}_0(w_i)]$$