Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application

Miguel Angel Luque-Fernandez* [1,2,3,4]
Michael Schomaker [5]
Daniel Redondo-Sanchez [1]
Maria Jose Sanchez Perez [1,6]
Anand Vaidya [7]
Mireille E. Schnitzer [8,9]

* **Corresponding author**
Miguel Angel Luque-Fernandez
Biomedical Research Institute of Granada (ibs.Granada),
Non-communicable disease and Cancer Epidemiology Group,
University of Granada, Spain.
Campus Universitario de Cartuja,
C/Cuesta del Observatorio 4, 18080 Granada, Spain.
Tel.: +34 958 027400.
miguel.luque.easp@juntadeandalucia.es

Abstract

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 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 DAGs terminology, a collider is the variable in the middle of an inverted fork (i.e., the variable C in A $\rightarrow$ C $\leftarrow$ Y). 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. We used an example from non-communicable disease epidemiology to contextualize and explain the effect of conditioning on a collider. We generated a dataset with 1,000 observations and ran Monte-Carlo simulations to estimate the effect of 24-hour dietary sodium intake on systolic blood pressure, controlling for age, which acts as a confounder, and 24-hour urinary protein excretion, which acts as a collider. We illustrate how adding a collider to a regression model introduces bias. Thus, to prevent paradoxical associations, epidemiologists estimating causal effects should be wary of conditioning on colliders. We provide R-code in easy-to-read boxes throughout the manuscript and a GitHub repository (https://github.com/migariane/ColliderApp) for the reader to reproduce our example. We also provide an educational web 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.
[2]Department of Non-Communicable Disease Epidemiology. London School of Hygiene and Tropical Medicine. London, U.K.
[3]Centre de Recherche en Epidemiologie, Biostatistique et Recherche Clinique Ecole de Sante Publique, Universite Libre de Bruxelles, Belgium.
[4]Department of Epidemiology. Harvard School of Public Health. Harvard University. Boston, MA, USA.
[5]Centre of Infectious Disease Epidemiology and Research. University of Cape Town, Cape Town, South Africa.
[6]Public Health and Epidemiology CIBER (CIBERESP), Madrid, Spain.
[7]Brigham and Women's Hospital. Harvard Medical School, Harvard University, Boston, MA, USA.
[8]Faculty of Pharmacy, University of Montreal, Montreal, Canada.
[9]Department Epidemiology, Biostatistics and Occupational Health. McGill University, Montreal, Canada.

Keywords

epidemiological methods, causality, noncommunicable disease epidemiology.

Key messages box

- Paradoxical associations between an outcome and exposure are common in epidemiological studies using observational data.
- A collider is a variable that is causally influenced by two other variables.
- Controlling for a collider in multivariable regression analyses can introduce a spurious association between its causes (exposure and outcome).
- Directed Acyclic Graphs based on existing subject matter knowledge can help to identify colliders.
- Whether or not it is adviseable to adjust for a collider depends on the main analytical interest. For instance, a predictive model may condition on a collider to increase prediction accuracy, while one should typically not condition on it when estimating causal effects to prevent bias.

1 Introduction

During the last 30 years, classical epidemiology has focused on the control of confounding [1]. It is only recently that epidemiologists have started to focus on the bias produced by colliders in addition to confounders [2, 3]. Directed acyclic graphs (DAGs) can help to visualize the assumed structural relationships between the variables under analysis. With this framework, we can distinguish between biases resulting from i) not conditioning on common causes of exposure and outcome (unadjusted confounding) or ii) conditioning on common effects (collider bias) [4, 5]. Epidemiologists use DAGs to determine the set of variables that are necessary to control for confounding and to summarize the subject-matter knowledge of the data-generating process. Using the DAGs terminology, variables, including A (exposure) and Y (outcome), are "nodes" connected by an arrow (a.k.a. directed edge) and a "path" is a way to get from one node to another, traveling along its arrows. The directed arrow ($\rightarrow$) from A to Y means that one does not exclude the possibility that A causes Y [6, 7].

A collider for a certain pair of variables (e.g. outcome and exposure) is a third variable that is caused by both of them. In DAG terminology, a collider is the variable in the middle of an inverted fork (i.e., variable C in A $\rightarrow$ C $\leftarrow$ Y) [6, 7].Using regression to control for a collider, or stratifying the analysis with respect to a collider, can introduce a spurious association between its causes, which can potentially introduce non-causal associations between the exposure and the outcome. This has been used to explain why the medical literature contains many paradoxical findings, where established risk factors appear protective for the outcome [8, 9, 10, 11]. For instance, numerous studies have reported a paradoxical protective effect of maternal cigarette smoking during pregnancy on preeclampsia, namely the pre-eclampsia smoking paradox. This paradox is due to gestational age at delivery, which is a collider between smoking (exposure) and pre-eclampsia (outcome) [8].

We hope that this methodological note will contribute to the increasing awareness and the general understanding of colliders among applied epidemiologists. The remainder of this note is structured as follows:

i. We review terminology related to DAGs and the rules one can follow to determine whether a causal effect is estimable;

ii. We demonstrate the statistical structure of collider bias using a simulated dataset;

iii. We illustrate the effect of conditioning on a collider using a realistic non-communicable disease epidemiology example (hypertension and dietary sodium intake);

iv. We provide R-code in easy-to-read boxes throughout the manuscript and in a GitHub repository: https://github.com/migariane/ColliderApp; and

v. 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/.

2 Statistical structure of confounding and a collider effects

2.1 Review of confounding

Traditionally a confounder has been defined as a third variable (W) associated with both the exposure (A) and the outcome (Y) that is not in the causal pathway between A and Y. Note that in Figure 1A, both the outcome (Y) and the exposure (A) share a common parent (direct cause). Y and A are both called descendants of W as they are both caused by W. The confounder wholly or partially accounts for the observed association of the exposure (A) on the outcome (Y). The presence of a confounder can lead to confounding bias and thus inaccurate estimates of the effect of A on Y. More precisely, bias means that the associational measure, for example the crude odds ratio, is different from the causal effect, such as the true marginal causal odds ratio (we give a clear definition of a marginal causal effect further below). Figure 1A gives an example of a confounding structure, where the path A $\rightarrow$ W $\leftarrow$ Y is called a back-door path which is defined as any path from A to Y that starts with an arrow into A. Without conditioning on variables, a path is open when it does not contain colliders. An open back-door path can be blocked, and confounding removed, by conditioning on non-colliders (via regression or stratification). In Figure 1, conditioning on the confounder W blocks the open back-door path. A path that is blocked by a collider can be opened by conditioning on the collider [11].

Causal effects are often formulated in terms of potential outcomes, as formalised by Rubin [12]. Let A denote a continuous exposure, W a pre-exposure vector of potential confounders, and Y a continuous outcome. Each individual has a potential outcome corresponding to any given level of the exposure, that is, the outcome they would have received had they been exposed to A=a, denoted Y(a). However, it is only possible to observe a single realisation of the outcome for an individual. We may observe Y(a) only for those who were exposed with A=a [12].

To sufficiently control for confounding, epidemiologists must identify a set of variables in the DAG that block all open back-door paths from the exposure (A) to the outcome (Y) by conditioning on variables along each path (i.e., using stratification or regression). In statistical terms, being able to block all back-door paths is known as conditional exchangeability or ignorability. If W is the set of confounding variables, then Y(a)$\perp$ A|W refers to conditional exchangeability, where the symbol $\perp$ means "independent". It implies that (within the strata of W) the expected potential outcome Y(a) is the same regardless of the exposure level, i.e. E(Y(a)|A, W) is the same regardless of the value of A that the individual actually received. We therefore have no systematic differences in how subjects would have performed under any given exposure that are not already explained by W.

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. The confounder W is generated as a standard normal random variable i.e. with mean 0 ($\mu = 0$) and variance 1 ($\sigma^2 = 1$). The generation of A depends on the value of W plus an error term and Y is generated depending on both A and W plus an error term, where both error terms have independent standard normal distributions. Note that the true simulated effect of A on Y is 0.3 (the coefficient in the linear regression model). Then, we fit unadjusted (fit1) and adjusted (fit2: adjusted for W) linear regression

models to estimate associations between A and Y. We visualize the fit of both models using the R software package visreg, where we used R version 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria).

Box 1

```
library(visreg) # load package to visualize regression output
library(ggplot2)# load package to visualize regression output
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
# visualize crude and adjusted models
visreg(fit1, "A", gg = TRUE, line = list(col = "blue"),
points = list(size = 2, pch = 1, col = "black")) + theme_classic()
visreg(fit2, "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 does not have parents in Figure 1A, i.e., it is not caused by any variable. Relatedly, in the code, it is the only variable that is generated independently of the other variables in the model. However, both A and Y depend on a common cause W (their parent) which is the source of the open back-door path between A and Y. As an illustration of the confounding bias due to W, Table 1 (columns 1, 2) shows the coefficients of A and W from the fitted regression models. The first regression does not condition on W and therefore has an upwards bias in the coefficient of A (0.471). However, the second regression closes the open back-door path by including the confounder W in the regression model. Thus, it estimates the causal effect as 0.289, close to the true coefficient (0.3) (Figure 2A, Table 1: columns 1, 2), the residual difference being entirely due to sampling variability.

2.2 Collider structure

Unlike in Figure 1A, where the causal arrows start from W, in Figure 1B they now point towards C from A and Y. If we condition on C (e.g. using regression or stratification), we will create a collider bias, i.e. a spurious association between A and Y. The common effect C is referred to as a collider on the path A $\rightarrow$ C $\leftarrow$ Y because two arrow heads collide on this node. For intuition, suppose that rain (A) and a sprinkler (Y) are the only two causes of a wet ground (C). We also assume that the sprinkler is on a daily timer, and not related to the weather. Then, if the ground is wet, knowing that it hasnt rained implies that the sprinkler must be on. If we ignore the colliding structure, we may conclude that rain has a negative effect on the sprinkler even when we know a priori that this is not the case.

Figure 1C gives another, more complex, collider structure, usually known as M-bias, in which the collider (C) is the effect of a common cause (W1) of the exposure (A) and a common cause (W2) of the outcome (Y). There is only one back-door path, and it is already blocked by the collider (C) thus we do not need to control for anything. This is the difference between confounders and colliders: a path will be open if one does not adjust for confounders, but blocked if adjustment is made; for colliders it is the other way around. However, some could consider C to be a classical confounder as it is associated with both A, via (A $\leftarrow$ W1 $\rightarrow$ C), and with Y, via a path that does not go through A (C $\leftarrow$ W2 $\rightarrow$ Y), and it is not in the causal pathway between A and Y. However, controlling for C will introduce a collider bias.

Classically, to describe collider bias we may use the expression association is not causation. As noted above, this means that measures of association, such as the conditional mean difference in the case of a binary A, E(Y|A=1,W) - E(Y|A=0,W), is not identical to its marginal causal counterpart, the average treatment effect: E(Y(1))-E(Y(0)).

The collider induces an association between the potential outcomes (Y(a)) and the exposure (A) given their common effect (C) and conditional ignorability (Y(a)⊥A|W) does not hold anymore. In other words: in Figures 1B and 1C, conditioning on the collider C opens the back-door path between A and Y which was previously blocked by the collider itself (A → C ← Y). Thus, the association between A and Y would be a mixture of the association due to the effect of A on Y and the association due to the open back-door path. Thus, association would not be causation anymore.

To simulate the scenario portrayed in Figure 1B, we generate data, again using a simple linear data generating mechanism (Box 2). First, we simulate A as a standard normally distributed variable. Y equals the value of A plus an error term and C is generated depending on both A and Y, plus error. Note that as shown in Figure 1B, now the exposure A and the outcome Y, are the parents of C (their common effect). We fit the unadjusted model excluding the collider (fit3) and then the model including the collider (fit4: collider model). The true coefficient of A is 0.3.

Box 2

```
library(visreg)  # load package to visualize regression output
library(ggplot2) # load package to visualize regression output
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
# visualize crude and adjusted models
visreg(fit3, "A", gg = TRUE, line = list(col = "red"),
points = list(size = 2, pch = 1, col = "black")) + theme_classic()
visreg(fit4, "A", gg = TRUE, line = list(col = "red"),
points = list(size = 2, pch = 1, col = "black")) + theme_classic()
```

Table 1 (columns 3, 4) shows the coefficient of A in the unadjusted model (fit3) and the coefficients of A and C in the model adjusting for the collider (fit4). Unlike in the previous section, the simpler regression without C approximately recovers the true coefficient of A (0.3) with an estimate of 0.326, while the regression adjusting for C is substantially biased (-0.416). The model which includes the collider (fit4) is not unequivocally inferior from a predictive point of view, where the main focus is to improve the models predictive performance. For instance, the model containing the collider has a much lower Akaike Information Critierion (AIC) than the one without the collider (Table 1). However, conditioning on the collider C has paradoxically changed the direction of the association between A and Y (Figure 2B, Table 1: column 3). Thus in this case, conditioning on the collider in theregression model introduces a bias while ignoring the collider does not add bias. The paradoxical negative association occurs when both A and Y are positively correlated with the collider.

From this demonstration, it is clear that subject-matter knowledge (i.e., plausible biological mechanisms clinical epidemiological settings) is necessary to perform causal estimation [13]. Thus, using DAGs to communicate causal structural relationships between variables helps in identifying variables that act as a collider, and identify where conditioning may create non-causal associations between the exposure (A) and outcome (Y) [13, 14, 15].

3 Motivating Example

3.1 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. Nearly 1 in 3 Americans suffer

from hypertension and more than half do not have it under control [16]. Increased levels of systolic blood pressure over time are associated with increased cardiovascular morbidity and mortality [17].

Summative evidence shows that exceeding the recommendations for 24-hour dietary sodium intake in grams (gr) is associated with increased levels of systolic blood pressure (SBP) in mmHg [18]. Furthermore, with advancing age, the kidney undergoes several anatomical and physiological changes that limit the adaptive mechanism responsible for maintaining the composition and volume of the extracellular fluid. These include a decline in glomerular filtration rate and the impaired ability to maintain water and sodium homeostasis in response to dietary and environmental changes [19].

Increasing age causes both high SBP and impaired sodium homeostasis. Thus, age acts as a confounder for the association between sodium intake and SBP (i.e. age is on the back-door path between sodium intake and SBP) as depicted in Figure 3. However, high levels of 24-hour excretion of urinary protein (proteinuria) are caused by sustained high SBP, advanced age and increased 24-hour dietary sodium intake. Therefore, as depicted in Figure 3, proteinuria acts as a collider (via the path SOD $\rightarrow$ PRO $\leftarrow$ SBP). Controlling for proteinuria (PRO) introduces collider bias.

We are interested in estimating the effect of 24-hour dietary sodium intake (in grams) on SBP, adjusting for 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 simulated data based on the structural relationship between the variables depicted in the DAG from 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 aimed to have a range of values of the simulated data which was biologically plausible and as close to reality as possible [20, 21]. Supplementary Table 1 shows the descriptive statistics (minimum, maximum, mean, median, first and third quartiles) of the generated data. Note that for educational purposes, we present the code and results for a single dataset simulated by our data-generating mechanism. However, at the end of the illustration, we also present the results of 1,000 Monte-Carlo simulations with a sample size of 10,000 patients aiming to quantify the bias associated with conditioning on a collider. The simulation assumes linear relationships 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 $= \beta_1 \times$ sodium $+ \beta_2 \times$ age $+ \varepsilon$; where $\beta_1 = 1.05$, $\beta_2 = 2.00$; and ε is a standard normally distributed error, where 1.05 is the true causal effect of sodium intake on SBP). Supplementary Figure 1 shows the functional form for each variable and the multivariable Spearmans correlation matrix.

Box 3

```
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 <- 0.90 * Age_years + 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)
```

We fit three different linear regression models (Box 4) to evaluate the effect of sodium intake on SBP: i) unadjusted model, ii) model adjusted for age, iii) model adjusted for age and the collider (proteinuria). The model specifications are shown here below; in Box 4 we show how to fit and visualize the corresponding models in R.

Models specification:

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

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

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

Box 4

```
library(broom) # load packages to visualize regression models output
library(visreg)
## 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")
```

We also fit three logistic regression models to evaluate the effect of sodium intake on hypertension defined as a binary outcome (SBP $\geq$140 mmHg = 1, SBP $<$140 mmHg = 0): i) an unadjusted model, ii) a model adjusted for age, and iii) a model adjusted for age and the collider (proteinuria). The model specifications are the same as described above but now with a binary outcome (hypertension); in Box 5 we show how to fit and visualize the corresponding models in R using a forest plot function.

Box 5

```
## Models fit on multiplicative scale
library(dplyr)
library(forestplot)
fit3 <- glm(hypertension ~ Sodium_gr, family=binomial(link='logit'), data=ObsData)
or <- round(exp(fit3$coef)[2],3) # conditional odds ratio from logistic model
ci95 <- exp(confint(fit3))[-1,]  # 95% CI of odds ratio

fit4 <- glm(hypertension ~ Sodium_gr + Age_years,
            family = binomial(link = "logit"), data = ObsData)
or <- round(exp(fit4$coef)[2],3)
ci95 <- exp(confint(fit4))[2,]

fit5 <- glm(hypertension ~ Sodium_gr + Age_years + Proteinuria_in_mg,
            family = binomial(link = "logit"), data = ObsData)
or <- round(exp(fit5$coef)[2],3)
ci95 <- exp(confint(fit5))[2,]

## Forest plot (see supplementary material for accessing the complete code)
fp <- rbind(result1, result2, result3); fp %>% or_graph()
```

3.2 Effect of conditioning on a collider

Table 2 shows the model coefficients and goodness of fit from the linear regression models, Figure 5 shows odds ratios from the logistic regression models. Figure 4 shows the regression line and 95% confidence

interval for the predicted level of SBP, illustrating the effect of conditioning on a collider. The adjusted regression line was derived as the predicted estimate of SBP, conditional on the median value of age for Figure 4B and age and proteinuria for Figure 4C [22]. As opposed to the unadjusted and bivariate models (Figures 4A and 4B), the collider model (Figure 4C) suggests a negative relationship between sodium intake and SBP (i.e., for one unit increase in sodium intake, the expected SBP decreases by 0.9 mmHg). The odds ratio for the effect of sodium on hypertension similarly suggests that it is protective (i.e., for one unit increase in sodium intake the risk of hypertension decreases by 98%) (Figure 5).

3.3 Monte-Carlo Simulation Results

Box 6 shows the code used to run the Monte-Carlo simulation on the additive scale. The true simulated causal effect of 24-hour sodium intake on SBP was 1.05 mmHg. The relative bias accrued from conditioning on proteinuria (the collider) in the regression model was 13.3%.

Box 6

```
# Monte Carlo 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"))
# Function to generate data
generateData <- function(n){
    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)
    Proteinuria_in_mg <- 0.90 * Age_years + 2.00 * sbp_in_mmHg + 2.80 * Sodium_gr + rnorm(n)
    data.frame(sbp_in_mmHg, Sodium_gr, Age_years, Proteinuria_in_mg)
}
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)
```

The code included in all of the boxes is provided in a supplementary file. We also provide the link to a web application http://watzilei.com/shiny/collider/ (Supplementary Figure 2) where users can dynamically modify the values of the beta coefficients in the data generation process of the collider model. The collider web application allows users to interactively modify the range of values of the slider input and

visualize the collider effect of the example. As shown in the web application the strength of the association of the collider with both the exposure and the outcome determines the strength of the paradoxical protective effect of 24-hour diary sodium intake in gr. on systolic blood pressure.

4 Conclusion

We investigated a situation where adding a certain type of variable to a linear regression model, called a "collider", led to bias with respect to the regression coefficient estimates while still improving the model fit. DAGs are based on subject matter knowledge and are vital for identifying colliders. Determining if a variable is a collider involves critical thinking about the true unobserved data generation process and the relationship between the variables for a given scenario [15, 23]. Then, the decision whether to include or exclude the variable in a regression model using observational data in epidemiology is based on whether the purpose of the study is prediction or explanation/causation. Under the structures we investigated here, adding a collider to a regression model is not advised when one is interested in the estimation of causal effects, as this may open a back-door path. However, if prediction is the purpose of the model, the inclusion of colliders in the models may be advisable if it reduces the models prediction error. Most research in epidemiology tries to explain how the world works (i.e., it is causal), thus to prevent paradoxical associations, epidemiologists estimating causal effects should be aware of such variables.

Competing Interests

The authors declare that they do not have any conflict of interest associated with this research and the content is solely the responsibility of the authors.

Funding

Miguel Angel Luque Fernandez is supported by the Spanish National Institute of Health, Carlos III Miguel Servet I Investigator Award (CP17/00206). Maria Jose Sanchez Perez is supported by the Andalusian Department of Health. Research, Development and Innovation Office project grant PI-0152/2017. 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.

Authors contributions

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.

5 Figures and Tables

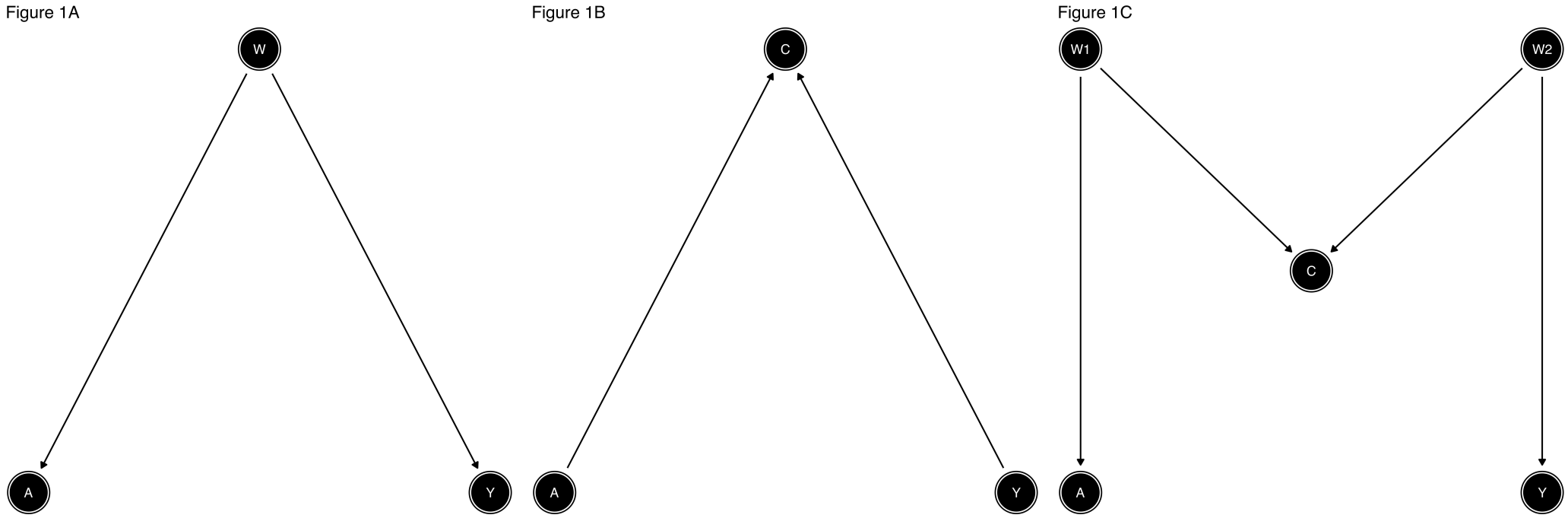


Figure 1: Basic structural associations between exposure and outcome: confounding (A), collider (B), and M-bias (C).

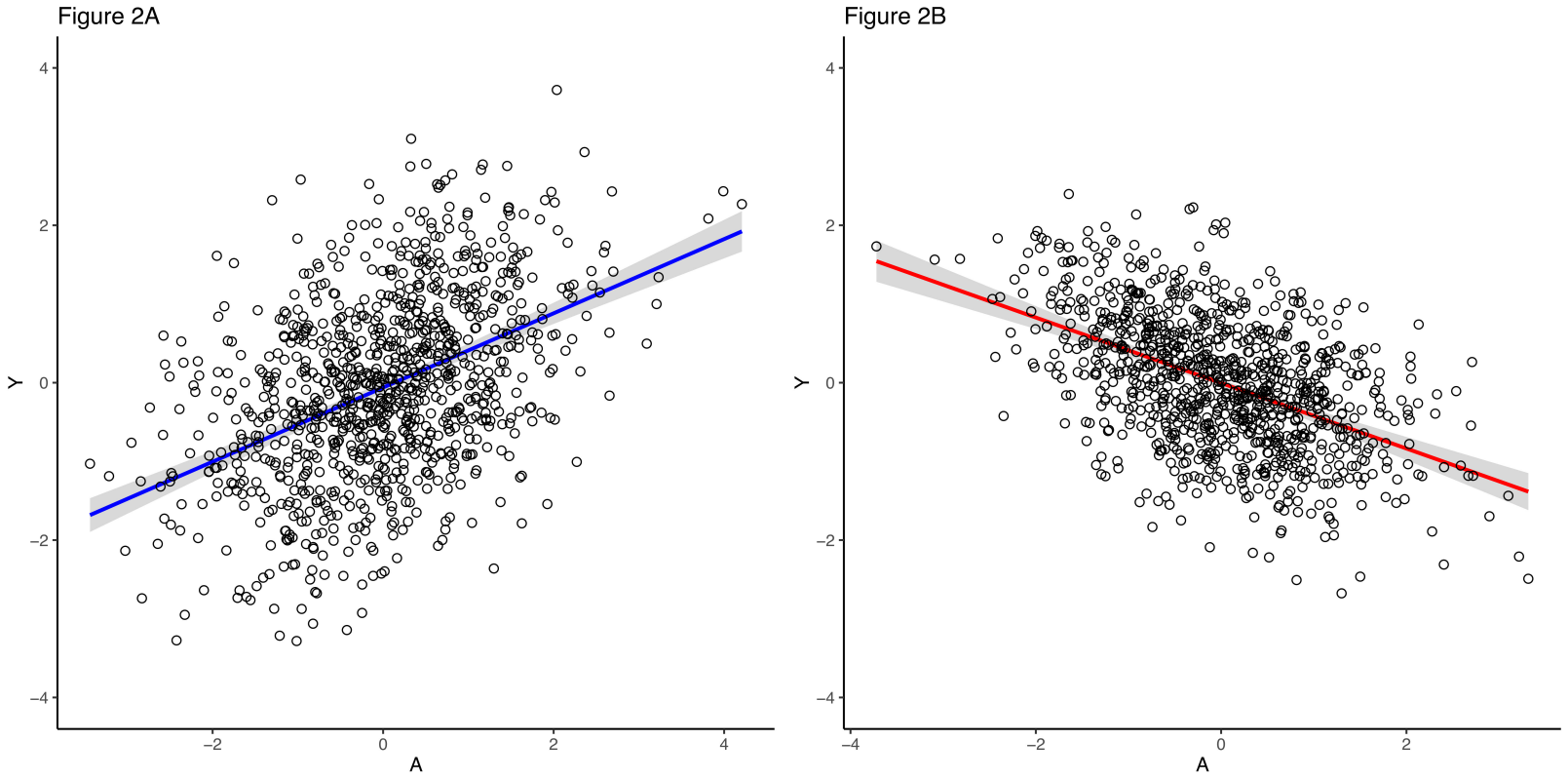


Figure 2: Visualization of the collider effect: Figure 2A model fit2 (Box 1) and Figure 2B model fit4 (Box 2).

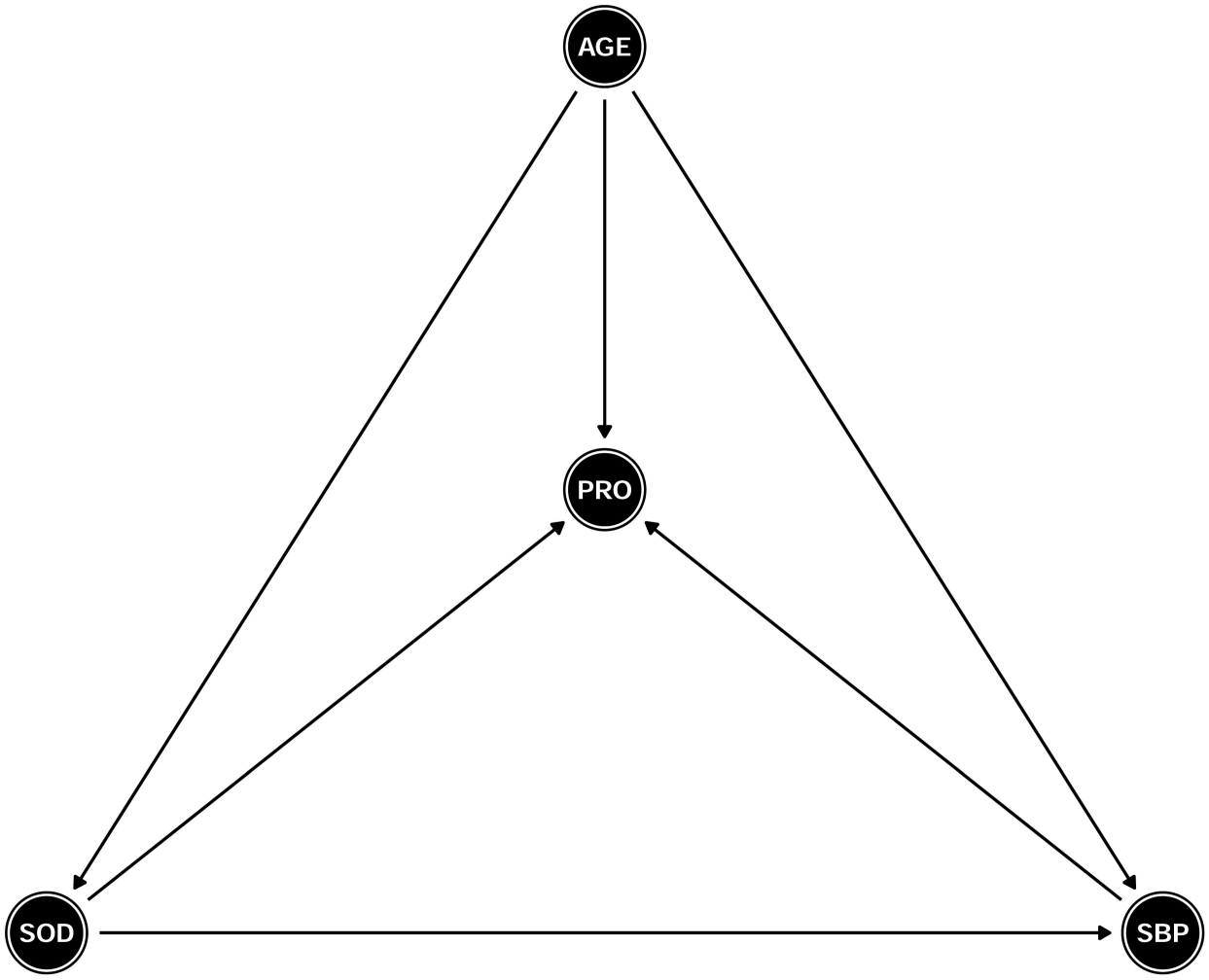


Figure 3: Directed acyclic graph depicting the structural causal relationship of the exposure and outcome, confounding and collider effects. Exposure: 24-hour sodium dietary intake in gr (SOD), outcome: systolic blood pressure in mmHg (SBP), confounder: age in years (AGE), collider: 24-hour urinary protein excretion, proteinuria (PRO).

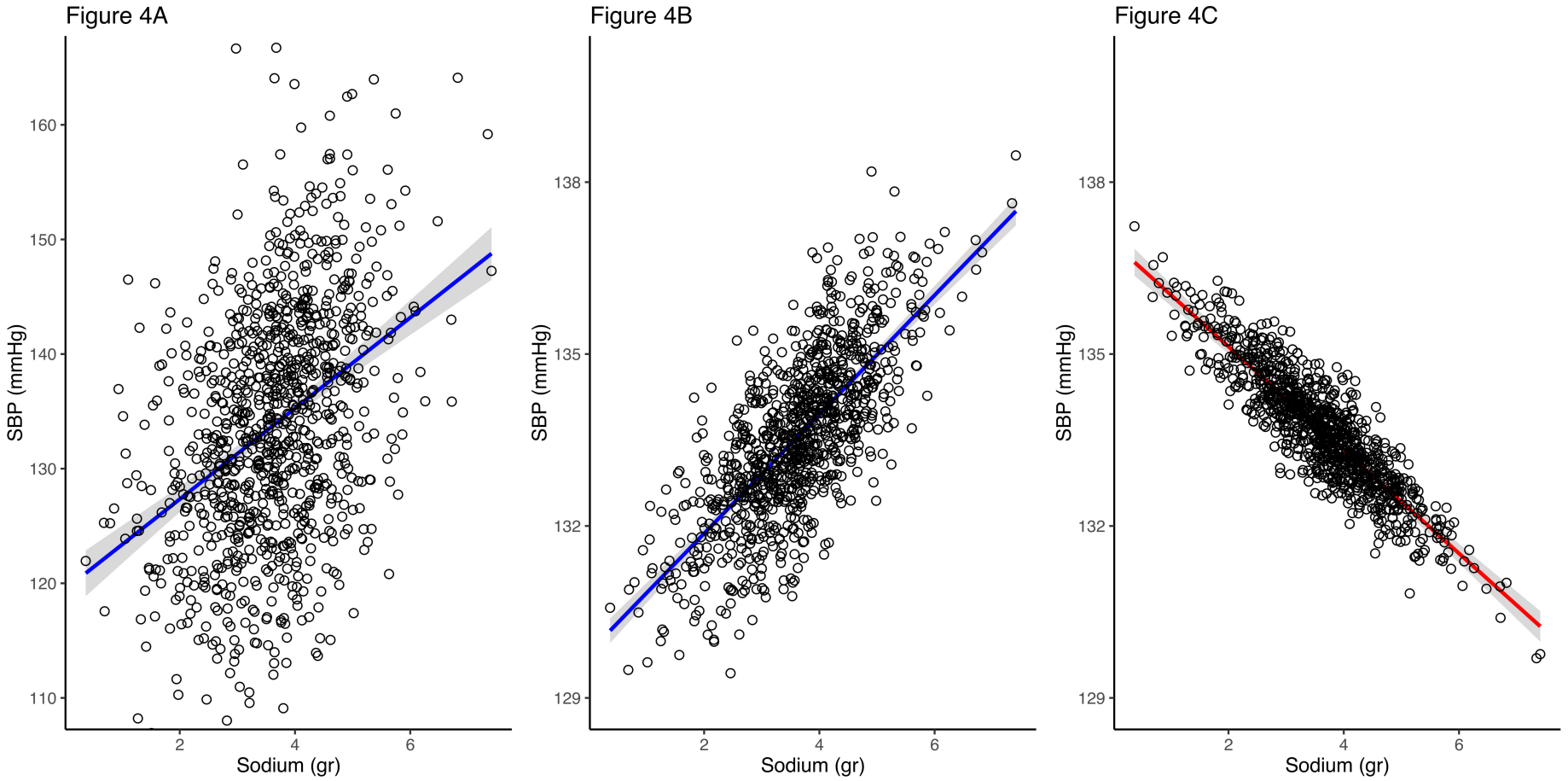


Figure 4: Collider effect for the illustration: Univariate (Figure 4A), bivariate (Figure 4B) and multivariate (Figure 4C) coefficients and standard errors for the linear association between systolic blood pressure and 24-hour sodium dietary intake adjusted for age acting as a confounder and proteinuria acting as a collider, n = 1,000.

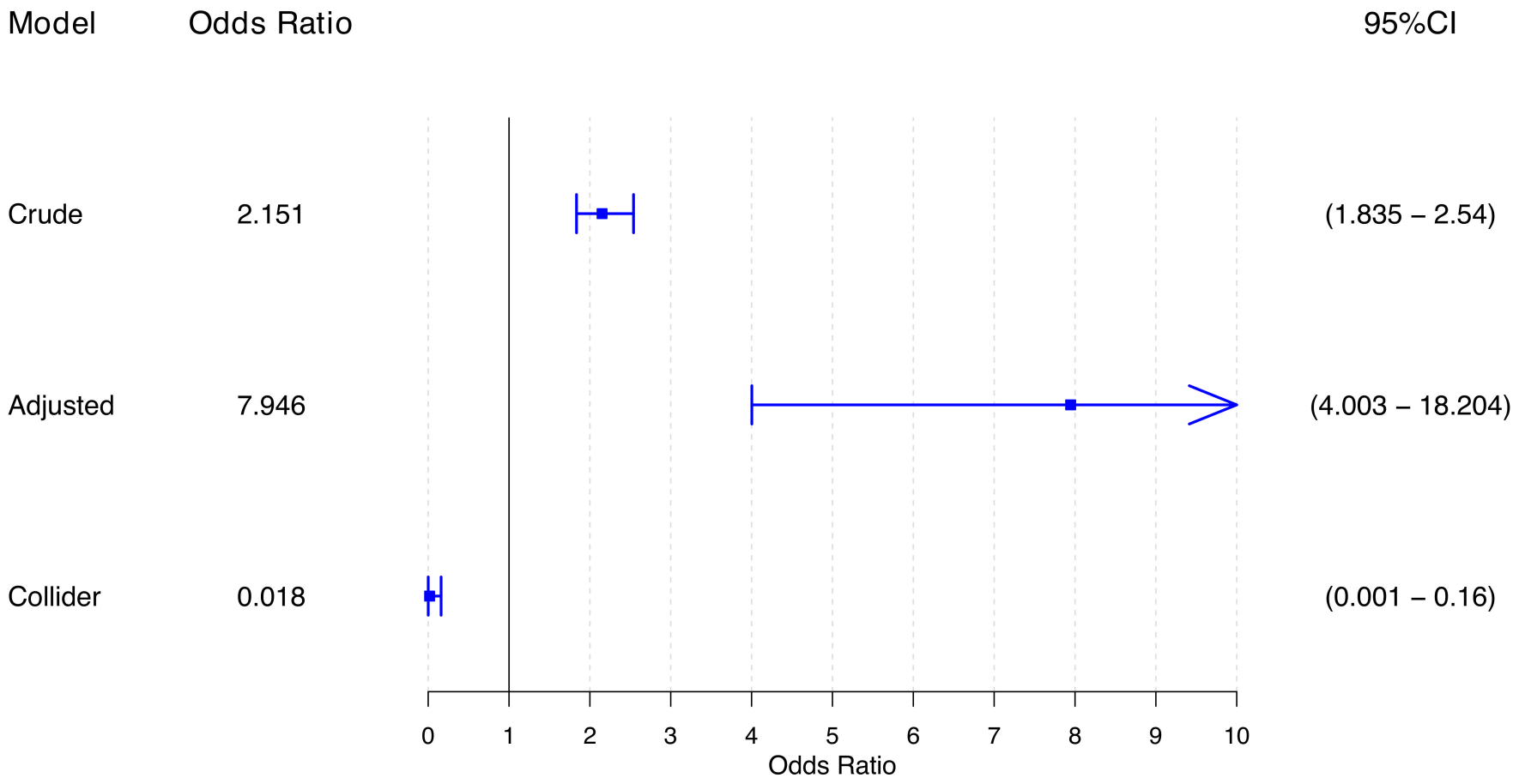
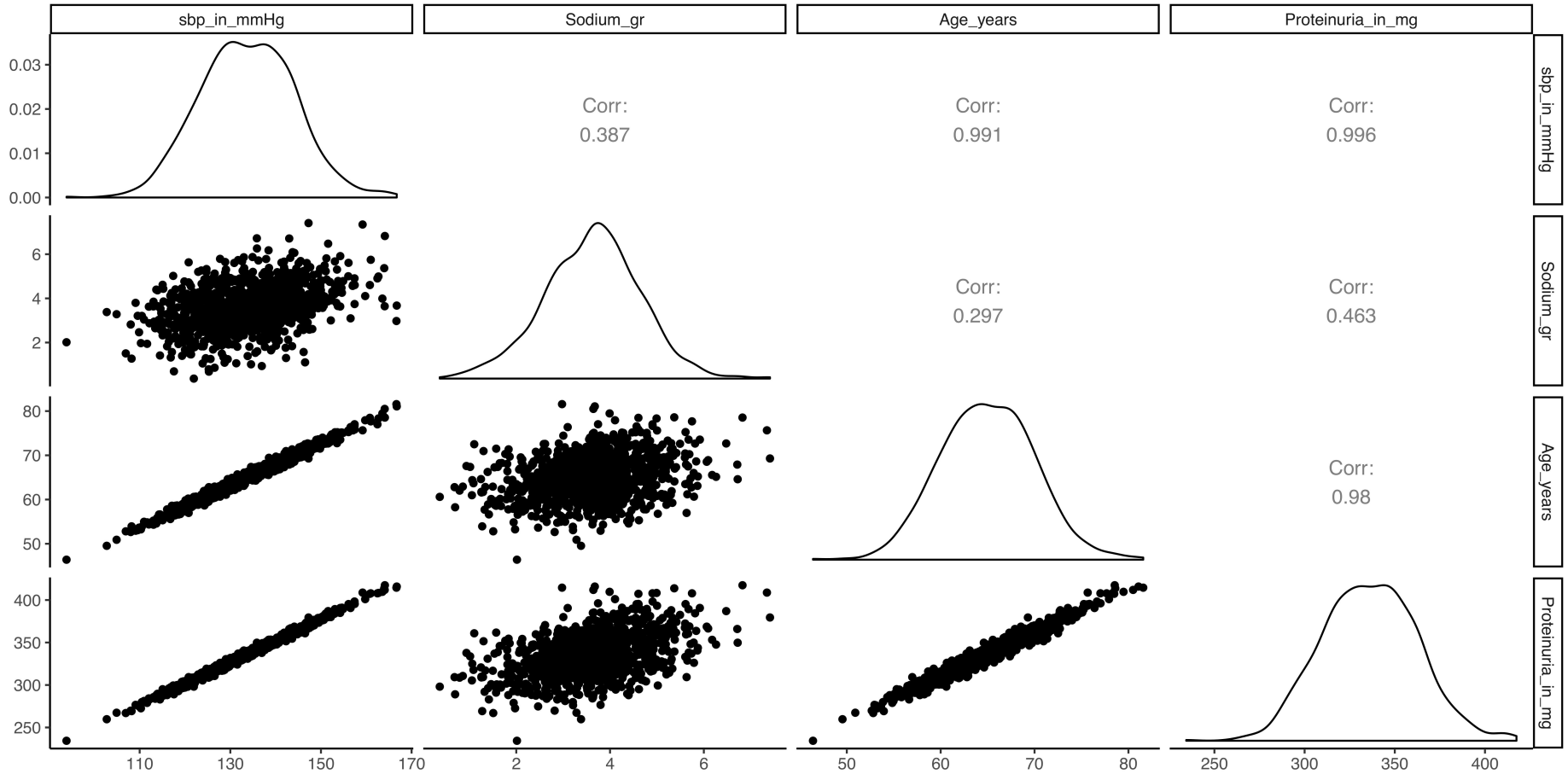
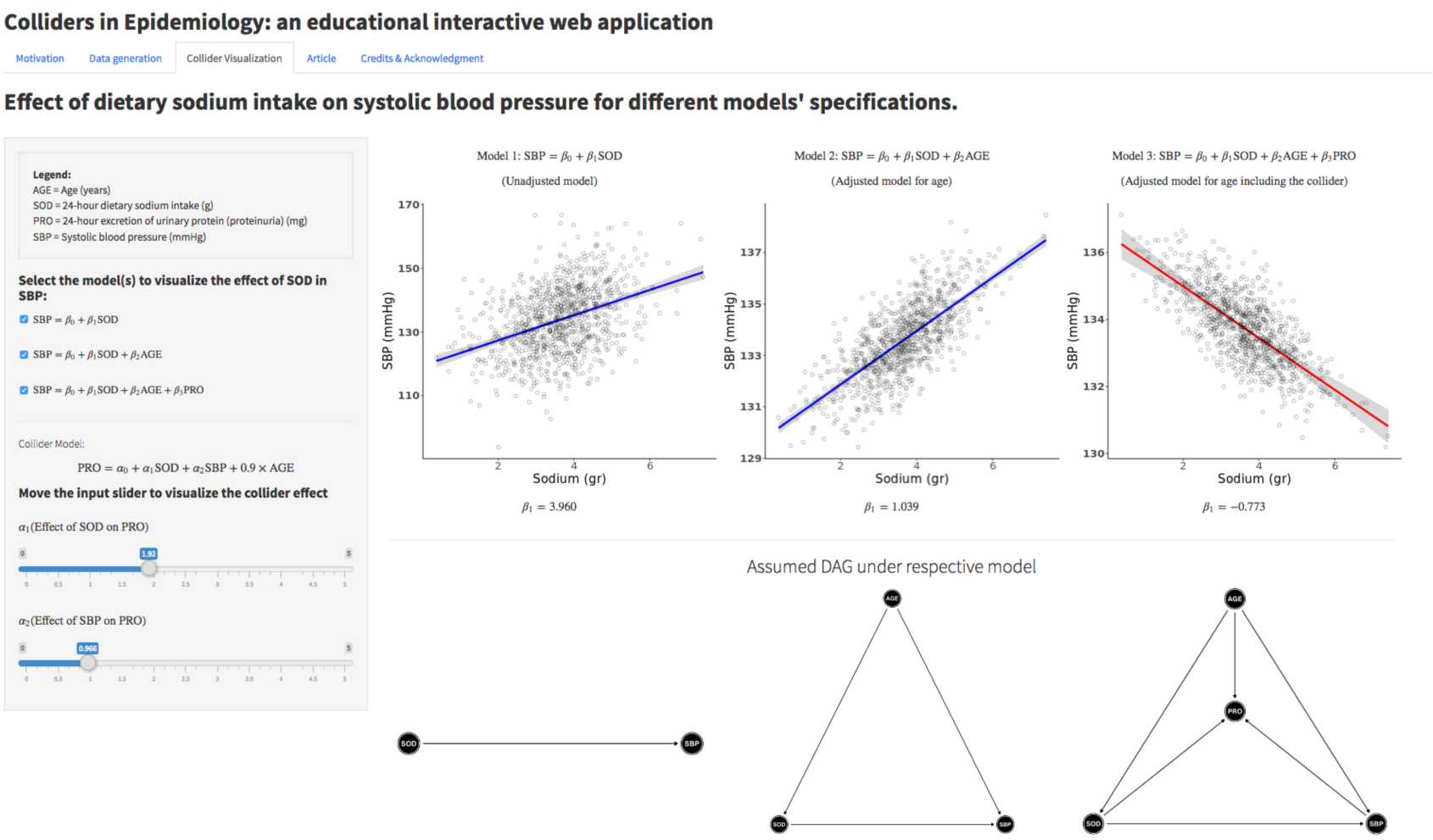


Figure 5: Collider effect for the illustration in a multiplicative scale for the effect of 24-hour sodium dietary intake on systolic blood pressure adjusted for age acting as a confounder and proteinuria acting as a collider, n = 1,000.



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



Supplementary Figure 2. Screenshot collider Shiny web application.

Table 1. Coefficients and standard errors of the linear association between Y (outcome) and A (exposure) illustrating confounding and collider effects, n = 1,000

	Dependent variable (Y)			
	W (confounder)		C (collider)	
	Unadjusted	Adjusted	Unadjusted	Adjusted
	Coefficient (Standard Error)	Coefficient (Standard Error)	Coefficient (Standard Error)	Coefficient (Standard Error)
	(Fit 1)	(Fit 2)	(Fit 3)	(Fit 4)
A	**0.471**	**0.289**	**0.326**	**-0.416**
	-0.03	-0.032	-0.031	-0.035
W		0.425	C	0.491
		-0.035		-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

Table 2. Univariate, bivariate and multivariate coefficients and standard errors for the linear 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: Systolic Blood Pressure in mmHg		
	Univariate Coefficient (Standard Error)	Bivariate Coefficient (Standard Error)	Multivariate collider Coefficient (Standard Error)
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	0.060
		(0.007)	(0.033)
Proteinuria in mg			0.396
			(0.007)
Intercept	119.420	-0.311	-0.091
	(1.122)	(0.407)	(0.192)
AIC	4523.571	-31.992	-1537.216

Note: Lower AIC is better

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

Systolic blood pressure in mmHg	Hypertension	Sodium in gr	Age in years	Proteinuria mg in 24h
Min. : 93.86	Min. : 0.00	Min. :0.37	Min. :46.40	Min. :234.40
1st Qu.:126.58	1st Qu.: 0.00	1st Qu.:2.95	1st Qu.:61.57	1st Qu.:317.40
Median :133.85	Median : 0.00	Median :3.66	Median :64.91	Median :336.00
Mean :133.76	Mean : 0.28	Mean :3.62	Mean :65.01	Mean :336.20
3rd Qu.:141.03	3rd Qu.: 1.00	3rd Qu.:4.26	3rd Qu.:68.35	3rd Qu.:354.30
Max. :166.73	Max. : 1.00	Max. :7.41	Max. :81.58	Max. :417.50

Hypertension: systolic blood pressure ≥140 mmHg

References

[1] Greenland S, Morgenstern H. Confounding in Health Research. Annual Review of Public Health. 2001 May;22(1):189–212. Available from: http://dx.doi.org/10.1146/annurev.publhealth.22.1.189.

[2] Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. International Journal of Epidemiology. 2009 Nov;39(2):417–420. Available from: http://dx.doi.org/10.1093/ije/dyp334.

[3] Vanderweele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. Statistics and Its Interface. 2009;2(4):457–468. Available from: http://dx.doi.org/10.4310/SII.2009.v2.n4.a7.

[4] Hernán MA, Hernández-Díaz S, Robins JM. A Structural Approach to Selection Bias. Epidemiology. 2004 Sep;15(5):615–625. Available from: http://dx.doi.org/10.1097/01.ede.0000135174.63482.43.

[5] Robins JM, Hernán MÁ, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology. 2000 Sep;11(5):550–560. Available from: http://dx.doi.org/10.1097/00001648-200009000-00011.

[6] Rohrer JM. Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. Advances in Methods and Practices in Psychological Science. 2017;.

[7] Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669–688. Available from: http://dx.doi.org/10.1093/biomet/82.4.669.

[8] Luque-Fernandez MA, Zoega H, Valdimarsdottir U, Williams MA. Deconstructing the smoking-preeclampsia paradox through a counterfactual framework. European Journal of Epidemiology. 2016 Jun;31(6):613–623. Available from: https://doi.org/10.1007/s10654-016-0139-5.

[9] Hernandez-Diaz S, Schisterman EF, Hernan MA. The Birth Weight Paradox Uncovered? American Journal of Epidemiology. 2006 Sep;164(11):1115–1120. Available from: http://dx.doi.org/10.1093/aje/kwj275.

[10] Banack HR, Kaufman JS. The "Obesity Paradox" Explained. Epidemiology. 2013 may;24(3):461–462.

[11] Whitcomb BW, Schisterman EF, Perkins NJ, Platt RW. Quantification of collider-stratification bias and the birthweight paradox. Paediatric and Perinatal Epidemiology. 2009 sep;23(5):394–402.

[12] Rubin DB. Causal Inference Using Potential Outcomes. Journal of the American Statistical Association. 2005 mar;100(469):322–331.

[13] Hernan MA. Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology. American Journal of Epidemiology. 2002 Jan;155(2):176–184. Available from: http://dx.doi.org/10.1093/aje/155.2.176.

[14] Greenland S, Pearl J, Robins JM. Causal Diagrams for Epidemiologic Research. Epidemiology. 1999 Jan;10(1):37–48. Available from: http://dx.doi.org/10.1097/00001648-199901000-00008.

[15] Pearce N, Richiardi L. Commentary: three worlds collide: Berkson's bias, selection bias and collider bias. International journal of epidemiology. 2014;43(2):521–524.

[16] Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, et al. Heart disease and stroke statistics-2017 update: a report from the American Heart Association. Circulation. 2017;135(10):e146–e603.

[17] Gu Q, Burt VL, Paulose-Ram R, Yoon S, Gillum RF. 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. 2008;18(4):302–309.

[18] Sacks FM, Svetkey LP, Vollmer WM, Appel LJ, Bray GA, Harsha D, et al. Effects on blood pressure of reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. New England journal of medicine. 2001;344(1):3–10.

[19] Tareen N, Martins D, Nagami G, Levine B, Norris KC. Sodium disorders in the elderly. Journal of the National Medical Association. 2005;.

[20] Van Horn L, Carson JAS, Appel LJ, Burke LE, Economos C, Karmally W, et al. 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. 2016 Nov;134(22):e505–e529.

[21] Carroll MF. Proteinuria in Adults: A Diagnostic Approach. American family physician. 2000;62(6).

[22] Breheny P, Burchett W. Visualization of Regression Models Using visreg. The R Journal. 2017;9(2):56–71. Available from: https://journal.r-project.org/archive/2017/RJ-2017-046/index.html.

[23] Pearce N, Lawlor DA. Causal inference—so much more than statistics. International journal of epidemiology. 2016;45(6):1895–1903.