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

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


**Published:** 14 December 2018  **Article history**
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.
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.
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 $A \rightarrow C \leftarrow Y$).
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
Colliders

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

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

<table>
<thead>
<tr>
<th></th>
<th>W (confounder)</th>
<th></th>
<th>C (collider)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unadjusted β (SE)</td>
<td>Adjusted β (SE)</td>
<td>Unadjusted β (SE)</td>
</tr>
<tr>
<td></td>
<td>(Fit 1)</td>
<td>(Fit 2)</td>
<td>(Fit 3)</td>
</tr>
<tr>
<td>A</td>
<td>0.471 (-0.030)</td>
<td>0.289 (-0.032)</td>
<td>A</td>
</tr>
<tr>
<td>W</td>
<td>0.425 (-0.035)</td>
<td>0.425</td>
<td>C</td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.061 (-0.033)</td>
<td>-0.06 (-0.031)</td>
<td>0.01</td>
</tr>
<tr>
<td>AIC</td>
<td>100.42</td>
<td>-31.992</td>
<td>-55.369</td>
</tr>
</tbody>
</table>

Note: Lower AIC is better

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

Collider Effect

Figure 2A

Figure 2B

Collider Effect

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

Model 1: $SBP = \beta_0 + \beta_1 \text{SOD}$
(Adjusted model for age including the collider)

Model 2: $SBP = \beta_0 + \beta_1 \text{SOD} + \beta_2 \text{AGE}$
(Adjusted model for age including the collider)

Model 3: $SBP = \beta_0 + \beta_1 \text{SOD} + \beta_2 \text{AGE} + \beta_3 \text{PRO}$
(Adjusted model for age including the collider)

Legend:
- AGE = Age (years)
- SOD = 24-hour dietary sodium intake (mg)
- PRO = 24-hour excretion of urinary protein (proteimuria) (mg)
- SBP = Systolic blood pressure (mmHg)

Select the model(s) to visualize the effect of SOD in SBP:
- $SBP = \beta_0 + \beta_1 \text{SOD}$
- $SBP = \beta_0 + \beta_1 \text{SOD} + \beta_2 \text{AGE}$
- $SBP = \beta_0 + \beta_1 \text{SOD} + \beta_2 \text{AGE} + \beta_3 \text{PRO}$

Assumed DAG under respective model
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).

### Data Generation

```r
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

R<-1000
ttrue <- 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, ","))
  ObsData <- generateData(n=10000)
  # True effect
ttrue[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(ttrue)
# 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.

### Unadjusted model

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

### Adjusted model (confounder)

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)

SBP = $\beta_0 + \beta_1 \times \text{Sodium} + \beta_2 \times \text{Age} + \beta_3 \times \text{Proteinuria} + \varepsilon$
## Collider and confounding effects

<table>
<thead>
<tr>
<th>Dependent variable: SBP in mmHg</th>
<th>Univariate (SE)</th>
<th>Bivariate (SE)</th>
<th>Multivariate (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sodium in gr</td>
<td>3.960 (0.298)</td>
<td>1.039 (0.032)</td>
<td>-0.902 (0.036)</td>
</tr>
<tr>
<td>Age in years</td>
<td>2.004 (0.007)</td>
<td>0.416 (0.027)</td>
<td></td>
</tr>
<tr>
<td>Proteinuria in mg</td>
<td></td>
<td>0.396 (0.007)</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>119.420 (1.122)</td>
<td>-0.311 (0.407)</td>
<td>-0.091 (0.192)</td>
</tr>
<tr>
<td>AIC</td>
<td>7363.45</td>
<td>2807.89</td>
<td>1302.66</td>
</tr>
</tbody>
</table>

**Note:** Lower AIC is better

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
¡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
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 $i$ $Y_i(1)$ and $Y_i(0)$ set to $A = a Y^{(a)}$, namely $A = 1$ and $A = 0$.

**Causal Effects**
- Average Treatment Effect: $E[Y(1) - Y(0)]$. 
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 \( Y_i(1) \) would have been non-treated \( Y_i(A = 0) = Y_i(0) \).

Controls \( Y_i(0) \) would have been treated \( Y_i(A = 1) = Y_i(1) \).

Identifiability

- How we can identify the effect of the potential outcomes \( Y^a \) 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**

\[(Y_i(1), Y_i(0)) \perp A_i \mid W_i\]

**POSITIVITY**

**(POSITIVITY):** \(P(A = a \mid W) > 0\) for all \(a, W\)

**SUTVA**

- We have assumed that there is only one version of the treatment (consistency) \(Y(1)\) if \(A = 1\) and \(Y(0)\) if \(A = 0\).
- 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**.
The **G-Formula** for the **identification** of the ATE with observational data is given by:

\[
E(Y^a) = \sum_y E(Y^a \mid W = w)P(W = w)
\]

\[
= \sum_y E(Y^a \mid A = a, W = w)P(W = w) \text{ by consistency}
\]

\[
= \sum_y E(Y = y \mid A = a, W = w)P(W = w) \text{ by ignorability}
\]

The **ATE** is defined as:

\[
\sum_w \left[ \sum_y P(Y = y \mid A = 1, W = w) - \sum_y P(Y = y \mid A = 0, W = w) \right] P(W = w)
\]

\[
P(W = w) = \sum_{y,a} P(W = w, A = a, Y = y)
\]
G-Formula, (Robins, 1986)

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

The $ATE = \sum_w \left[ \sum_y P(Y = y | A = 1, W = w) - \sum_y P(Y = y | A = 0, W = w) \right] P(W = w)$

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

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

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

\[ m_A(w_i) = E(Y_i | A_i = A, W_i) \]

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