Comparing covariate adjustment in interventional and observational studies

Markus Kalisch, Seminar für Statistik, ETH Zürich

What is the total causal effect ?

Treatment
X

→

Outcome
Y

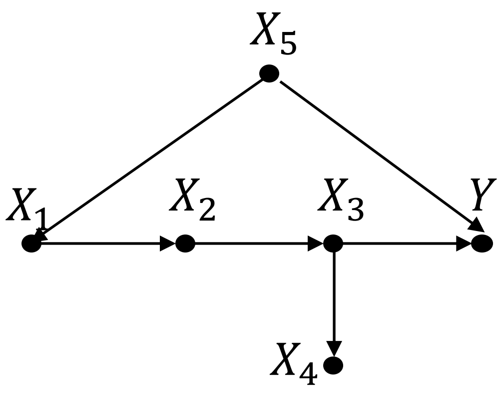
- If we apply treatment X, how will outcome Y change ?
- Data collection:
  - observational study
  - interventional study (RCT)

Outline for the rest of the talk

- Total causal effect and covariate adjustment
- Issues in observational studies
- Issues in interventional studies
- Insights from recent theoretical developments

Causal Model: How the real world might look like

- We use directed acyclic graphs (DAG) – no feedback loops
- Example: DAG G



- Terminology:
  Set of all variables: $\boldsymbol{X} = \{X_1, X_2, \dots, X_5, Y\}$
  Path: (X_1, X_2, X_3, Y)
  Directed path = "causal-path": (X_1, X_2, X_3)
  Not directed path = Non-causal path: (X_4, X_3, Y)
  Parents $\mathrm{pa}(X_3) = \{X_2, X_4\}$, Children $ch(X_1) = \{X_2\}$
  Ancestor an, Descendant de, Non-descendants nd

Think of family tree

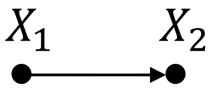
More details: Structural Equation Model (SEM)

- Example of SEM:

$$X_1 = N_1$$
$$X_2 = 4X_1 + N_2$$
$$N_1, N_2 \sim N(0,1) \;\; iid$$

Causal interpretation

- Visualization of **causal structure**:



- Difference to arbitrary hierarchical system of equations: Due to causal interpretation, solving for a variable on the RHS is not meaningful in SEM.

Quantifying the total causal effect

Define intervention distribution by replacing (some) structural equations

- do-Operator
  Reference: Pearl, J. (2009). Causality: Models, Reasoning and Inference. 2nd edition. Cambridge Univ. Press.

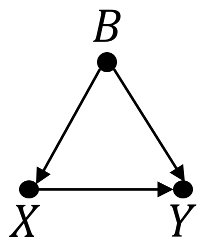
E.g. «intervention on X»:

- Old SEM: S with equation $X = 2 + X_5 + N_X$
- New SEM: $\hat{S}$ with equation $X = 4$
- New SEM generates new distribution:
  $P_{\hat{S}}(\boldsymbol{X}) = P_S(\boldsymbol{X}|do(X = 4))$ and in particular $\mathrm{P}(Y|do(X = 4))$

- Final goal: Estimate intervention distribution given observational data
- Oftentimes: Expectation is enough – e.g. $E(Y|do(X = 4))$

Covariate adjustment: Adjustment set

- Idea: Identify intervention effects by only using conditional probabilities / expectations



«do» No «do»

$$P(Y = y|do(X = x)) = \sum_{b \in B} P(Y = y|X = x, B = b)P(B = b)$$

Adjustment set

- Practice: Often interested in $E(Y = y|do(X = x))$
- Can show for multivariate Gaussian density:

$$E(Y|do(X = x)) = \alpha + \gamma x + \beta^T E(B)$$

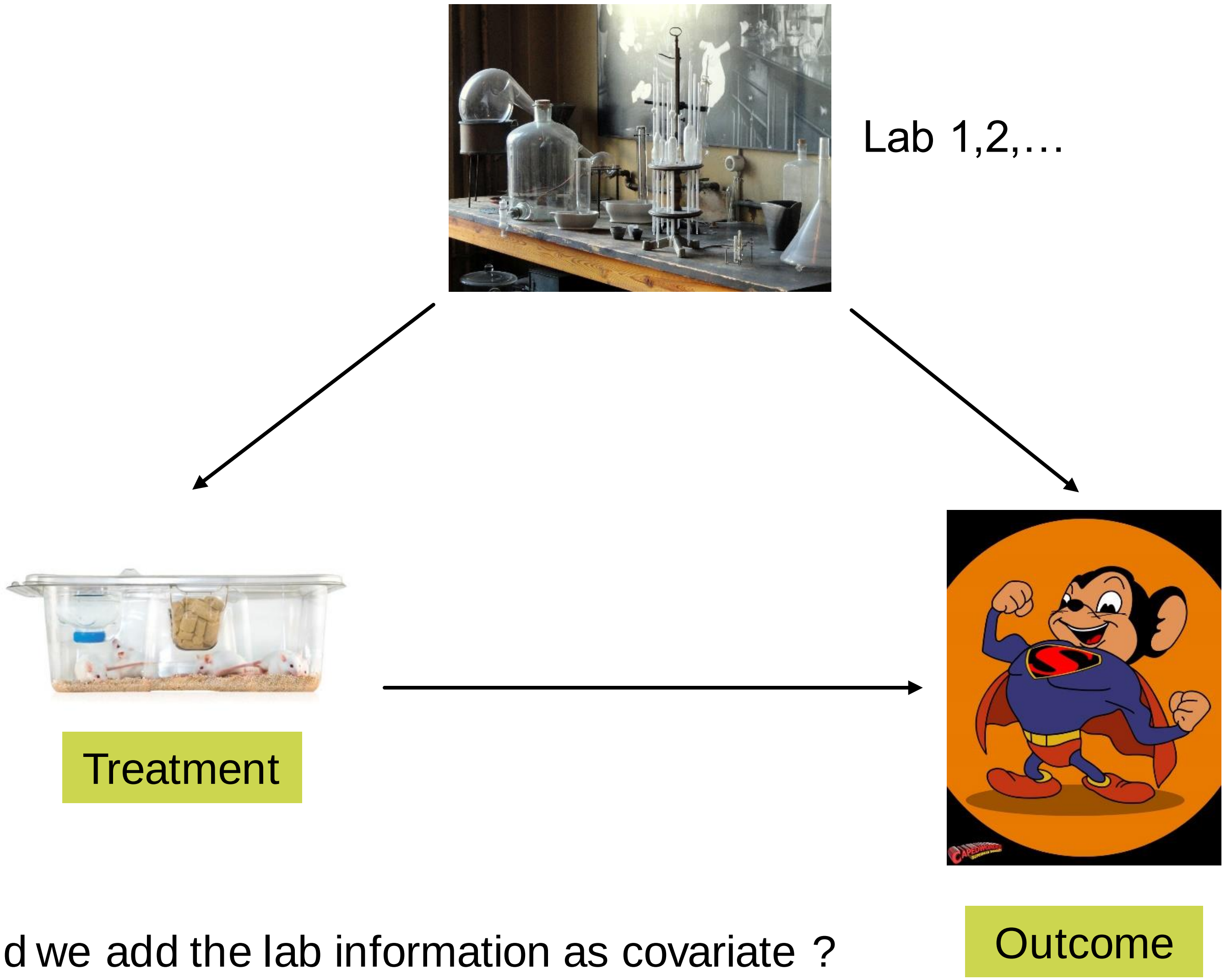
- Total Causal Effect: $\frac{\mathrm{d}}{dx} E(Y|do(X = x)) = \gamma$

  This is the regression coefficient of X in the regression of Y on X and B

Outline for the rest of the talk

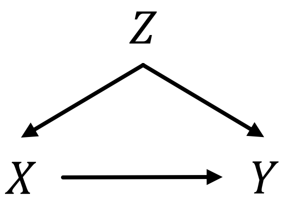
- Total causal effect and covariate adjustment
- Issues in observational studies
- Issues in interventional studies
- Insights from recent theoretical developments

Causal Diagram: Example 1 - confounder



Should we add the lab information as covariate ?

Example 1 in numbers



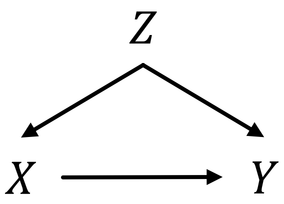
- $\varepsilon_X \sim N(0,1)$, $\varepsilon_Z \sim N(0,1)$, $\varepsilon_Y \sim N(0,1)$ independent
- True causal system:

$$Z = \varepsilon_Z$$
$$X = 0.7 * Z + \varepsilon_X$$
$$Y = 1 * X + 0.5 * Z + \varepsilon_Y$$

```
set.seed(123)
n <- 1000
z <- rnorm(n)
x <- 0.7*z + rnorm(n)
y <- 1*x + 0.5*z + rnorm(n)
```

- True causal effect of X on Y: 1
  If we increase X by one unit, Y will also increase by one unit
- Can we estimate the true causal effect with a linear regression ?

Example 1 in numbers



- **True causal effect** of X on Y: 1
- Simple Regression: $lm(Y \sim X)$

```
> confint(lm(y~x))
                  2.5 %     97.5 %
(Intercept) -0.09005941 0.03942158
x            1.19606286 1.29767266
```

Incorrect

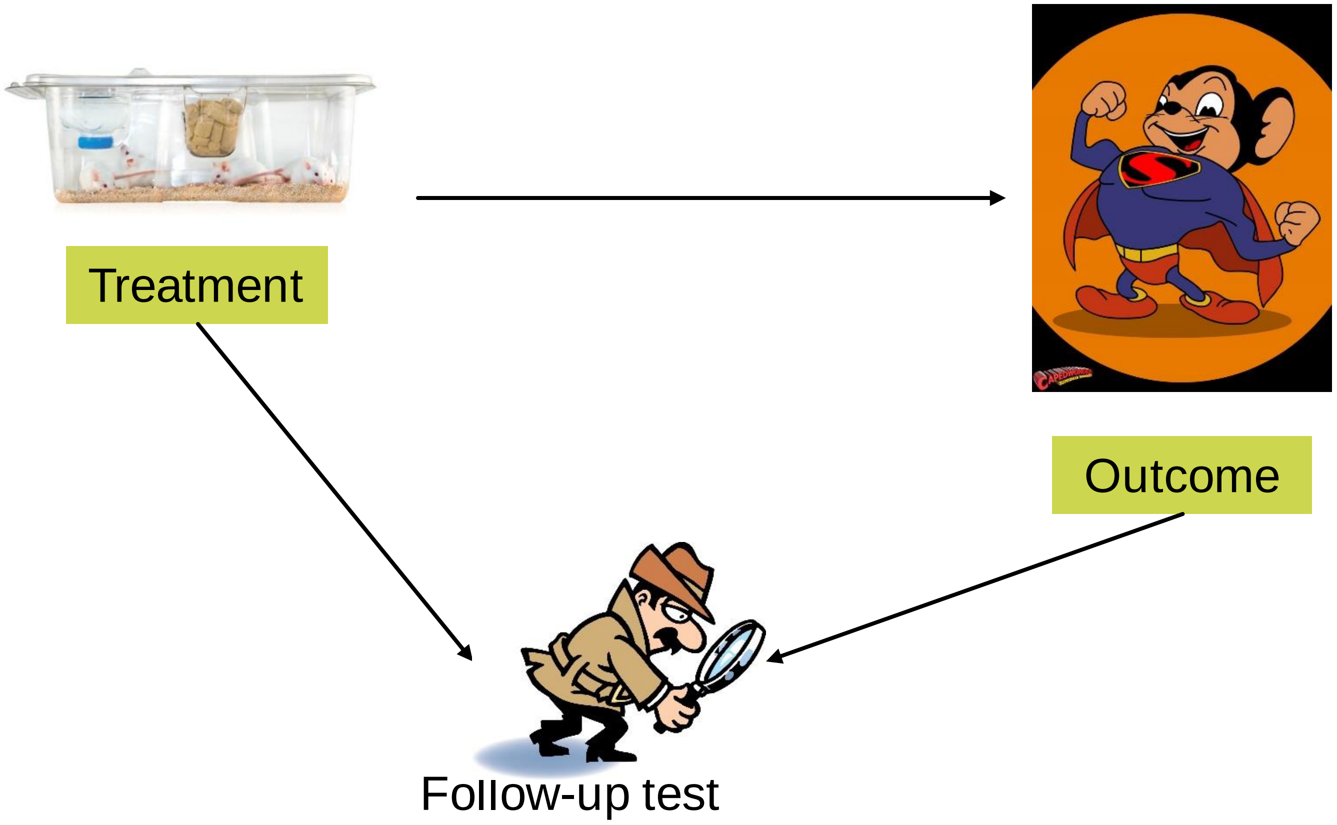
Missing the confounder introduced a **bias**!

- Multiple Regression: $lm(Y \sim X + Z)$

```
> confint(lm(y~x+z))
                  2.5 %     97.5 %
(Intercept) -0.08172964 0.03986164
x            0.96709528 1.08791825
z            0.38165727 0.53685033
```

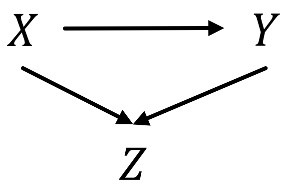
Correct

Causal Diagram: Example 2 – selection variable



Should we add the info of the follow-up test as covariate ?

Example 2 in numbers



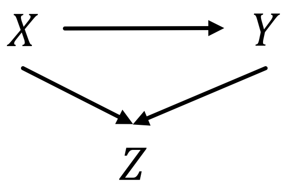
- $\varepsilon_X \sim N(0,1)$, $\varepsilon_Z \sim N(0,1)$, $\varepsilon_Y \sim N(0,1)$ independent
- True causal system:

$$\mathrm{X} = \varepsilon_X$$
$$\mathrm{Y} = 0.7 * X + \varepsilon_Y$$
$$\mathrm{Z} = 0.8 * X + 0.5 * Y + \varepsilon_Z$$

```
set.seed(124)
n <- 1000
x <- rnorm(n)
y <- 0.7*x + rnorm(n)
z <- 0.8*x + 0.5*y + rnorm(n)
```

- True causal effect of X on Y: 0.7
  If we increase X by one unit, Y will also increase by 0.7 units
- Can we estimate the true causal effect with a linear regression ?

Example 2 in numbers



- **True causal effect** of X on Y: 0.7
- Simple Regression: $lm(Y \sim X)$

```
                  2.5 %     97.5 %
(Intercept) -0.06766214 0.06193577
x            0.61398524 0.74623873
```

Correct

- Multiple Regression: $lm(Y{\sim}X + Z)$

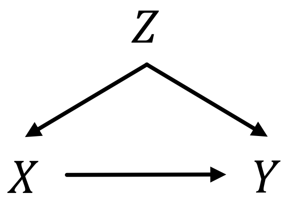
```
                  2.5 %     97.5 %
(Intercept) -0.06560545 0.05044087
x            0.13182538 0.29761022
z            0.35627606 0.45774568
```

Incorrect

Including the selection variable introduced a **bias**!

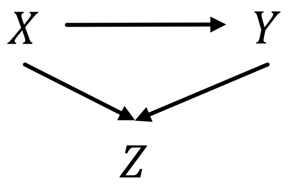
“Parent Criterion” (PC)

- Take parents of X as adjustment set (special case of Pearl’s back-door criterion)
- Sufficient but not complete
- **Example 1:**



PC: Z is a valid adjustment set; would $\{\}$ be a valid adjustment set, too → ???
(perhaps we can not measure Z although we know it exists)

- **Example 2:**



PC: $\{\}$ is a valid adjustment set; would Z be a valid adjustment set, too → ???

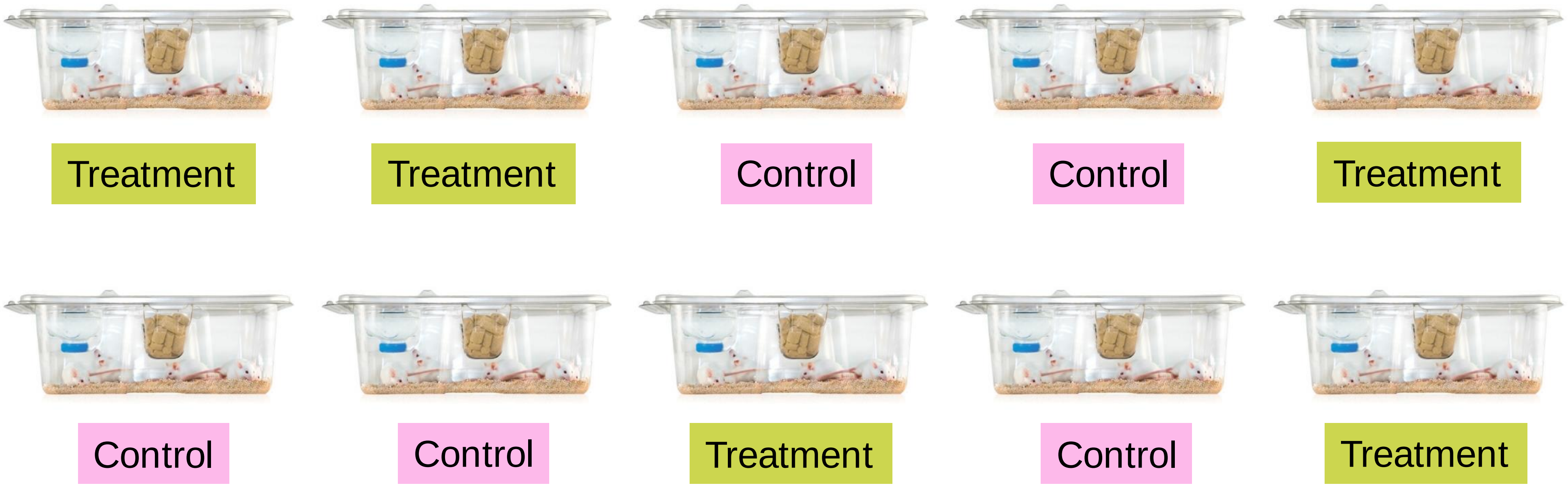
Conclusion 1

In observational studies: Judging if an adjustment set is valid is not trivial

Outline for the rest of the talk

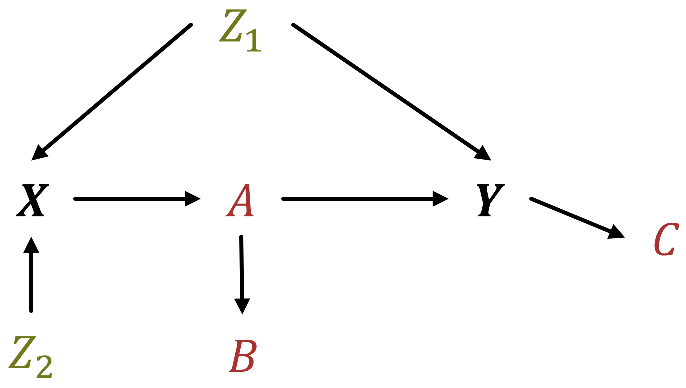
- Total causal effect and covariate adjustment
- Issues in observational studies
- Issues in interventional studies
- Insights from recent theoretical developments

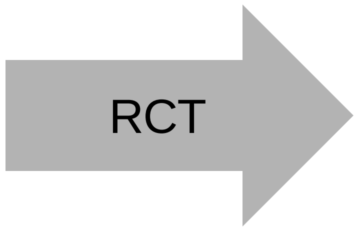
RCT: Evaluation

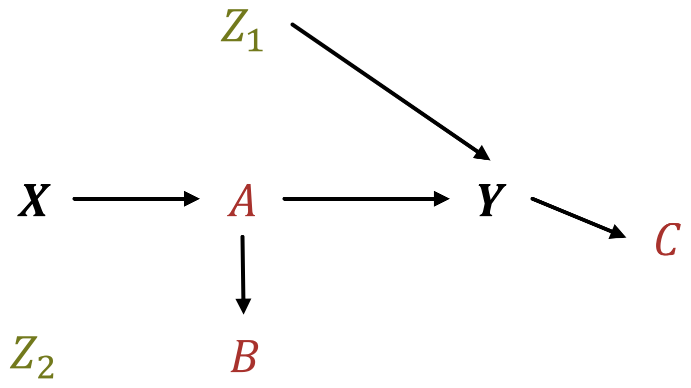


- Cage: Experimental Unit
- 5 cages with treatment ($X = 1$), 5 cages with control ($X = 0$)
- Randomize allocation: In causal diagram think of “deleting all incoming edges to X”

RCT in causal diagram





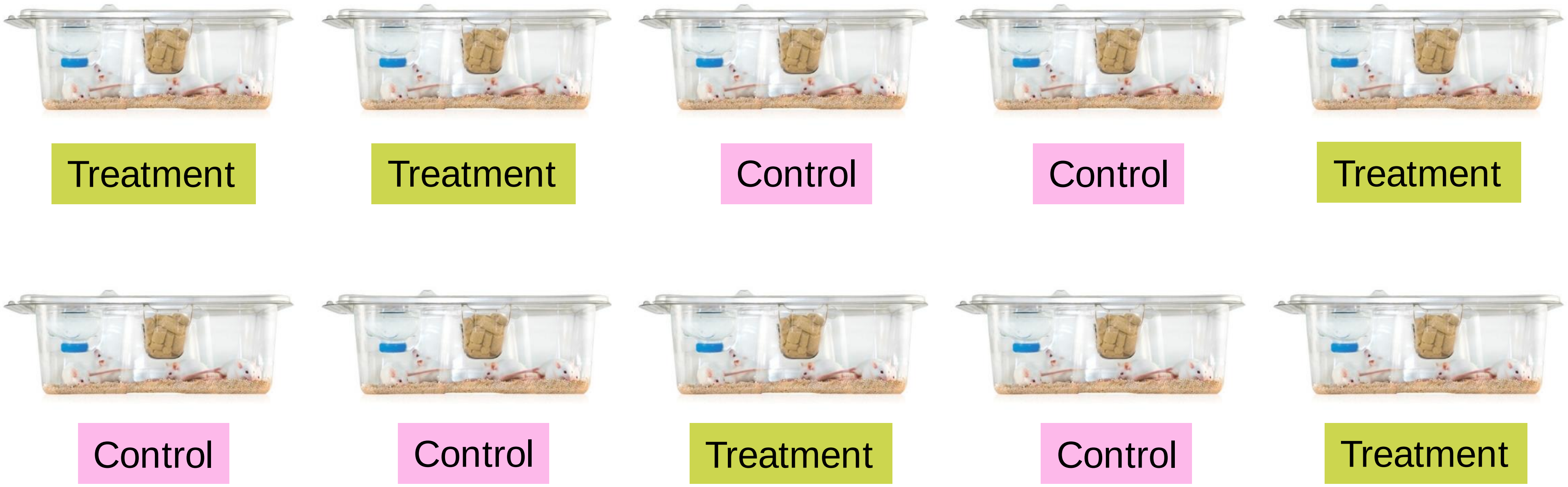


PC:
Valid adjustement set
is $\{Z_1, Z_2\}$

PC:
Valid adjustment set
is $\{\}$

PC: $\{\}$ is always valid adjustment set after randomization

RCT: Evaluation



- Given a proper design, we can do a **two-sample t-test** with two groups (i.e. empty adjustment set).
- What if we have **more covariates** (sex, age, intermediate blood test, follow-up information, …) ?
- Is it always **better to add covariates** to the analysis ?

Messing up the evaluation of a randomized controlled trial (RCT)

- You can bias (“mess up”), the analysis by adding the “wrong” covariates.
- **RCT**: It is always **safe** to add **no covariates** to the analysis.
- Adding the “right” covariates might increase precision.

Causal Diagram: Example 1

Treatment

Intermediate
Blood Test

Outcome

Should we add the intermediate blood test as covariate ?

Example 1 in numbers

$$X \longrightarrow Z \longrightarrow Y$$

- $\varepsilon_X \sim N(0,1)$, $\varepsilon_Z \sim N(0,1)$, $\varepsilon_Y \sim N(0,1)$ independent
- True causal system:

$$X = \varepsilon_X$$
$$Z = 2 * X + \varepsilon_Z$$
$$Y = 0.5 * Z + \varepsilon_Y$$

```
set.seed(123)
n <- 1000
x <- rnorm(n)
z <- 2*x + rnorm(n)
y <- 0.5*z + rnorm(n)
```

- True causal effect of X on Y: $2 * 0.5 = 1$
  If we increase X by one unit, Y will also increase by one unit
- Can we estimate the true causal effect with a linear regression ?

Example 1 in numbers

$$X \longrightarrow Z \longrightarrow Y$$

- **True causal effect** of X on Y: $2 * 0.5 = 1$
- Simple Regression: $lm(Y \sim X)$

```
> confint(lm(y~x))
                  2.5 %     97.5 %
(Intercept) -0.06836077 0.06979605
x            0.95527153 1.09463662
```

Correct

- Multiple Regression: $lm(Y \sim X + Z)$

```
> confint(lm(y~x+z))
                  2.5 %     97.5 %
(Intercept) -0.08172964 0.03986164
x           -0.21674264 0.06373265
z            0.46709528 0.58791825
```

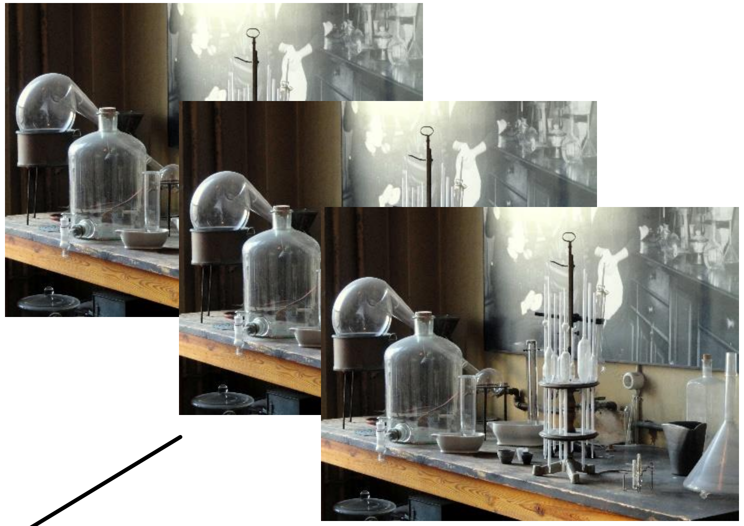
Incorrect

Adding a covariate introduced a **bias**!

Causal Diagram: Example 2

Treatment

Outcome



Lab 1,2,…

Should we add the lab information as covariate ?

Example 2 in numbers

$$X \longrightarrow Y \longleftarrow Z$$

- $\varepsilon_X \sim N(0,1)$, $\varepsilon_Z \sim N(0,1)$, $\varepsilon_Y \sim N(0,1)$ independent
- True causal system:

$$X = \varepsilon_X$$
$$Z = \varepsilon_Z$$
$$Y = 1 * X + 0.5 * Z + \varepsilon_Y$$

```
set.seed(123)
n <- 1000
x <- rnorm(n)
z <- rnorm(n)
y <- 1*x + 0.5*z + rnorm(n)
```

- True causal effect of X on Y: 1
  If we increase X by one unit, Y will also increase by one unit
- Can we estimate the true causal effect with a linear regression ?

Example 2 in numbers

$$X \longrightarrow Y \longleftarrow Z$$

- True causal effect of X on Y: 1
- Simple Regression: $lm(Y \sim X)$

```
> confint(lm(y~x))
                  2.5 %     97.5 %
(Intercept) -0.06836077 0.06979605
x            0.95527153 1.09463662
```

Correct

- Multiple Regression: $lm(Y \sim X + Z)$

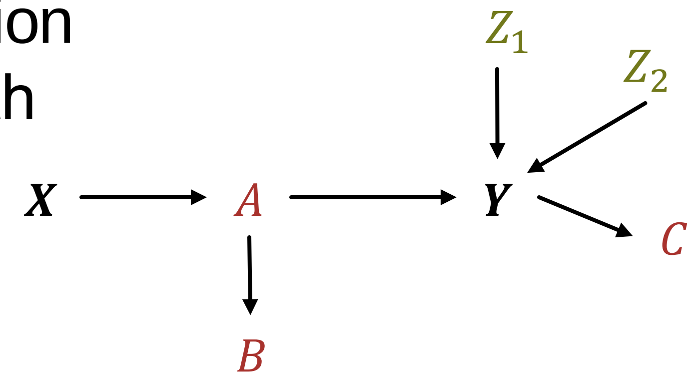
```
> confint(lm(y~x+z))
                  2.5 %     97.5 %
(Intercept) -0.08172964 0.03986164
x            0.91700180 1.04001526
z            0.46709528 0.58791825
```

Correct

- Adding a covariate did ***not*** introduce a bias
- Confidence interval with covariate is slightly smaller (0.12 vs 0.14)

Summary

- Adding the wrong variable will introduce a bias
  "Wrong variable": On causal path from X to Y or «descendants» of those nodes (*post-intervention*)
- Adding the right variables might increase precision
  "Right variable": Parents of nodes on causal path from X to Y (*pre-intervention*)



- Problem in practice:
  Usually **don't know true causal** structure!
  What are "right" and "wrong" variables ?
- If in doubt, don't use covariate !
- Safe variables: Things that clearly "preceded" X (e.g. gender)

Outline for the rest of the talk

- Total causal effect and covariate adjustment
- Issues in observational studies
- Issues in interventional studies
- Insights from recent theoretical developments

Adjustment Criteria

Getting the “right estimate”:

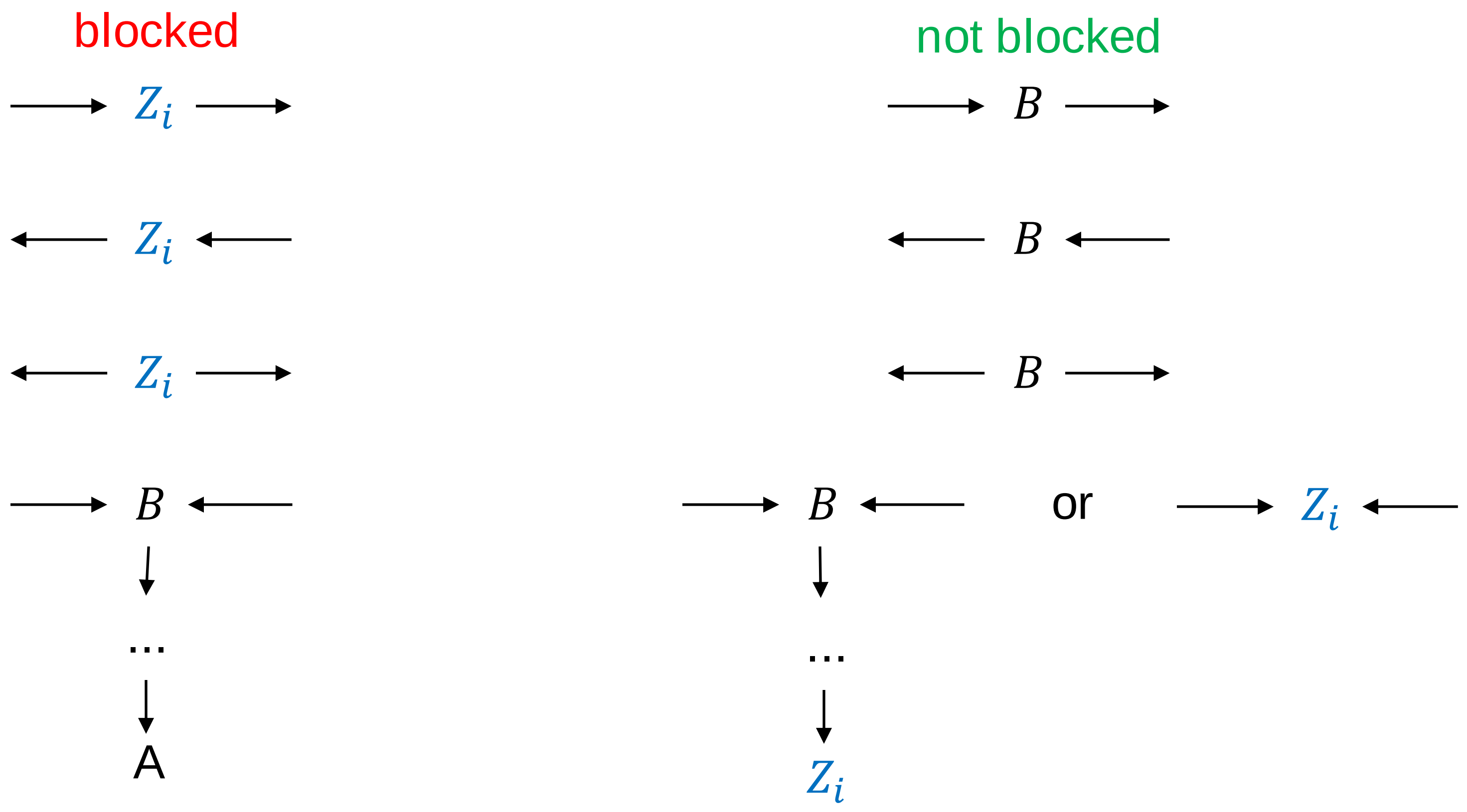
- given causal structure, criterion to check if a set is a valid adjustment set
- assuming causal structure is a strong assumption in practice
- discussion can shift to discussing reasonable causal structures

- Pearl’s back-door criterion
- Generalized Adjustment criterion

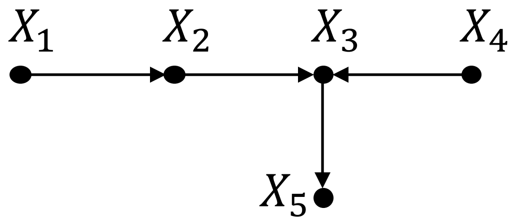
DAG

Background: d-separation

- Given a DAG G: X and Y are d-separated («blocked») by $\{Z_1, \ldots, Z_p\}$ if you can not walk from X to Y.
- Rules for walking from X to Y:



d-separation: Example



- X_1 and X_3 are d-sep by X_2
- X_1 and X_3 are not d-sep by $\{\}$

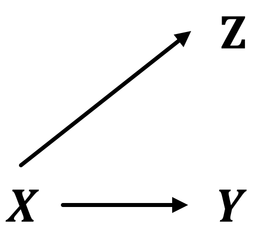
- X_2 and X_4 are d-sep by $\{\}$
- X_2 and X_4 are not d-sep by X_3
- X_2 and X_4 are not d-sep by X_5

Pearl's back-door criterion (PBC)

- Improvement on Parent Criterion
- PBC: Set Z satisfies **back-door criterion** relative to (X, Y) if
  - No node in Z is a descendant of X and
  - Z d-separates every path between X and Y that contains an arrow into X
- Example: Parents of X always satisfy the back-door criterion
- Result (Pearl): **If** a set of variables Z satisfies the back-door criterion relative to (X, Y), **then** Z is a valid adjustment set.

$$P(Y = y | do(X = x)) = \sum_z P(Y = y | X = x, Z = z) P(Z = z)$$

Pearls back-door criterion is not complete



- Empty set satisfies back-door criterion

- Z does not satisfy back-door criterion,
  but Z is a valid adjustment set !

- → Pearl's back-door criterion is not complete

```r
## counter example PBC
set.seed(123)
n <- 1000
x <- rnorm(n)
z <- 0.5*x + rnorm(n)
y <- 1*x + rnorm(n)
```

```
> confint(lm(y~x+z))
                  2.5 %     97.5 %
(Intercept) -0.08172964 0.03986164
x            0.89392580 1.03558450
z           -0.03290472 0.08791825
```

Correct

Improvements: Generalized Adjustment Criterion (GAC) & asymptotic variance

Getting the "right estimate":

- "Sound and complete" (= correct and does not miss anything)
- We will simplify and **show results only for DAGs and single node interventions**
- GAC is general:
  - DAGs, PDAGs, CPDAGs
  - MAGs, PAGs
  - sets and not only single variables

See also Shpitser, 2012 for the DAG case

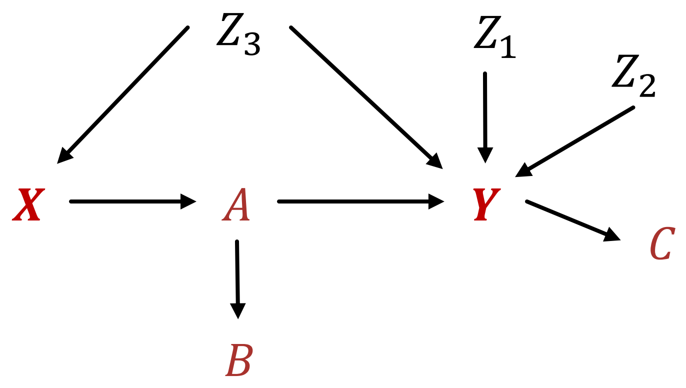
GAC for DAGs: Preliminaries

E. Perković, J. Textor, M. Kalisch and M.H. Maathuis (2018). Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research* **18** (220): 1-62. (published version)

- **Causal nodes** $Cn(X,Y,G)$ relative to X and Y in G:
  All nodes on a causal path from X to Y (excluding X but including Y)
- **Forbidden set** $Forb(X,Y,G)$ relative to nodes X and Y in DAG G:
  All nodes on causal paths from X to Y (excluding X but including Y) and all descendants of those nodes together with X.

$$Forb(X,Y,G) = De\big(Cn(X,Y,G)\big) \cup X$$

- Example



$Cn(X,Y,G) = \{A,Y\}$
$Forb(X,Y,G) = \{A,Y,B,C,X\}$

"post-treatment"

GAC for DAGs

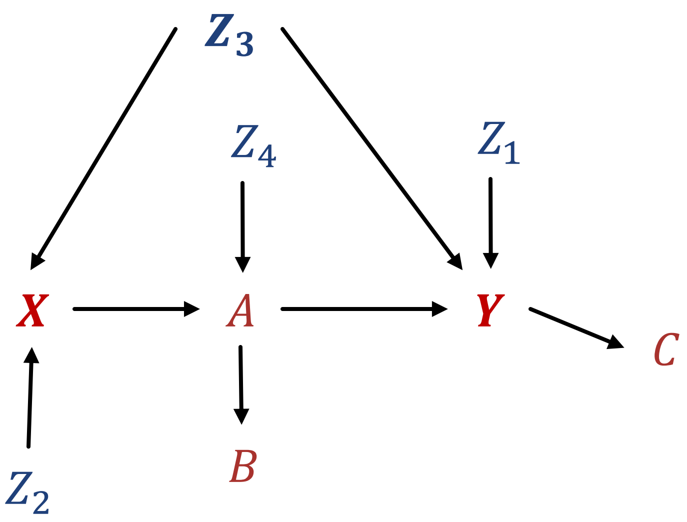
E. Perković, J. Textor, M. Kalisch and M.H. Maathuis (2018). Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research* **18** (220): 1-62. (published version)

Z is an adjustment set relative to (X, Y) in G if and only if

- no node in Z is in the **forbidden set** relative to X and Y in G and
- all non-causal paths from X to Y are **blocked** by Z in G.

Example:

- R package dagitty
- Online tool dagitty



Possible choices for blocking:
$\{Z_3\} \cup$ any subset of $\{Z_1, Z_2, Z_4\}$
→ 8 possible valid adjustment sets

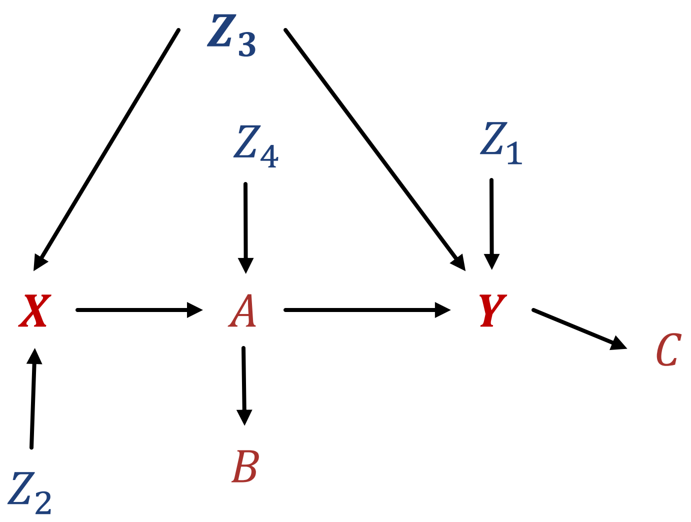
"pre-treatment"

$Cn(X, Y, G) = \{A, Y\}$
$Forb(X, Y, G) = \{A, Y, B, C, X\}$

"post-treatment"

Getting more precision

(For linear structural equation models with Gaussian errors)



Possible choices for blocking:
$\{Z_3\} \cup$ any subset of $\{Z_1, Z_2, Z_4\}$
→ 8 possible valid adjustment sets

"pre-treatment"

$Cn(X, Y, G) = \{A, Y\}$
$Forb(X, Y, G) = \{A, Y, B, C, X\}$

"post-treatment"

- All 8 adjustment sets have no bias but which one has lowest (asymptotic) variance ?
- Optimal set $O(X, Y, G) = Pa(Cn(X, Y, G), G) \setminus Forb(X, Y, G)$
- In example: $Cn(X, Y, G) = \{A, Y\}, Pa(Cn(X, Y, G), G) = \{X, Z_1, Z_3, Z_4\}$
  Of those, X is in $Forb(X, Y, G)$. Thus, $O(X, Y, G) = \{Z_1, Z_3, Z_4\}$

Summary

- Total causal effect and covariate adjustment
  → find the “right” *adjustment set* → linear regression
- Issues in observational studies
  → not easy to find right adjustment set; bigger ≠ better
- Issues in interventional studies
  → can “mess up” RCT by using “wrong” adjustment set;
  if in doubt, use empty set after RCT
- Insights from recent theoretical developments
  → GAC is sound and complete for finding adjustment set **given causal structure** (strong assumption)
  → discussion can shift to discussing reasonable causal structures
  → RCT remains gold standard