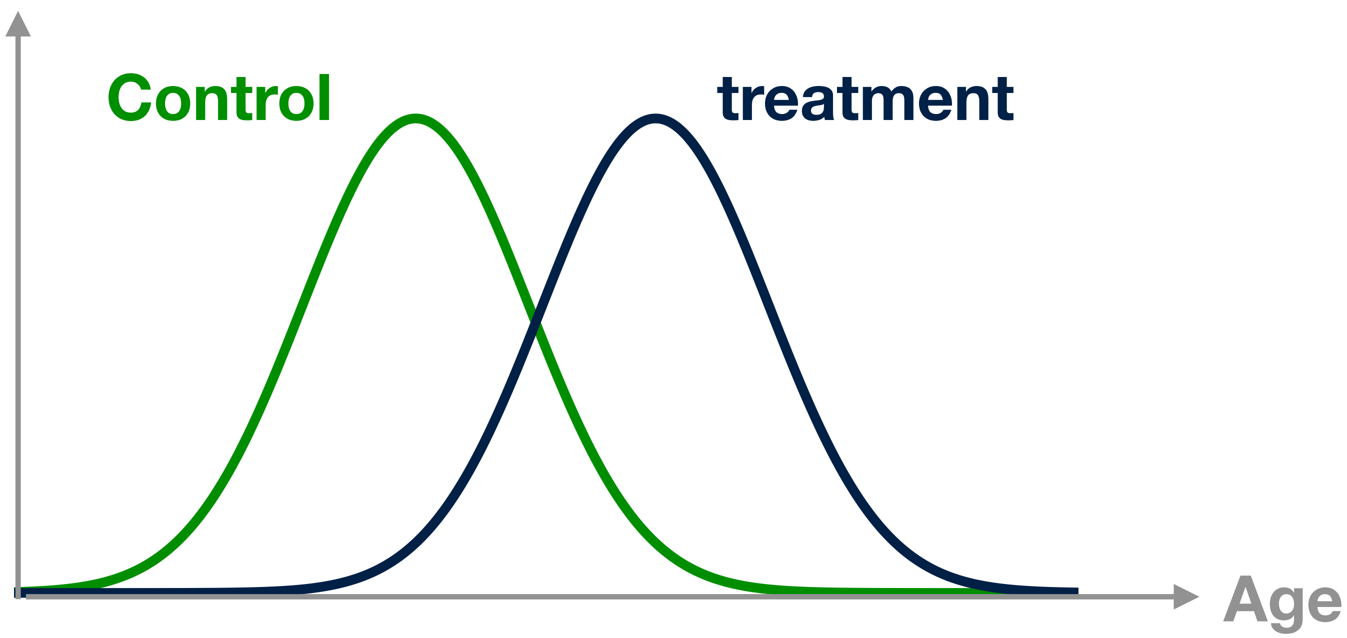
Causality in Biomedicine Lecture Series: Lecture 2

Ava Khamseh

29 Jan, 2020

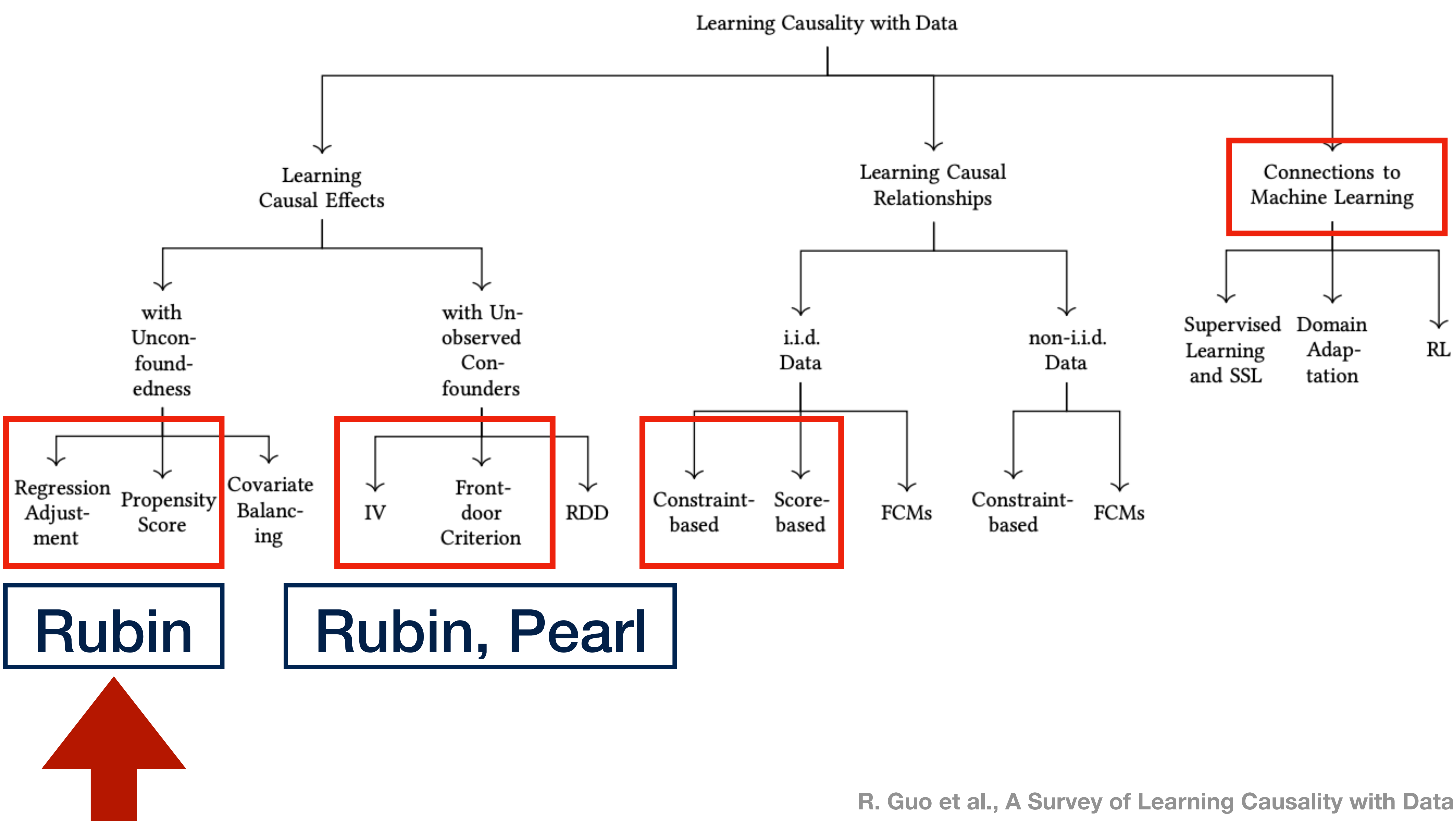
Last time: **Observational data, w**hat goes wrong?

$$p(x|t=1) \neq p(x|t=0)$$



$$\left(\int y_1(x) p(x|t=1) dx \; - \int y_0(x) p(x|t=0) dx \right) \neq \int \big(y_1(x) - y_0(x)\big) p(x) dx$$

Overview of the field



R. Guo et al., A Survey of Learning Causality with Data

Potential Outcomes Framework (Rubin)

- **Definition:** Given treatment, t, and outcome, y, the **potential outcome** of instance/individual (i) is denoted by $y_t^{(i)}$ is the value y *would have* taken if individual (i) had been under treatment t.

- $y_0^{(i)}$ and $y_1^{(i)}$ are not **observed**, but **potential** outcomes
- $t^{(i)}$ is the observed treatment applied to individual (i), 0 or 1
- **Observed** outcomes: $y_{\underline{0}}^{(i)}$ **OR** $y_{\underline{1}}^{(i)}$ depend on treatment (**fundamental problem of causal inference**):

$$y_{obs}^{(i)} = t^{(i)} y_1^{(i)} + (1 - t^{(i)}) y_0^{(i)}$$

- **Individual treatment effect**: $\tau^{(i)} = y_1^{(i)} - y_0^{(i)}$

- **Average treatment effect:**

$$\tau = \hat{\mathbb{E}}[\tau^{(i)}] = \hat{\mathbb{E}}[y_1^{(i)} - y_0^{(i)}] = \frac{1}{N} \sum_{i=0}^{N} \left(y_1^{(i)} - y_0^{(i)} \right)$$

Potential Outcomes Assumptions (Rubin)

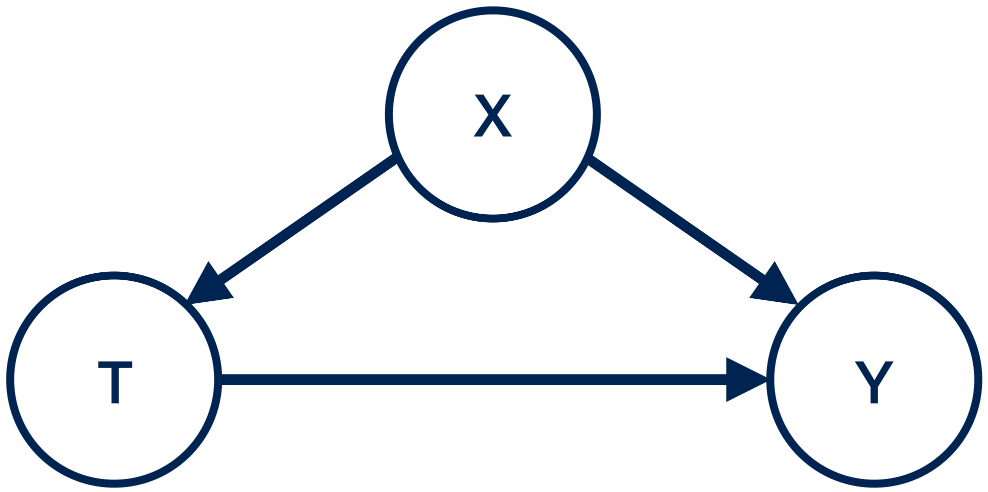
- **SUTVA:** Stable Unit Treatment Value Assumption
  - Well-defined treatment (no different versions)
  - No interference: Different individuals (units) within a population do not influence each other (e.g. does not work in social behavioural studies, care must be taken for time series data when defining the units)

- **Consistency:** The observed outcome is independent of how the treatment is assigned

- **Unconfoundedness (ignorability):** Treatment assignment is random, given X:

$$y_1^{(i)}, y_0^{(i)} \perp\!\!\!\perp t^{(i)} \mid x$$

Causal inference with observed confounders



Regression Adjustment

- X is a sufficient set of confounders if conditioning on X, there would be no confounding bias: $p(y|t,x) = p(y|t, do(x))$

- For individual (i) there is only one observed outcome: $y_{t_i}^{(i)}$
- Would like to estimate (infer) **counterfactual**: $\hat{y}_{1-t}^{(i)} = \hat{\mathbb{E}}\left[y^{(i)}|1-t_i, x^{(i)}\right]$

- Using a design matrix, fit: $Y = \beta_X X + \beta_T T + \epsilon$

$$T = \begin{array}{c} \begin{matrix} \textbf{Ctrl} & \textbf{Drug} \end{matrix} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ .. & .. \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array} \qquad X = \begin{array}{c} \begin{matrix} \textbf{Young} & \textbf{Old} \end{matrix} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ .. & .. \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \quad \Longrightarrow \quad \begin{pmatrix} y^{(1)} \\ y^{(1)} \\ .. \\ y^{(N-1)} \\ y^{(N)} \end{pmatrix} = \begin{pmatrix} \beta_{t=0} + \beta_{x=\text{young}} \\ \beta_{t=0} + \beta_{x=\text{old}} \\ .. \\ \beta_{t=1} + \beta_{x=\text{young}} \\ \beta_{t=1} + \beta_{x=\text{old}} \end{pmatrix}$$

- Assumptions: Overlap and additivity

Matching

- **Idea:** Blind ourselves to the outcomes, try to get as similar to a randomised experiment as possible ('correct for confounding')
- Reveals **lack of overlap** in treatment vs control distributions: individuals in the treatment group that have no chance of having an 'equivalent' in control group, ie, parts of the distribution with:

$$p(t=1|x)=0,\ p(t=0|x)=0$$

- **Mahalanobis distance**: Difference scaled by variance

$$D(x^{(i)},x^{(j)})=\sqrt{\left(x^{(i)}-x^{(j)}\right)^T S_{ij}^{-1}\left(x^{(i)}-x^{(j)}\right)},\ S=\mathrm{Cov}(X)$$

- Issues: Outliers. Use a calliper: maximum acceptable distance, to avoid violating the positivity (strong ignorability) assumption. But the populations becomes harder to define.

- See papers on **anomaly detection:** When in fact, we are interested in the outliers

Propensity Score

- In a **randomised** trial: p(t=1|x)=p(t=1)=0.5

- In an **observational study**, p(t=1|x) can be **estimated**, since it involves **observational data** at a t and x (hence identifiable).

- A **balancing score** is any function b(x) such that:

$$x \perp\!\!\!\perp t | b(x)$$

i.e., distribution of confounders is independent of treatment given b(x):

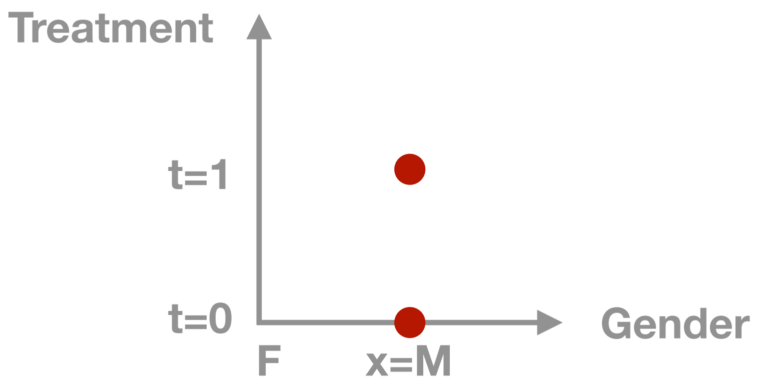
$$p(X = x | b(x), t = 1) = p(X = x | b(x), t = 0)$$

$$y_1^{(i)}, y_0^{(i)} \perp\!\!\!\perp t^{(i)} \; | b(x)$$

Rosenbaum, Rubin 1983

Propensity Score

- Candidate b(x) = x, trivially satisfies:

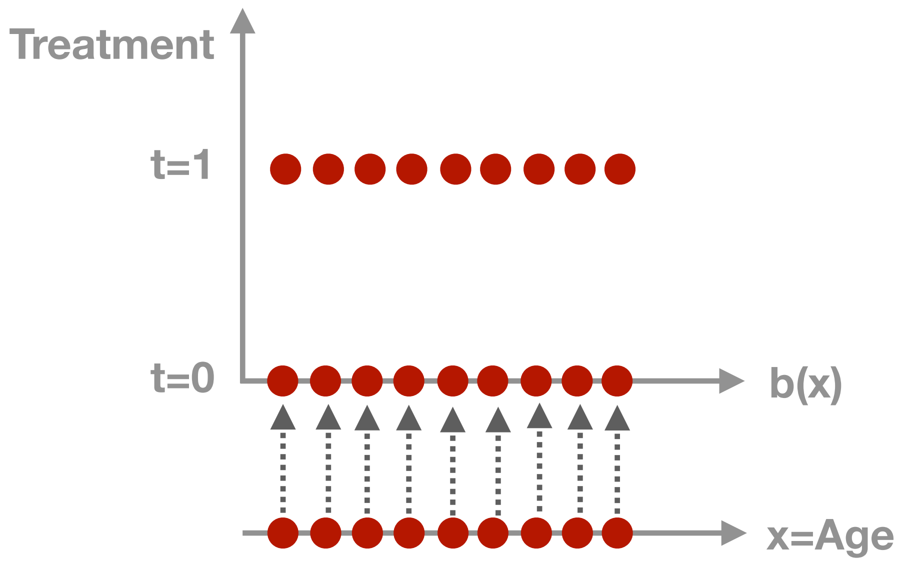


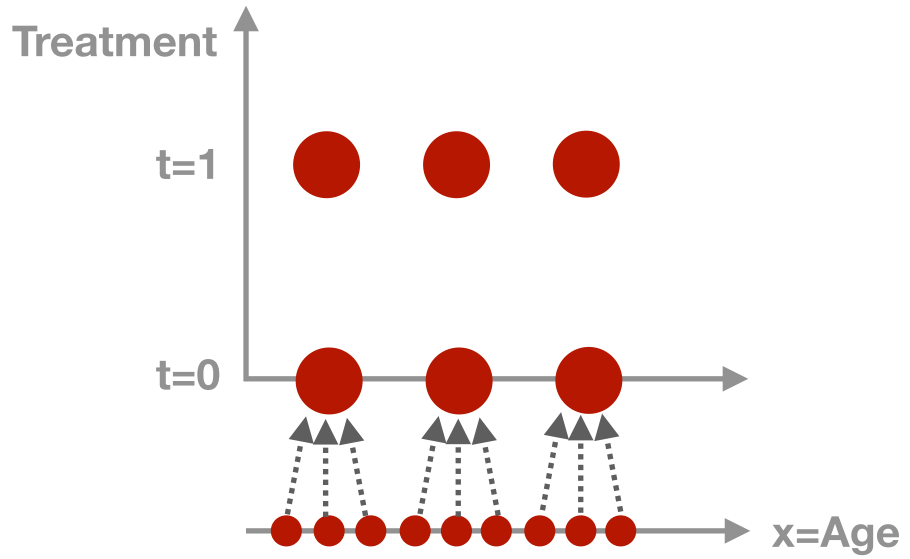
$$p(X = x|x, t = 1) = p(X = x|x, t = 0) = 1$$

- b(x) = x is the **finest** such function: OK for e.g. binary confounders, but only gives point estimates for (almost) continuous confounders!

- **Propensity score** is the **coarsest** such function (i.e. more data points, leading to better estimates):

$$\boxed{e(x) = p(t = 1|x)}$$





Propensity Score Matching

- Let the distribution of covariates follow an exponential family of distributions ($P_{t^*}(x)$ polynomial of degree k):

$$p(x|t=t^*) = h(X)\exp(P_{t^*}(x)) \;,\; \text{ for } t = 0 \text{ or } 1$$

- Estimate propensity score e(x)=p(t=1|x):

$$\log\left(\frac{e(x)}{1-e(x)}\right) = \log\left(\frac{p(t=1|x)}{p(t=0|x)}\right) = \log\left(\frac{p(x|t=1)p(t=1)}{p(x|t=0)p(t=0)}\right) = \log\left(\frac{p(t=1)}{p(t=0)}\right) + P_1(x) - P_0(x)$$

- If we consider k=1, linear exponential family (e.g. Bernoulli),

$$\log\left(\frac{e(x)}{1-e(x)}\right) = wx + w_0 \;\Rightarrow\; e(x) = \frac{1}{1+e^{-wx-w_0}}$$

- Fit parameters by maximising log-likelihood: $LL = \frac{1}{N}\sum_{i=0}^{N} \log p(t^{(i)}|x^{(i)})$

Rosenbaum, Rubin 1983

Propensity Score Matching Algorithms

- Match control and treatment individuals based on their propensity score
- Greedy matching:
  - Randomly order list of control and treated.
  - Start with the first individual from e.g. treated and match to control with the smallest distance (i.e. obtains the **local** minimum)
  - Remove individuals from control and matched treated
  - Move to the next treated subject

- Optimal matching: Minimises the **global** distance, computationally demanding

- **ATE**: Since the number of treatment and control can be different, match twice (N = treatment + control):

Inverse Probability of Treatment Weighting (IPTW)

- Inflate the weight for under represented-subjects due to missing data
- Based on propensity score
- Weight: inverse probability of receiving observed treatment, for individual i:

$$w_i = \frac{t_i}{e_i} + \frac{1 - t_i}{1 - e_i}$$

- Use these weights when averaging over individuals for ATE
- Weights may be inaccurate/unstable for subjects with a very low probability of receiving the observed treatment
- Other variations to stabilise the above

Rosenbaum 1987

Sensitivity Analysis

- **Randomised** trials are unconfounded **by design** (flipping a coin)
- **Observational data** may have possible hidden bias/unobserved confounder that is not controlled for
- No guarantee that matching leads to balance on variables we did **not** match for!
- **People who look comparable may differ**
- Violates ignorability (unconfoundedness) assumption
- Unconfoundedness is fundamentally (directly) unverifiable

Rosenbaum Design of Observational Studies, Springer, 2010

Sensitivity Analysis

- “This difference in the unobserved covariate u, the critic continues, is the real reason outcomes differ in the treated and control groups: it is not an effect caused by the treatment, but rather a failure on the part of the investigators to measure and control imbalances in u. Although not strictly necessary, the critic is usually aided by an air of superiority: “This would never happen in my laboratory.””

- “It is important to recognize at the outset that our critic may be, but need not be, on the side of the angels. The tobacco industry and its (sometimes distinguished) consultants criticized, in precisely this way, observational studies linking smoking with lung cancer.”

Rosenbaum Design of Observational Studies, Springer, 2010

Sensitivity Analysis

- If there is hidden bias, how severe is it:
  - Does the conclusion change from statistically significant to not?
  - Does it change the direction of effect?

- i.e., how sensitive are our conclusions to minor violation of our keys assumption

- If very sensitive: change strategy (see **Causal Inference with Unobserved Confounders**)

Sensitivity Analysis

- Take individuals (i) and (j), such that their observed covariates are the same: $X^{(i)} = X^{(j)}$
- Then, if the propensity scores are the same, $e^{(i)} = e^{(j)}$, there is no hidden bias
- Consider e.g., the odds ratio:

$$\frac{1}{\Gamma} \leq \frac{\frac{e^{(i)}}{1-e^{(i)}}}{\frac{e^{(j)}}{1-e^{(j)}}} \leq \Gamma$$

- Then, if $\Gamma \approx 1$, there is no overt bias, otherwise there is a hidden bias. If, e.g., $\Gamma = 2$, one subject is twice as likely to receive treatment because of unobserved pre-treatment feature

- Γ quantifies degree of bias.

Sensitivity Analysis Computations: An example

- S pairs, s = 1,...,S of two subjects, one treated, one control, **matched** for observed covariates

- Statistical test: **Wilcoxon's signed rank test** (non-parametric), W is the sum of the ranks of the positive differences between treatment and control

- In a moderately large randomized experiment, under the **null hypothesis of no effect**, W is approximately normally distributed

$$\mathbb{E}[W] = S(S+1)/4 \quad , \quad \mathrm{Var}[W] = S(S+1)(2S+1)/24$$

Rosenbaum Sensitivity Analysis in Observational Studies, 2005

Sensitivity Analysis Computations: An example

- Example: W=300, S=25 pairs in a randomised experiment
- In a randomised experiment ($\Gamma \approx 1$, well-matched):

$$\mathbb{E}[W] = 162.5 \;,\;\; \mathrm{Var}[W] = 1381.25 \;,\;\; \text{deviate } Z = (300 - 162.5)/\sqrt{1381.25} = 3.70$$

- Compared to a normal distribution: p-value = 0.0001
- In a moderately large observational study, under the null hypothesis of no effect, the distribution of W is approximately bounded between two Normal distributions (notice: $\Gamma \approx 1$)

$$\mu_{\max} = \lambda S(S+1)/2 \quad , \quad \mu_{\min} = (1-\lambda)S(S+1)/2$$

$$\sigma^2 = \lambda(1-\lambda)S(S+1)(2S+1)/6$$

$$\lambda = \Gamma/(1+\Gamma)$$

Rosenbaum Sensitivity Analysis in Observational Studies, 2005

Sensitivity Analysis Computations: An example

- Example: W=300, S=25 pairs in a randomised experiment
- For $\Gamma = 2, \;\; \lambda = \Gamma/(1+\Gamma) = 2/3$

$$\mu_{\max} = \lambda S(S+1)/2 = 216.67 \quad , \quad \mu_{\min} = (1-\lambda)S(S+1)/2 = 108.33$$

$$\sigma^2 = \lambda(1-\lambda)S(S+1)(2S+1)/6 = 1227.78$$

$$Z_1 = 5.47 \;\; \Rightarrow \;\; p = 0.00000002$$

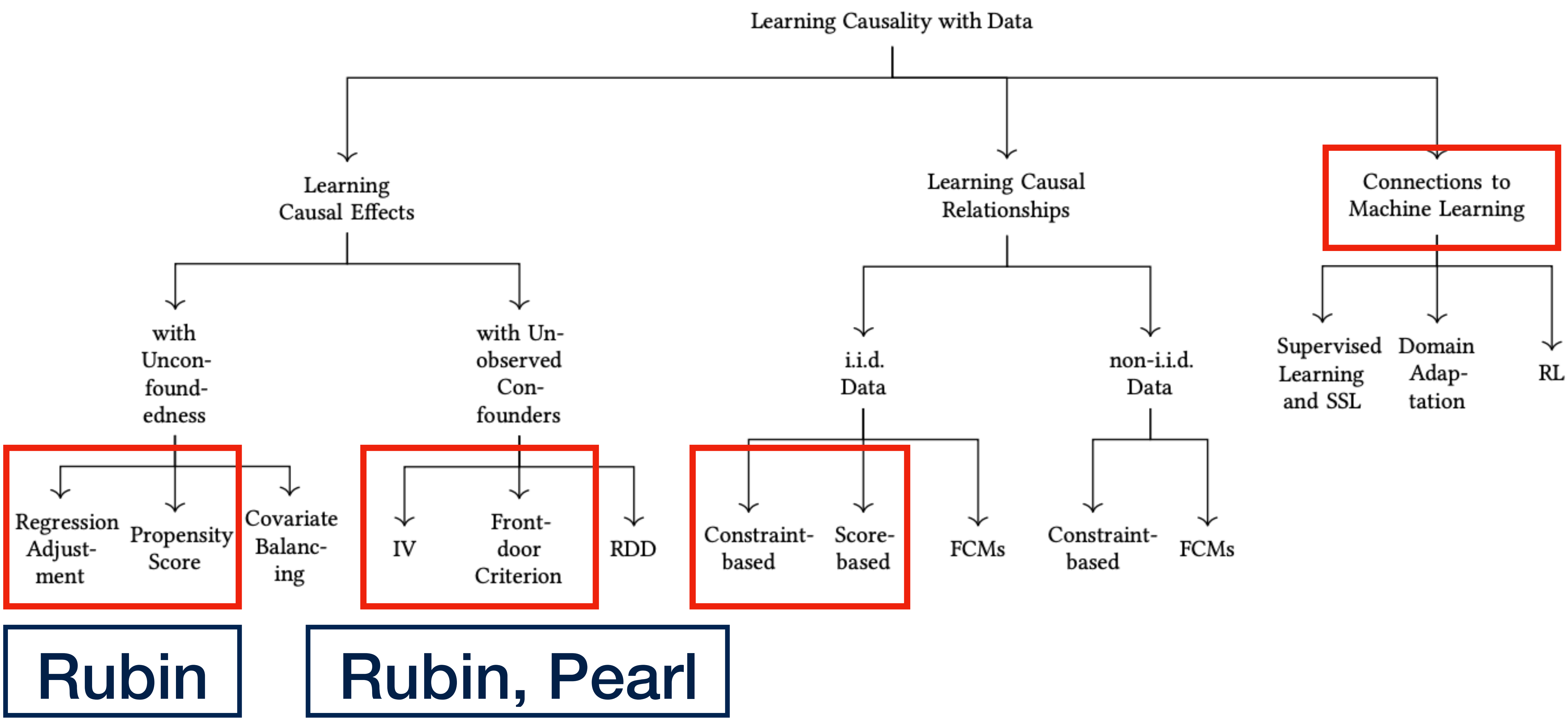
$$Z_2 = 2.38 \;\; \Rightarrow \;\; p = 0.009$$ **still significant, even with $\Gamma = 2$**

- For the tobacco and lung cancer example, $\Gamma = 6$.

Notice: There are two sources of uncertainty (confidence interval):
1) Due to the causal estimates
2) Due to sensitivity analysis

Causal inference with observed confounders:
Simulations

Overview of the field

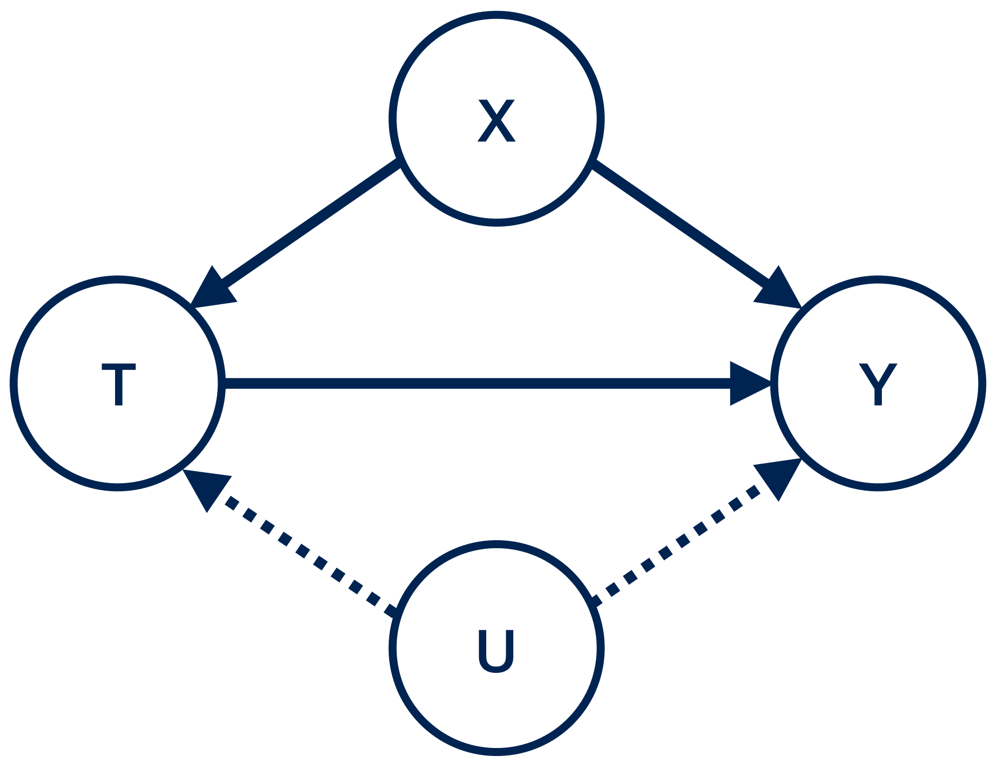


R. Guo et al., A Survey of Learning Causality with Data

Causal inference with unobserved confounders

Instrumental Variable (Originally due to Rubin)

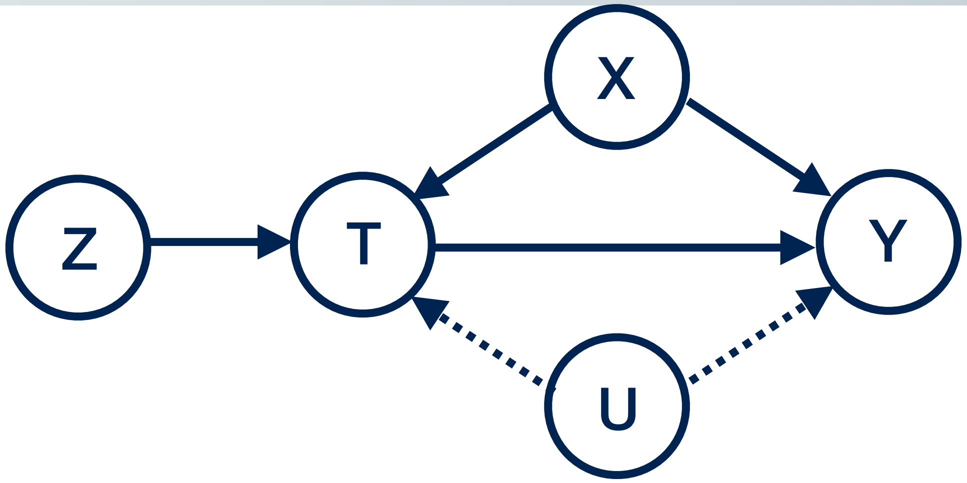
- Unobserved confounders (U), **violates unconfoundedness (ignorability),** i.e. conditioning on X alone, would not results in a randomised treatment assignment
- Unconfoundedness is fundamentally unverifiable



Rubin 1996

Instrumental Variable example

- Example 1:
  - T: smoking during pregnancy
  - Y: birthweight
  - X: parity, mother's age, weight, …
  - U: Other unmeasured confounders



- Randomise Z (intention-to-treat): either receive encouragement to stop smoking (Z=1), or receive usual care (Z=0)
- Intention-to-treat analysis gives causal effect estimator of encouragement z on outcome y:

$$\mathbb{E}(y|z=1) - \mathbb{E}(y|z=0)$$

- What can we say about the causal effect of smoking itself?

Causality in Biomedicine Lecture Series: Lecture 2

Ava Khamseh

29 Jan, 2020