

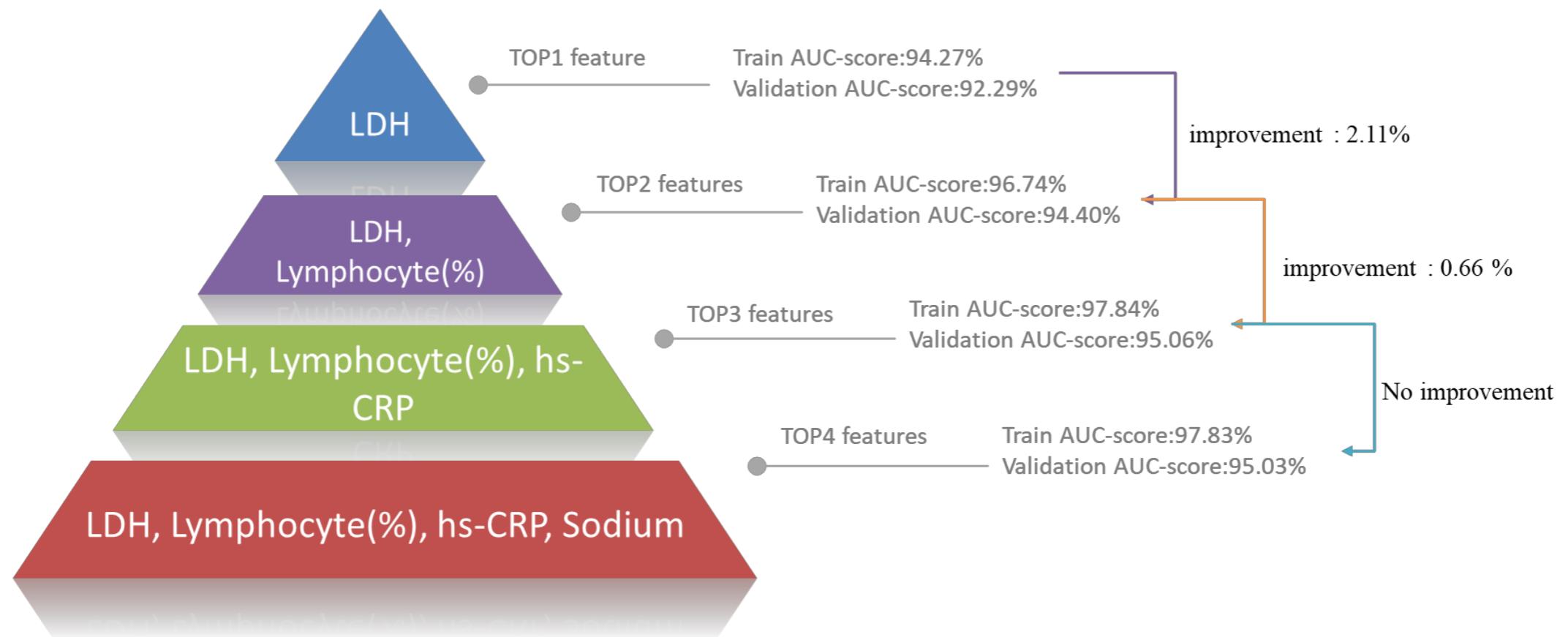
STATS 604

Lecture 8

Causal inference

Jonathan Terhorst

Predicting COVID mortality



Effectiveness of mitigation strategies

Table 1 | Comparison of effectiveness rankings on L2

L2 category	Score (%)	Consensus	ΔR_t^{CC}	$\Delta R_t^{\text{LASSO}}$	Importance (RF)	ΔR_t^{TF}
Small gathering cancellation	83	4	-0.35 (2)	-0.22 (5)	0.020 (2)	-0.327 (3)
Closure of educational institutions	73	4	-0.16 (2)	-0.21 (4)	0.028 (2)	-0.146 (2)
Border restriction	56	4	-0.23 (2)	-0.12 (2)	0.017 (2)	-0.057 (2)
Increased availability of PPE	51	4	-0.11 (2)	-0.13 (2)	0.012 (1)	-0.062 (2)
Individual movement restrictions	42	4	-0.13 (2)	-0.08 (3)	0.017 (2)	-0.121 (2)
National lockdown	25	4	-0.14 (3)	-0.09 (2)	0.0020 (9)	-0.008 (3)
Mass gathering cancellation	53	3	-0.33 (2)	0	0.012 (1)	-0.127 (2)
Educate and actively communicate with the public	48	3	-0.18 (4)	0	0.018 (2)	-0.276 (2)
The government provides assistance to vulnerable populations	41	3	-0.17 (3)	-0.18 (4)	0.009 (1)	0.090 (3)
Actively communicate with managers	40	3	-0.15 (2)	-0.20 (4)	0.004 (2)	-0.050 (2)
Measures for special populations	37	3	-0.19 (2)	0	0.008 (1)	-0.100 (2)
Increase healthcare workforce	35	3	-0.17 (20)	-0.13 (3)	0.030 (8)	0.011 (2)
Quarantine	30	3	-0.28 (2)	-0.2 (1)	0.0023 (9)	0.023 (2)
Activate or establish emergency response	29	3	-0.13 (2)	0	0.0037 (9)	-0.121 (2)
Enhance detection system	25	3	-0.19 (3)	0	0.0032 (9)	-0.106 (2)
Increase in medical supplies and equipment	25	3	-0.13 (3)	-0.004 (3)	0.003 (2)	-0.200 (3)
Police and army interventions	23	3	-0.16 (2)	0	0.003 (2)	-0.091 (2)
Travel alert and warning	20	3	-0.13 (3)	0.0 (1)	0.002 (1)	-0.159 (3)
Public transport restriction	13	3	0.020 (4)	-0.01 (7)	0.004 (1)	-0.023 (3)
Actively communicate with healthcare professionals	11	3	0	-0.08 (4)	0.003 (1)	-0.003 (2)

Out of the 46 NPI categories, all four methods show significant results for six NPIs (consensus 4) while three methods agree on 14 further NPIs (consensus 3). We report the average normalized score, the observed reduction in R_t for the various methods and NPI importance for RF. Numbers in parentheses denote half of the amount by which the last digit of the corresponding number outside the parentheses fluctuates within the 95% confidence interval.

Does smoking cause lung cancer?

- Randomized controlled trial is impossible.
- What about comparing $P(\text{cancer} \mid \text{smoker})$ vs. $P(\text{cancer} \mid \text{nonsmoker})$?
- Cannot naively use observational data due to confounding.

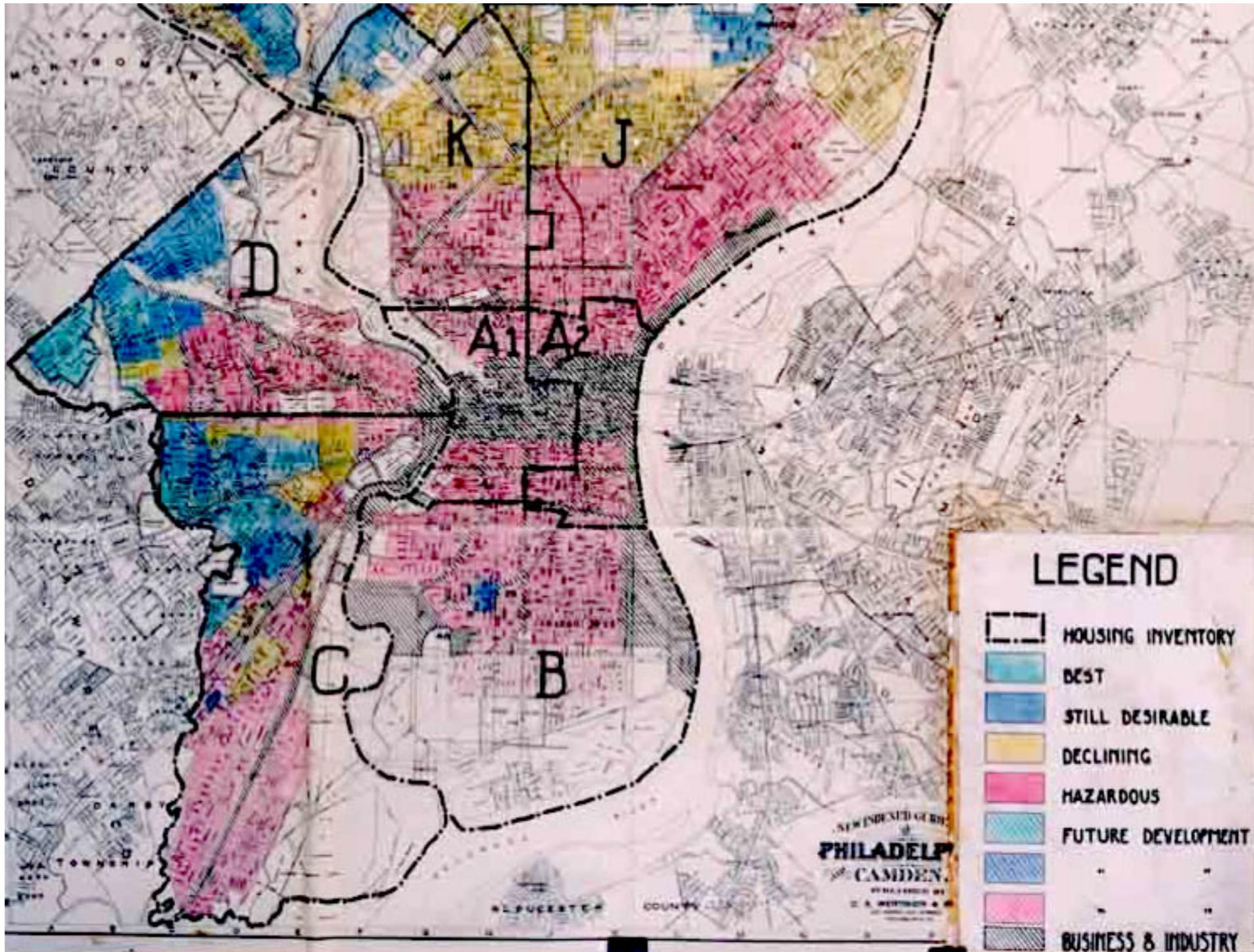


Durante et al.

Abstract

Each month, many women experience an ovulatory cycle that regulates fertility. Although research has found that this cycle influences women's mating preferences, we proposed that it might also change women's political and religious views. Building on theory suggesting that political and religious orientation are linked to reproductive goals, we tested how fertility influenced women's politics, religiosity, and voting in the 2012 U.S. presidential election. In two studies with large and diverse samples, ovulation had drastically different effects on single women and women in committed relationships. Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led women in committed relationships to become more conservative, more religious, and more likely to vote for Mitt Romney. In addition, ovulation-induced changes in political orientation mediated women's voting behavior. Overall, the ovulatory cycle not only influences women's politics but also appears to do so differently for single women than for women in relationships.

Example: redlining in Philadelphia



Involact vs. race

```
lm(formula = involact ~ race, data = chicago)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.7496	-0.2479	-0.1487	0.3129	1.1724

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.129218	0.096611	1.338	0.188
race	0.013882	0.002031	6.836	1.78e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4488 on 45 degrees of freedom

Multiple R-squared: 0.5094, Adjusted R-squared: 0.4985

F-statistic: 46.73 on 1 and 45 DF, p-value: 1.784e-08

Prediction vs. causation

- Causation involves predicting the effect of an intervention:
 - Prediction: predict Y after *observing* $X = x$.
 - Causation: predict Y after *setting* $X = x$.
- Example:
 - Prediction: predict the health of a person who takes vitamin C.
 - Causation: predict health *if I give* a person vitamin C.
 - This is a much harder question because the person might ordinarily never take vitamin C!

“Correlation is not causation”

- Every intro stats course teaches you $P(Y \in A \mid X = x)$: the probability that $Y \in A$ *given that you (passively) observe $X = x$.*
- This is not the same as the probability that $Y \in A$ **given that you actively set $X = x$.**
- We will denote this by $P(Y \in A \mid \text{set } X = x)$.
- Correlation is not causation can be mathematically expressed as

$$P(Y \in A \mid X = x) \neq P(Y \in A \mid \text{set } X = x).$$

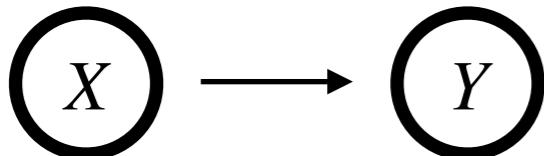
Example: sleep and health

- We've all heard that getting enough sleep is important for our health, and indeed many studies have shown that these are positively associated.
- $P(\text{healthy} \mid \text{avg. nightly hours of sleep} = 4)$ is likely to be low.
- But this is **not the same** as saying
 $P(\text{healthy} \mid \text{set avg. nightly hours of sleep} = 4)$ is low.
- Why?

Prediction vs. causality

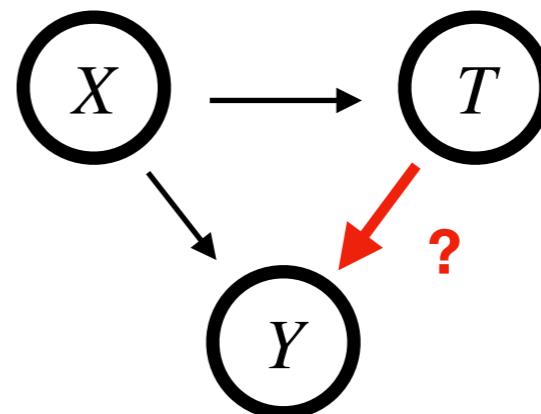
Predictive modeling

- Inputs X , outputs Y
- Learn $P(Y|X)$.
- Graphical model:



Causal inference

- Inputs X , outputs Y , *interventions* T .
- Graphical model



The potential outcomes framework

(Rubin-Neyman model)

- Each individual (“unit”) x_i has two potential outcomes:
 - $Y_0(x_i)$ is the potential outcome if the unit is not treated (**control outcome**)
 - $Y_1(x_i)$ is the potential outcome if the unit is treated (**treated outcome**)
- The conditional average treatment effect for unit i :

$$\text{CATE}(x_i) = \mathbb{E}_{Y_1 \sim p(Y_1|x_i)}(Y_1 | x_i) - \mathbb{E}_{Y_0 \sim p(Y_1|x_i)}(Y_0 | x_i).$$

Observed and counterfactual outcomes

- We only observe the **factual outcome**

$$y_i = t_i Y_1(x_i) + (1 - t_i) Y_0(x_i).$$

- We never observe the **counterfactual outcome**

$$y_i^{\text{CF}} = t_i Y_0(x_i) + (1 - t_i) Y_1(x_i).$$

- The counterfactual outcome is “what would have happened if the individual received the treatment that we did not observe.”

Important technical point

- In general,

$$\mathbb{E}(Y_1) \neq E(Y \mid T = 1)$$

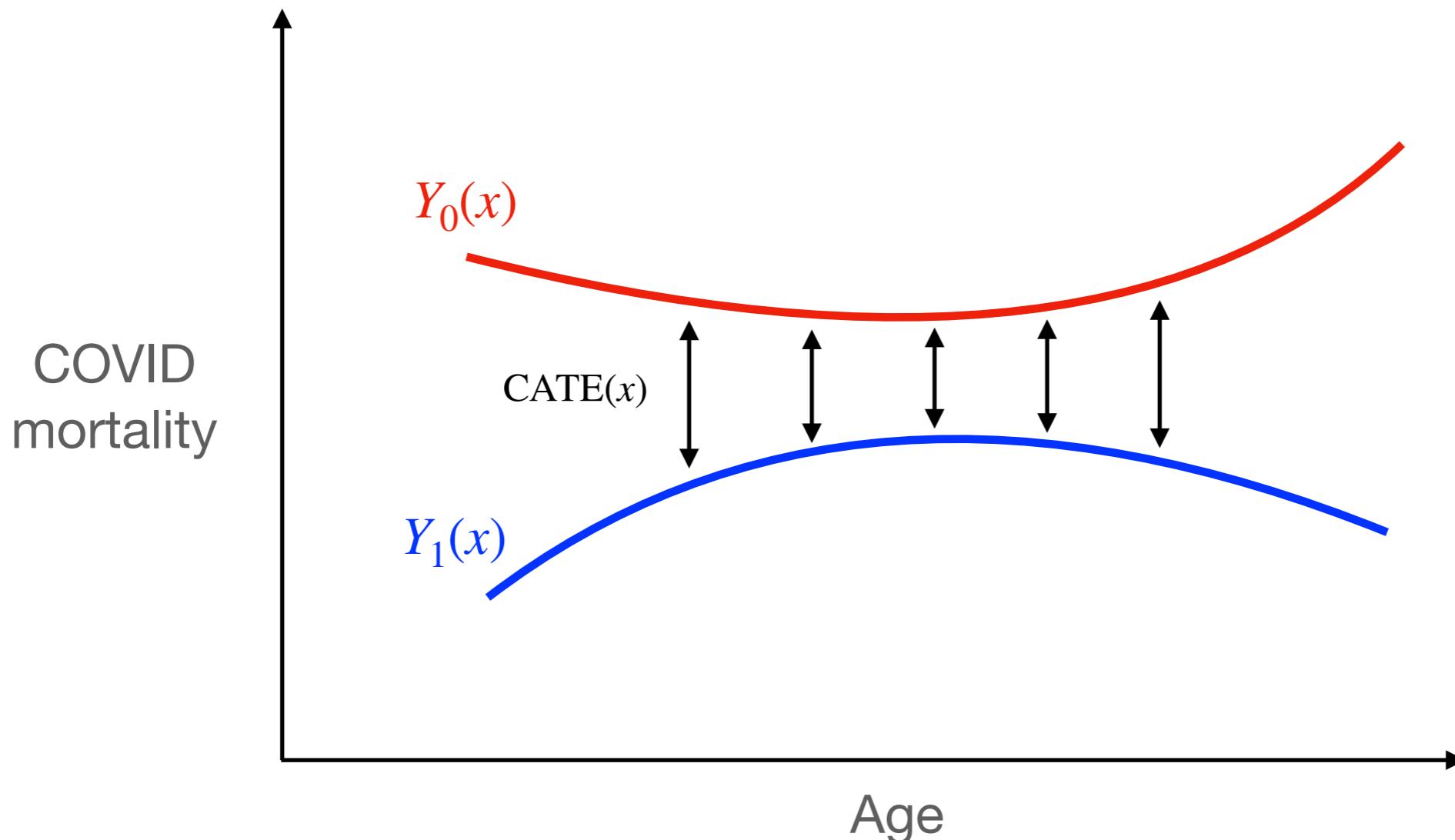
and similarly for Y_0 .

- (Why?)

Fundamental problem of causal inference:

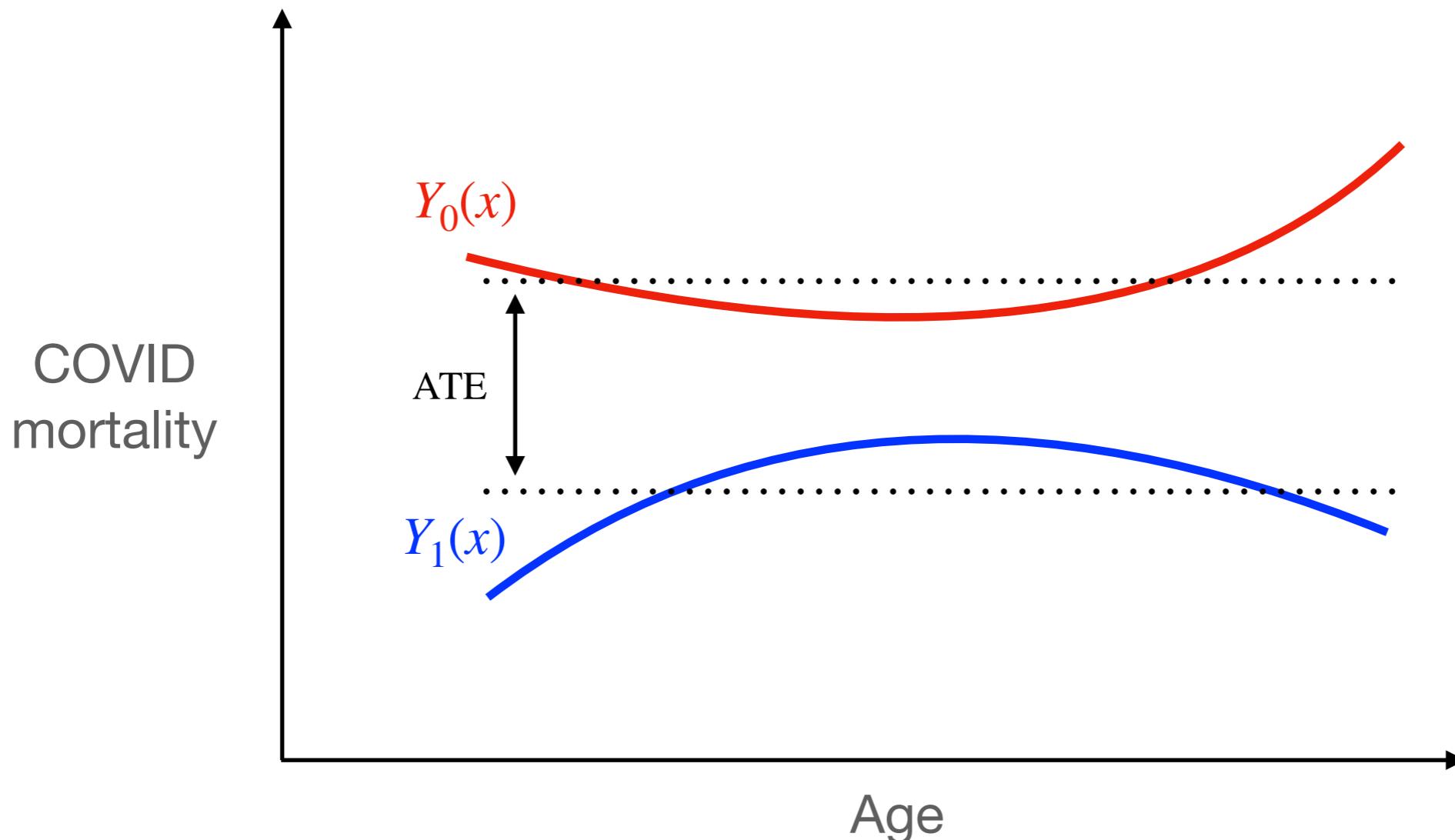
We only ever observe one of the two outcomes.

Example: COVID mortality and age



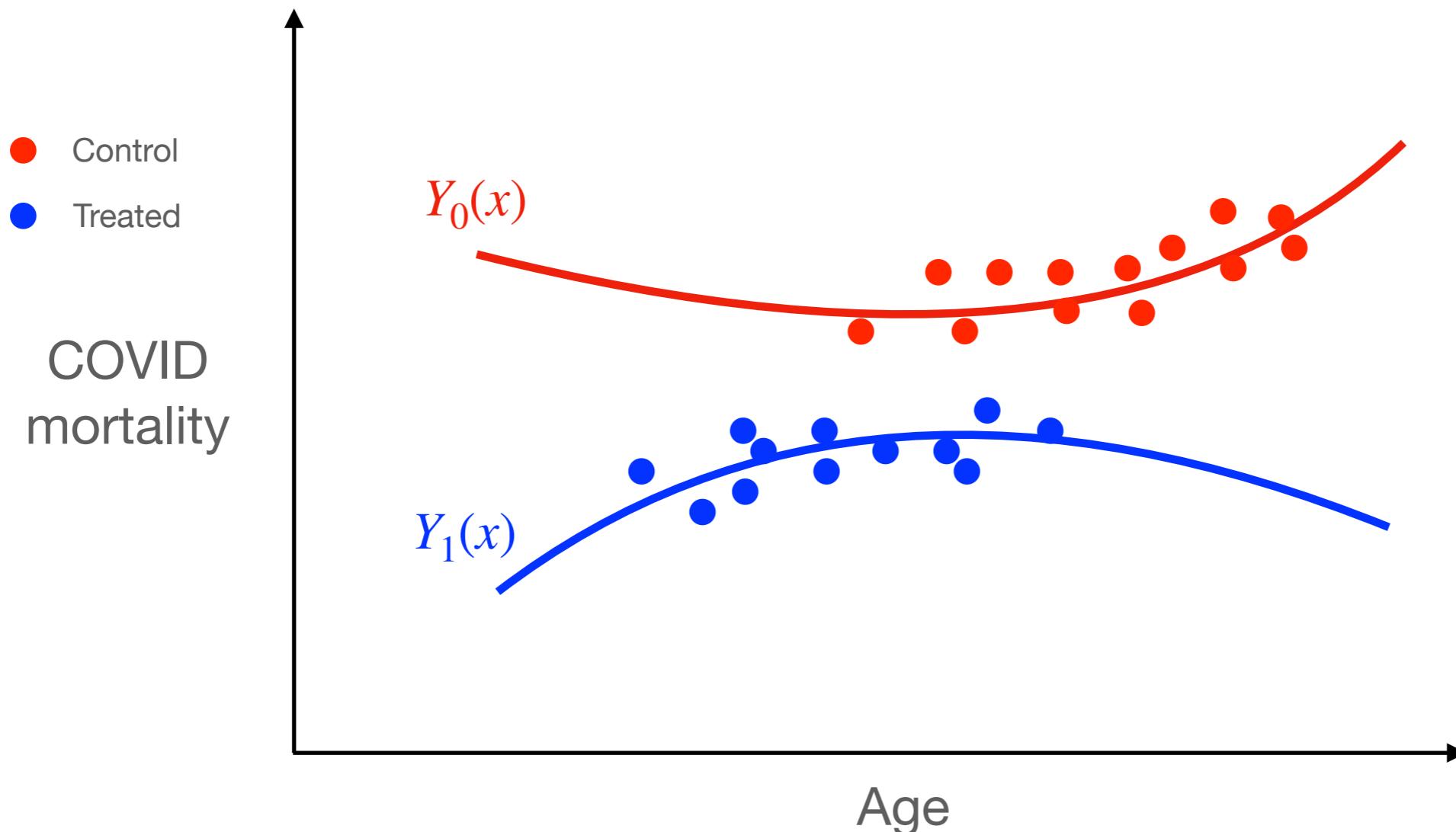
Example: COVID mortality and age

Average treatment effect



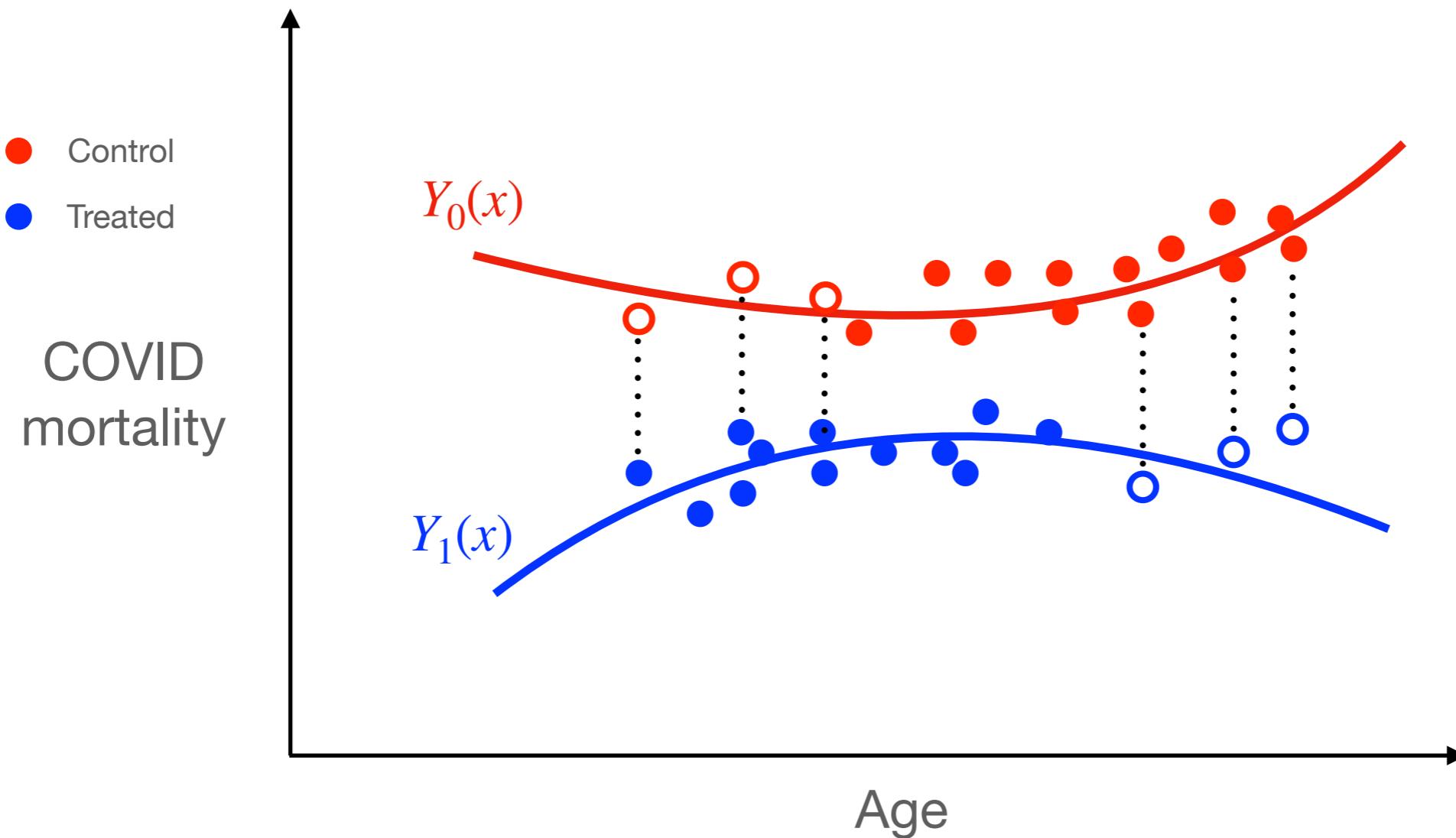
Example: COVID mortality and age

Observed data



Example: COVID mortality and age

Counterfactual outcomes



1=death; 0=no death

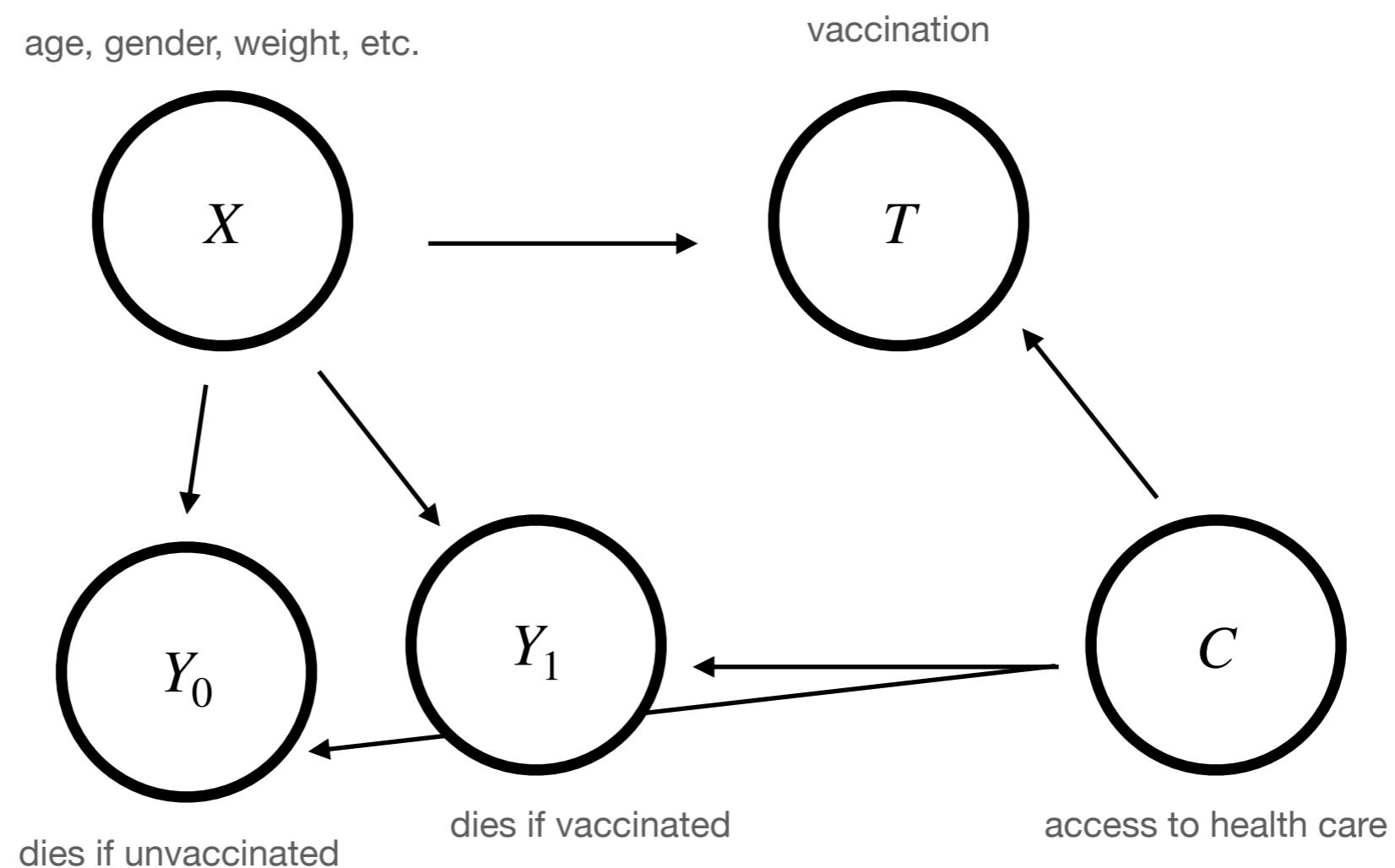
Age	Gender	Vaccinated?	Y0	Y1	Observed
25	F		0	0	0
26	M		0	0	0
26	M		1	0	0
30	F		0	0	0
33	F		0	0	1
33	M		1	1	0
44	M		0	1	0
45	F		0	1	0
51	M		0	1	0
56	F		1	1	1
58	F		1	1	1
68	M		1	1	1
72	F		1	1	1
74	F		1	1	1
81	F		1	1	1
84	M		1	1	0

Key assumptions

Ignorability

- Y_0, Y_1 potential outcomes.
- x unit covariates (features).
- T : treatment assignment.
- **Key assumption:** $(Y_0, Y_1) \perp T \mid x$.
 - The potential outcomes are independent of the treatment assignment, *conditional* on x .
 - This cannot be checked from data!

Ignorability



Key assumptions

Common support

- Y_0, Y_1 potential outcomes.
- x unit covariates (features).
- T : treatment assignment.
- **Key assumption:** $P(T = t \mid X = x) > 0$ for all t, x .
 - Every individual has a positive probability of receiving every treatment.
 - Can be checked in data.

Key assumptions

SUTVA

- “Stable unit treatment value assumption”
- The response of a particular unit only depends on the treatment that was assigned to them, and not on the treatments of others around them.
- What are some possible violations?

Average treatment effect

(AKA g -formula, adjustment formula)

- How can we compute $\text{ATE} = \mathbb{E}(Y_1 - Y_0)$?
- By the tower property,

$$\mathbb{E}Y_1 = \mathbb{E}_{x \sim p(x)} \left[\mathbb{E}_{Y_1 \sim p(Y_1|x)} (Y_1 \mid x) \right].$$

- By ignorability,

$$\mathbb{E}_{Y_1 \sim p(Y_1|x)} (Y_1 \mid x) = \mathbb{E}_{Y_1 \sim p(Y_1|x)} (Y_1 \mid x, T = 1).$$

- Hence,

$$\text{ATE} = \mathbb{E}_{x \sim p(x)} [\mathbb{E}(Y_1 \mid T = 1, x) - \mathbb{E}(Y_0 \mid T = 0, x)].$$

Average treatment effect

- $\mathbb{E}(Y_1 \mid T = 1, x), \mathbb{E}(Y_0 \mid T = 0, x)$ can be estimated from data.
- But we still need to estimate $p(x)$:
 - Covariate adjustment;
 - Propensity score reweighting;
 - Instrumental variables.
 - ...

Covariate adjustment

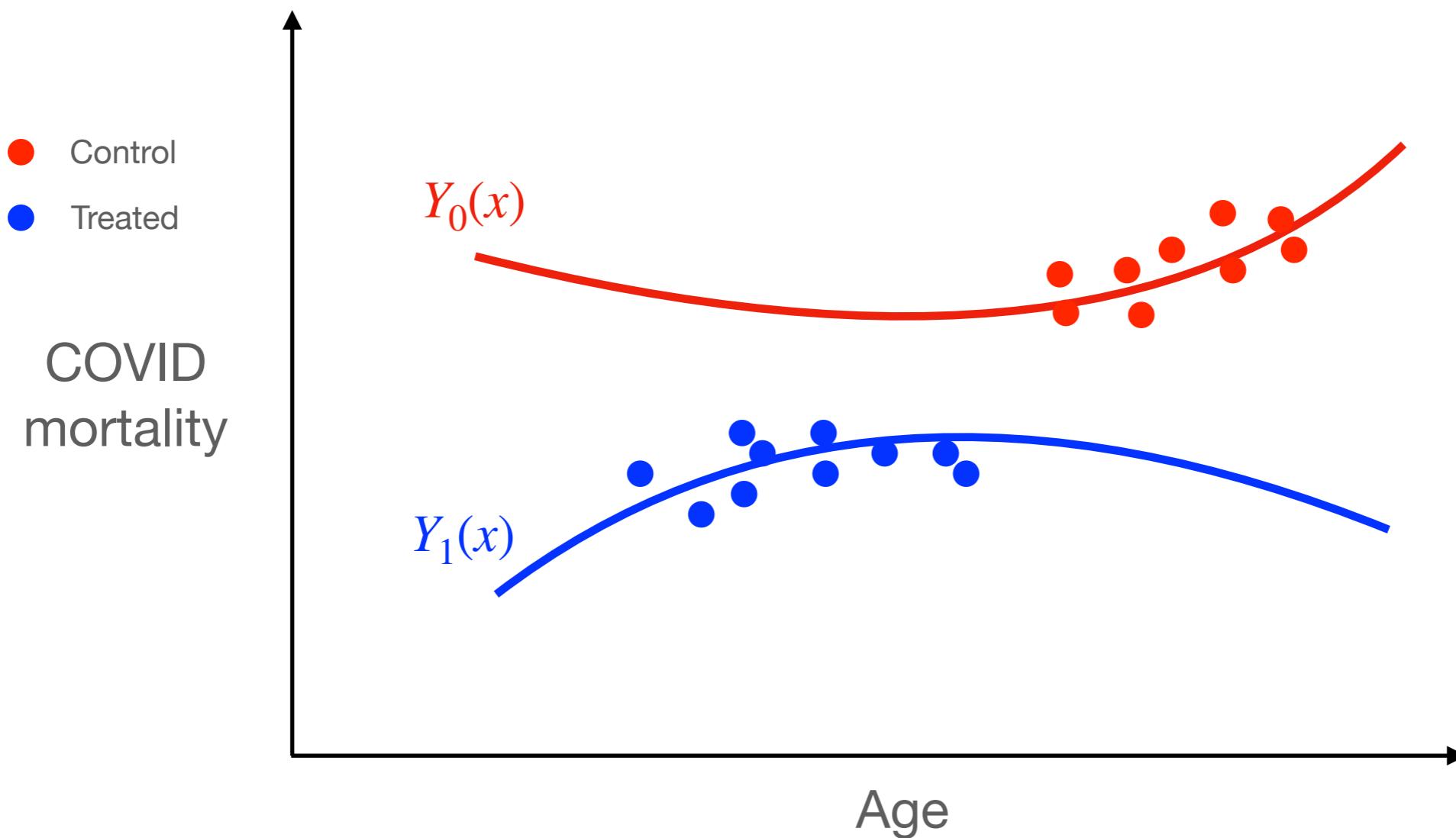
Parametric g -formula

- Goal: learn a function $f(x, T)$ that predicts Y .
- I.e., fit a model $\mathbb{E}(Y_t \mid T = t, x)$.
- Then,

$$\widehat{\text{ATE}} = \frac{1}{n} \sum_{i=1}^n f(x_i, 1) - f(x_i, 0).$$

- Additionally, $\widehat{\text{CATE}}(x_i) = f(x_i, 1) - f(x_i, 0)$.

What if no overlap?



Covariate adjustment

- Let's assume that

$$Y_t(x) = \underbrace{\beta x}_{\text{age}} + \underbrace{\gamma t}_{\text{medication}} + \epsilon_t.$$

- $\mathbb{E}\epsilon_t = 0$.
- Then $\text{CATE}(x) = \mathbb{E}(Y_1(x) - Y_0(x)) = \gamma$.
- Consequently, $\text{ATE}(x) = \gamma$ too.

If the true model is not linear

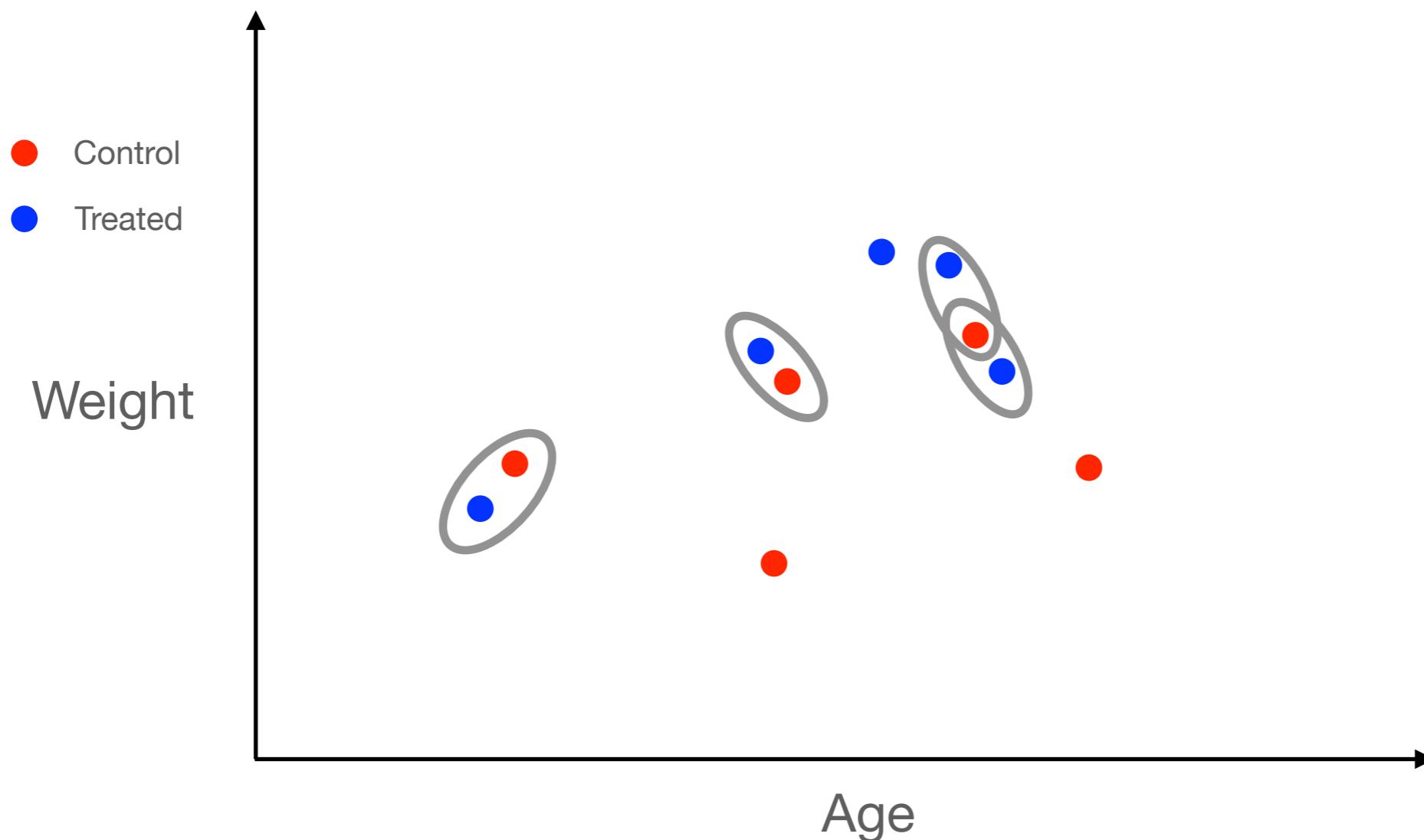
- Suppose the true model is $Y_t(x) = \beta x + \gamma t + \delta x^2$.
- If we were to fit $\hat{Y}_t(x) = \hat{\beta}x + \hat{\gamma}t$, we would find that

$$\hat{\gamma} = \gamma + \delta \frac{\mathbb{E}(xt)\mathbb{E}(x^2) - \mathbb{E}(t^2)\mathbb{E}(x^2t)}{\mathbb{E}(xt)^2 - \mathbb{E}(x^2)\mathbb{E}(t^2)}.$$

- Our estimate of γ could be arbitrarily bad based on the magnitude of δ .
- This is what makes covariate adjustment dangerous.

Matching

- For each unit, match them with their counterfactual twin.



1-NN matching

- Assume $d(\cdot, \cdot)$ is a distance metric on the x_i .
- For each i let $j(i) = \arg \min_{t_j \neq t_i} (x_i, x_j)$. Thus, $j(i)$ is the closest neighbor to i that received a different treatment.
- If $t_i = 1$ then define $\widehat{\text{CATE}}(x_i) = y_i - y_{j(i)}$.
- If $t_i = 0$ then define $\widehat{\text{CATE}}(x_i) = y_{j(i)} - y_i$.
- (Equivalently, $\widehat{\text{CATE}}(x_i) = (2t_i - 1)(y_i - y_{j(i)})$.)

Pros and cons

- Pros:
 - Interpretable, especially with small sample size.
 - Nonparametric.
- Cons:
 - Results strongly depend on metric $d(\cdot, \cdot)$.
 - Potentially misled by features which are not related to the outcome.

Matching and covariate adjustment

- Matching is equivalent to covariate adjustment using 1-NN classifiers:
 - $\hat{Y}_1(x) = Y_{\text{NN}(1)}(x)$
 - $\hat{Y}_0(x) = Y_{\text{NN}(0)}(x)$
- 1-NN matching is inconsistent, but the bias is small.

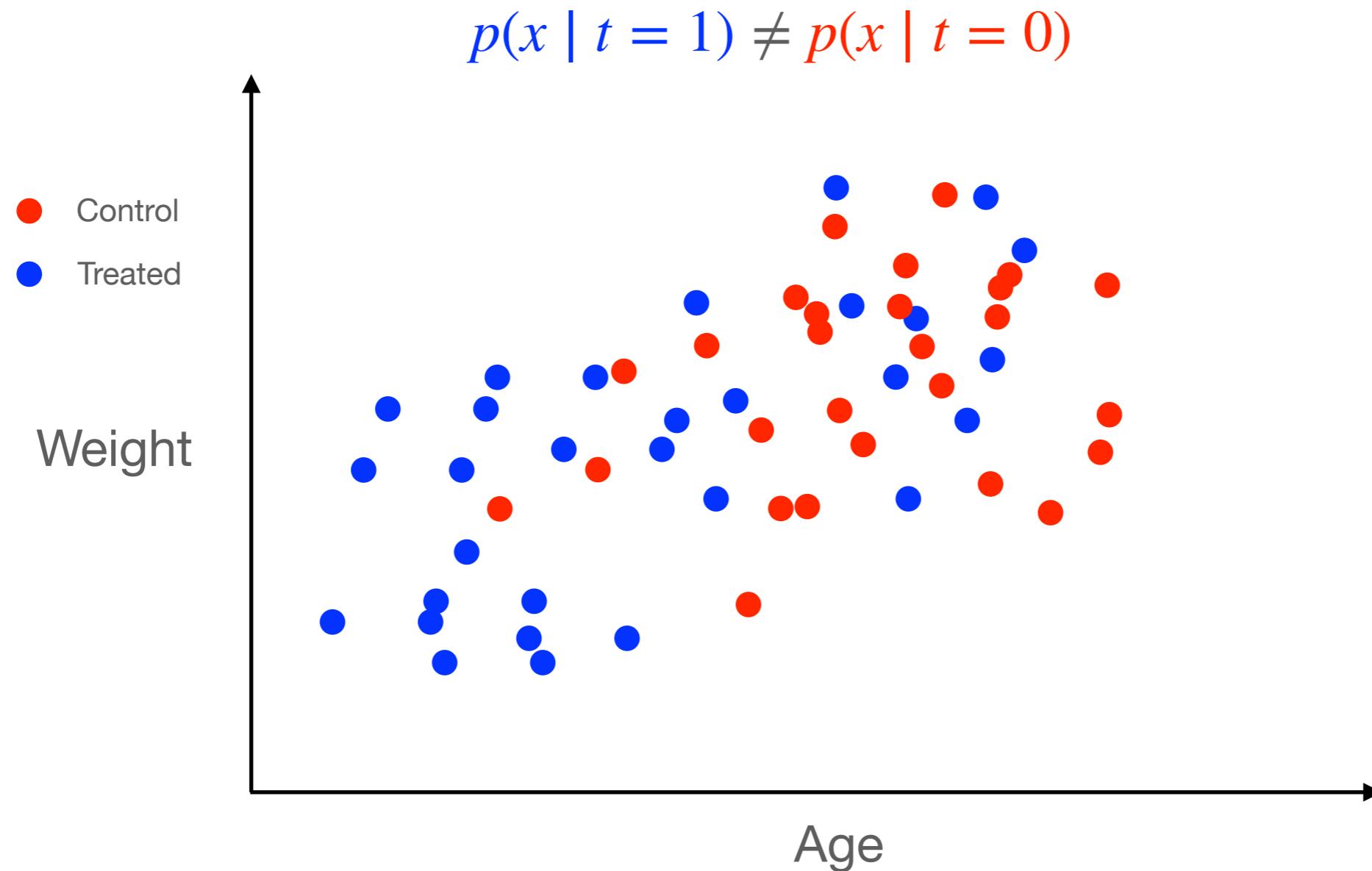
Propensity scoring (IPW)

- Imagine we had data from a randomized controlled trial (RCT).
- We could simply estimate

$$\text{ATE} = \frac{1}{n_1} \sum_{i:T_i=1} Y_i - \frac{1}{n_0} \sum_{i:T_i=0} Y_i.$$

- As we saw, we cannot conduct RCTs in many settings of interest.
- Idea: turn observational study into a pseudo-RCT by reweighting the observations.

Propensity scoring



Propensity score

- Propensity score $P(T = 1 \mid x)$, the propensity for being treated given covariates x . (Learn this using logistic regression, ML, etc.)
- Note this is different than covariate adjustment, where we learned $P(Y_t \mid x, t)$.
- Samples are then re-weighted by 1/propensity score of their treatment.

Propensity score algorithm

1. Use some method to estimate $P(T = 1 \mid x)$.

2. $\widehat{\text{ATE}} = \frac{1}{n} \sum_{t_i=1} \frac{y_i}{\hat{P}(t_i = 1 \mid x_i)} - \frac{1}{n} \sum_{t_i=0} \frac{y_i}{\hat{P}(t_i = 0 \mid x_i)}$.

Example: randomized controlled trial

- In an RCT, $p(T_i = 1) = 1/2$ irrespective of x .
- So,

$$\widehat{\text{ATE}} = \frac{2}{n} \sum_{t_i=1} y_i - \frac{2}{n} \sum_{t_i=0} y_i.$$

- Since each sum consists of about $n/2$ terms, we get that this is a consistent estimator of $\mathbb{E}(Y_1 - Y_0)$.

Propensity scores

Derivation

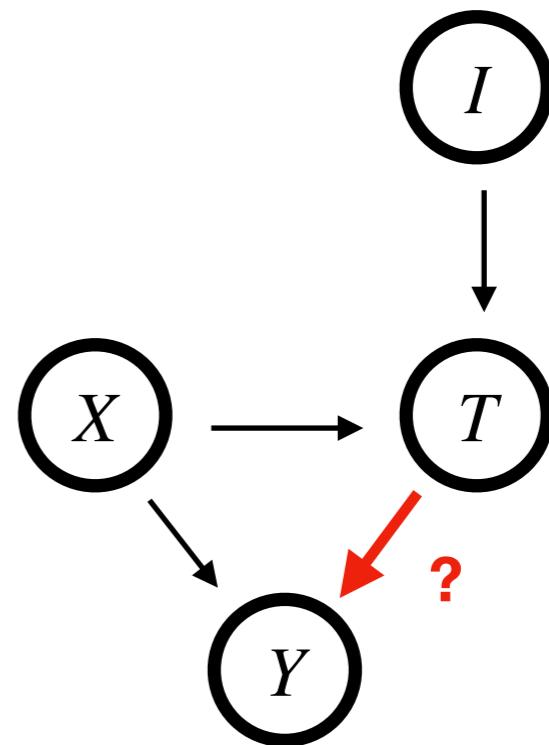
- Recall $\text{ATE} = \mathbb{E}_{x \sim p(x)}[Y_1(x) - Y_0(x)]$.
- $p(x) = p(x \mid T = 1) \frac{p(T = 1)}{p(T = 1 \mid x)}$
- $\mathbb{E}_{x \sim p(x)} Y_1(x) = \mathbb{E}_{x \sim p(x|T=1)} \left[\frac{p(T = 1)}{p(T = 1 \mid x)} Y_1(x) \right]$
- Plug in $p(T = 1) = n_1/n$ and our propensity score estimator $\hat{p}(T = 1 \mid x)$.

Problems with IPW

- Need to estimate the propensity score.
- High-dimensional non-parametric regression 😬
- Without much overlap, propensity scores become non informative.
- Weighting by $1/\text{prob}$ can result in large variances.

Instrumental variables

- An **instrument** is a variable which affects treatment assignment but not the outcome.
- Example: smoking and health.
 - As noted, we cannot do a RCT on smoking.
 - The tax rate on cigarettes is a possible instrument.



Estimating IV models

Wald's estimator

- Suppose that $y = x\beta + \epsilon$ but $\mathbb{E}(\epsilon x) \neq 0$.
- OLS estimates of β are biased.
- Now let z be an **instrument**: $\mathbb{E}(xz) \neq 0$ and $\mathbb{E}(\epsilon z) = 0$.
- Then

$$\frac{\text{cov}(y, z)}{\text{cov}(x, z)} = \frac{\beta \text{cov}(x, z) + \text{cov}(\epsilon, z)}{\text{cov}(x, z)} = \beta .$$

Natural experiments

- A **natural experiment** is particularly powerful form of IV analysis where a subgroup is exposed to a condition due to some exogenous natural phenomenon.
- (The condition needs to be correlated with the intervention we are interested in studying.)



Military service and lifetime earnings

- How does having served in the military affect lifetime earnings?
- Both may be affected by other factors like education, family wealth, ...
- During the Vietnam war, induction into the military was randomly assigned based on birth date.
- Thus, birth date become as instrument for military service.
- Findings:
 - White men suffered a 15 percent earnings penalty in the 1980s for being drafted, while black men experienced no penalty.
 - If you already faced discrimination, losing two years of civilian experience didn't make a difference.

Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records

By JOSHUA D. ANGRIST*

The randomly assigned risk of induction generated by the draft lottery is used to construct estimates of the effect of veteran status on civilian earnings. These estimates are not biased by the fact that certain types of men are more likely than others to serve in the military. Social Security administrative records indicate that in the early 1980s, long after their service in Vietnam was ended, the earnings of white veterans were approximately 15 percent less than the earnings of comparable nonveterans. (JEL 824)

A central question in the debate over military manpower policy is whether veterans are adequately compensated for their service. The political process clearly reflects the desire to compensate veterans: since World War II, millions of veterans have enjoyed benefits for medical care, education and training, housing, insurance, and job placement. Recent legislation provides additional benefits for veterans of the Vietnam era. Yet, academic research has not shown conclusively that Vietnam (or other) veterans are worse off economically than nonveterans. Many studies find that Vietnam veterans earn less than nonveterans, but others find positive effects, or effects that vary with age and

schooling. Regarding the general position of veterans, a member of the Twentieth Century Fund's Task Force on Policies Toward Veterans concludes that "Within any age group, veterans have higher incomes, more education, and lower unemployment rates than their nonveteran counterparts."¹

The goal of this paper is to measure the long-term labor market consequences of military service during the Vietnam era. Previous research comparing civilian earnings by veteran status may be biased by the fact that certain types of men are more likely to serve in the armed forces than others. For example, men with relatively few civilian opportunities are probably more likely to enlist. Estimation strategies that do not control for differences in civilian earnings potential will incorrectly attribute lower civilian earnings of veterans to military service. The research reported here overcomes such statistical problems by using the Vietnam era draft

*Department of Economics, Harvard University, Cambridge, MA 02138. Grateful thanks go to Warren Buckler, Cresson Smith, Ada Enis, and Bea Matsui for their assistance in producing the Social Security data; to Chester Bowie for his help in producing the SIPP data; and to Mike Dove for providing DMDC administrative records. Special thanks also go to David Card and Whitney Newey, from whose instruction and comments I have benefited greatly, and to Alan Krueger and an anonymous referee, whose careful reviews of an earlier draft led to substantial improvement. Data collection for this project was funded by the Princeton Industrial Relations Section. Funds for computation and financial support of the author were provided by the Industrial Relations Section, the Princeton Department of Economics, the Sloan Foundation, and the Olin Foundation.

¹The quote is from Michael Taussig (1974, p. 51). Legislation pertaining to veterans benefits is outlined in Veterans Administration (1984) and in other annual reports of the Veterans Administration. Studies by Sherwin Rosen and Paul Taubman (1982), Saul Schwartz (1986), and Jon Crane and David Wise (1987) find that Vietnam veterans earn less than nonveterans. Dennis DeTray (1982) and Mark Berger and Barry Hirsch (1983) find some positive effects for different age and schooling classes, and Veterans Administration (1981a) researchers find an overall positive effect.

Childbearing and labor supply

- Does having children cause women to become less likely to work?
- Difficult to establish causality because both fertility and labor supply are jointly determined by other things.
- We cannot randomly assign the number of children people have, **but ...** people have a strong preference for gender mix.
- An indicator of whether the first and second child have the same gender is therefore a valid instrument for the overall number of children.

Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size

By JOSHUA D. ANGRIST AND WILLIAM N. EVANS *

Research on the labor-supply consequences of childbearing is complicated by the endogeneity of fertility. This study uses parental preferences for a mixed sibling-sex composition to construct instrumental variables (IV) estimates of the effect of childbearing on labor supply. IV estimates for women are significant but smaller than ordinary least-squares estimates. The IV are also smaller for more educated women and show no impact of family size on husbands' labor supply. A comparison of estimates using sibling-sex composition and twins instruments implies that the impact of a third child disappears when the child reaches age 13. (JEL J13, J22)

An understanding of the relationship between fertility and labor supply is important for a number of theoretical and practical reasons. First, economists and demographers have developed a variety of models linking the family and the labor market. Empirical studies of childbearing and labor supply are sometimes seen as tests of these models (e.g., Reuben Gronau, 1973; Mark R. Rosenzweig and Kenneth I. Wolpin, 1980b; T. Paul Schultz, 1990). Second, the link between fertility and labor supply might partly explain the postwar increase in women's labor-force participation rates if having fewer children causes an increase in labor-force attachment (Mary T. Coleman and John Pencavel, 1993). Evidence for this thesis includes Claudia Goldin's (1995) study, which shows that few women in the 1940's and 1950's birth cohorts

were able to combine childbearing with strong labor-force attachment. Other researchers have also drawn a link between fertility-induced withdrawals from the labor force and lower wages of women (e.g., Gronau, 1988; Sanders Korenman and David Neumark, 1992). So perhaps childbearing keeps women from developing their careers.

Any success in disentangling the causal mechanisms linking fertility and labor supply should shed light on other substantive issues as well. For example, reductions in female labor supply could increase the total time parents devote to child care, making at least some children better off (see, e.g., Frank P. Stafford, 1987; Francine Blau and Adam J. Grossberg, 1992). Some theories of family behavior also suggest that changes in wives' earnings affect marital stability (Becker et al., 1977; Becker, 1985).

Not surprisingly, given the wide and long-standing interest in the connection between childbearing and labor supply, hundreds of empirical studies report estimates of this relationship. The vast majority of these studies find a negative correlation between fertility (or family size) and female labor supply.¹ As

* Angrist: Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02139; Evans: Department of Economics, University of Maryland, College Park, MD 20742. The authors thank seminar participants at the Council of Economic Advisors, Brown University, the University of Maryland, MIT, the University of Michigan, Northwestern University, Princeton University, the University of Virginia, and the July 1996 NBER Labor Studies meetings, as well as Susan Athey, Peter Diamond, Jon Gruber, Judy Hellerstein, Jim Poterba, David Weil, and three anonymous referees for helpful discussions or comments. John Johnson and Amanda Honeycutt provided excellent research assistance. Special thanks go to Duncan Thomas who stimulated our interest in the subject of parental sex preferences. The authors bear sole responsibility for the content of this paper.

¹ There is less work on the effects of children on husbands' labor supply. See Pencavel (1986 Table 1.17) for a few estimates, which suggest a positive association between fathers' labor supply and the number of children. The relationship between husbands' and wives' labor sup-

Schooling and wages

- Effect of schooling on lifetime wages. Quarter of birth is an instrument, because:
 - You can only enter school if 5 years old by December 31;
 - You can only (legally) drop out if you are older than 16.
 - Thus, being born in the beginning of the year, you start school at an older age, and can drop out after completing less schooling.
- (Consider being born on January 1 vs December 31.)

Chapter 30

THE CAUSAL EFFECT OF EDUCATION ON EARNINGS

DAVID CARD*

Department of Economics, University of California at Berkeley

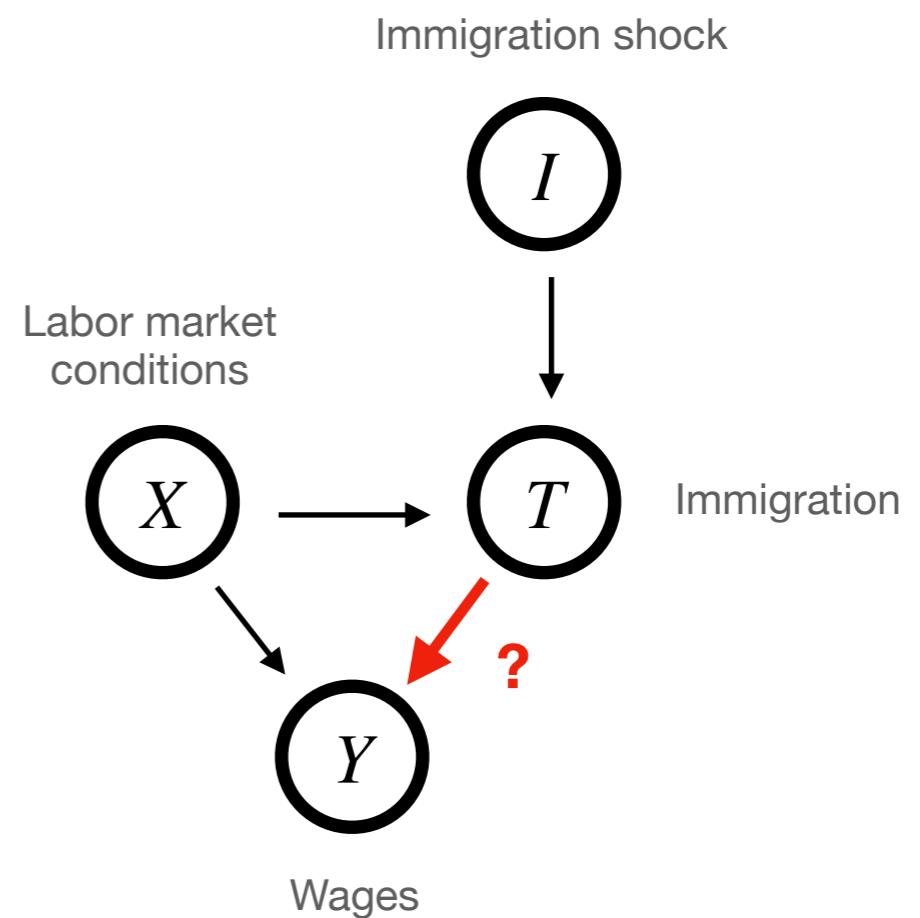
Contents

Abstract	1802
JEL codes	1802
1 Introduction and overview	1802
2 The human capital earnings function	1803
2.1 Functional form	1804
2.2 Measurement of education	1806
2.3 Which measure of earnings?	1808
2.4 Summary	1809
3 Causal modelling of the return to education	1810
3.1 Theoretical issues	1810
3.2 Observed schooling and earnings outcomes	1813
3.3 Measurement error	1815
3.4 Instrumental variables estimates of the return to schooling	1817
3.5 Limitations of instrumental variables	1819
3.6 Family background	1822
3.7 Models for siblings and twins	1826
3.8 Summary	1831
4 A selective review of recent empirical studies	1834
4.1 Instrumental variables based on institutional features of the school system	1834
4.2 Estimators using family background as a control or instrument	1842
4.3 Studies of education and earnings using twins	1846
4.4 Direct evidence on the heterogeneity in returns to education	1852
5 Conclusions	1855
Appendix A	1856
A.1 OLS estimation of a random coefficients model	1856
A.2 Estimation of a random coefficients model	1857
A.3 Measurement error in a bivariate regression model	1858
References	1859

* I am grateful to David Lee and Gena Estes for research assistance, to Orley Ashenfelter, Alan Krueger, and James Powell for helpful discussions, and to Michael Boozer, Ken Chay, Andrew Hildreth and Gary Solon for comments on earlier drafts that substantially improved the chapter. This research was funded in part by a grant from the NICHD.

Example: does immigration depress wages?

- Some politicians and economists claim that immigration drives down wages.
- Empirical studies consistently show no effect.
 - Labor market mobility?
 - Cannot do an RCT. Best we could hope for is something that exogenously altered the labor supply.



THE MARIEL BOATLIFT



No effect on wages

- The boatlift increased the Miami labor force by 7% in a matter of months.
- “Nevertheless, the Mariel influx appears to have had virtually no effect on the wages or unemployment rates of less-skilled workers, even among Cubans who had immigrated earlier.”

THE IMPACT OF THE MARIEL BOATLIFT ON THE MIAMI LABOR MARKET

DAVID CARD*

Using data from the Current Population Survey, this paper describes the effect of the Mariel Boatlift of 1980 on the Miami labor market. The Mariel immigrants increased the Miami labor force by 7%, and the percentage increase in labor supply to less-skilled occupations and industries was even greater because most of the immigrants were relatively unskilled. Nevertheless, the Mariel influx appears to have had virtually no effect on the wages or unemployment rates of less-skilled workers, even among Cubans who had immigrated earlier. The author suggests that the ability of Miami's labor market to rapidly absorb the Mariel immigrants was largely owing to its adjustment to other large waves of immigrants in the two decades before the Mariel Boatlift.

ONE of the chief concerns of immigration policy-makers is the extent to which immigrants depress the labor market opportunities of less-skilled natives. Despite the presumption that an influx of immigrants will substantially reduce native wages, existing empirical studies suggest that the effect is small. (See the survey by Greenwood and McDowell [1986] and studies by Grossman [1982], Borjas [1987], and Lalonde and Topel [1987].) There are two leading explanations for this finding. First, immigrants have, on average, only slightly lower skills than the native population. Thus, econometric studies based on the distribution of the existing stock of immigrants probably underestimate the effect of unskilled immigration on less-skilled

* The author is Professor of Economics, Princeton University. He thanks George Borjas, Alan Krueger, Bruce Meyer, and seminar participants at Princeton University for their comments.

A data appendix with copies of the computer programs used to generate the tables in this paper is available from the author at the Industrial Relations Section, Firestone Library, Princeton University, Princeton, NJ 08544.

natives. Second, the locational choices of immigrants and natives presumably depend on expected labor market opportunities. Immigrants tend to move to cities where the growth in demand for labor can accommodate their supply. Even if new immigrants cluster in only a few cities (as they do in the United States), inter-city migration of natives will tend to offset the adverse effects of immigration.

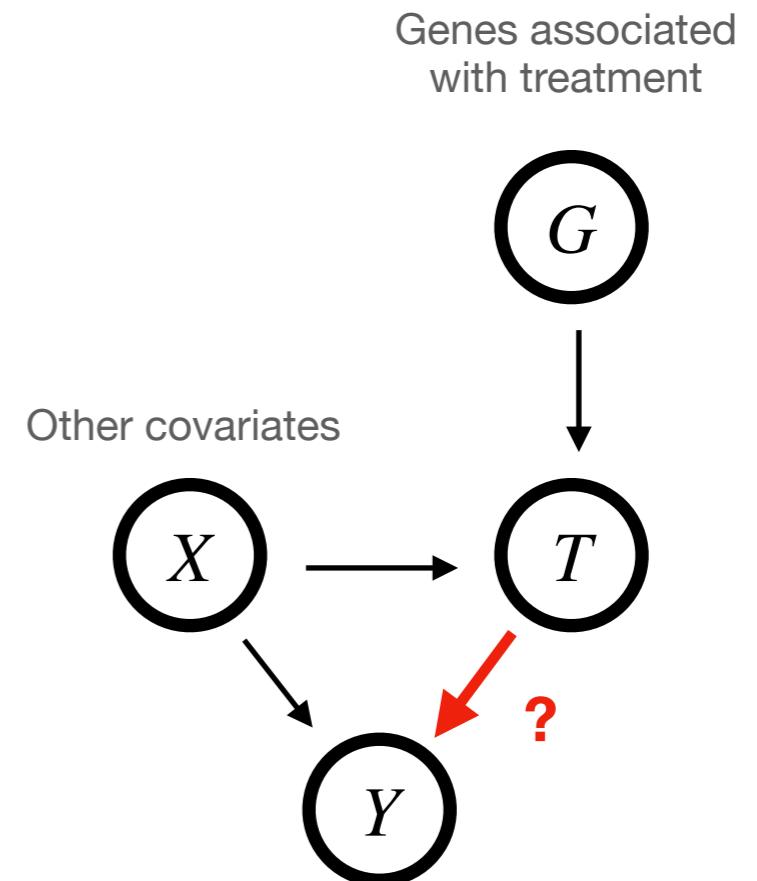
These considerations illustrate the difficulty of using the correlation across cities between wages and immigrant densities to measure the effect of immigration on the labor market opportunities of natives. They also underscore the value of a natural experiment that corresponds more closely to an exogenous increase in the supply of immigrants to a particular labor market.

The experiences of the Miami labor market in the aftermath of the Mariel Boatlift form one such experiment. From May to September 1980, some 125,000 Cuban immigrants arrived in Miami on a flotilla of privately chartered boats. Their arrival was the consequence of an unlikely sequence of events culminating in Castro's

Industrial and Labor Relations Review, Vol. 43, No. 2 (January 1990). © by Cornell University.
0019-7939/90/4302 \$01.00

Mendelian randomization

- A form of natural experiment arising in genetics.
- Genes are randomly assigned at birth and cannot be modified, so there is no reverse causality.
- Genes which predispose you to a certain treatment act as a sort of natural randomized trial.
- (What could go wrong?)



MR to assess risks of smoking

- Smoking causes lung cancer, but causal links to other cancers is unclear.
- Alcohol consumption is positively associated with head and neck, esophageal, stomach, liver, and breast cancers, but inversely associated with kidney cancer and non-Hodgkin lymphoma.
- Genetic instruments explain 2.3% and 0.2%–0.3% of the variation in smoking initiation and alcohol consumption, respectively.
- Genetic predisposition to smoking is associated with a significant increased risk of cancer of the lung, oesophagus, cervix, and bladder and with a nonsignificant increased risk of head and neck and stomach cancer.

RESEARCH ARTICLE

Smoking, alcohol consumption, and cancer: A mendelian randomisation study in UK Biobank and international genetic consortia participants

Susanna C. Larsson  *, Paul Carter³, Siddhartha Kar   ⁴, Mathew Vithayathil   ⁵, Amy M. Mason   ^{6,7}, Karl Michaëlsson¹, Stephen Burgess  

¹ Department of Surgical Sciences, Uppsala University, Uppsala, Sweden, ² Unit of Cardiovascular and Nutritional Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden, ³ Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom, ⁴ MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, United Kingdom, ⁵ MRC Cancer Unit, University of Cambridge, Cambridge, United Kingdom, ⁶ British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom, ⁷ National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge and Cambridge University Hospitals, Cambridge, United Kingdom, ⁸ MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom

* susanna.larsson@ki.se



OPEN ACCESS

Citation: Larsson SC, Carter P, Kar S, Vithayathil M, Mason AM, Michaëlsson K, et al. (2020) Smoking, alcohol consumption, and cancer: A mendelian randomisation study in UK Biobank and international genetic consortia participants. PLoS Med 17(7): e1003178. <https://doi.org/10.1371/journal.pmed.1003178>

Academic Editor: Konstantinos K. Tsilidis, University of Ioannina Medical School, GREECE

Received: October 10, 2019

Accepted: June 25, 2020

Published: July 23, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pmed.1003178>

Copyright: © 2020 Larsson et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Primary data from the UK Biobank resource are accessible upon application (<https://www.ukbiobank.ac.uk/>).

Abstract

Background

Smoking is a well-established cause of lung cancer and there is strong evidence that smoking also increases the risk of several other cancers. Alcohol consumption has been inconsistently associated with cancer risk in observational studies. This mendelian randomisation (MR) study sought to investigate associations in support of a causal relationship between smoking and alcohol consumption and 19 site-specific cancers.

Methods and findings

We used summary-level data for genetic variants associated with smoking initiation (ever smoked regularly) and alcohol consumption, and the corresponding associations with lung, breast, ovarian, and prostate cancer from genome-wide association studies consortia, including participants of European ancestry. We additionally estimated genetic associations with 19 site-specific cancers among 367,643 individuals of European descent in UK Biobank who were 37 to 73 years of age when recruited from 2006 to 2010. Associations were considered statistically significant at a Bonferroni corrected *p*-value below 0.0013. Genetic predisposition to smoking initiation was associated with statistically significant higher odds of lung cancer in the International Lung Cancer Consortium (odds ratio [OR] 1.80; 95% confidence interval [CI] 1.59–2.03; *p* = 2.26 × 10⁻²¹) and UK Biobank (OR 2.26; 95% CI 1.92–2.65; *p* = 1.17 × 10⁻²²). Additionally, genetic predisposition to smoking was associated with statistically significant higher odds of cancer of the oesophagus (OR 1.83; 95% CI 1.34–2.49; *p* = 1.31 × 10⁻⁴), cervix (OR 1.55; 95% CI 1.27–1.88; *p* = 1.24 × 10⁻⁵), and bladder