CS224C: NLP for CSS

Casual Inference for CSS

Diyi Yang

Stanford CS

Announcements

Homework 4 is out; due May 30th

Sharing your course project on our website [optional]

Poster session: 4-6pm on 6/6; in front of Bytes, in the grassy area

Final report is due on June 7th

Lecture Overview

- Prediction vs. Understanding
- Randomized controlled trail (RCT)
- Observation data and studies
- Propensity score methods
- Case studies
- Mediation analysis

Prediction vs. Understanding

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

Two main uses of statistical models:

Prediction: inferring the most likely values (+ prediction intervals) for data where you don't know the answer

Understanding: estimating the relationship between a predictor variable and some outcome (+ quantifying uncertainty about that relationship)

Starting with Regression

Logistic regression

$$P(y = 1 \mid x, \beta) = \frac{\exp(\sum_{i=1}^{F} x_i \beta_i)}{1 + \exp(\sum_{i=1}^{F} x_i \beta_i)}$$

Linear regression

$$y = \sum_{i=1}^{F} x_i \beta_i + \epsilon$$

Slides Credit to David Bamman

Features and Coefficients

x_i refers to each feature

such as “speaking English”, “mentioning Clinton on Twitter”

β_i refers to the coefficient associated with x_i

A Simple Example

$$P(y = 1 \mid x, \beta) = \frac{\exp(x_0\beta_0 + x_1\beta_1)}{1 + \exp(x_0\beta_0 + x_1\beta_1)}$$

x_0: whether the user speaks English

x_1: how many times the user mentions Clinton on Twitter

y: 1 if the user votes for Clinton, otherwise 0

A Simple Example

$$P(y = 1 \mid x, \beta) = \frac{\exp(x_0\beta_0 + x_1\beta_1)}{1 + \exp(x_0\beta_0 + x_1\beta_1))}$$

$$\frac{P(y = 1 \mid x, \beta)}{1 - P(y = 1 \mid x, \beta)} = \exp(x_0\beta_0 + x_1\beta_1) = \exp(x_0\beta_0) \cdot \exp(x_1\beta_1)$$

If x_1 increases by 1,

$$\exp(x_0\beta_0) \cdot \exp((x_1 + 1)\beta_1) = \exp(x_0\beta_0) \cdot \exp(x_1\beta_1 + \beta_1) = \exp(x_0\beta_0) \cdot \exp(x_1\beta_1) \cdot \exp(\beta_1)$$

$\exp(\beta)$ **refers to the factor by which the odds change with a 1-unit increase in** x

Slides Credit to David Bamman

Interpreting the coefficient for "explanation"

We can assess how significant is the relationship between a predictor and its outcome (aka correlations) with a hypothesis test

But are these reliable?

Can we add control variables?

Refined correlations!

Correlation vs. Causation

Understand the causal relationship of a treatment Z on some outcome Y

Treatment	Outcome
take a drug	cured of disease
graduate high school	earnings
cast John Goodman	box office
living in Berkeley	political preference

Slides Credit to David Bamman

Terminology

Treatment: $Z(0), Z(1)$

The predictor variable whose casual relationship we're interested in

Potential outcomes: Y=0, Y=1

The dependent variable

We're interested in the causal relationship between the treatment Z and Y

Slides Credit to David Bamman

Counterfactual

John doesn't brush his teeth (Z=0) and developed heart disease (Y=1)

What would have happened if he did brush his teeth (Z=1)?

For any data point, we only ever get to observe one outcome. We never observe the counterfactual.

Slides Credit to David Bamman

Observational Data

Hypothesis tests for observational data assess the relationship between variables but don't establish causality

Examples: if we intervened and relocated someone to Palo Alto, would they become liberal?

Slides Credit to David Bamman

Experimental Data

Data that allows you to perform an intervention and determine the value of some variable

- **Clinical data**: treatment vs. placebo
- **Web design:** one or two homepage designs
- **Political email campaigns:** one of two (differently worded) solicitations

A potential confound exists if any other variable is correlated with your intervention decision:

- E.g., users volunteering to receive a drug (and not the placebo)

Slides Credit to David Bamman

Randomization Experiments

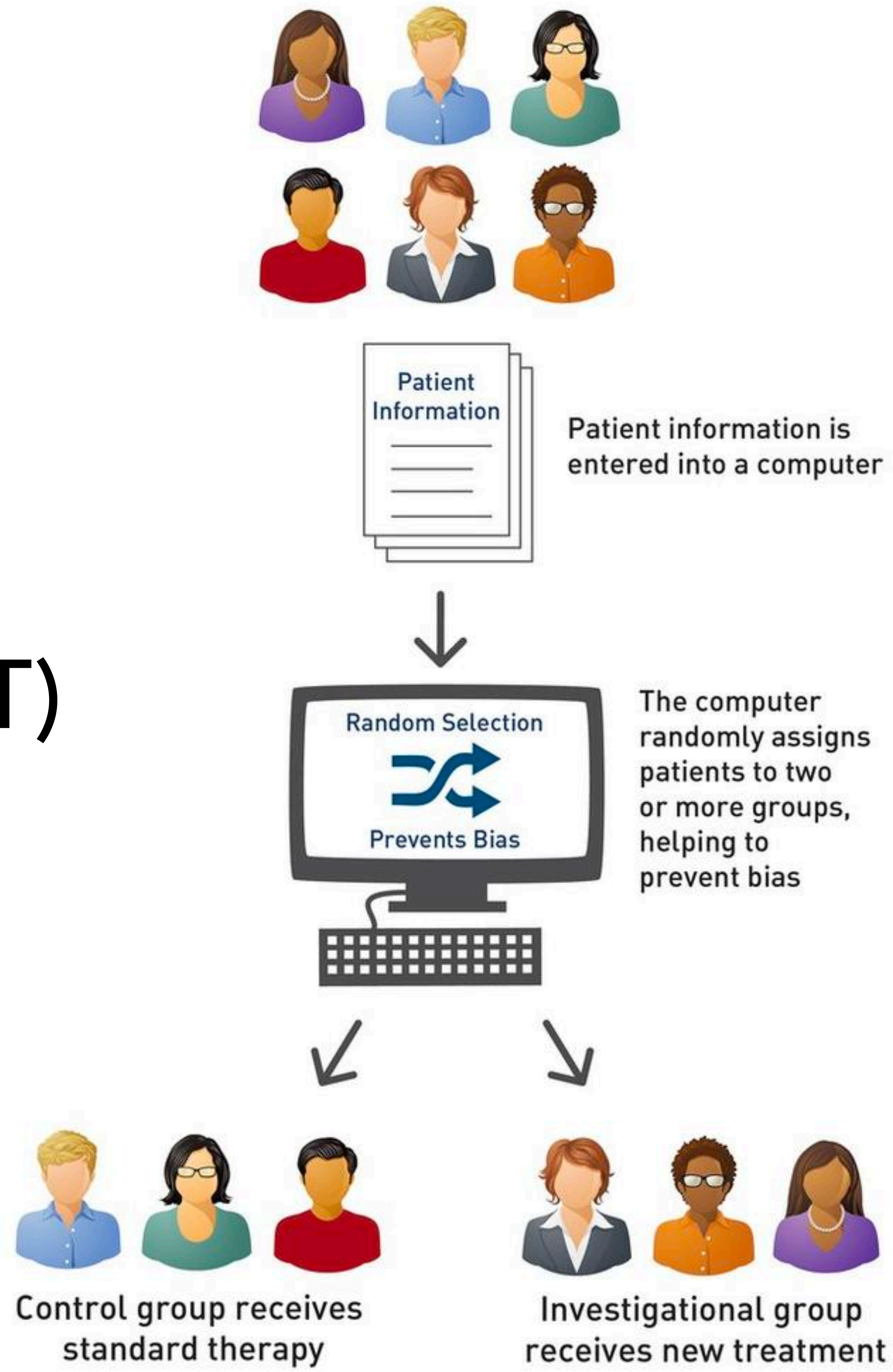
Users are randomly assigned an outcome (which web page), which allows us to better establish causality

A/B testing = significance test in randomized experiment with two outcomes

We can run a standard regression, but now if the β_{design_A} is significant, we can interpret it causally. By randomly assigning the treatment, we are ensuring that its value is uncorrelated with any other variable

Slides Credit to David Bamman

Randomized Control Trail (RCT)



https://www.cancer.gov/about-cancer/treatment/clinical-trials/what-are-trials/randomization/clinical-trial-randomization-infographic

RCT Estimation

$$E[Y(1) - Y(0)] = E[Y|Z = 1] - E[Y|Z = 0]$$

RCT gives an unbiased estimate of the average effect of the treatment

Randomization May Not Be Feasible

- Ethical Issues
- Controlled or the treatment conditions may be harmful

Observational Data

Observational data can’t be intervened to establish an causal relationship

Instead, we could:

- Accounting for confounding variables
- **Assume** there is a randomization experiment **lurking** in the data

Propensity Score

Propensity score: the probability of treatment assignment conditional on observed baseline covariates - also called a ***balancing score***

$$e_i = \boldsymbol{Pr}(Z_i = 1 \mid \mathbf{X_i})$$

In RCTs, propensity score is known and defined by the study design.
In observational studies, the true propensity score is not known, but can be estimated using the study data

Lecture Overview

- Prediction vs. Understanding
- Randomized controlled trail (RCT)
- Observation data and studies
- **Propensity score methods**

Four Propensity Score Methods

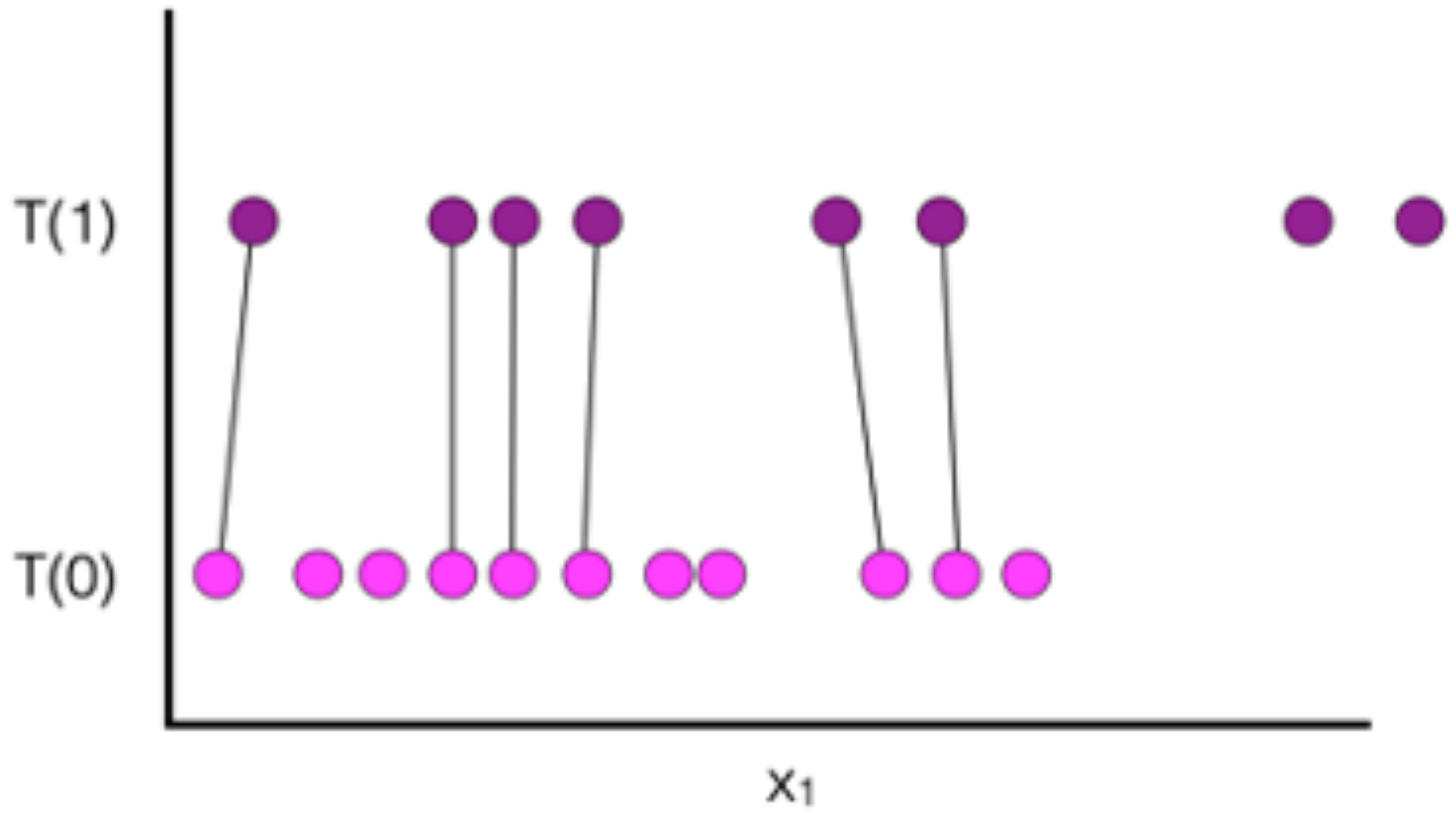
1. Propensity score matching
2. Stratification on the propensity score
3. Inverse probability of treatment weighting
4. Covariate adjustment using the propensity score

1 Propensity Score Matching

"Form matched sets of treated and untreated subjects who share a similar value of propensity score"

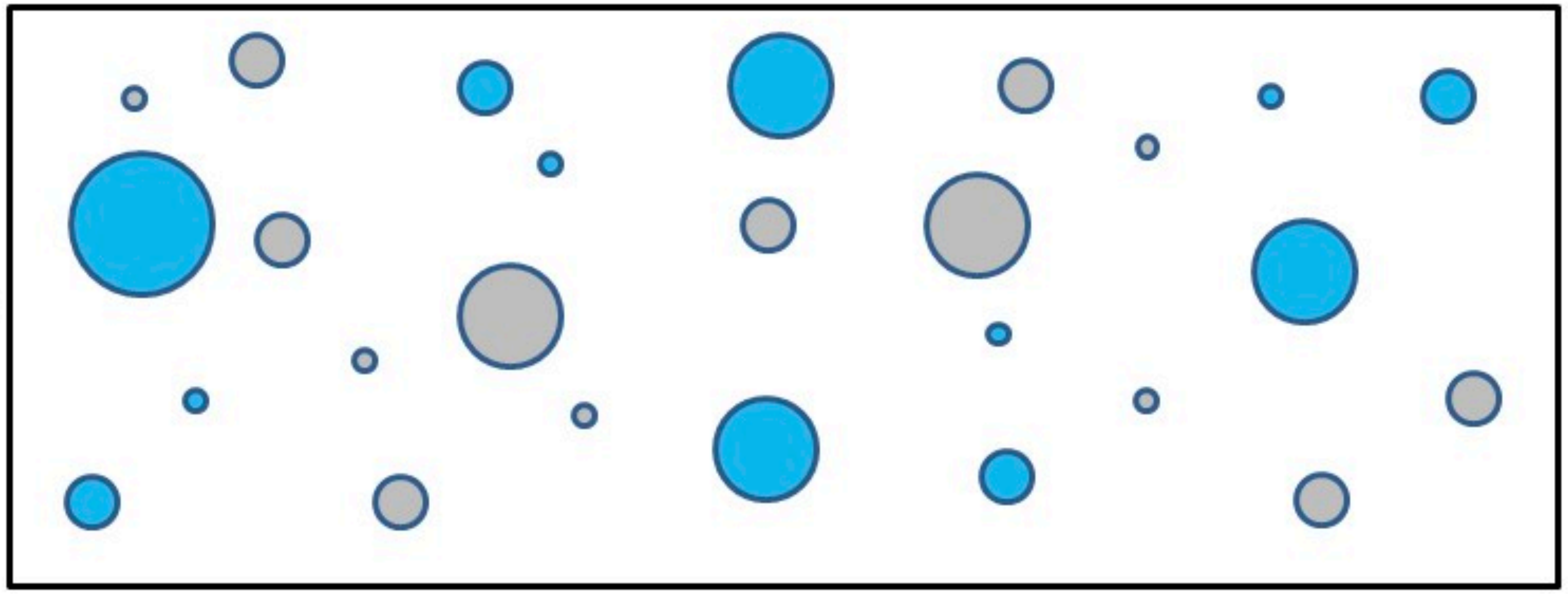
Common approach: **one-to-one or pair matching**

One can directly compare outcomes between treated and untreated subjects within the propensity score matched sample



Austin, Peter C. "An introduction to propensity score methods for reducing the effects of confounding in observational studies." Multivariate behavioral research 46, no. 3 (2011): 399-424.

Population with varying characteristics



https://www.summitllc.us/propensity-score-matching

Decisions on how to form matched pairs

1. Choose between matching ***without replacement*** and ***with replacement***

2. Go with ***greedy*** or ***optimal matching***

    - **Greedy**: a treated subject is *first selected* at random, and the untreated subject whose propensity score is closet to that is chosen for matching

    - **Optimal:** matches are formed so as to minimize the total within-pair difference of the propensity score

Decisions on how to determine the "close"

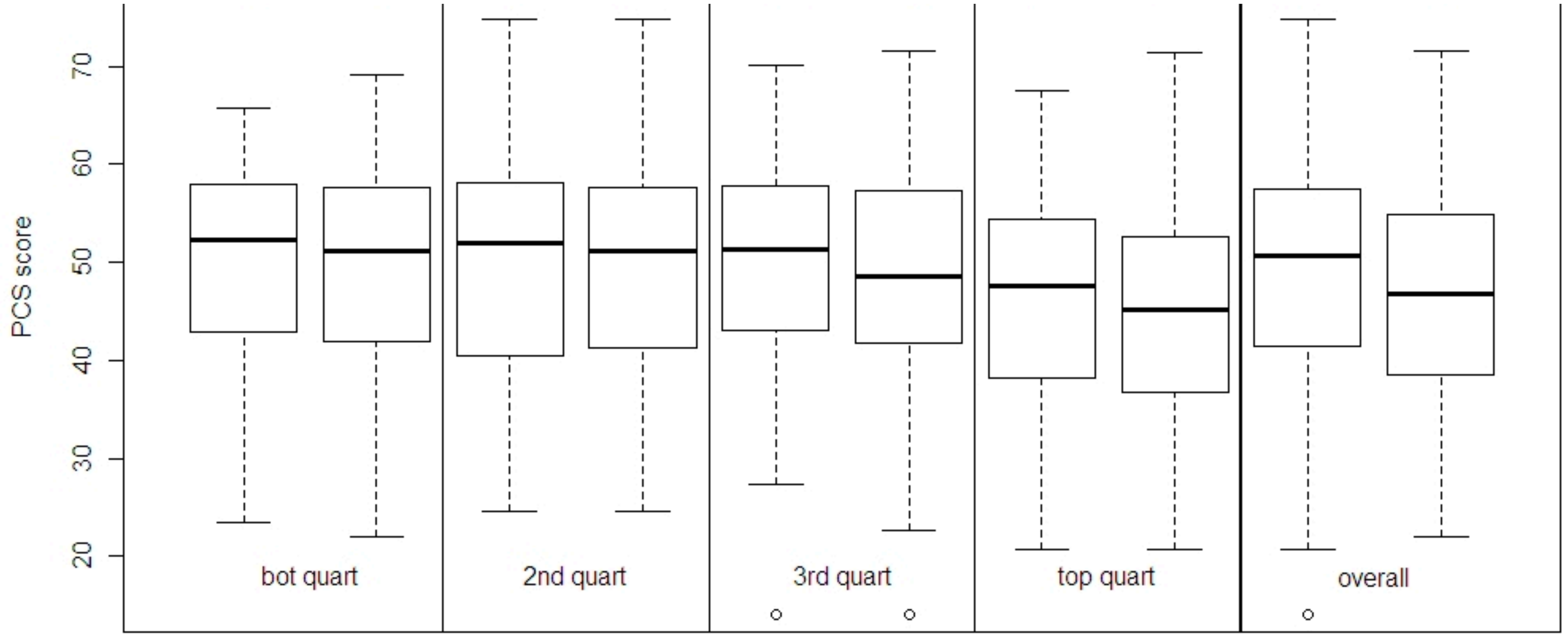
Two primary methods for selecting untreated subjects whose propensity score is "close" to that of a treated subject

- Nearest neighbor matching
- Nearest neighbor matching within a specified caliper distance

2. Stratification on the Propensity Score

- Stratify subjects into mutually exclusive subsets based on the rank-ordered propensity score.

- Overall treatment effect is pooled over stratum-specific treatment effects – *a meta-analysis of a set of quasi-RCTs*



http://sas-and-r.blogspot.com/2010/05/example-736-propensity-score.html

3 Inverse Probability of Treat Weighting (IPTW)

*IPTW using the propensity score c*reates weights based on the probability score to create a synthetic dataset

$$w_i = \frac{Z_i}{e_i} + \frac{(1 - Z_i)}{1 - e_i}$$

Aka, the *inverse probability* of the treatment received.

4 Regression Adjustment using Propensity Score

Outcome is regressed on an indicator of the treatment status and the estimated propensity scores.

Continuous outcome: linear models

Dichotomous outcome: logistic regression

The effect of treatment is determined using the estimated regression coefficient from the fitted regression model.

Comparing Different Propensity Score Methods

The shared goal

To remove confounding so that the treatment condition is independent of baseline characteristics between treated and untreated subjects

Differences:

Matching, stratification and weighting separate the design of the study from the analysis of the study, while regression requires both the propensity score and the outcome to be in the same model

Different tolerance to sensitivity

Primary study analysis method	Pros	Cons
Traditional covariate adjustment	• Performed well • Provides prognostic model for outcome of interest	• May not be suitable with many covariates in smaller studies
Propensity score (PS) stratification	• Retains data from all study participants • Opportunity to explore interactions between treatment and PS on outcome risk • Provides effect estimates for every stratum	• Performs less well in datasets with few outcomes, particularly when the number of strata is large • May not account for strong confounding
PS matching	• Reliable; provides excellent covariate balance in most circumstances • Simple to analyze, present and interpret	• Some patients are unmatched leading to information excluded from the analysis • Less precise
PS inverse probability weighting	• Retains data from all study participants • Easy to implement • Creates a pseudo population with perfect covariate balance	• Can be unstable when extreme weights occur
PS covariate adjustment (use of PS as a covariate)	• Performed well	• Adding the PS as an additional covariate produced results very similar (and not necessarily superior) to traditional covariate adjustment

Elze, Markus C., John Gregson, Usman Baber, Elizabeth Williamson, Samantha Sartori, Roxana Mehran, Melissa Nichols, Gregg W. Stone, and Stuart J. Pocock. "Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies." Journal of the American College of Cardiology 69, no. 3 (2017): 345-357.

Balance Diagnostics

"The true propensity score is a balancing score"

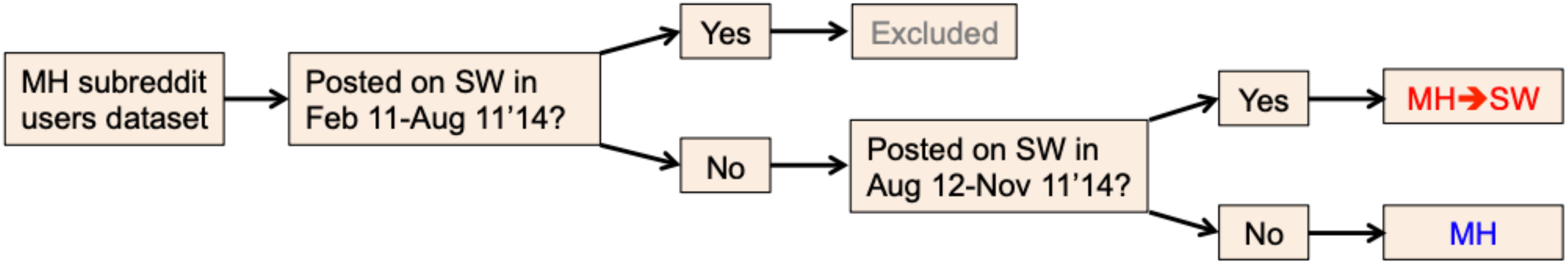
Standardized differences to compare the similarity of treated and untreated subjects in the matched samples

For continuous variables: $$d = \frac{(\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}})}{\sqrt{\frac{s^2_{\text{treatment}} + s^2_{\text{control}}}{2}}}$$

Lecture Overview

- Prediction vs. Understanding
- Randomized controlled trail (RCT)
- Observation data and studies
- Propensity score methods
- **Case studies**

Case Study 1: Discovering Shifts to Suicidal Ideation



De Choudhury, Munmun, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. "Discovering shifts to suicidal ideation from mental health content in social media." In Proceedings of the 2016 CHI conference on human factors in computing systems, pp. 2098-2110. 2016.

Case Study 1: Discovering Shifts to Suicidal Ideation

Understanding the possible casual factors in users' transitions from posting in MH to posting in SW by

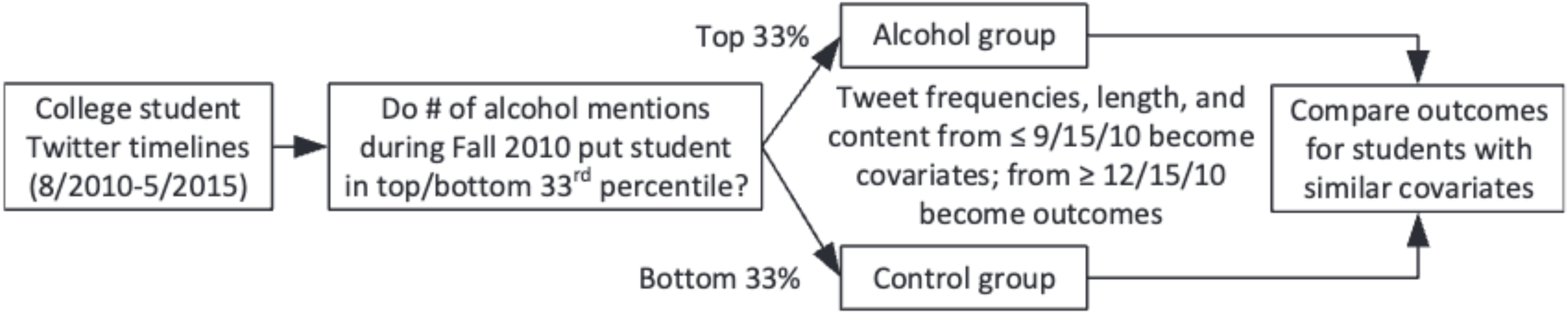
Estimate the effect of specific treatment *(e.g., the use of certain words in an MH post)* on a measured outcome *(e.g., the likelihood of transitioning to post in SW)* conditioned on confounding variables

Stratified propensity score matching achieves this by subdividing the treatment group and the control group into comparable groups based on the individuals' estimated propensity to use the token.

Differences btw MH-> SW and MH User Classes

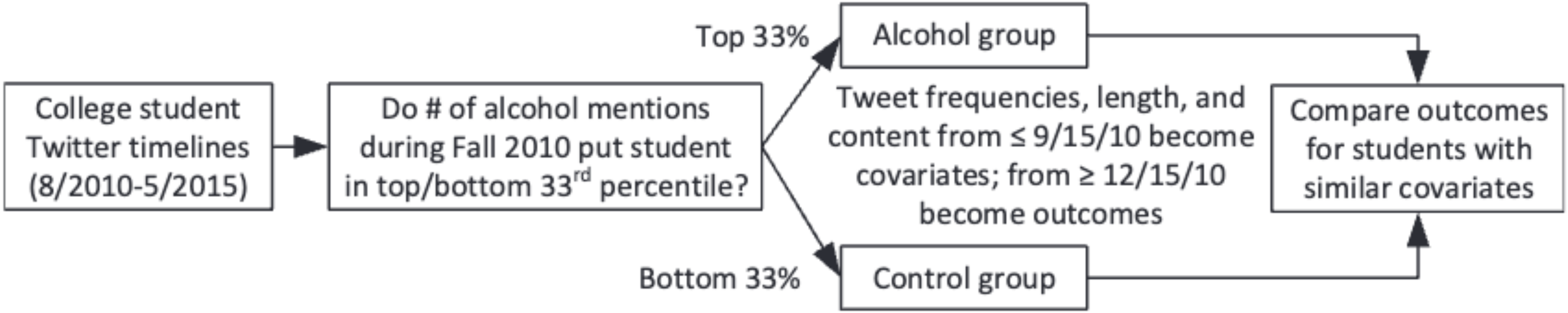
	MH	MH → SW	z	p
Linguistic Structure				
nouns	0.294	0.125	6.51	***
verbs	0.045	0.107	2.19	**
abverbs	0.048	0.099	4.87	***
readability index	0.609	0.232	5.51	***
accommodation	0.857	0.487	5.46	**
Interpersonal Awareness				
1st person singular	0.018	0.086	-10.6	***
1st person plural	0.093	0.078	4.53	*
2nd person	0.058	0.031	8.01	*
3rd person	0.087	0.042	6.32	***
Interaction				
posts authored	18.97	10.31	2.53	*
post length	215.62	443.73	-15.4	***
comments authored	122.42	106.22	0.95	-
comments received	19.862	13.414	1.05	*
comment length authored	63.417	87.116	-1.88	*
comment length received	42.323	26.362	5.44	**
response velocity (mins)	7.746	6.966	0.84	-
vote difference	28.788	7.681	7.18	***

Case Study 2: Understand the Effects of Early College Alcohol Use



Kiciman, Emre, Scott Counts, and Melissa Gasser. "Using longitudinal social media analysis to understand the effects of early college alcohol use." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 12, no. 1. 2018.

Case Study 2: Understand the Effects of Early College Alcohol Use

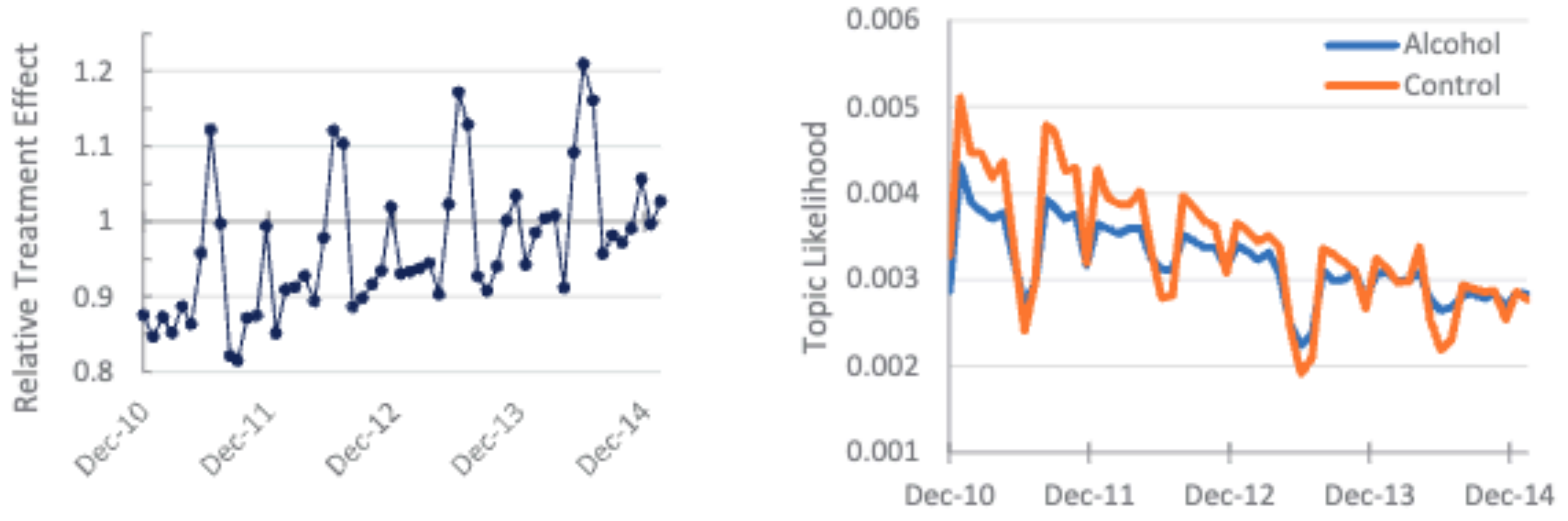


"The stratified propensity score analysis estimates missing counterfactual outcomes by identifying matching sub populations of individuals with similar distributions of covariates, but with differing treatment status"

Case Study 2: Understand the Effects of Early College Alcohol Use

Matching of groups is predicted via a propensity score model, which infers the likelihood of an individual being in the Alcohol group as a function of a set of covariates

Individuals with similar propensity scores are grouped into state



(a) Study Habits RTE (b) Study Habits Outcomes

Figure 2: Academic effects: People in the Alcohol group were significantly less likely ($p<.05$; effect size = .65) to mention studying over the next two years, and somewhat less likely ($p=.12$;effect size = .30) over the entire time period.

Lecture Overview

- Prediction vs. Understanding
- Randomized controlled trail (RCT)
- Observation data and studies
- Propensity score methods
- Case studies
- **Mediation analysis**

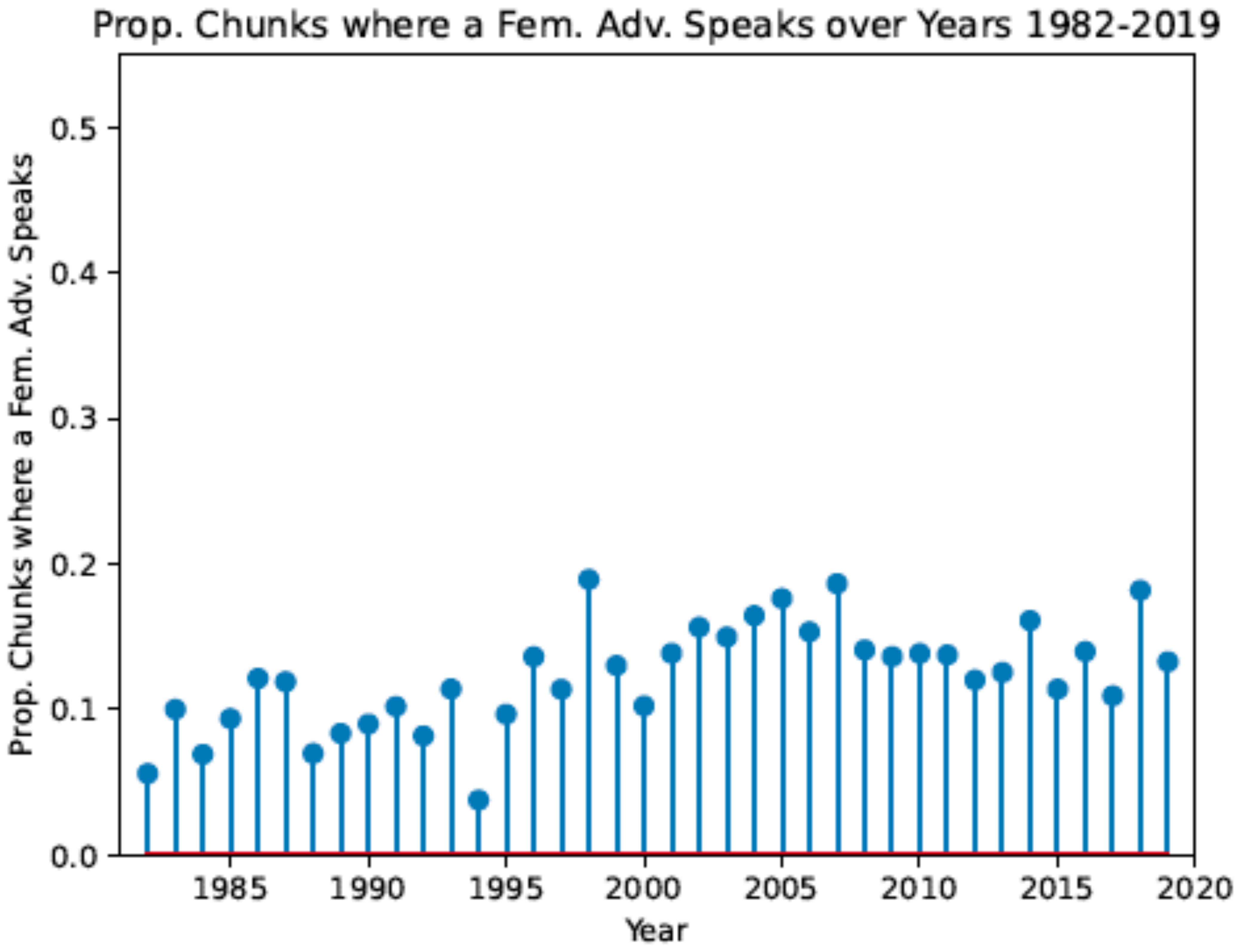
Estimating Gender Effects in Supreme Court Oral Arguments

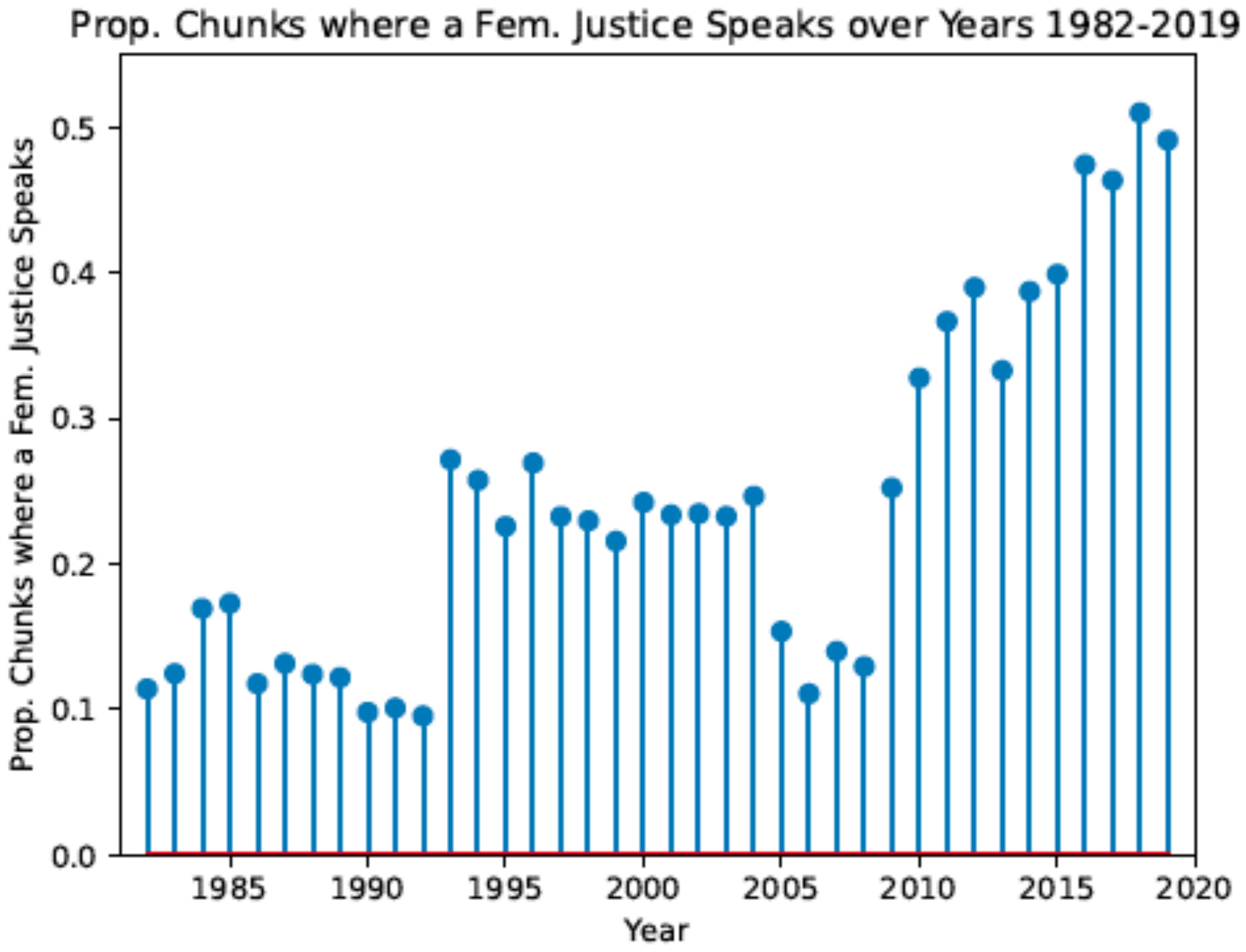
“If everything else in an oral argument had remained the same, but we swapped a female advocate for a male advocate, would judges have behaved differently? “

What would an **ideal experiment** look like?

Cai, Erica, Ankita Gupta, Katherine Keith, Brendan O'Connor, and Douglas R. Rice. ""Let Me Just Interrupt You": Estimating Gender Effects in Supreme Court Oral Arguments." SocArxiv (2023).

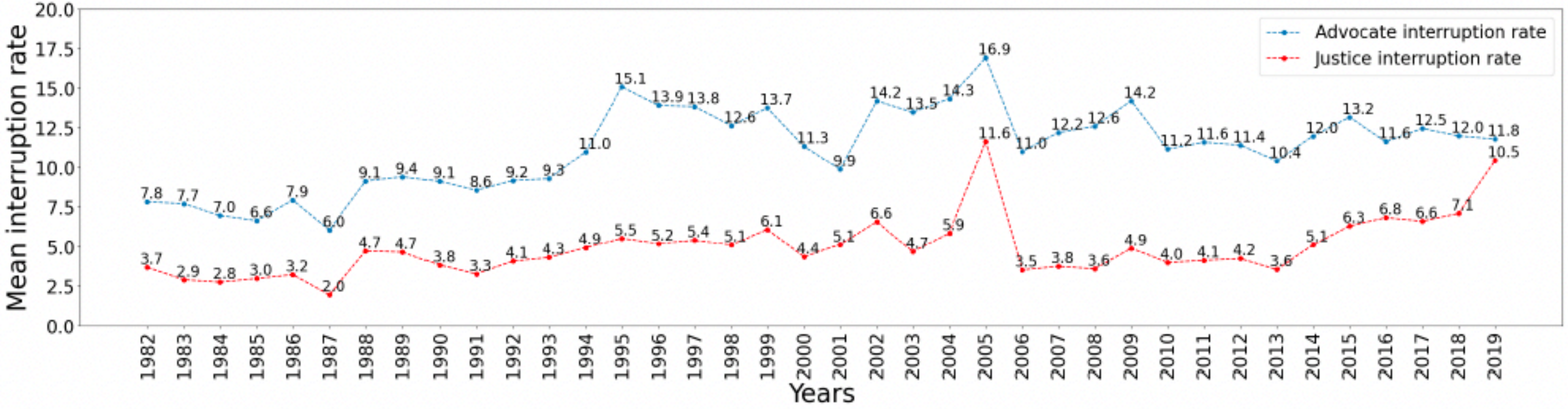
Participation of Women in Oral Argument Exchange on the US Supreme Court





Cai, Erica, Ankita Gupta, Katherine Keith, Brendan O'Connor, and Douglas R. Rice. ""Let Me Just Interrupt You": Estimating Gender Effects in Supreme Court Oral Arguments." SocArxiv (2023).

Interruption Rate Over time



Cai, Erica, Ankita Gupta, Katherine Keith, Brendan O'Connor, and Douglas R. Rice. ""Let Me Just Interrupt You": Estimating Gender Effects in Supreme Court Oral Arguments." SocArxiv (2023).

Formulation of Interruption Rate

Gender signal of an advocate (T)
Ideological alignment of an advocate and justice (A)
Token-normalized interruption rates (Y)

$$Y_{i|j} = \frac{\text{number of advocate utterances interrupted by justice } j \text{ in chunk } i}{(\text{number of advocate tokens in chunk } i)/1000}$$

For a given justice j and chunk i, define the unit-specific quantity of interruption rate given counterfactual genders as: $Y_{i|j}(T_i = F) - Y_{i|j}(T_i = M)$

Effects on Advocate Interruption Rate

Justices	θ_{Gender}	$\theta_{\text{Ideological Alignment}}$	$\frac{\theta_{\text{Gender}}}{\theta_{\text{Ideological Alignment}}}$
All	0.90±0.19	-0.25±0.12	3.60
Male	1.06±0.22	-0.20±0.13	5.30
Female	0.43±0.36	-0.39±0.24	1.10

$\theta_{IdeologicalAlignment}$ indicates justices interrupt ideologically opposed advocates more often than they interrupt ideologically aligned advocates

Effects on Advocate Interruption Rate

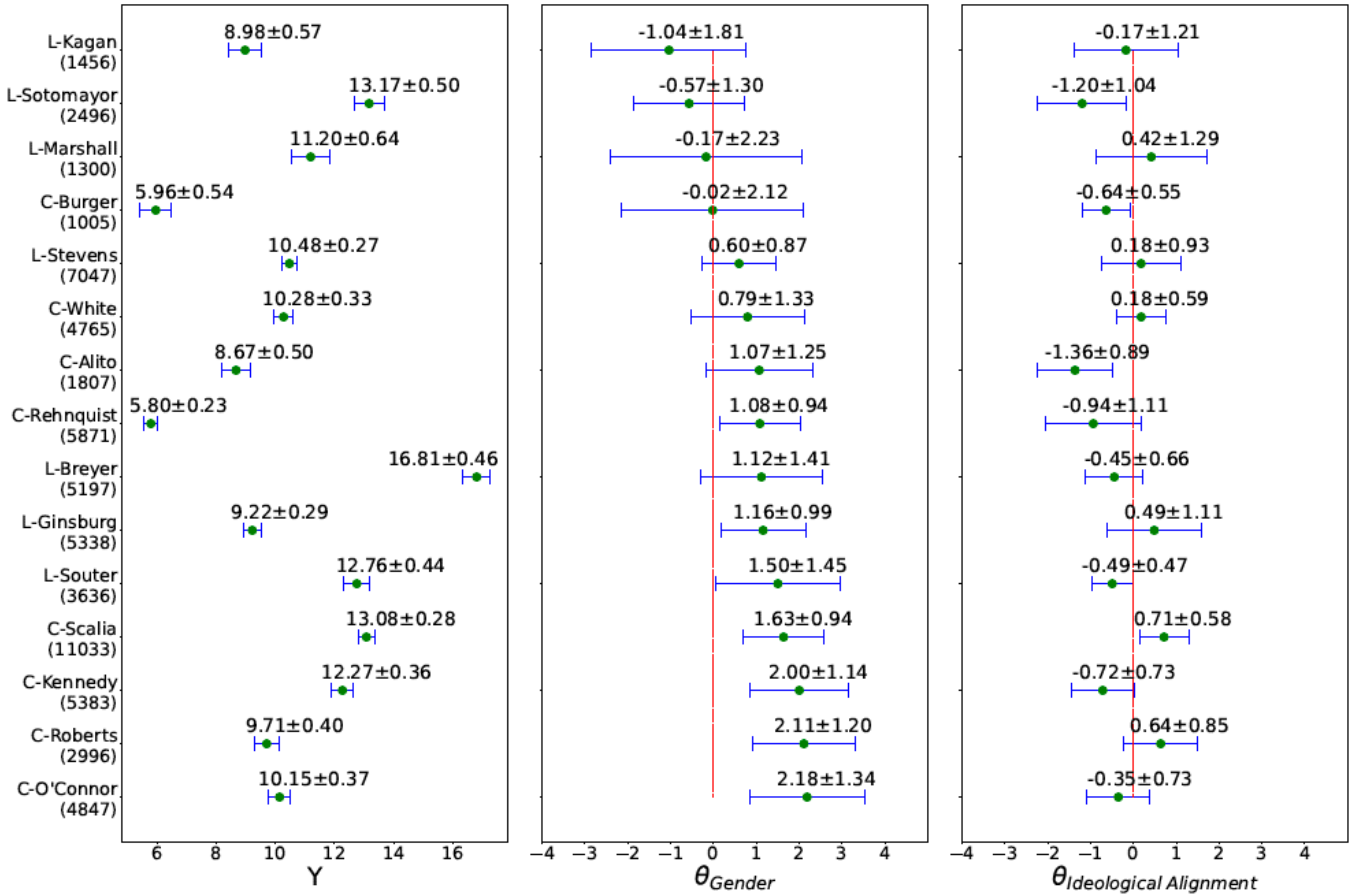
Justices	θ_{Gender}	$\theta_{\text{Ideological Alignment}}$	$\frac{\theta_{\text{Gender}}}{\theta_{\text{Ideological Alignment}}}$
All	0.90±0.19	-0.25±0.12	3.60
Male	1.06±0.22	-0.20±0.13	5.30
Female	0.43±0.36	-0.39±0.24	1.10

θ_{Gender} is equal to $E[Y|Gender = F] - E[Y|Gender = M]$

Positive values indicating higher interruption rates for female advocates

Negative values indicating higher interruption rates for male advocates

Justice-level Interruption Rates, Effect of Gender and Ideological Alignment



Causal Mediation Analysis

Two pathways to the gender effect:

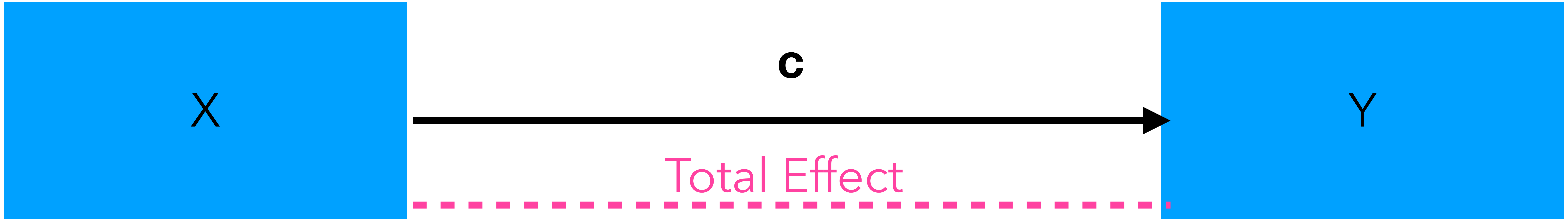
- Differences in the ideological orientation of advocates and justices
- Differences in quality or style of arguments
  speaking fluidity
  advocate experience

Causal Mediation Analysis

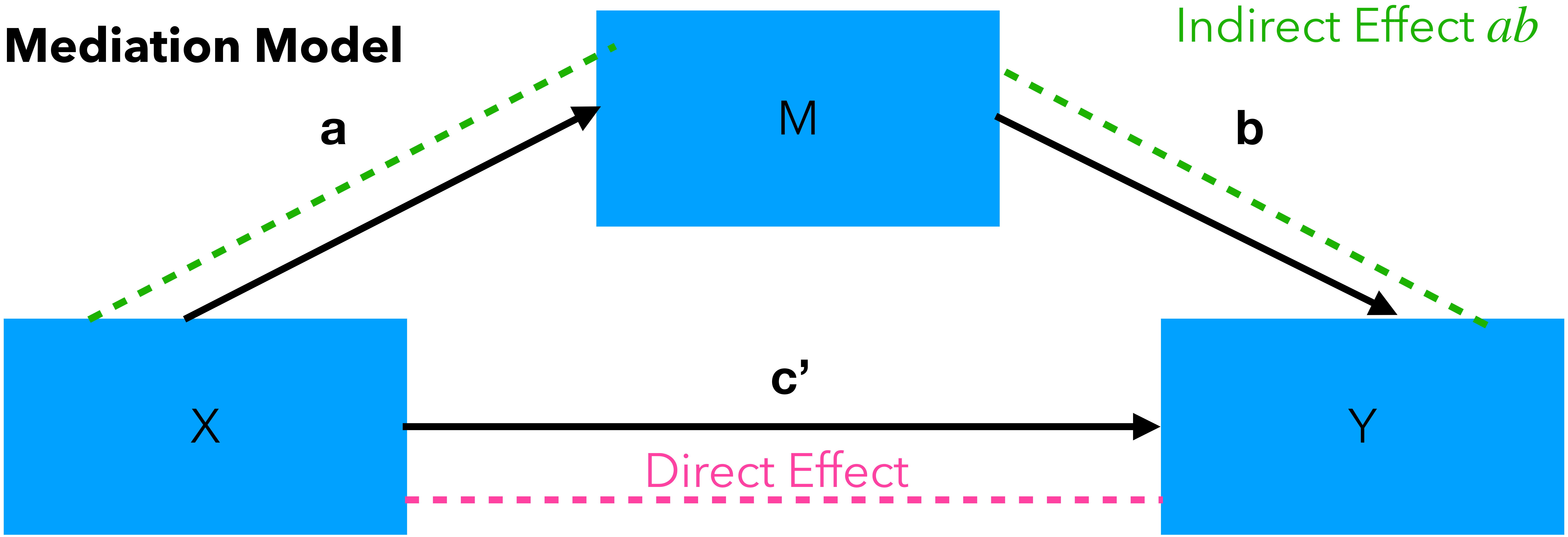
All justices (n=36,633)	NDE	NIE
Speech disfluencies as mediator	0.39±0.34	-0.02±0.12
Ideological alignment as mediator	0.41±0.39	0.01±0.03
Advocate experience as mediator	0.31±0.36	-0.01±0.03
Male justices (n=27,703)	NDE	NIE
Speech disfluencies as mediator	0.61±0.44	0.02±0.15
Ideological alignment as mediator	0.66±0.50	0.02±0.04
Advocate experience as mediator	0.56±0.47	0.02±0.03
Female justices (n=9,560)	NDE	NIE
Speech disfluencies as mediator	-0.22±0.36	-0.10±0.10
Ideological alignment as mediator	-0.20±0.37	0.00±0.02
Advocate experience as mediator	-0.36±0.37	-0.07±0.05

Causal mediation estimates of the natural direct effect (NDE) from gender to interruption and the natural indirect effect (NIE) from gender through the mediator speech disfluencies, ideological alignment, or advocate experience to interruption, aggregated across justices

Total Effect Model



Mediation Model



Causal Mediation Analysis

All justices (n=36,633)	NDE	NIE
Speech disfluencies as mediator	0.39±0.34	-0.02±0.12
Ideological alignment as mediator	0.41±0.39	0.01±0.03
Advocate experience as mediator	0.31±0.36	-0.01±0.03
Male justices (n=27,703)	NDE	NIE
Speech disfluencies as mediator	0.61±0.44	0.02±0.15
Ideological alignment as mediator	0.66±0.50	0.02±0.04
Advocate experience as mediator	0.56±0.47	0.02±0.03
Female justices (n=9,560)	NDE	NIE
Speech disfluencies as mediator	-0.22±0.36	-0.10±0.10
Ideological alignment as mediator	-0.20±0.37	0.00±0.02
Advocate experience as mediator	-0.36±0.37	-0.07±0.05

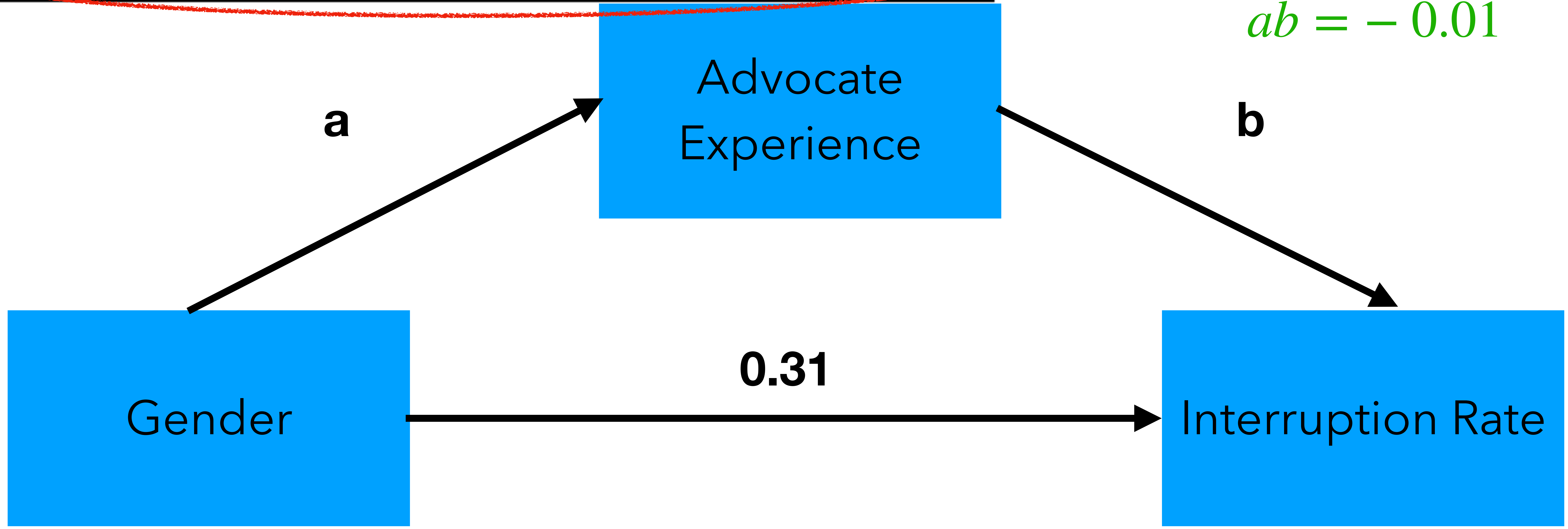
Causal mediation estimates of the natural direct effect (NDE) from gender to interruption and the natural indirect effect (NIE) from gender through the mediator speech disfluencies, ideological alignment, or advocate experience to interruption, aggregated across justices

Causal Mediation Analysis

All justices (n=36,633)	NDE	NIE
Speech disfluencies as mediator	0.39±0.34	-0.02±0.12
Ideological alignment as mediator	0.41±0.39	0.01±0.03
Advocate experience as mediator	0.31±0.36	-0.01±0.03

Indirect Effect

$ab = -0.01$



Lecture Overview

- Prediction vs. Understanding
- Randomized controlled trail (RCT)
- Observation data and studies
- Propensity score methods
- Case studies
- Mediation analysis