State-of-the-art Strategies for Addressing Selection Bias When Comparing Two or More Treatment Groups

Beth Ann Griffin
Daniel McCaffrey

RAND CORPORATION

Acknowledgements

- This work has been generously supported by NIDA grant 1R01DA034065
- Our colleagues

Daniel Almirall

Craig Martin

James Gazis

Andrew Morral

Rajeev Ramchand

Lisa Jaycox

Natalia Weil

Greg Ridgeway

Goals

- To increase your understanding of
  - How to define and estimate causal effects
  - How to use propensity scores when estimating causal effects
- To provide you with step-by-step instructions for using propensity score weights when you have:
  - 2 groups (binary treatment)
  - More than 2 groups (multiple treatments)

BACKGROUND ON CAUSAL EFFECTS ESTIMATION

Why causal effects?

- Causal effects are of great interest in many fields
  - Interested in causal effects, not merely correlations
- Causal effects describe how individuals will change their behaviors (substance use, mental health, criminal involvement) in response to
  - A new treatment
  - A prevention program
  - A new policy
  - A new exposure

Motivating example

- **Case study objective:** To estimate the causal effect of MET/CBT5 versus "usual care" among adolescent clients
  - Data from 2 SAMSHA CSAT discretionary grants

MET/CBT5	"Usual Care"
• Longitudinal, observational data	• Longitudinal, observational data
• MET/CBT5 at 37 sites from EAT	• "Usual care" at 4 sites from ATM

Motivating example

- **Case study objective:** To estimate the causal effect of MET/CBT5 versus "usual care" among adolescent clients
  - Data from 2 SAMSHA CSAT discretionary grants

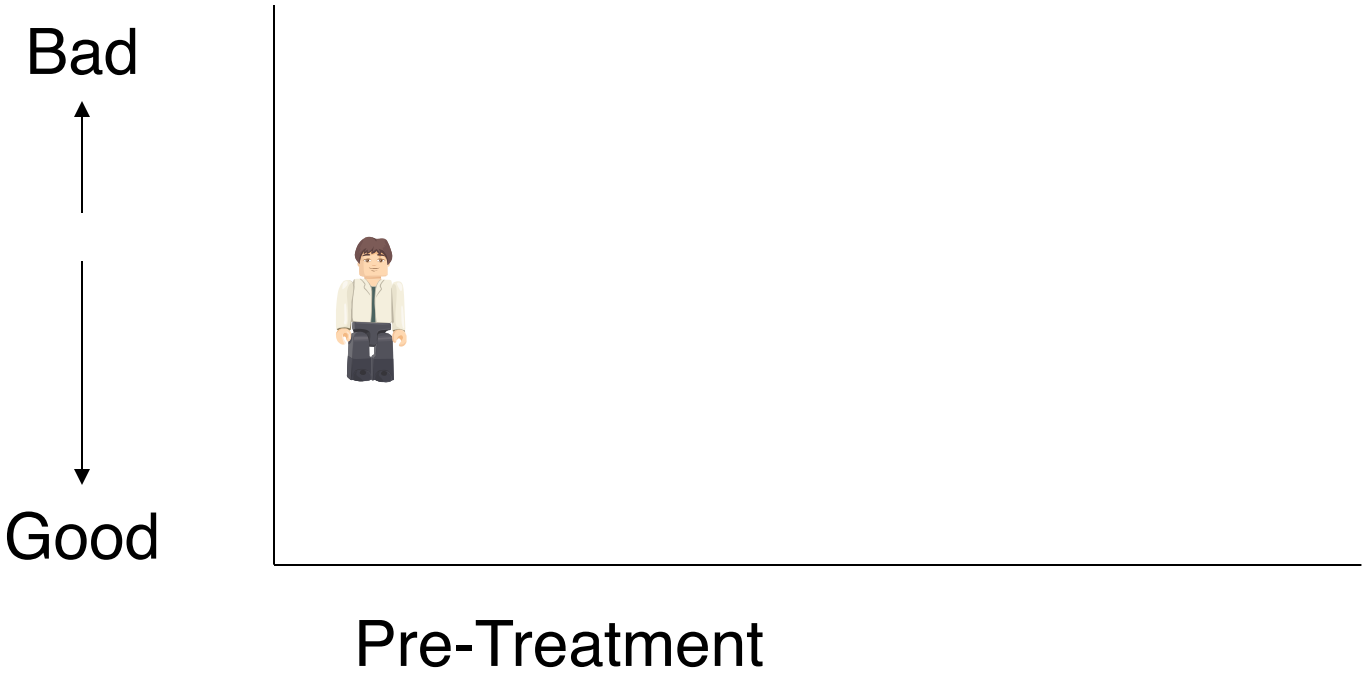
MET/CBT5

"Usual Care"

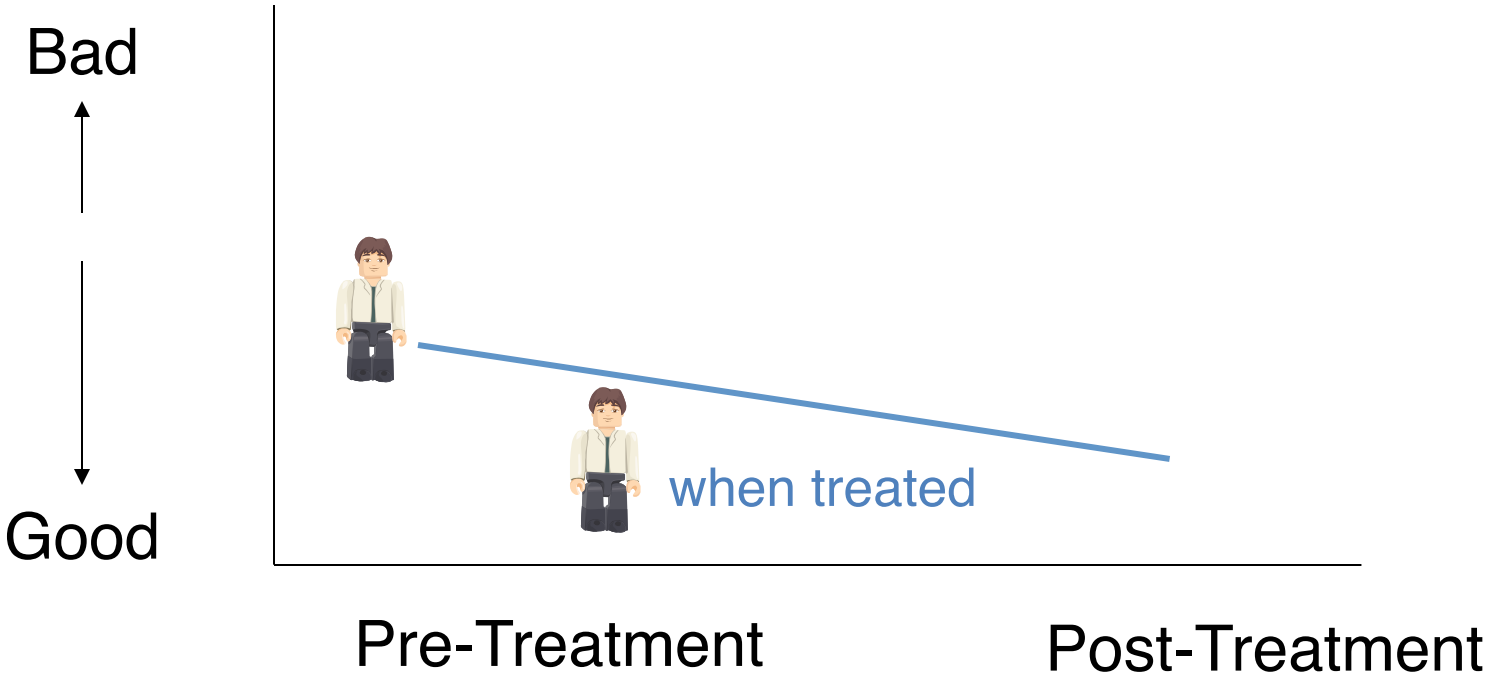
Various situations can benefit from use of causal inference methods

- Longitudinal observational studies (like our case study)
- Administrative databases with outcomes (i.e., in a health plan, who gets what treatment and what is their subsequent service use)
- Aggregating across studies to maximize power and look at moderation or mediation
  - E.g., different RCTs done over time that compared A to B then A+C to B and you want to compare A to A +C

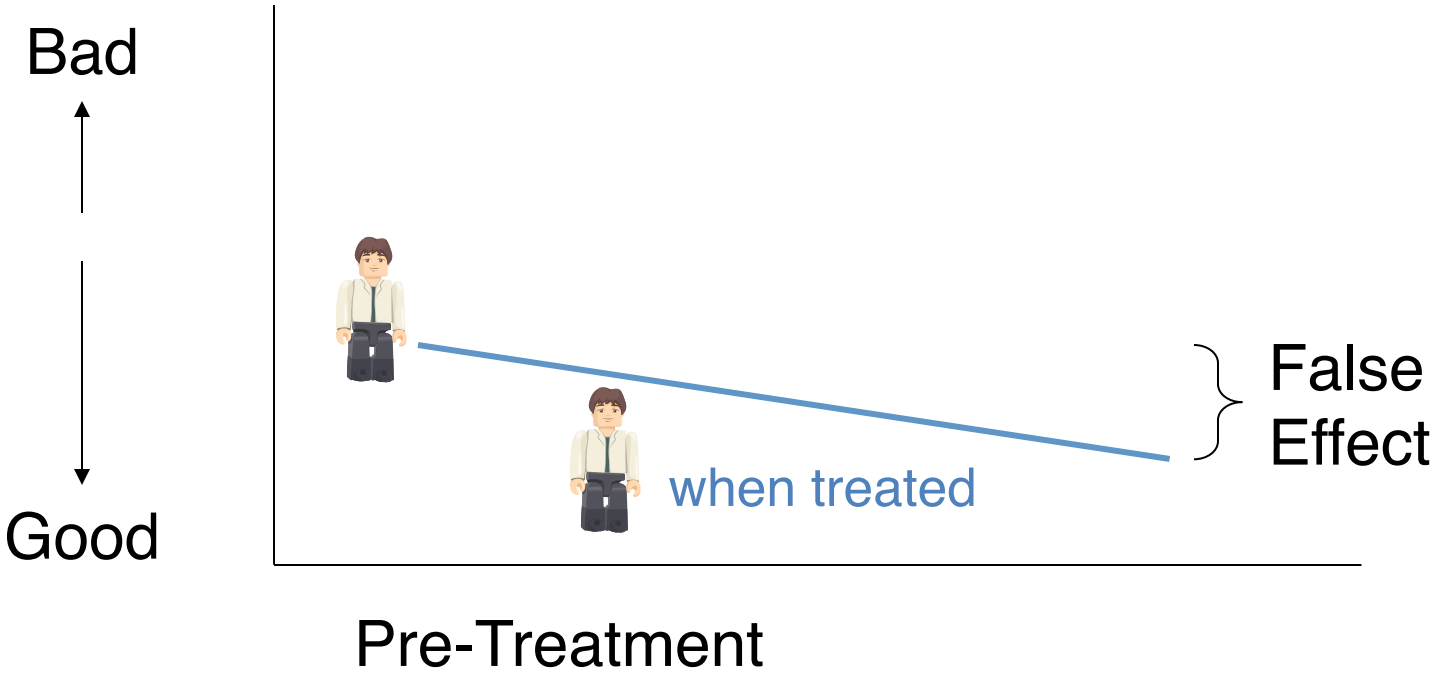
How do we get a causal effect estimate?



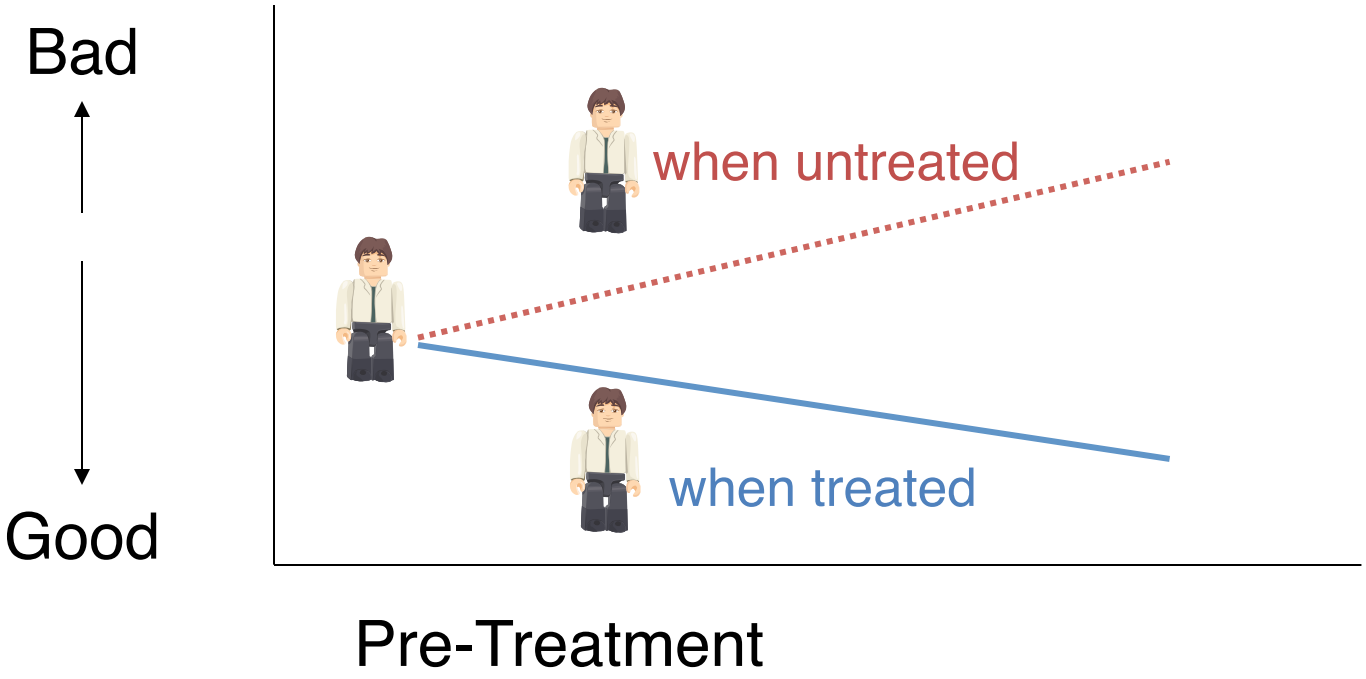
How do we get a causal effect estimate?



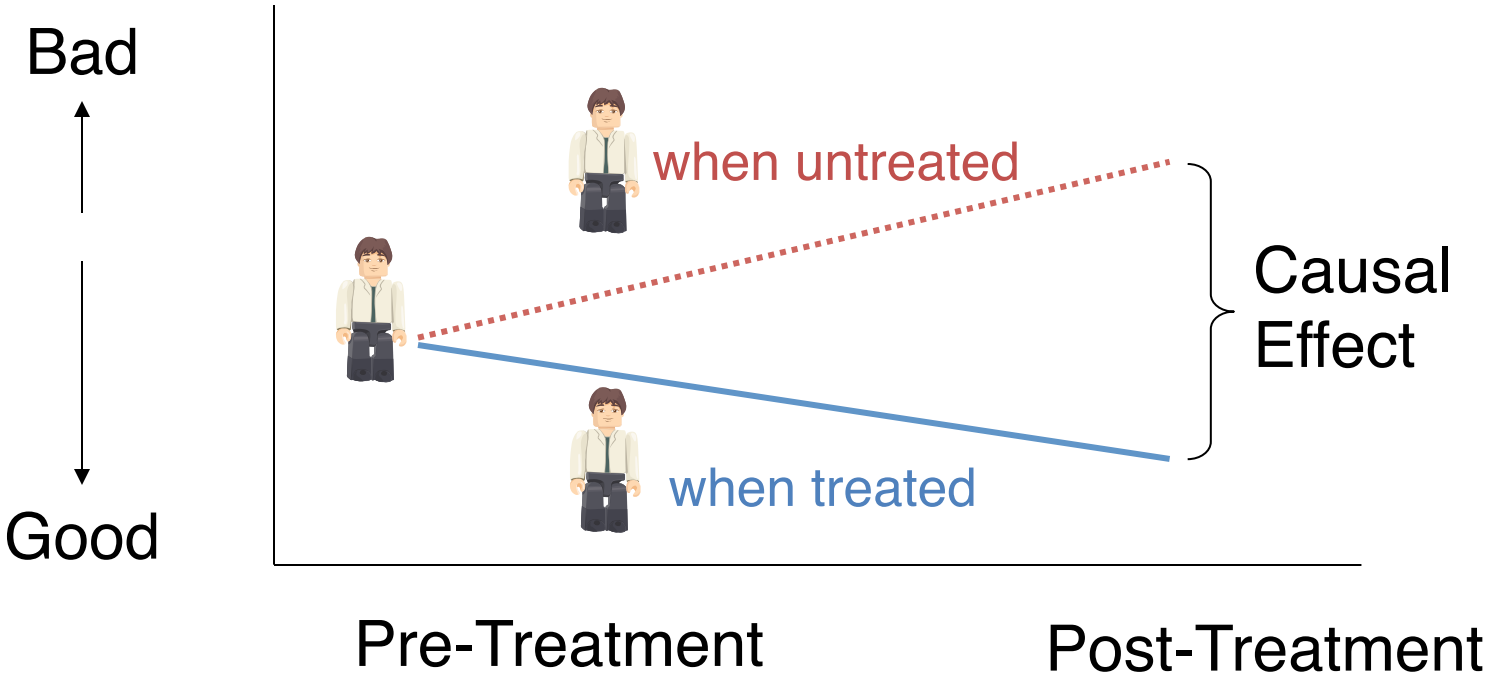
How do we get a causal effect estimate?



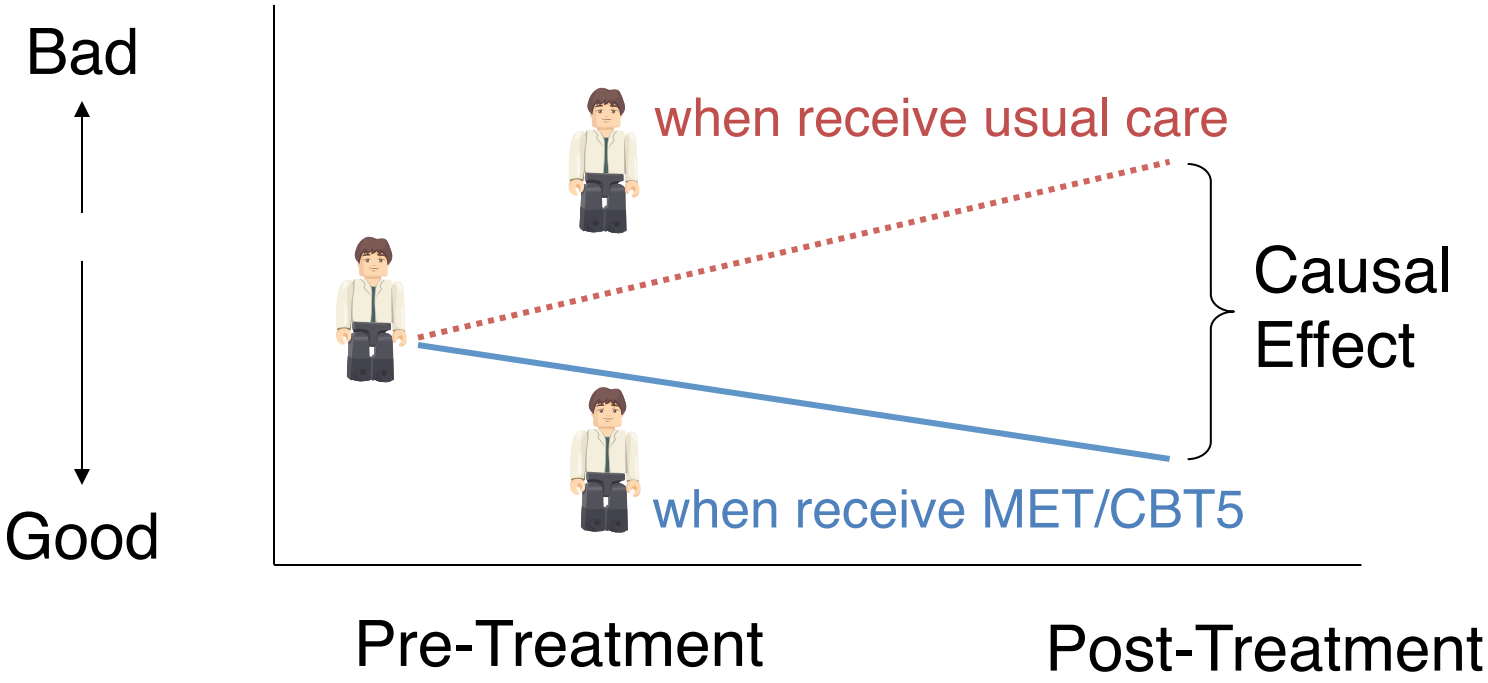
How do we get a causal effect estimate?



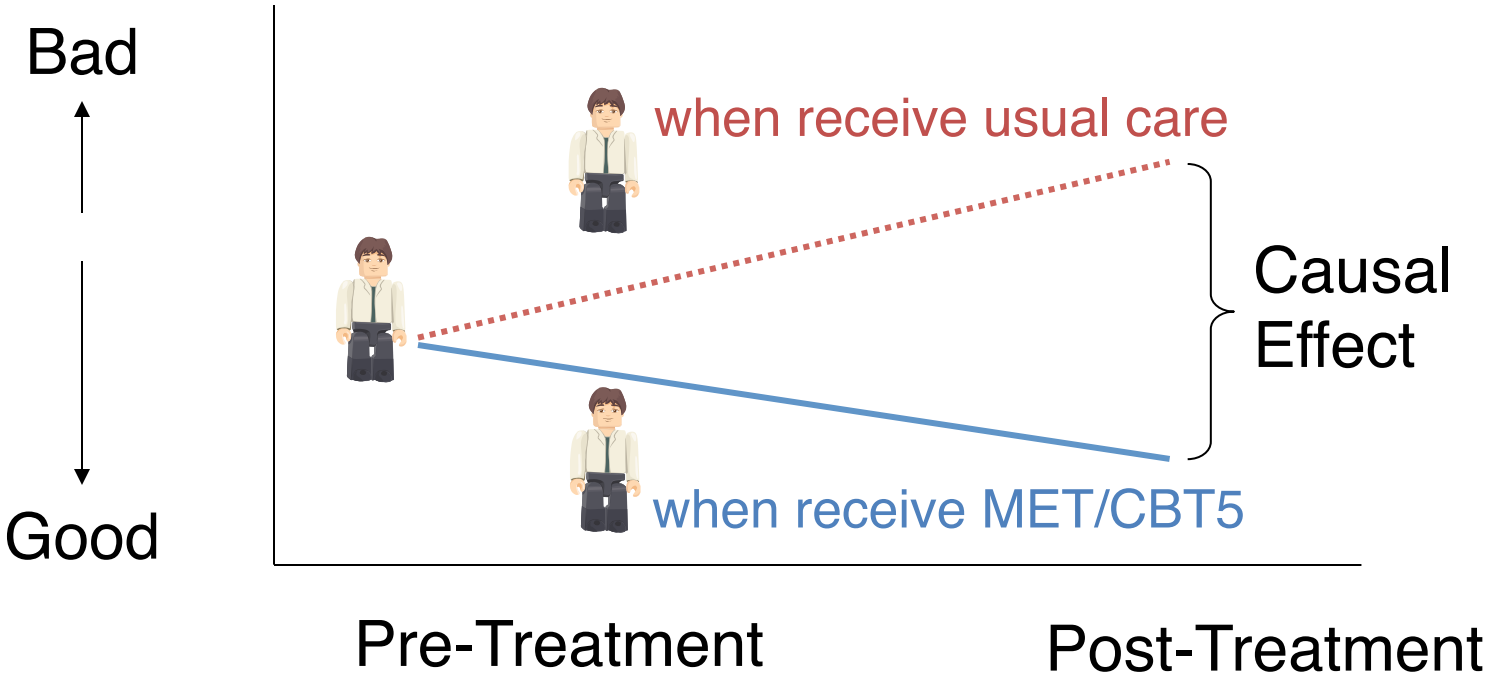
Causal effect = difference between two potential outcomes



Causal effect of MET/CBT5 vs. Usual Care



Causal effect of MET/CBT5 vs. Usual Care

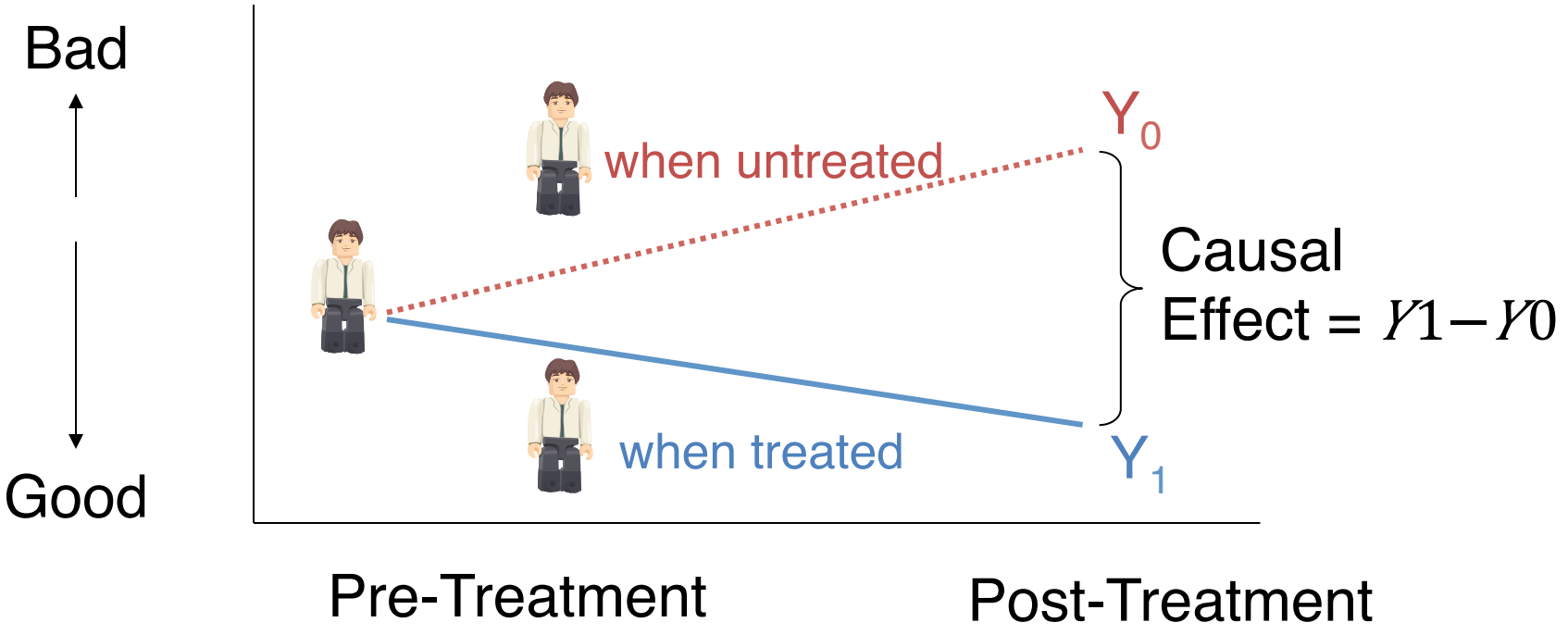


Fundamental problem for causal inference: We only observe ONE of these potential outcomes/counterfactuals

Potential outcomes framework

- Two potential outcomes for each study participant:
  - Outcome after receiving treatment = $Y1$
  - Outcome after receiving comparison condition = $Y0$ (can be either another treatment or no treatment control)
- $Y1$ and $Y0$ exist for all individuals in the population regardless of whether the individual received treatment ($Z=1$) or comparison condition ($Z=0$)
- Observe only one of these outcomes for each participant

Causal effect = difference between two potential outcomes



Primary types of causal effects

- Average treatment effect in the population (ATE)
  - Answers the question:
    - How effective is the treatment in the population? (if you have a *control condition*)
    - What is the relative effectiveness of two treatments on average in the population? (if you have *2 treatments*)
  - $E(Y1-Y0)$
- Average treatment effect in the treated population (ATT)
  - Answers the question:
    - How would those who received treatment have done had they received the comparison condition?
  - $E(Y1-Y0|Z=1)$

Primary types of causal effects: case study

- Average treatment effect in the population (ATE)
  - Answers the question:
    - What is the relative effectiveness of MET/CBT5 versus usual care on average in the population?
  - $E(Y1-Y0)$
- Average treatment effect in the treated population (ATT)
  - Answers the question:
    - How would those who had received usual care have done had they received MET/CBT5?
  - $E(Y1-Y0|Z=1)$

Experiments vs. observational studies

- Random experiments = gold standard for estimating causal effects
  - Randomization (if it works) makes groups being compared balanced on baseline characteristics
  - Treatment assignment is unrelated to potential outcomes (strong ignorability)
- Randomization is not always feasible
- Observational studies provide another way to get at causal effects
  - Treatment assignment is not controlled by the researcher
  - Groups being compared are usually imbalanced
  - Can use causal inference methods to try to replicate what a randomized study does

Biggest challenge to causal estimation is selection effects

- Selection occurs when the people getting the treatments being compared differ

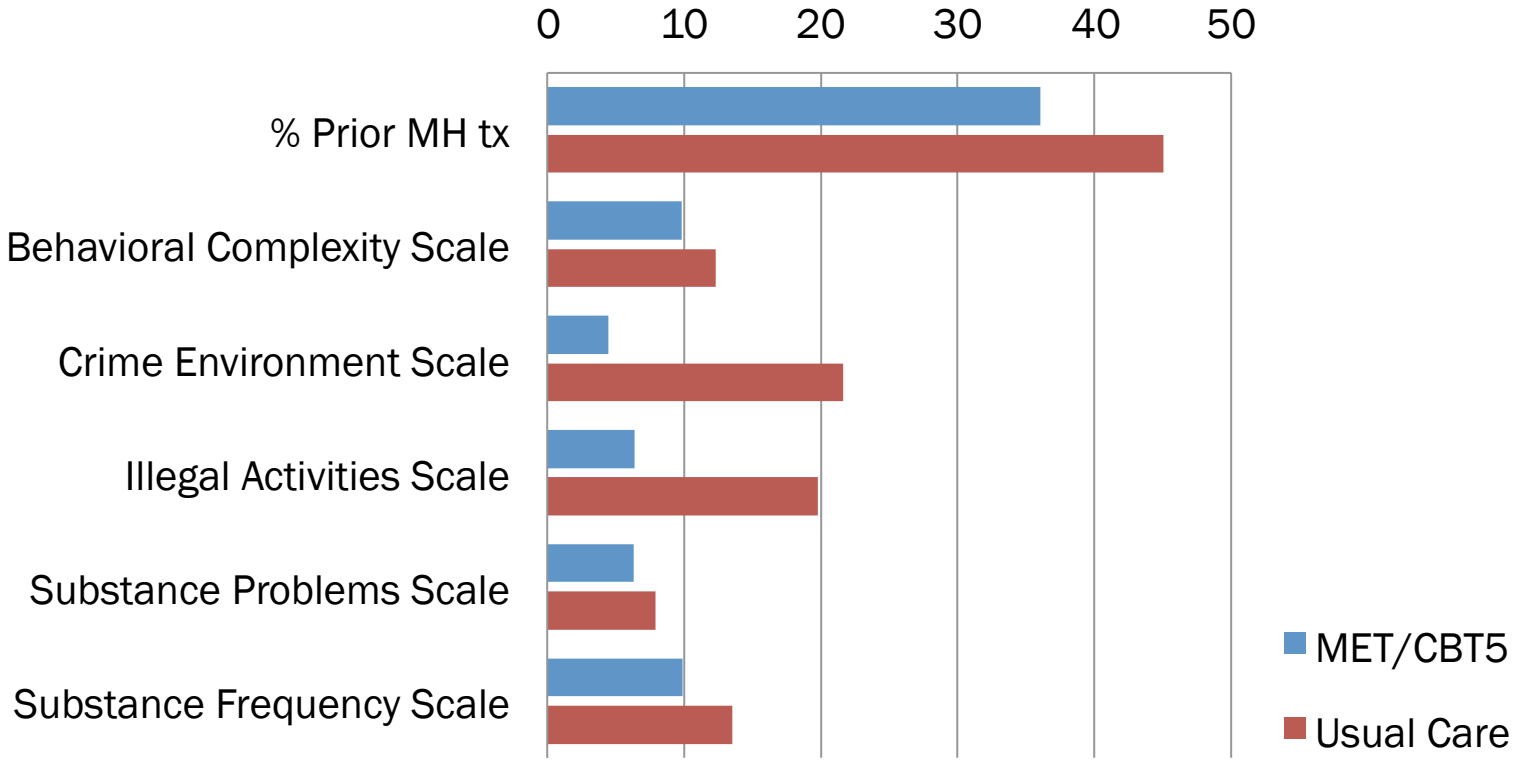


Illustration: The potential impact of selection on our case study

- Mean substance frequency scale (SFS) at 12-months =

  0.11 (~10 days of use per month) for usual care

  0.07 (~ 6 days of use per month) for MET/CBT5

  → Effect size (ES) difference of 0.41

Illustration: The potential impact of selection on our case study

- Mean substance frequency scale (SFS) at 12-months =

  0.11 (~10 days of use per month) for usual care

  0.07 (~ 6 days of use per month) for MET/CBT5

  → Effect size (ES) difference of 0.41

- I am going to report that usual care results in 0.41 standard deviations greater SFS than MET/CBT5
  - Is this a good idea?

Illustration: The potential impact of selection on our case study

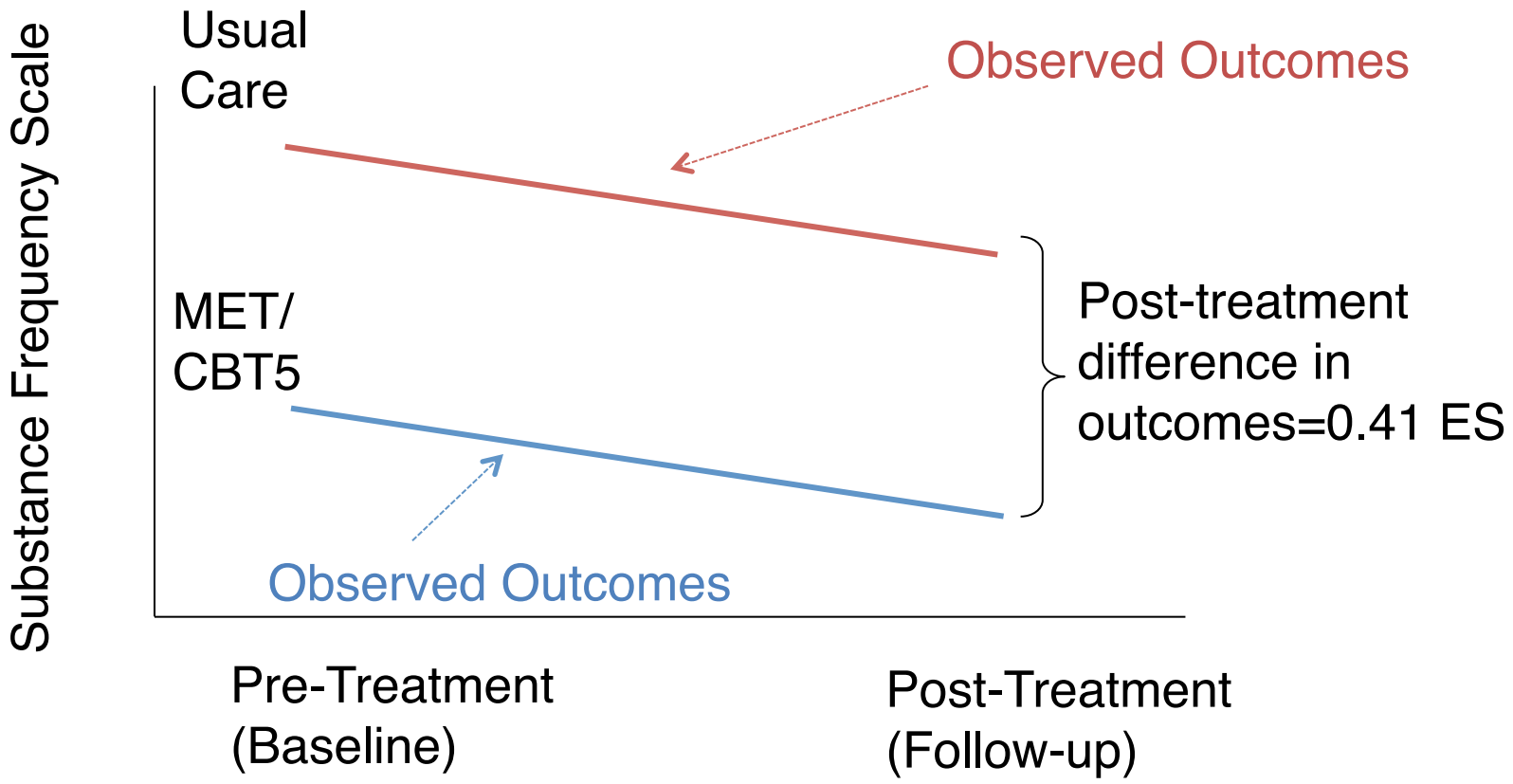


Illustration: The potential impact of selection on our case study

- Mean substance frequency scale (SFS) at 12-months =

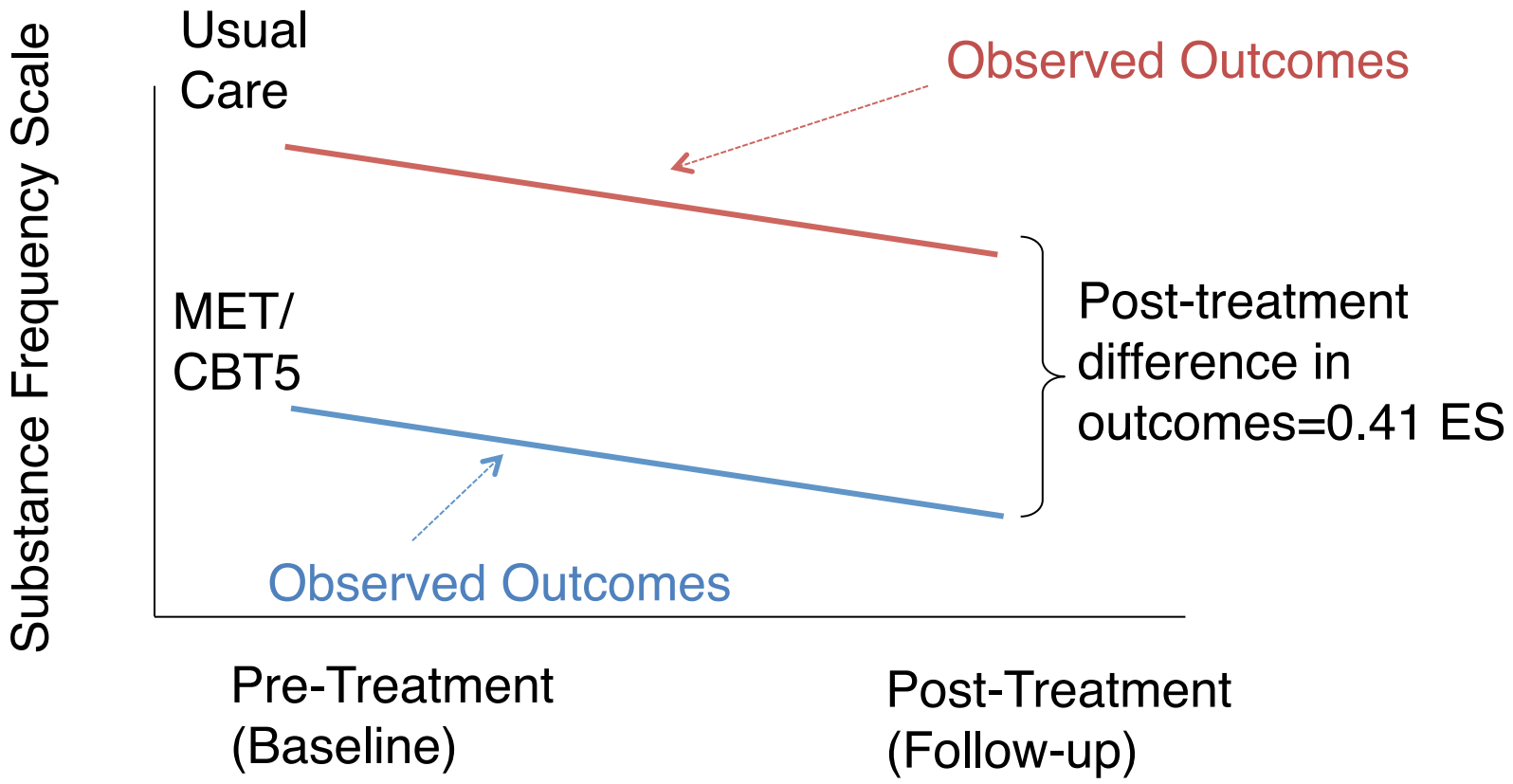
  0.11 (~10 days of use per month) for usual care

  0.07 (~ 6 days of use per month) for MET/CBT5

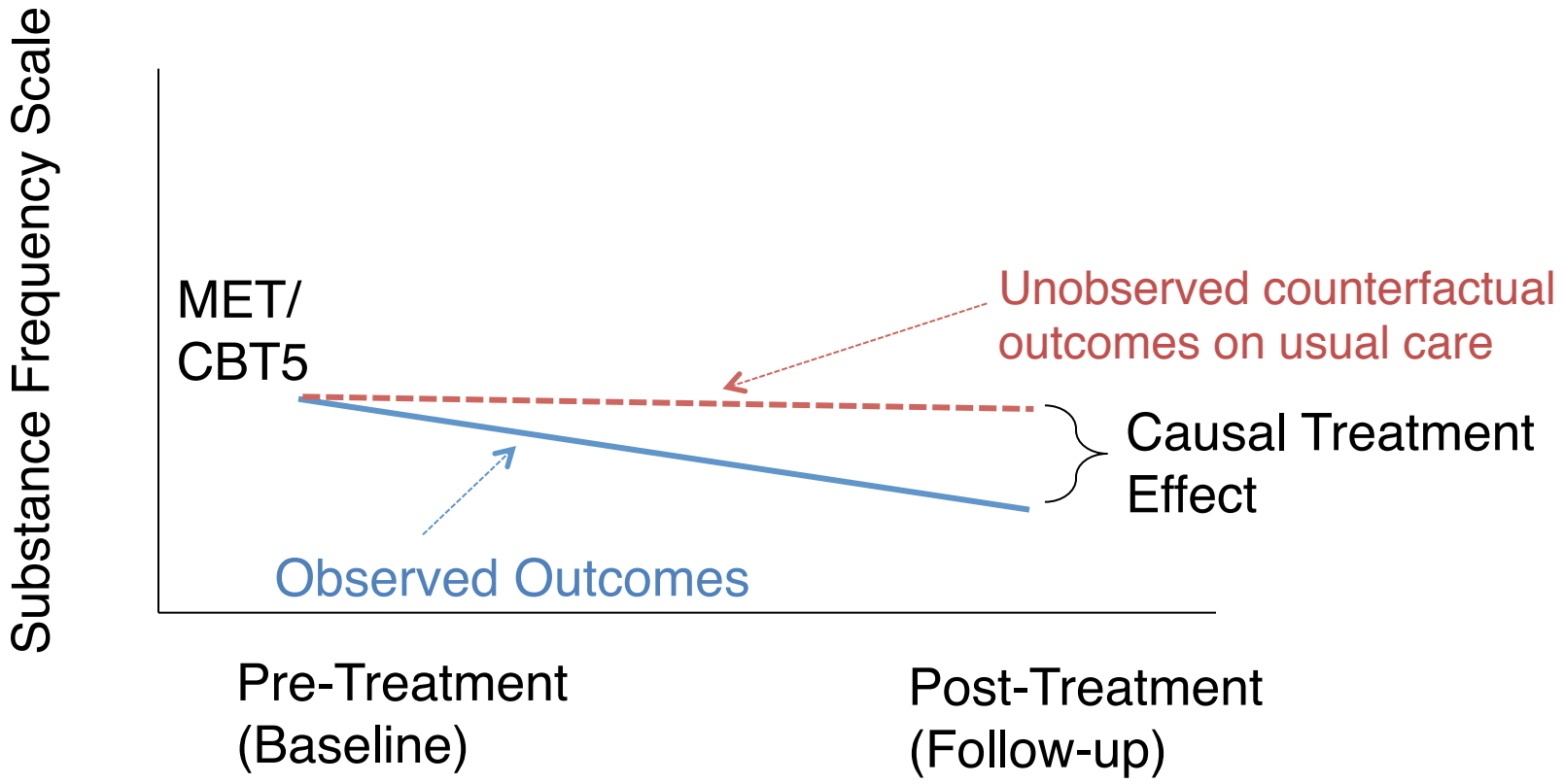
  → Effect size (ES) difference of 0.41

- I am going to report that usual care results in 0.41 standard deviations greater SFS than MET/CBT5
  - Is this a good idea?
  - No!!! Groups might differ on pretreatment characteristics that influence SFS at 12-months such that even in the absence of the treatment the groups might differ on outcomes at 12-months

Different kids, different treatments



WANT: Same kids, different treatments



Propensity scores help us deal with the challenge of selection

- The propensity score is an individual's probability of receiving treatment given pretreatment characteristics

$$p(X)=\Pr(Z=1|X)$$

  - Used to create balance between treatment and comparison conditions
  - Balancing on $p(X)$ can balance distribution of X between treatment and comparison groups

Propensity scores help us deal with the challenge of selection

- The propensity score is an individual's probability of receiving treatment given pretreatment characteristics

$$p(X)=\Pr(Z=1|X)$$

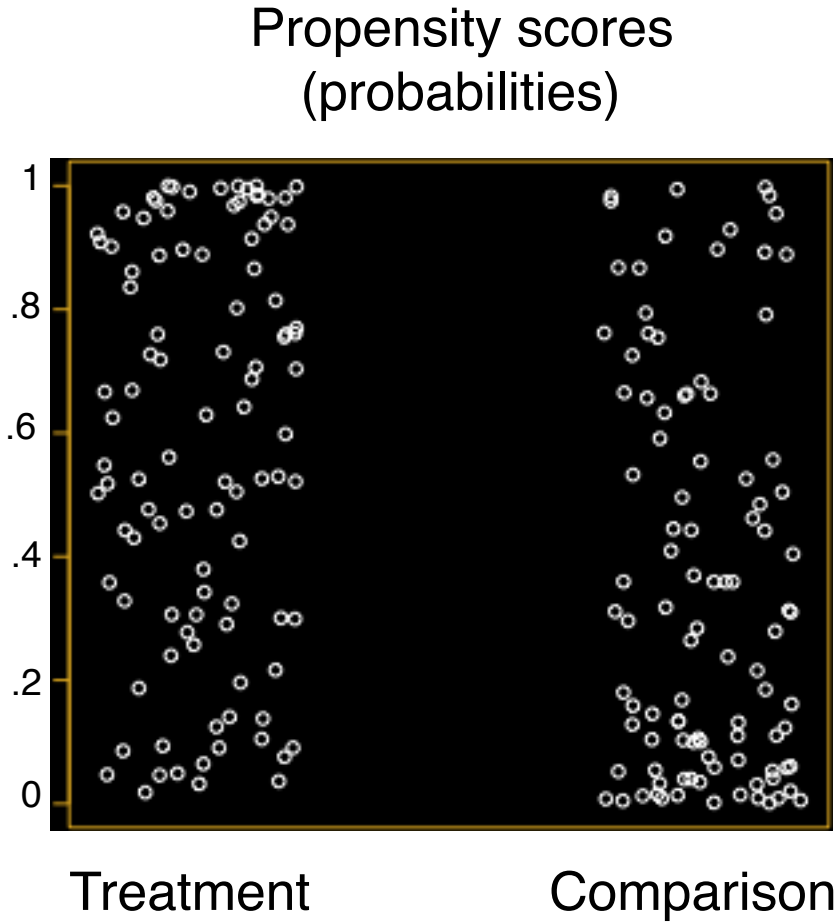
  - Used to create balance between treatment and comparison conditions
  - Balancing on $p(X)$ can balance distribution of X between treatment and comparison groups

- If strong ignorability holds, propensity score is all that is required to control for observed, pretreatment differences between groups
  - Treatment assignment is strongly ignorable if

    (1) $(Y_1, Y_0) \perp Z|X$ – e.g, no unobserved confounders

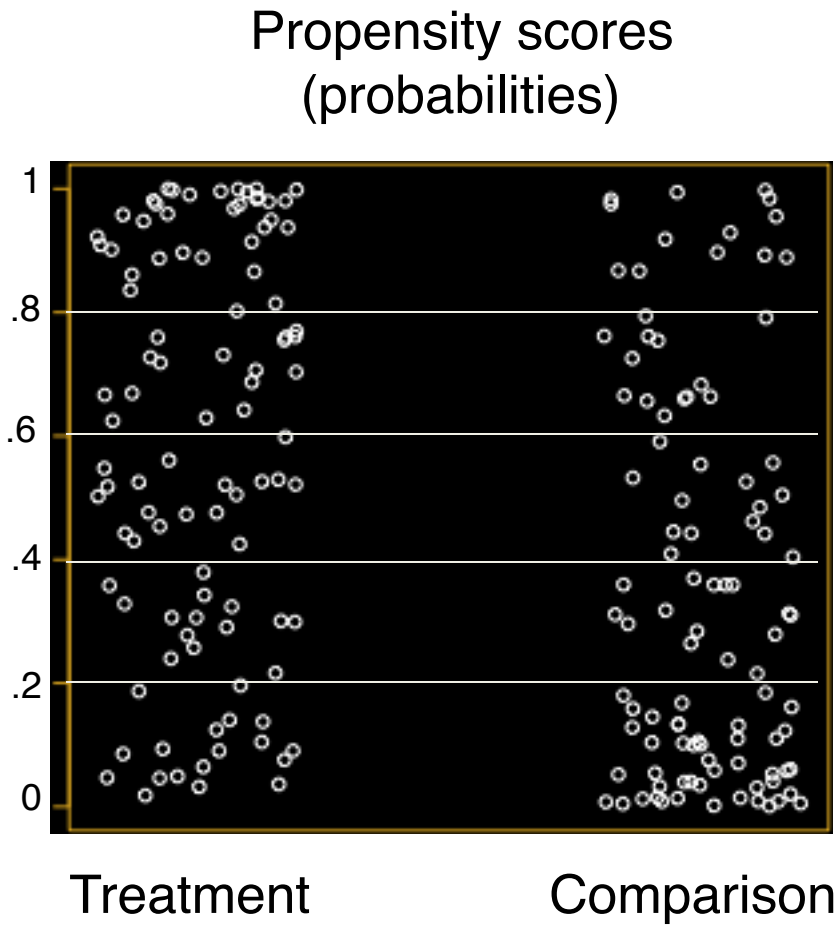
    (2) $0 < p(X) < 1$ – e.g, overlap between groups

Various ways to use propensity scores to estimate causal effects



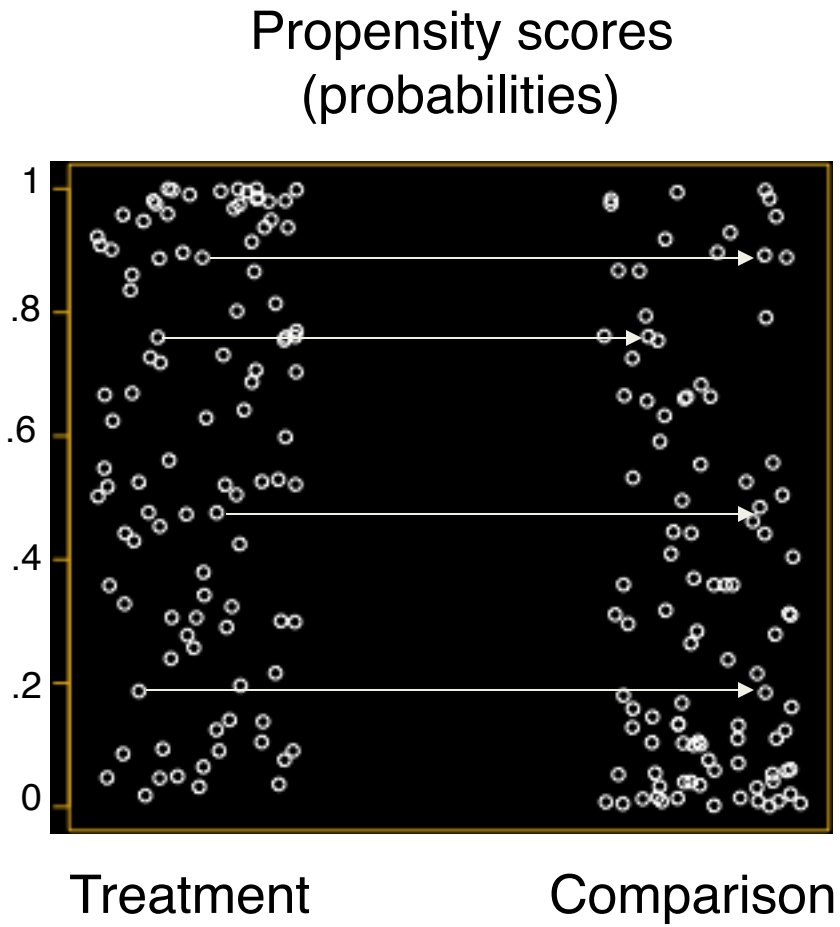
Various ways to use propensity scores to estimate causal effects

- Stratification



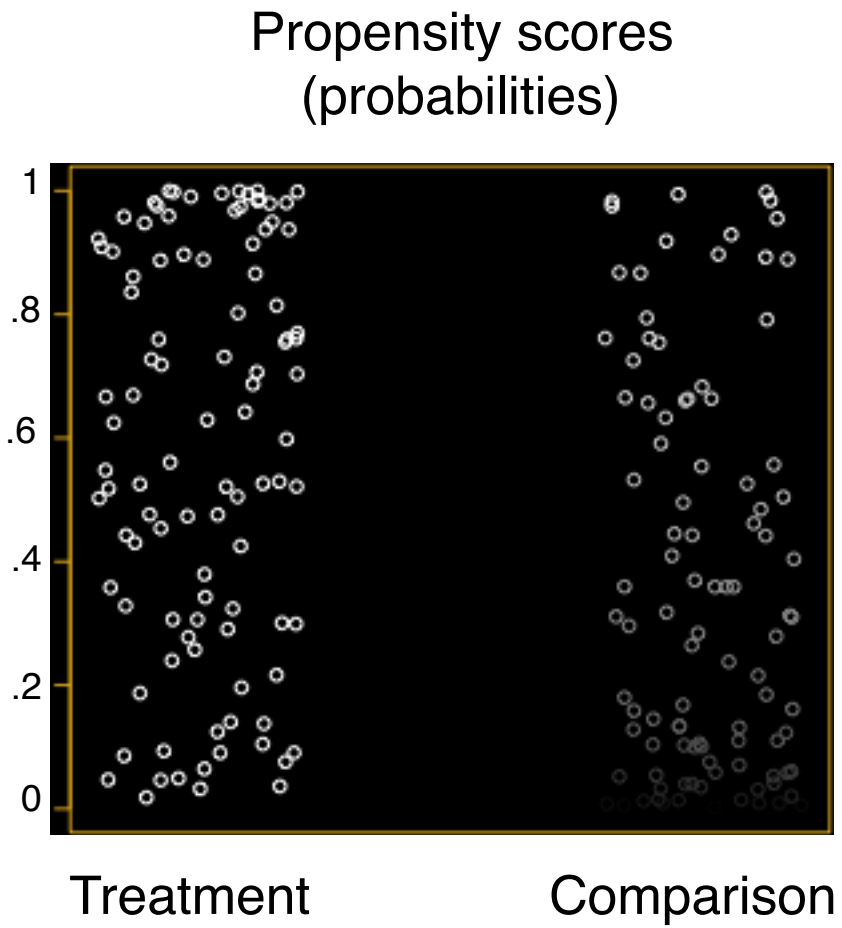
Various ways to use propensity scores to estimate causal effects

- Stratification
- Matching



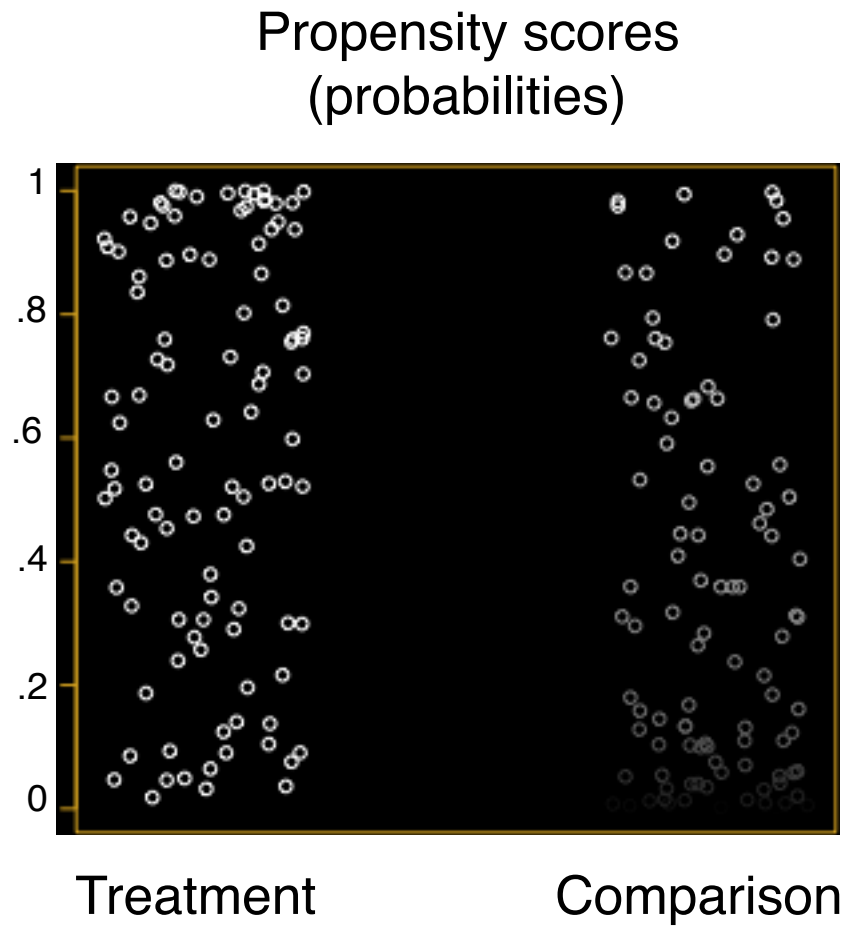
Various ways to use propensity scores to estimate causal effects

- Stratification
- Matching
- Weighting



Various ways to use propensity scores to estimate causal effects

- Stratification
- Matching
- Weighting
  - Today, we will focus on how to use propensity score weights to correct for selection effects on observed pretreatment characteristics



How do we get good propensity score weights?

- Everything above has assumed that propensity scores are known
  - Not so unless we have a randomized study
- Typically have to estimate the propensity scores from data
- Standard approach has been to use logistic regression model with treatment indicator as outcome
  - Can be challenging
  - Need to find the right combination higher order terms, interactions, variable selection, etc.

How do we get good propensity score weights? (cont.)

- Machine learning methods have been shown to perform better than logistic regression
  - Achieve better balance between treatment and comparison group on pretreatment covariates
  - Reduce bias in treatment effect estimates
  - Produce more stable propensity score weights (thereby also improving precision)
- Field constantly coming up with new methods
  - Covariate Balancing Propensity Score (CBPS; Imai and Ratkovic 2013)
  - Super Learning (van der Laan 2014)
  - High Dimensional Propensity Score (HDps) (Schneeweiss et al 2009)
  - Entropy Balance (Hainmueller 2012)

Today we focus on using GBM to estimate propensity score weights

- GBM = generalized boosted regression models
  - Data adaptive, nonparametric model
  - Combines many piecewise-constant functions of the covariates to estimate unknown propensity score
  - Can include nominal, discrete, and continuous covariates and covariates with missing values
  - Can include large number of covariates
  - Automatically selects which covariates to include and best functional form including interactions
  - "optimal" iteration chosen as that which balances treatment and comparison group best

Use GBM via TWANG package

- TWANG = Toolkit for weighting and analysis of nonequivalent groups
  - Uses GBM to estimate propensity score weights
  - Provides various tools for assessing the quality of the propensity score weights
  - Handles both binary and more than 2 treatment group settings
  - Available in R, SAS and STATA

Assessing the quality of propensity score weights

- Primarily focuses on how well propensity score weights balance the two groups
- Various metrics available to assess balance
  - Commonly used methods include:
    - Standardized Effect Size (ES) differences between the two groups
    - Kolmogorov-Smirnov (KS) statistic
    - *T*-test *p*-values

Balance measures: ES

- The standardized effect size (ES) is the difference in mean of a pretreatment variable in the treated group versus the comparison group, divided by the standard deviation

$$(X\downarrow 1 - X\downarrow 0)/\sigma$$

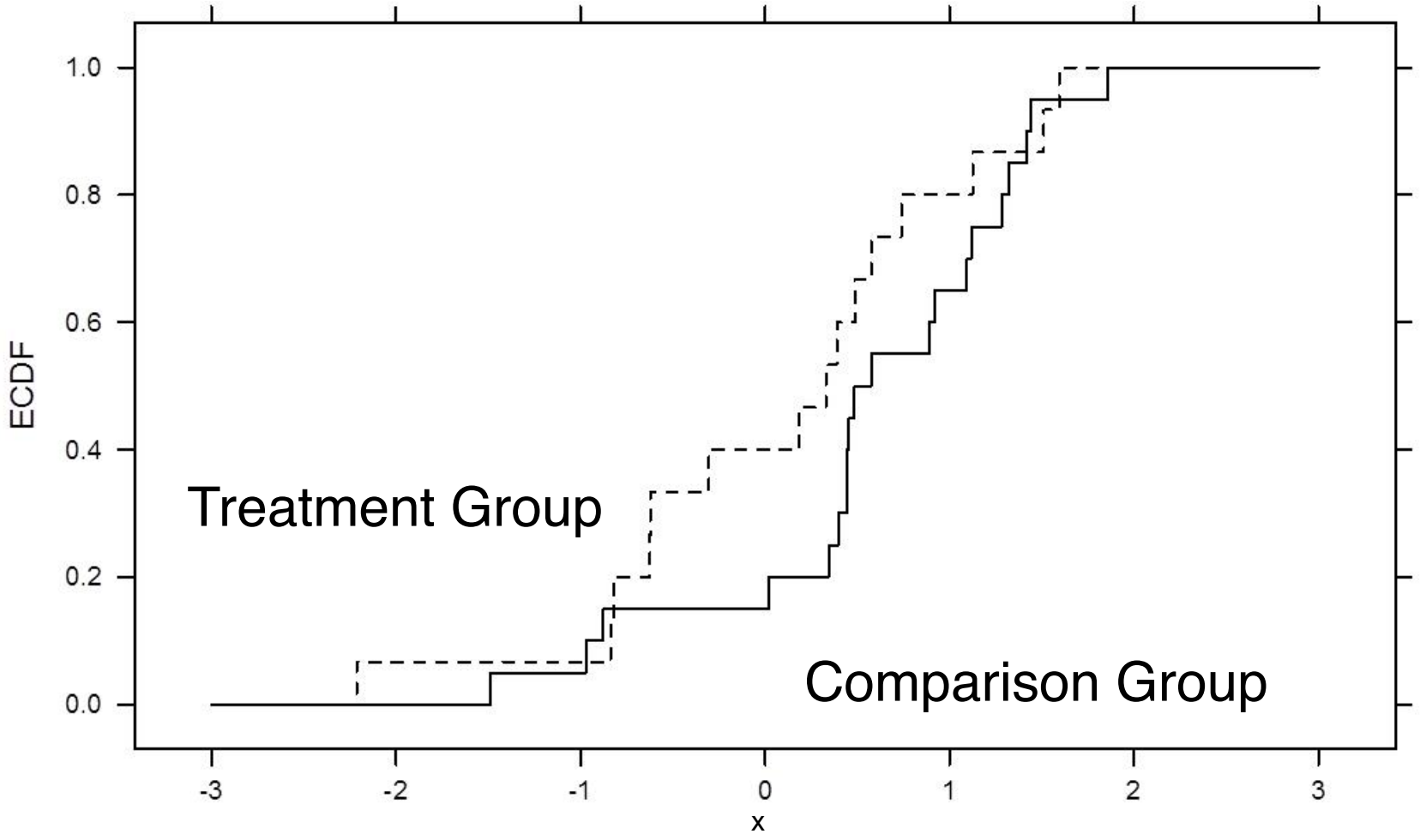
  - In ATE analyses, σ is the pooled standard deviation
  - In ATT analyses, σ is from the treated group
  - Also called Standardized Mean Difference (SMD)

- Our group uses the rule of thumb that an absolute ES under 0.2 is indicative of good balance
  - Some other authors have suggested that under 0.25 is OK

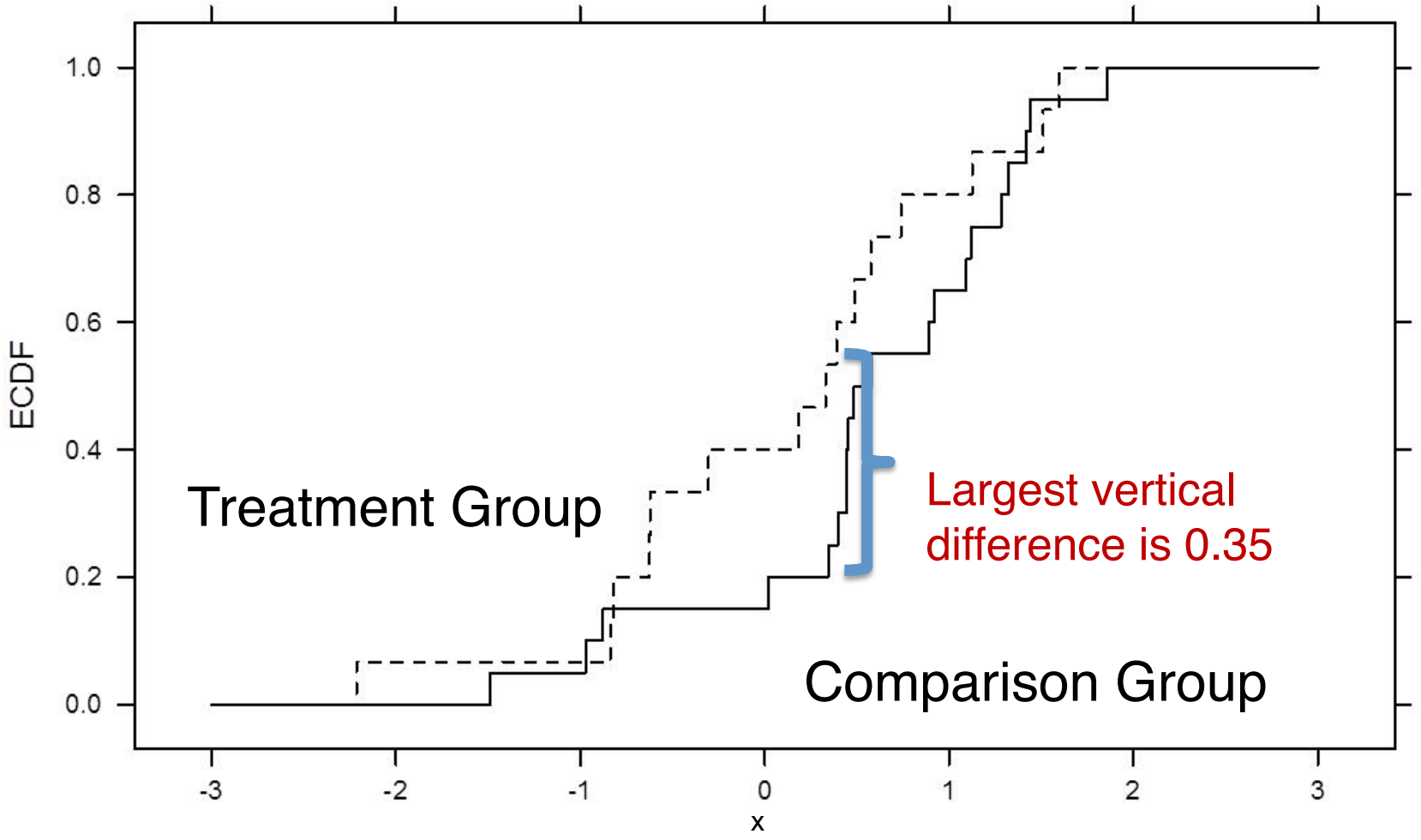
Balance measures: KS

- Another measure that we use is the Kolmogorov-Smirnof (KS) statistic
- KS statistic is the maximum vertical distance between the empirical cumulative distribution functions of two samples

KS example


ECDF
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
x
Treatment Group
Comparison Group

KS example


ECDF
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
x
Treatment Group
Comparison Group
Largest vertical
difference is 0.35

Propensity score weights for different causal effects

- For ATE:
  - weight treatment group by $1/p\left(X\downarrow i\right)$
  - weight comparison group by $1/(1-p\left(X\downarrow i\right))$
- For ATT:
  - weight comparison group by $p\left(X\downarrow i\right)/(1-p\left(X\downarrow i\right))$
- Both can provide unbiased estimates of the causal treatment effect in question assuming
  1. covariates are balanced in treatment and control groups after weighting
  2. there are no unobserved confounders and overlap (strong ignorability)

Final remarks

- Causal effects are important
- Biggest challenges =
  - Can't estimate directly because don't observe all potential outcomes we need
  - Selection: pretreatment differences between groups might bias our treatment effect estimates
- Propensity scores can help address these challenges
  - Assuming no unobserved confounders
- When estimating propensity scores, inferences depend on how well they are estimated
  - Use twang to get good estimates
  - Go doubly robust

JUSTICE, INFRASTRUCTURE, AND ENVIRONMENT

Please note this video is part of a three-part series. We encourage viewers to watch all three segments.

Please check out our website

http://www.rand.org/statistics/twang.html

RAND CORPORATION
OBJECTIVE ANALYSIS. EFFECTIVE SOLUTIONS.

About | Support RAND | Newsroom | Events | 中文(简体) | Sign In

RESEARCH | LATEST INSIGHTS | POLICY EXPERTS | CAPABILITIES | GRADUATE SCHOOL

- TWANG Home
- Frequently Asked Questions (FAQ)
- Downloads
- Tutorials
- Publications
- Project Members

RAND > Statistics Group >

Toolkit for Weighting and Analysis of Nonequivalent Groups (TWANG)

photo by Iz Marty/flickr.com

The Toolkit for Weighting and Analysis of Nonequivalent Groups, or **TWANG**, contains a set of functions to support causal modeling of observational data through the estimation and evaluation of propensity score weights. The TWANG package was first developed in 2004 by RAND researchers for the R statistical computing language and environment. The R version of the package contains functions for creating high-quality propensity score weights which can be used to estimate treatment effects with two or more treatment groups.

In 2014, TWANG macros were developed for SAS to support the use of these tools without requiring researchers and analysts to learn R. At this time, the SAS TWANG macros can support estimation of propensity scores and their associated weights for comparisons involving two treatment groups. SAS macros will be made available shortly for handling the case of three or more treatment groups.

Recommend 0 | Tweet 0 | g+1 0

History

Project Researchers

Beth Ann Griffin
Senior Statistician

Beth Ann Griffin is a senior statistician at the RAND Corporation, where she codirects the RAND Center for Causal Inference. Her statistical research focuses on causal effects estimation when using observational data and innovative techniques for the analysis of time-to-event data. Her...

Explore Our People »

Stay Informed

Be notified when new studies, tools, and updates are released, or contact the TWANG team to ask a question about the effort.

Contact Form

Please check out our website

http://www.rand.org/statistics/twang.html

RAND CORPORATION
OBJECTIVE ANALYSIS. EFFECTIVE SOLUTIONS.

About | Support RAND | Newsroom | Events | 中文(简体) | Sign In

RESEARCH | LATEST INSIGHTS | POLICY EXPERTS | CAPABILITIES | GRADUATE SCHOOL

- TWANG Home
- Frequently Asked Questions (FAQ)
- Downloads
- Tutorials
- Publications
- Project Members

RAND > Statistics Group >

Toolkit for Weighting and Analysis of Nonequivalent Groups (TWANG)

The Toolkit for Weighting and Analysis of Nonequivalent Groups, or **TWANG**, contains a set of functions to support causal modeling of observational data through the estimation and evaluation of propensity score weights. The TWANG package was first developed in 2004 by RAND researchers for the R statistical computing language and environment. The R version of the package contains functions for creating high-quality propensity score weights which can be used to estimate treatment effects with two or more treatment groups.

photo by Iz Marty/flickr.com

In 2014, TWANG macros were developed for SAS to support the use of these tools without requiring researchers and analysts to learn R. At this time, the SAS TWANG macros can s[upp]ort estimation of propensity scores and their associated weights for comparisons involvin[g t]wo treatment groups. SAS macros will be made available shortly for handling the case o[f th]ree or more treatment groups.

Recommend 0 | Tweet 0 | +1 0

Project Researchers

Beth Ann Griffin
Senior Statistician

Beth Ann Griffin is a senior statistician at the RAND Corporation, where she codirects the RAND Center for Causal Inference. Her statistical research focuses on causal effects estimation when using observational data and innovative techniques for the analysis of time-to-event data. Her...

Explore[...]ple »

Stay Informed

Be notified when new studies, tools, and updates are released, or contact the TWANG team to ask a question about the effort.

Contact Form

Click on "Stay Informed" to keep up-to-date on tools available

Data acknowledgements

The development of these slides was funded by the National Institute of Drug Abuse grant number 1R01DA01569 (PIs: Griffin/McCaffrey) and was supported by the Center for Substance Abuse Treatment (CSAT), Substance Abuse and Mental Health Services Administration (SAMHSA) contract #270-07-0191 using data provided by the following grantees: Adolescent Treatment Model (Study: ATM: CSAT/SAMHSA contracts #270-98-7047, #270-97-7011, #277-00-6500, #270-2003-00006 and grantees: TI-11894, TI-11874; TI-11892), the Effective Adolescent Treatment (Study: EAT; CSAT/SAMHSA contract #270-2003-00006 and grantees: TI-15413, TI-15433, TI-15447, TI-15461, TI-15467, TI-15475,TI-15478, TI-15479, TI-15481, TI-15483, TI-15486, TI-15511, TI-15514,TI-15545, TI-15562, TI-15670, TI-15671, TI-15672, TI-15674, TI-15678, TI-15682, TI-15686, TI-15415, TI-15421, TI-15438, TI-15446, TI-15458, TI-15466, TI-15469, TI-15485, TI-15489, TI-15524, TI-15527, TI-15577, TI-15584, TI-15586, TI-15677), and the Strengthening Communities-Youth (Study: SCY; CSAT/SAMHSA contracts #277-00-6500, #270-2003-00006 and grantees: TI-13305, TI-13308, TI-13313, TI-13322, TI-13323, TI-13344, TI-13345, TI-13354). The authors thank these grantees and their participants for agreeing to share their data to support these secondary analyses. The opinions about these data are those of the authors and do not
reflect official positions of the government or individual grantees.

CHILDREN AND FAMILIES

EDUCATION AND THE ARTS

ENERGY AND ENVIRONMENT

HEALTH AND HEALTH CARE

INFRASTRUCTURE AND TRANSPORTATION

INTERNATIONAL AFFAIRS

LAW AND BUSINESS

NATIONAL SECURITY

POPULATION AND AGING

PUBLIC SAFETY

SCIENCE AND TECHNOLOGY

TERRORISM AND HOMELAND SECURITY

The RAND Corporation is a nonprofit institution that helps improve policy and decisionmaking through research and analysis.

This electronic document was made available from www.rand.org as a public service of the RAND Corporation.

Support RAND

Browse Reports & Bookstore

Make a charitable contribution

For More Information

Visit RAND at www.rand.org

Explore the RAND Corproration

View document details

Presentation

RAND presentations may include briefings related to a body of RAND research, videos of congressional testimonies, and a multimedia presentation on a topic or RAND capability. All RAND presentations represent RAND's commitment to quality and objectivity.

Limited Electronic Distribution Rights

This document and trademark(s) contained herein are protected by law as indicated in a notice appearing later in this work. This electronic representation of RAND intellectual property is provided for non-commercial use only. Unauthorized posting of RAND electronic documents to a non-RAND Web site is prohibited. RAND electronic documents are protected under copyright law. Permission is required from RAND to reproduce, or reuse in another form, any of our research documents for commercial use. For information on reprint and linking permissions, please see RAND Permissions.