Exploring finite-sample bias in propensity score weights

LUCY D'AGOSTINO MCGOWAN

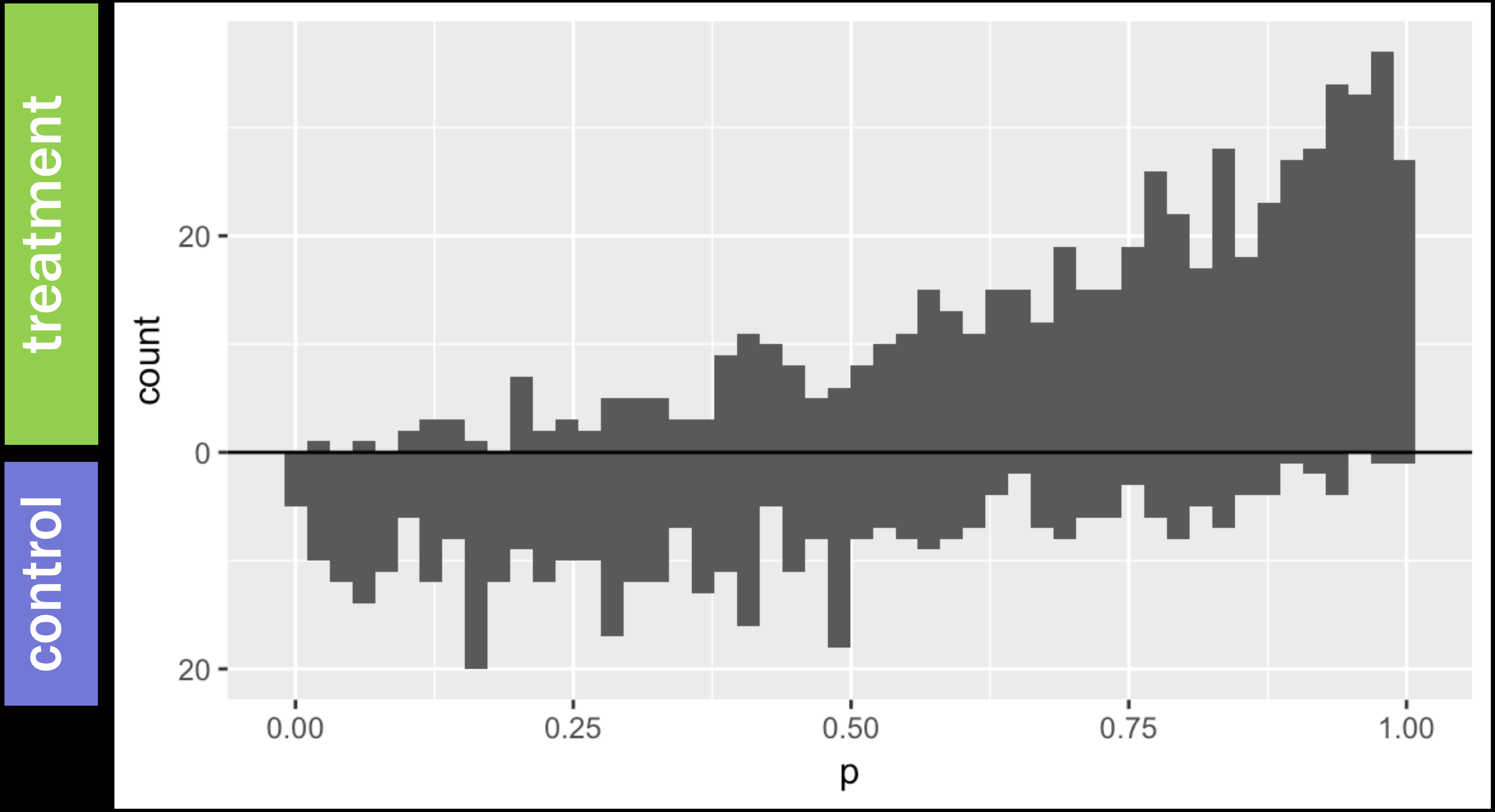
&

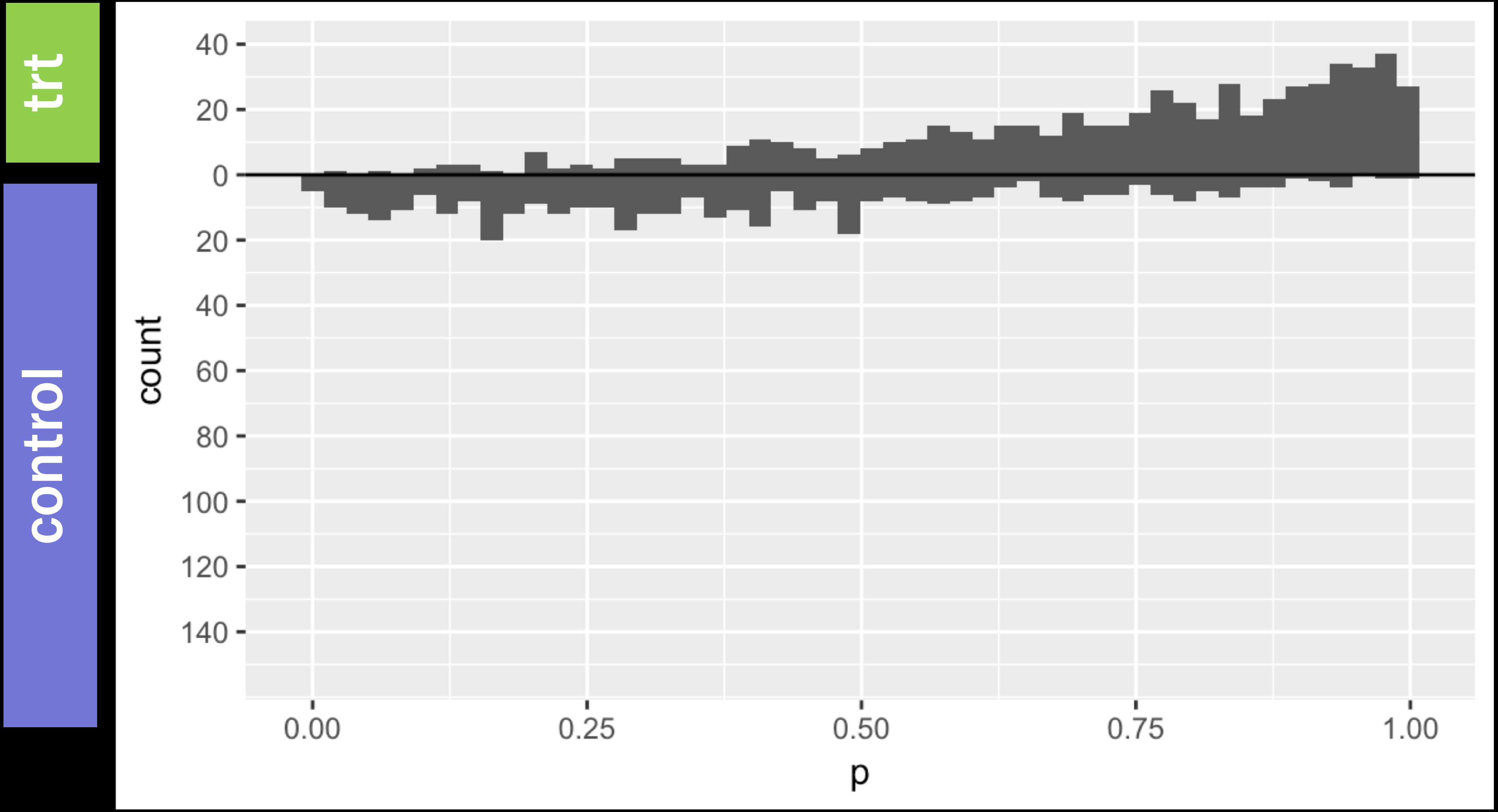
ROBERT GREEVY, JR.

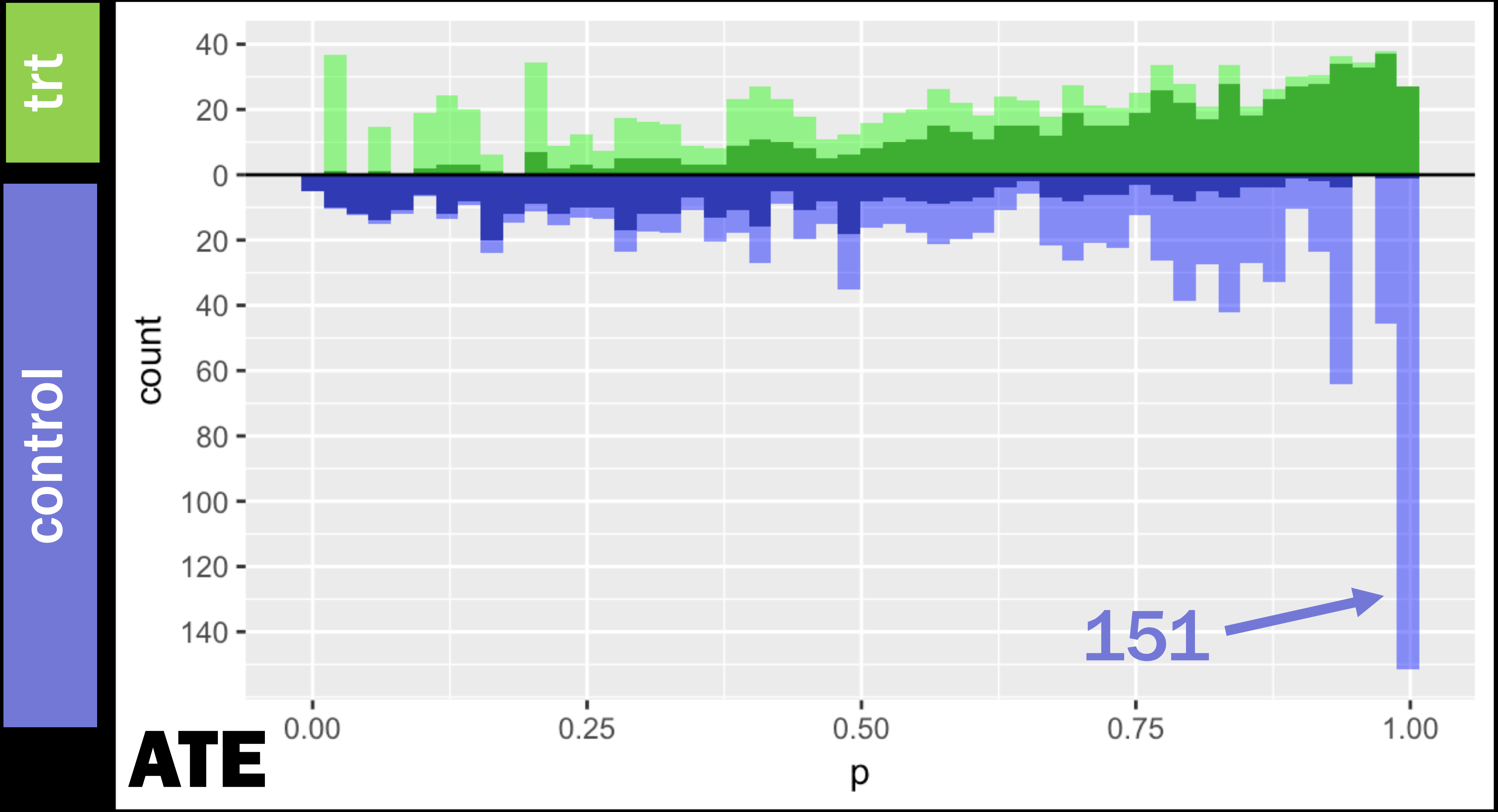
ENAR 2018

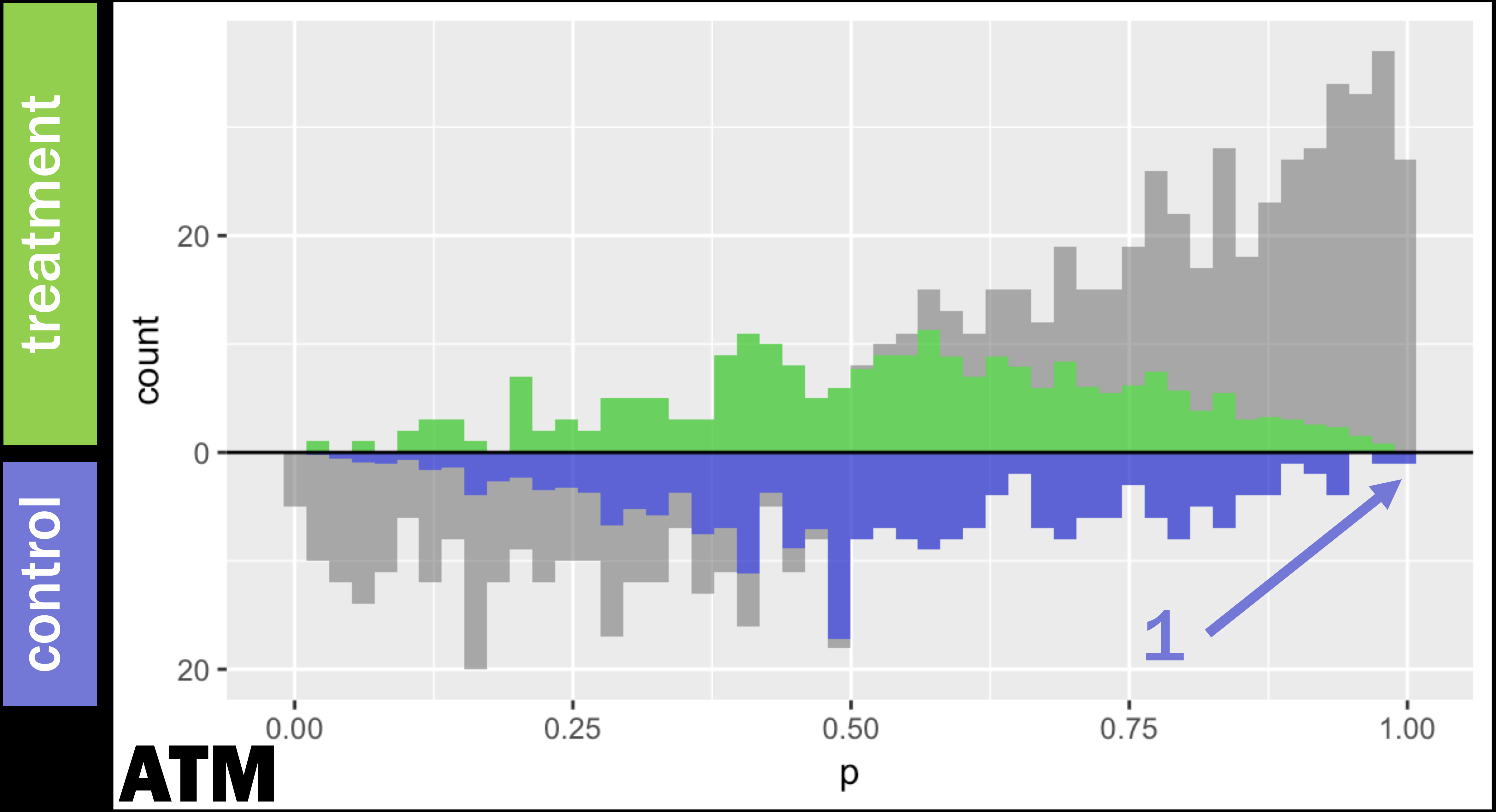
Overview

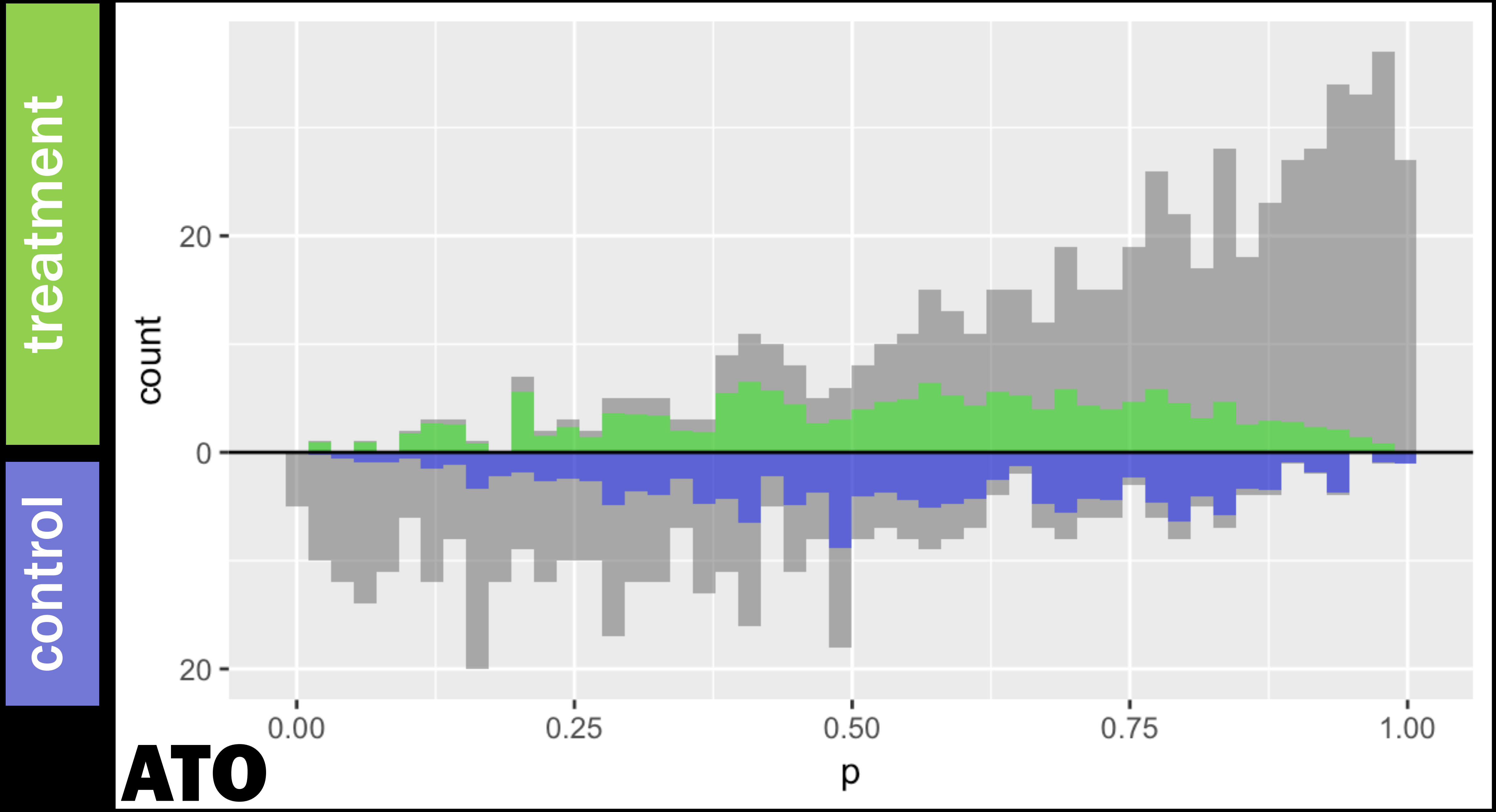
1. Finite sample bias
2. Unmeasured confounding – the problem
3. Unmeasured confounding – a solution


treatment
control
count
20
0
20
0.00
0.25
0.50
0.75
1.00
p


trt
control
count
40
20
0
20
40
60
80
100
120
140
0.00
0.25
0.50
0.75
1.00
p

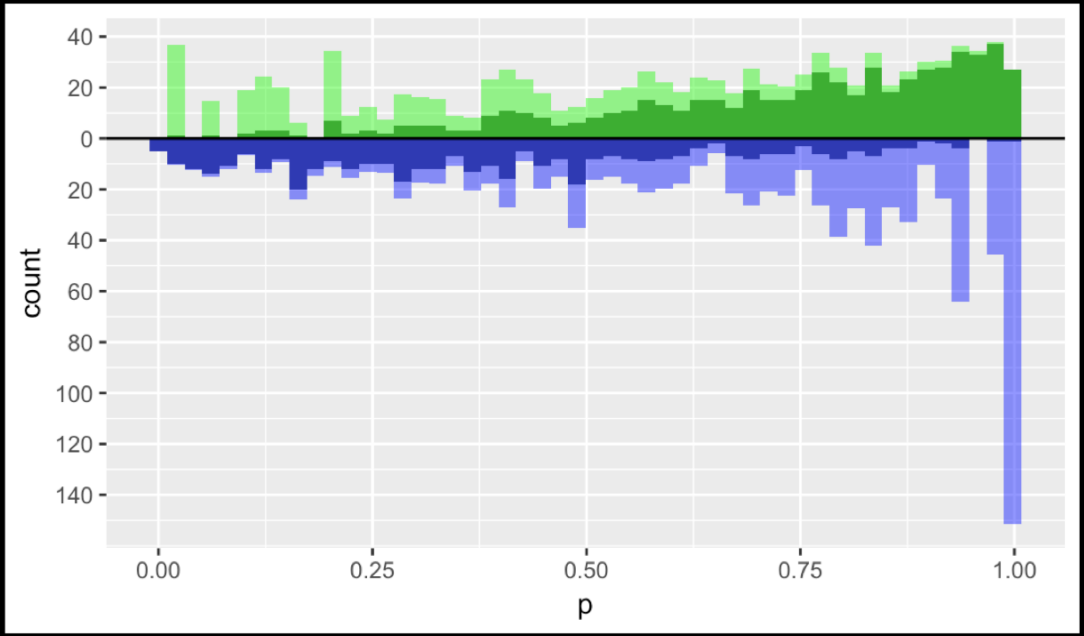

trt
control
40
20
0
20
40
60
80
100
120
140
count
151
ATE
0.00
0.25
0.50
0.75
1.00
p

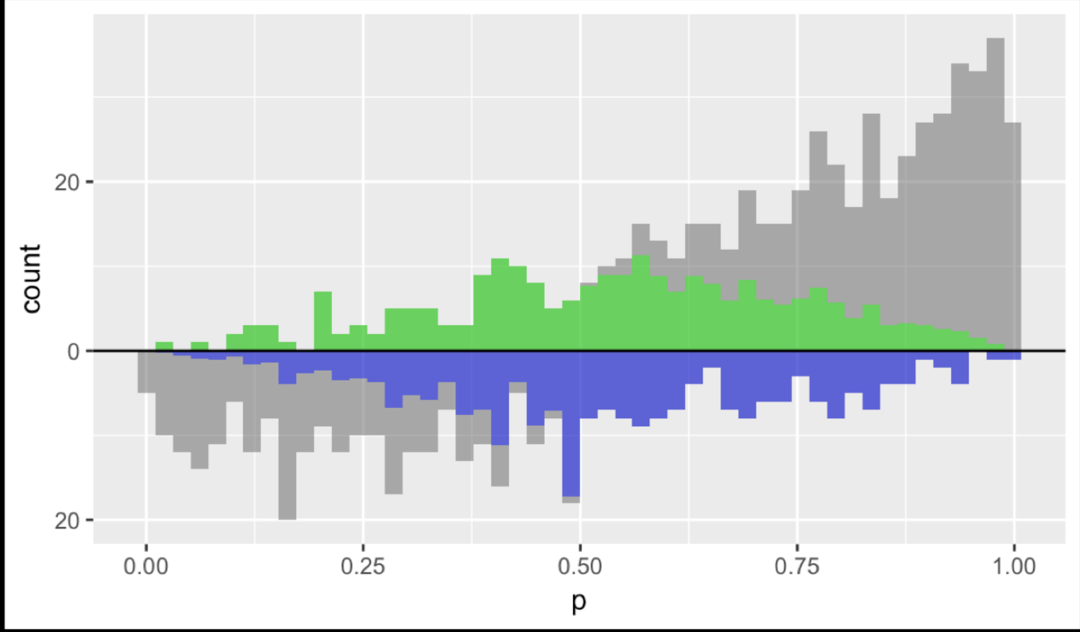

treatment
control
count
20
0
20
0.00
0.25
0.50
0.75
1.00
p
1
ATM



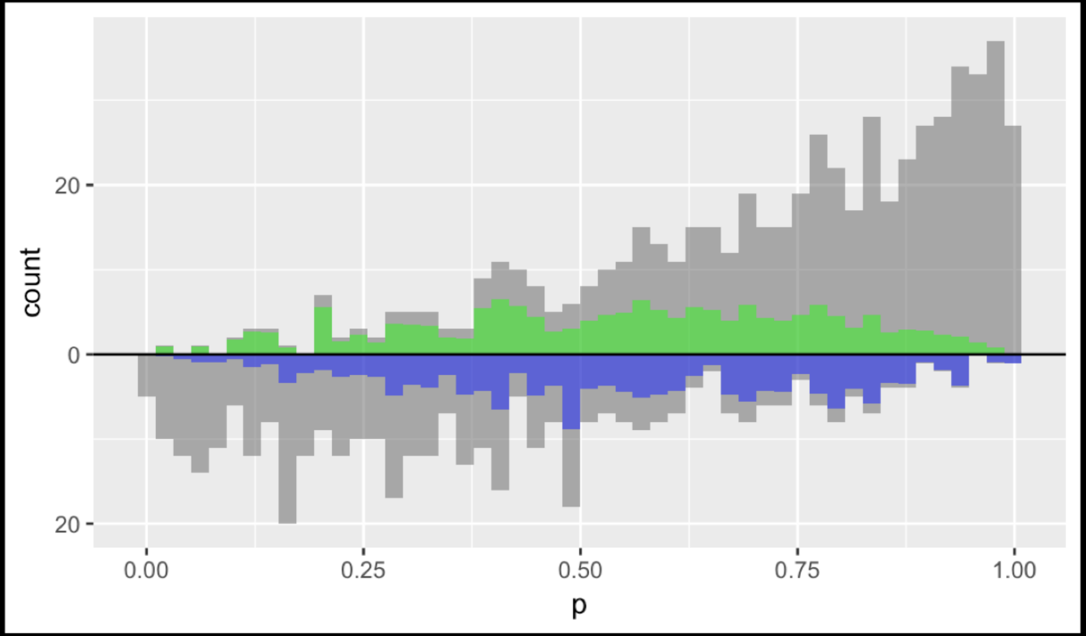
treatment
control
count
20
0
20
0.00
0.25
0.50
0.75
1.00
p
ATO

ATE



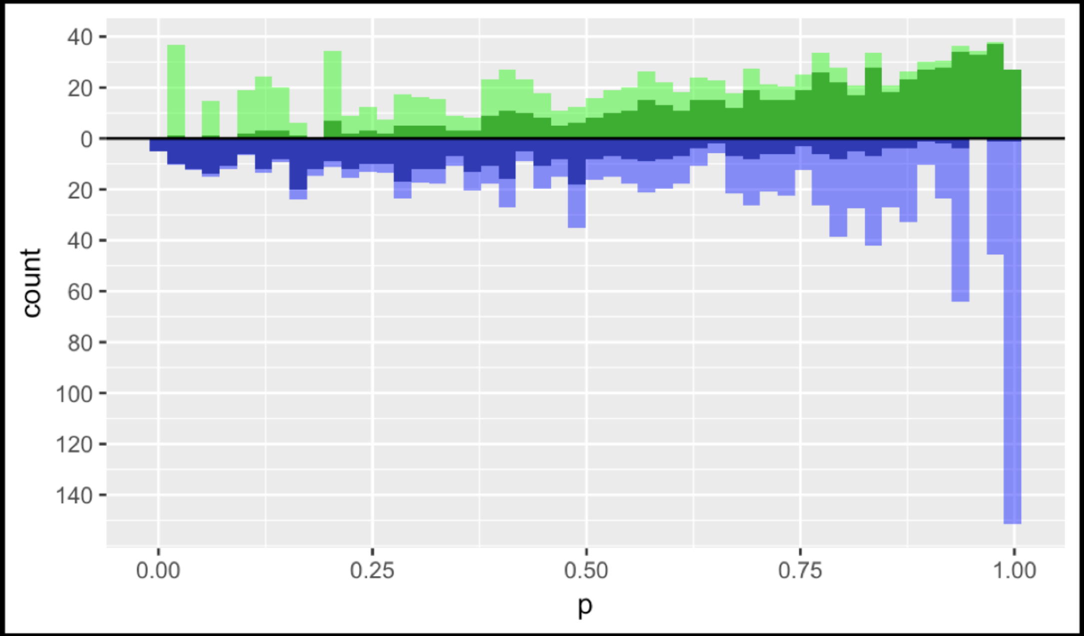


ATM

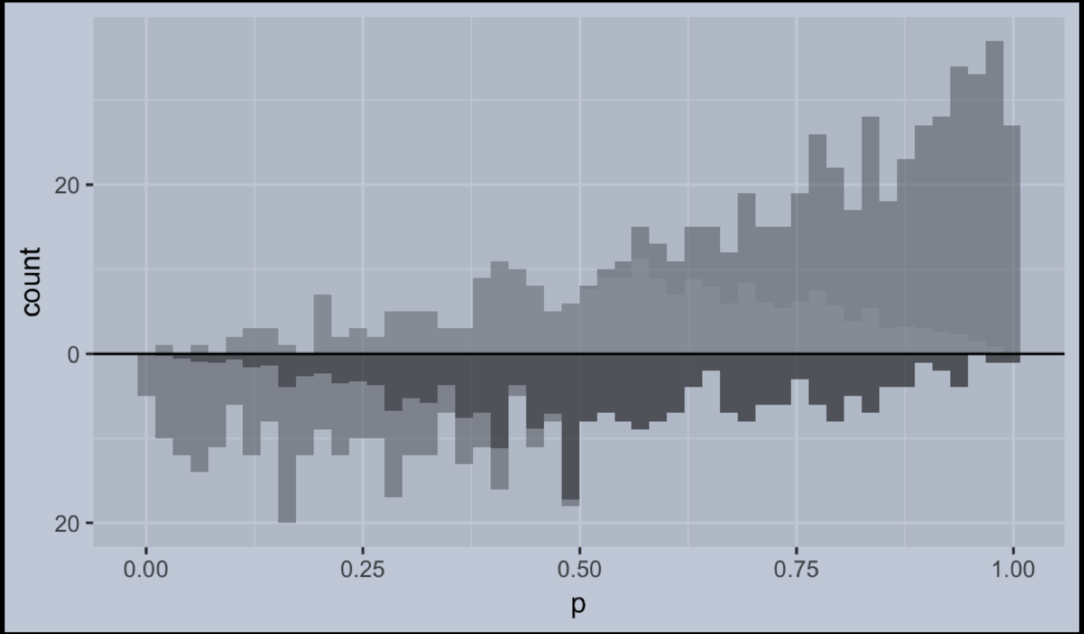


ATO

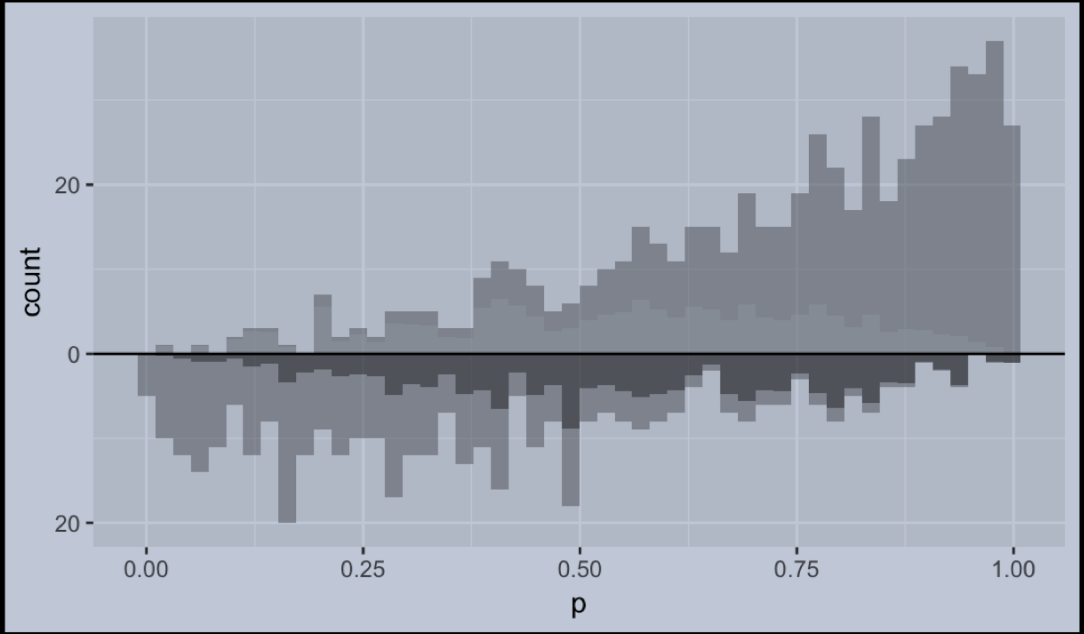
ATE



40
20
0
20
40
60
80
100
120
140
count
0.00
0.25
0.50
0.75
1.00
p



20
0
20
count
0.00
0.25
0.50
0.75
1.00
p

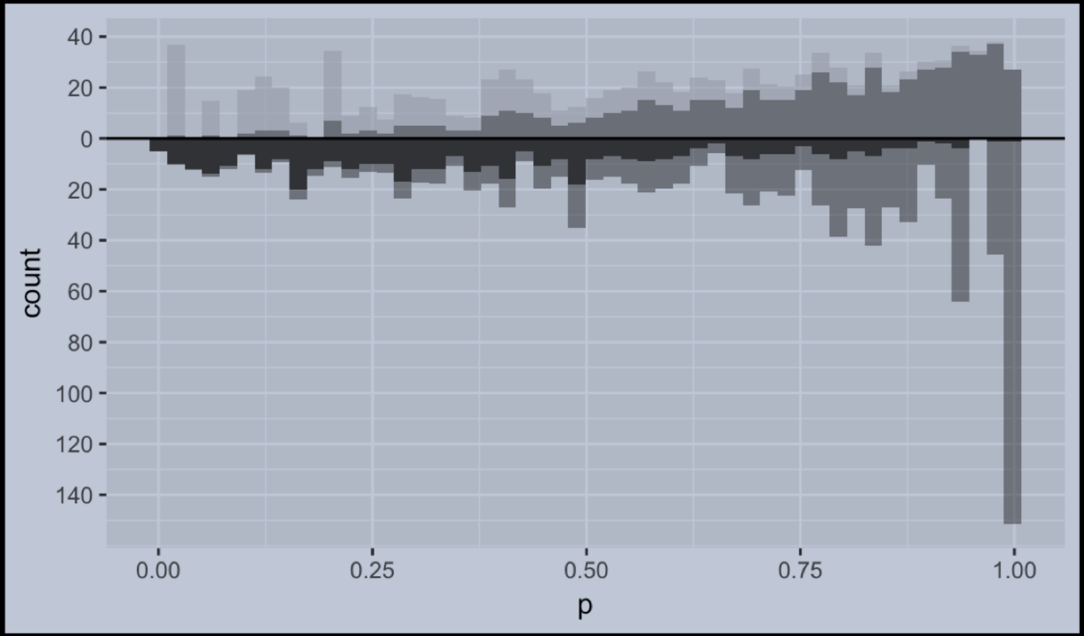


20
0
20
count
0.00
0.25
0.50
0.75
1.00
p

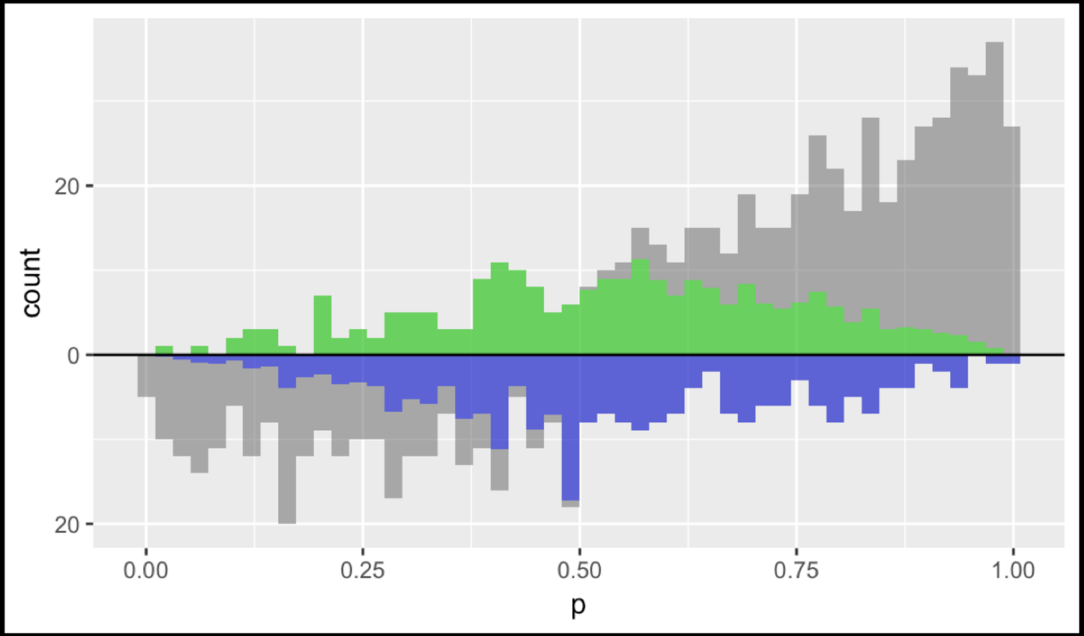
ATM

ATO

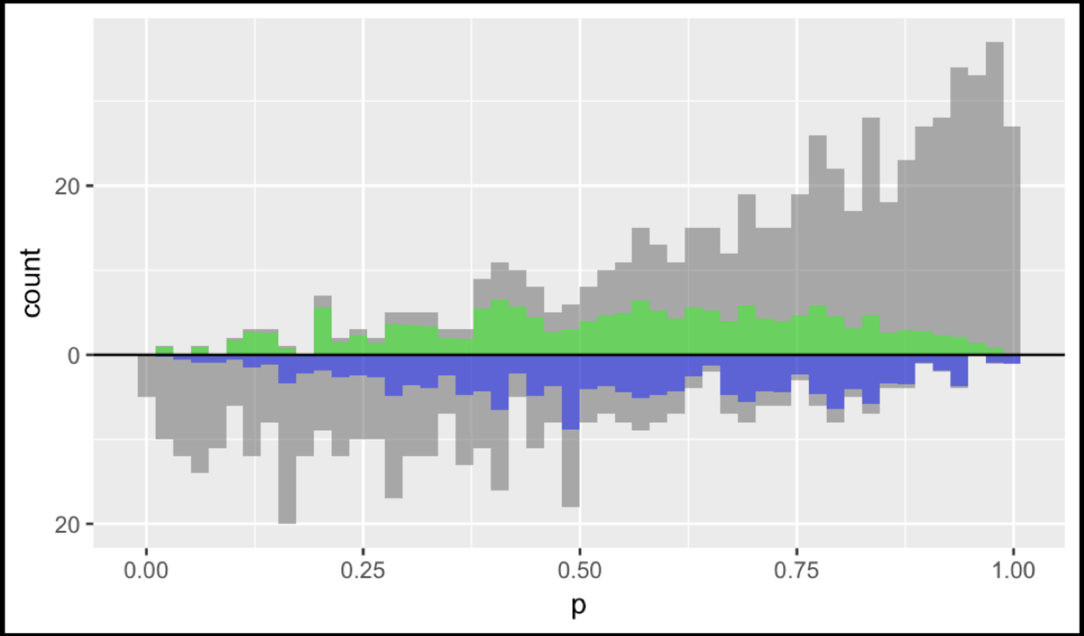
ATE



count
40
20
0
20
40
60
80
100
120
140
0.00
0.25
0.50
0.75
1.00
p



count
20
0
20
0.00
0.25
0.50
0.75
1.00
p



count
20
0
20
0.00
0.25
0.50
0.75
1.00
p

ATM

ATO

Finite-sample bias

Evaluation Review
Volume 32 Number 4
August 2008 392-409

10.1177/0193841X08317586
http://erx.sagepub.com
hosted at
http://online.sagepub.com

Weighting Regressions by Propensity Scores

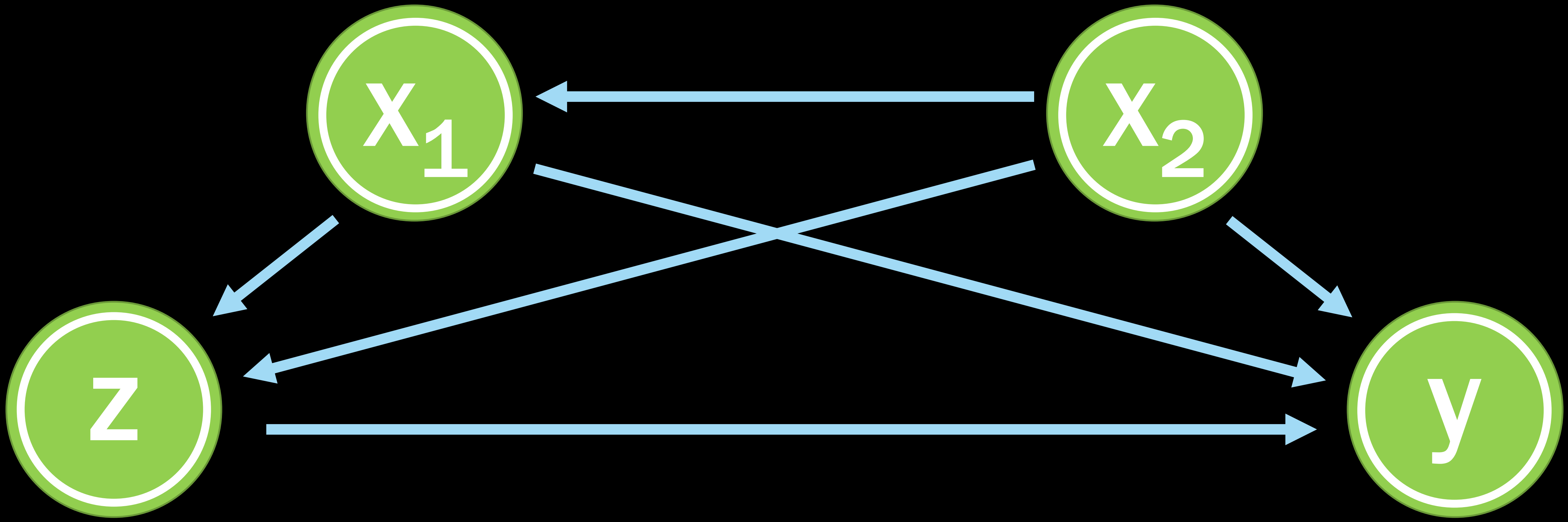
David A. Freedman
University of California, Berkeley
Richard A. Berk
University of Pennsylvania

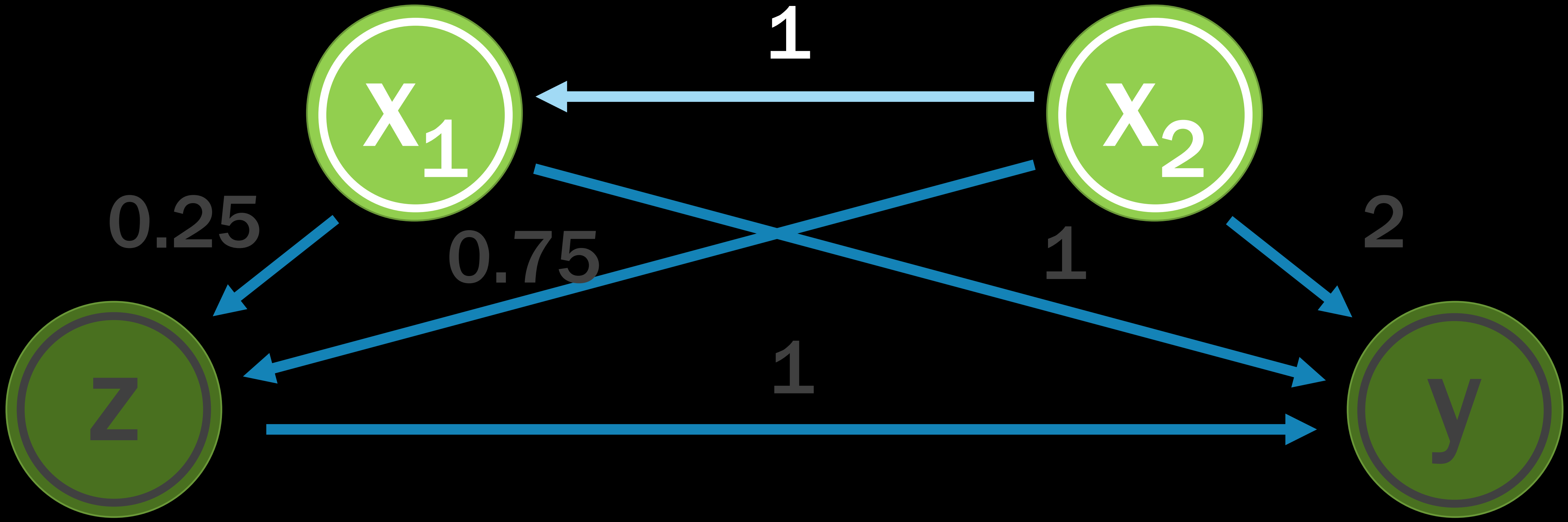
Regressions can be weighted by propensity scores in order to reduce bias. However, weighting is likely to increase random error in the estimates, and to bias the estimated standard errors downward, even when selection mechanisms are well understood. Moreover, in some cases, weighting will increase the bias in estimated causal parameters. If investigators have a good causal model, it seems better just to fit the model without weights. If the causal model is improperly specified, there can be significant problems in retrieving the situation by weighting, although weighting may help under some circumstances.

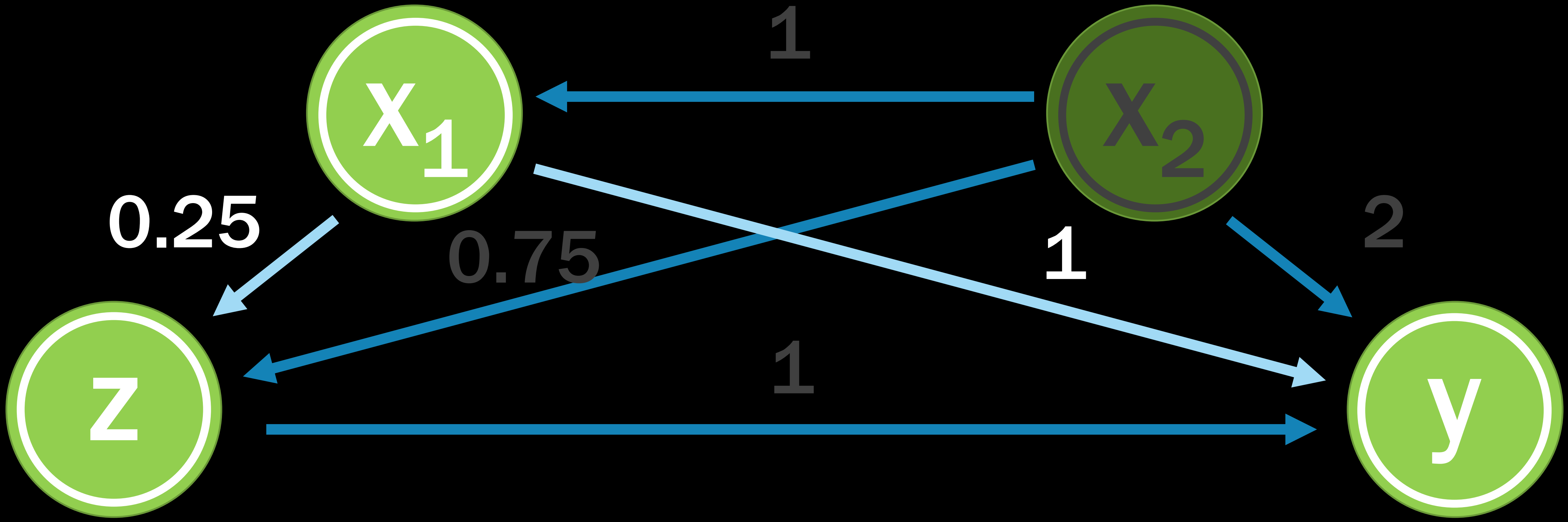
Keywords: *causation; selection; models; experiments; observational studies; regression; propensity scores*

Estimating causal effects is often the key to evaluating social programs, but the interventions of interest are seldom assigned at random. Observational data are therefore frequently encountered. In order to

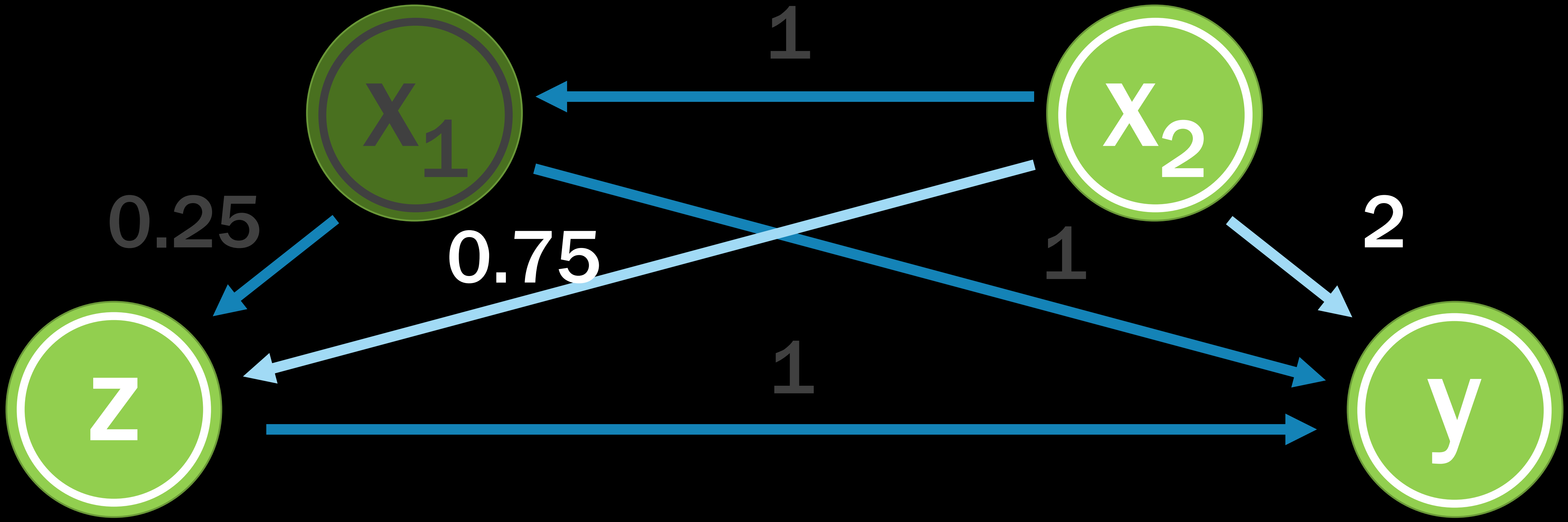
Freedman & Berk Simulation


x1
x2
z
y

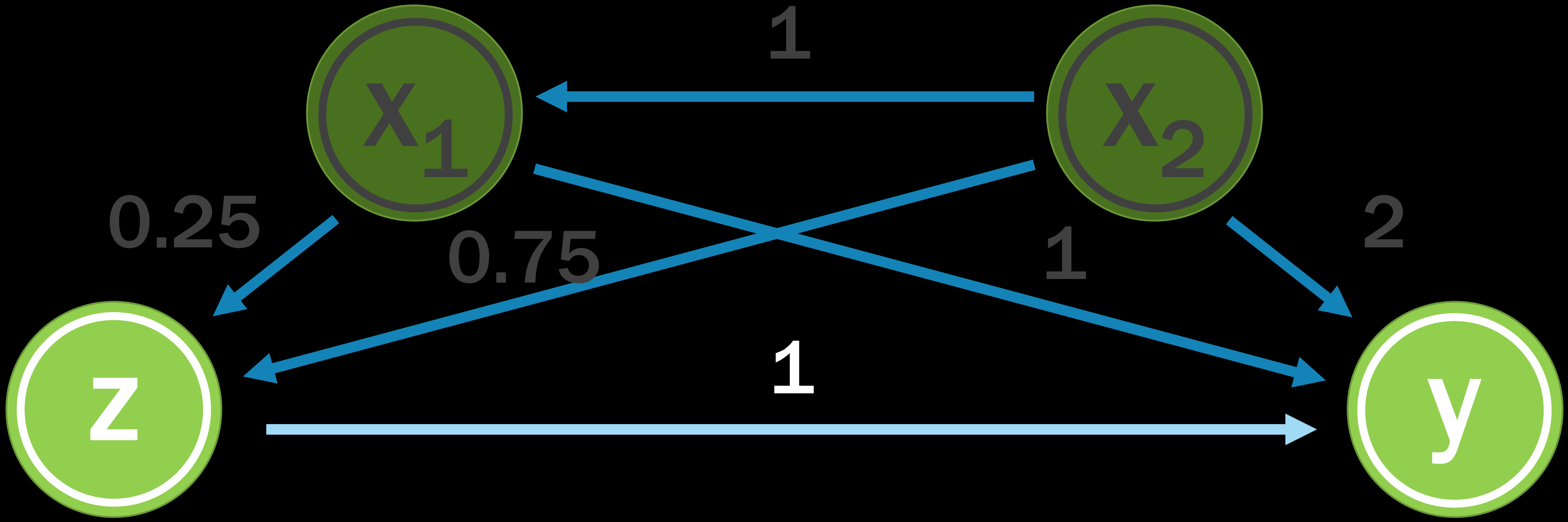

1
x_1
x_2
0.25
0.75
1
2
z
1
y



x_1
x_2
z
y
1
0.25
0.75
1
2
1

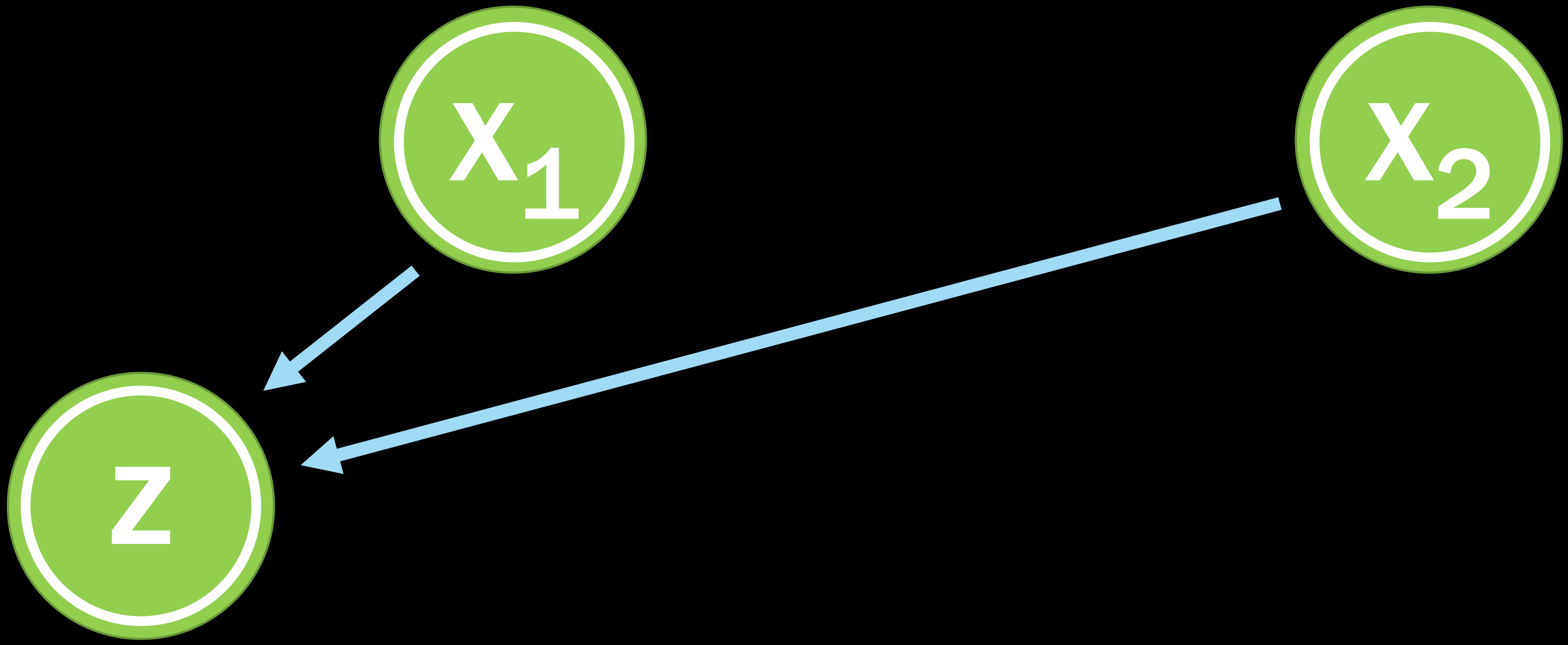


1
x_1
x_2
0.25
0.75
1
2
1
z
y

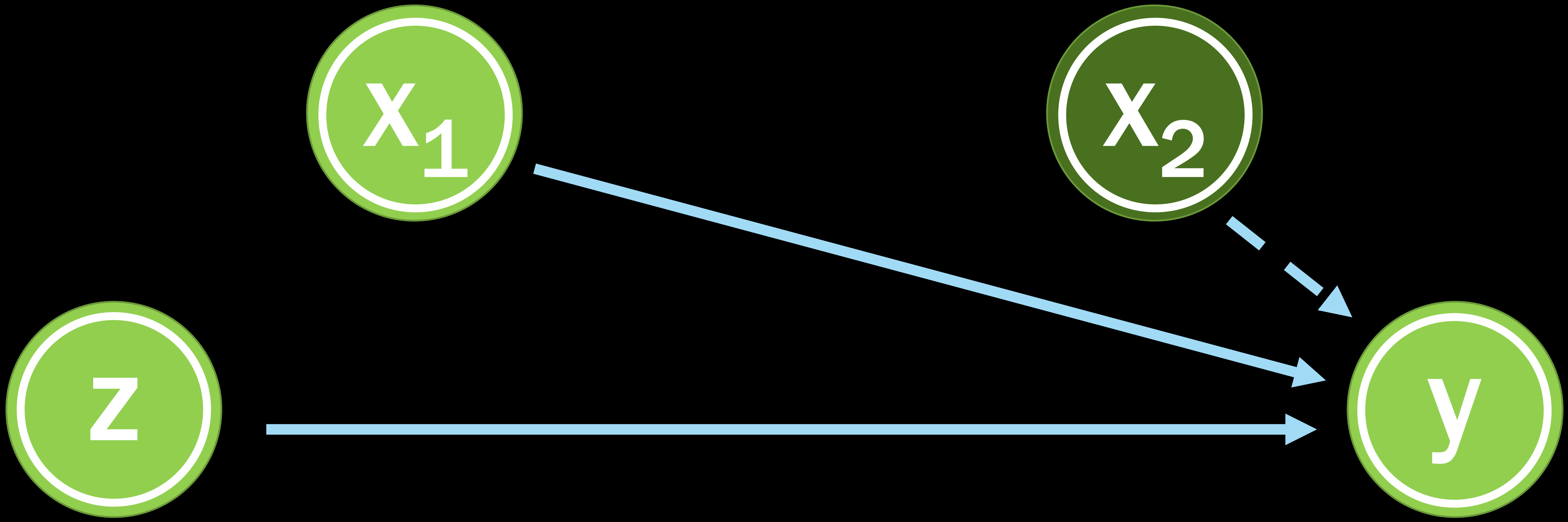


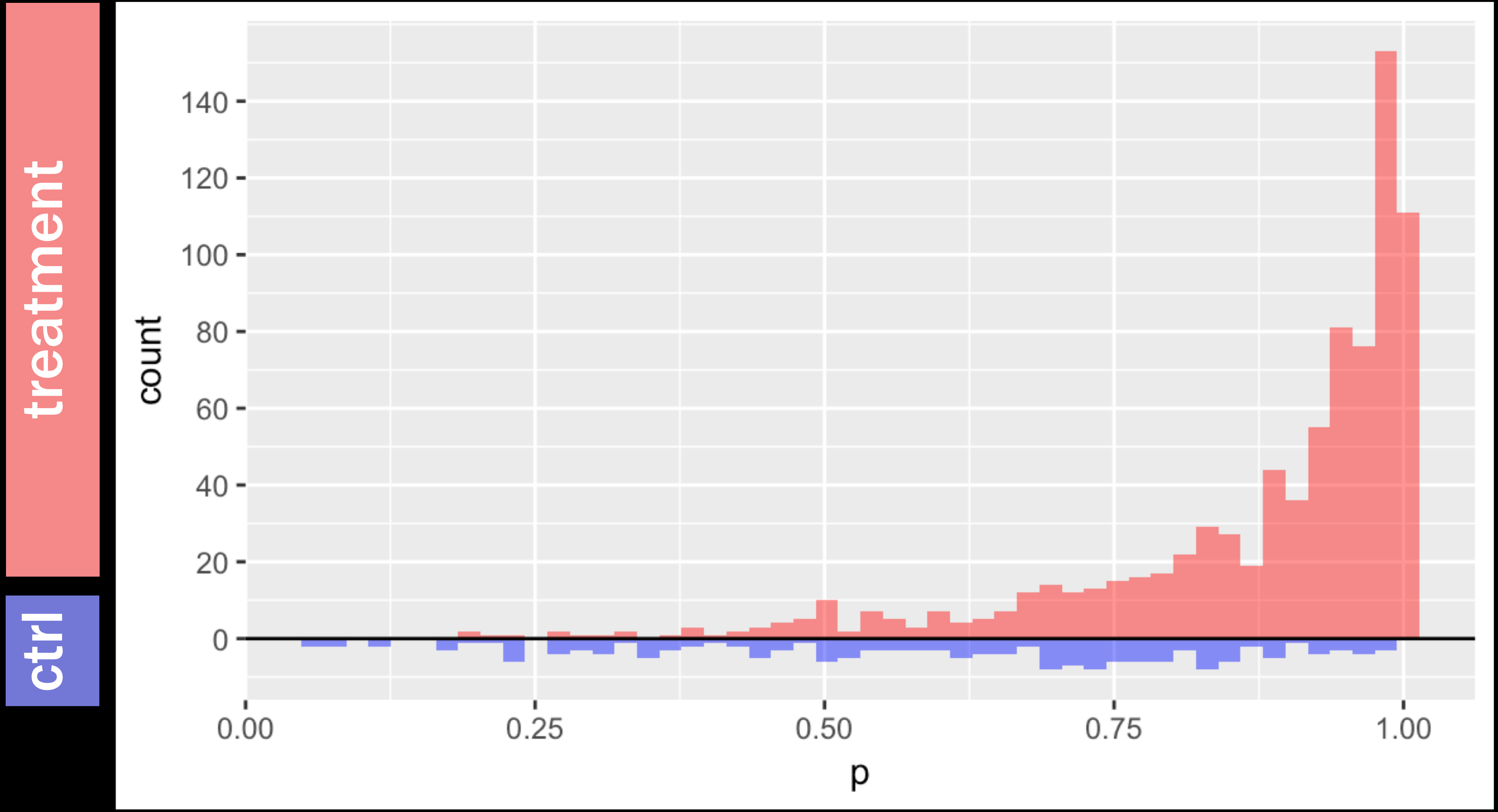
1
x_1
x_2
0.25
0.75
1
2
z
1
y

Propensity score model

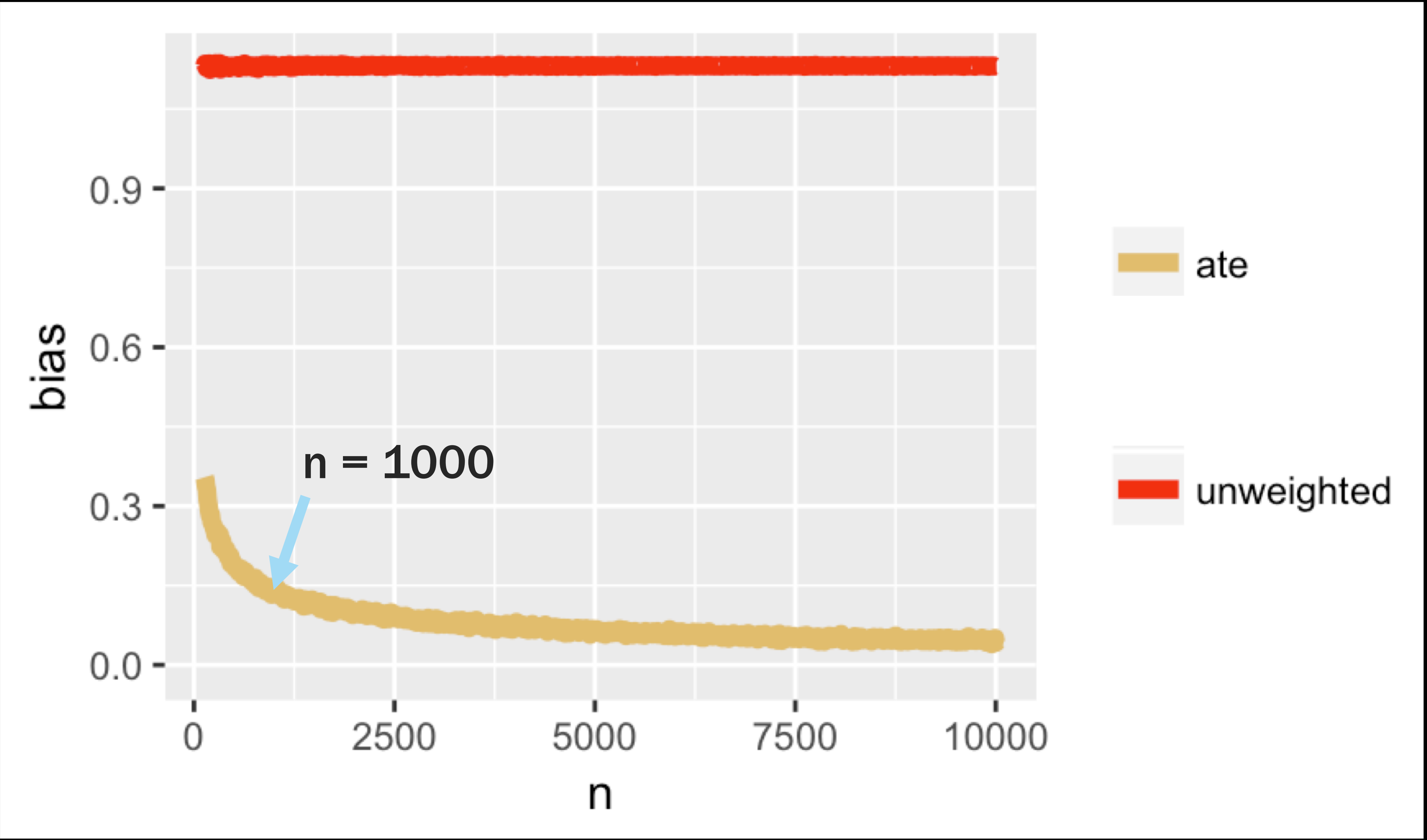


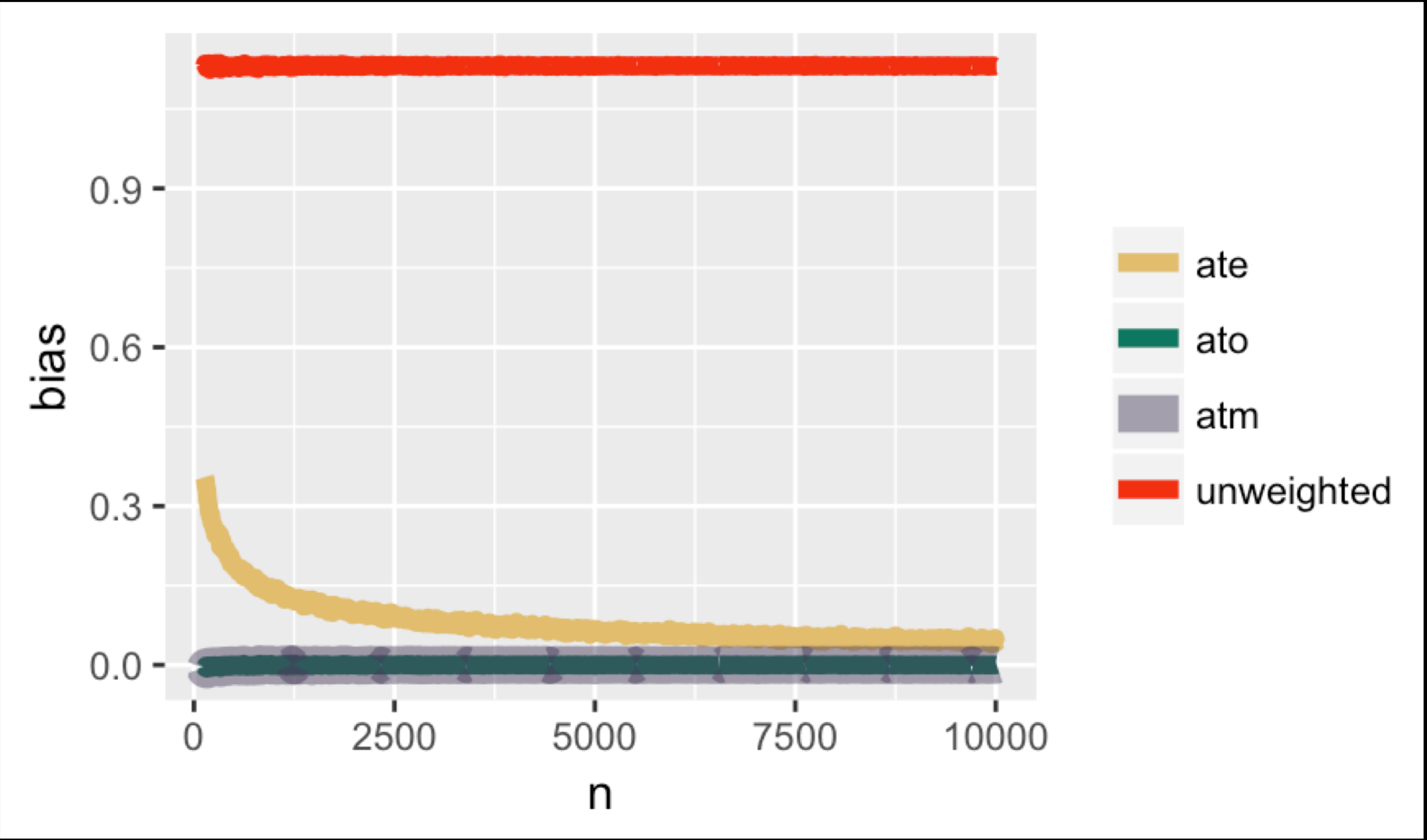
Outcome model





treatment
ctrl
count
140
120
100
80
60
40
20
0
0.00
0.25
0.50
0.75
1.00
p

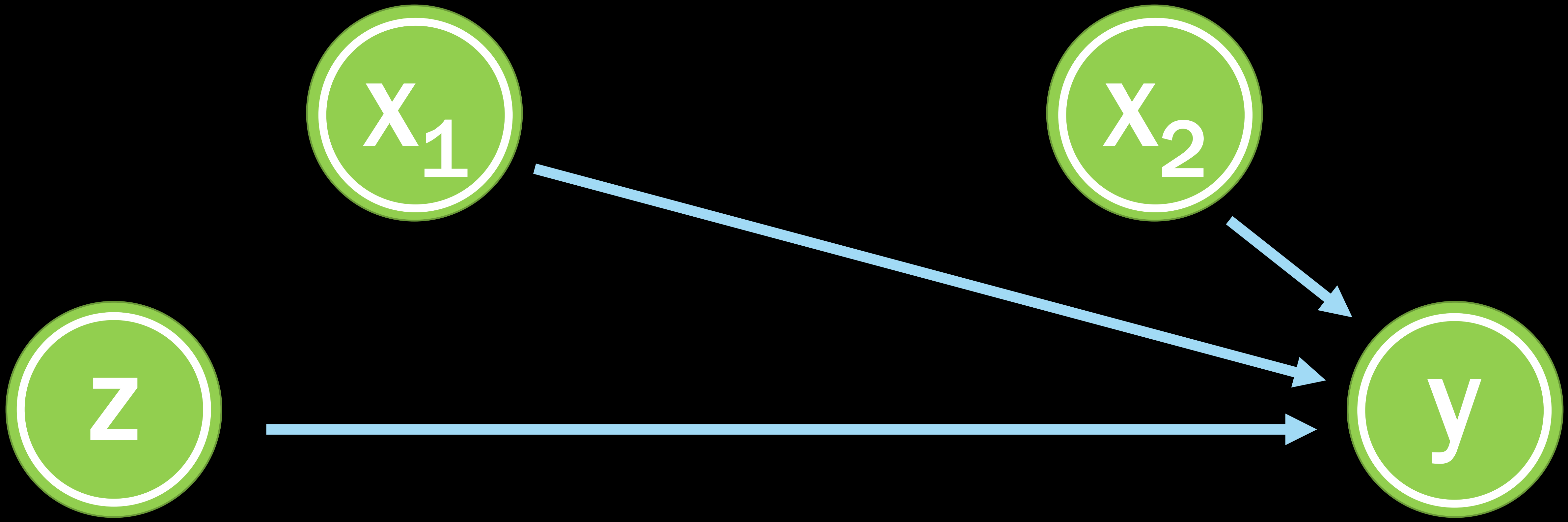

0.9
0.6
0.3
0.0
bias
0
2500
5000
7500
10000
n
n = 1000
ate
unweighted

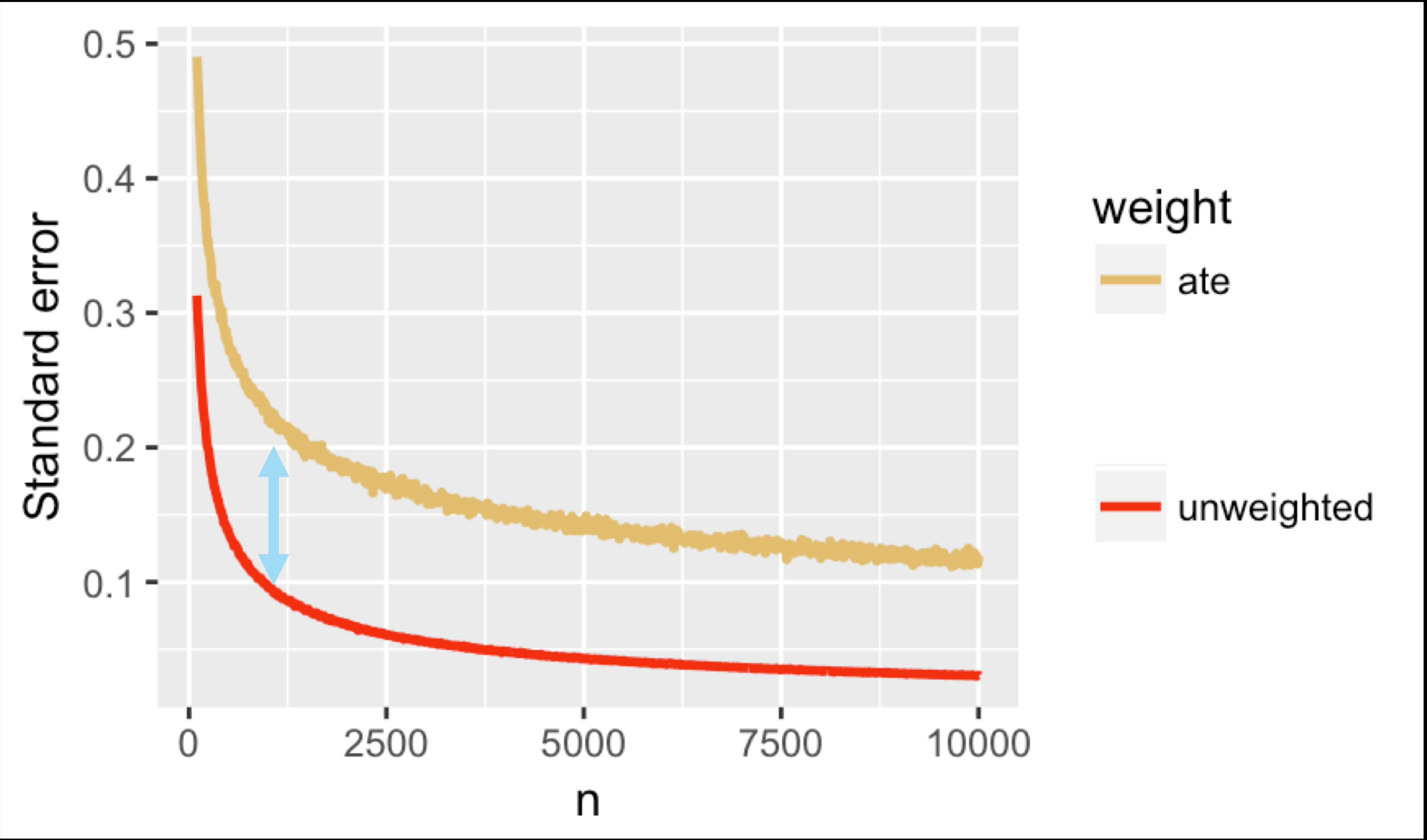


bias
0.9
0.6
0.3
0.0
0
2500
5000
7500
10000
n
ate
ato
atm
unweighted

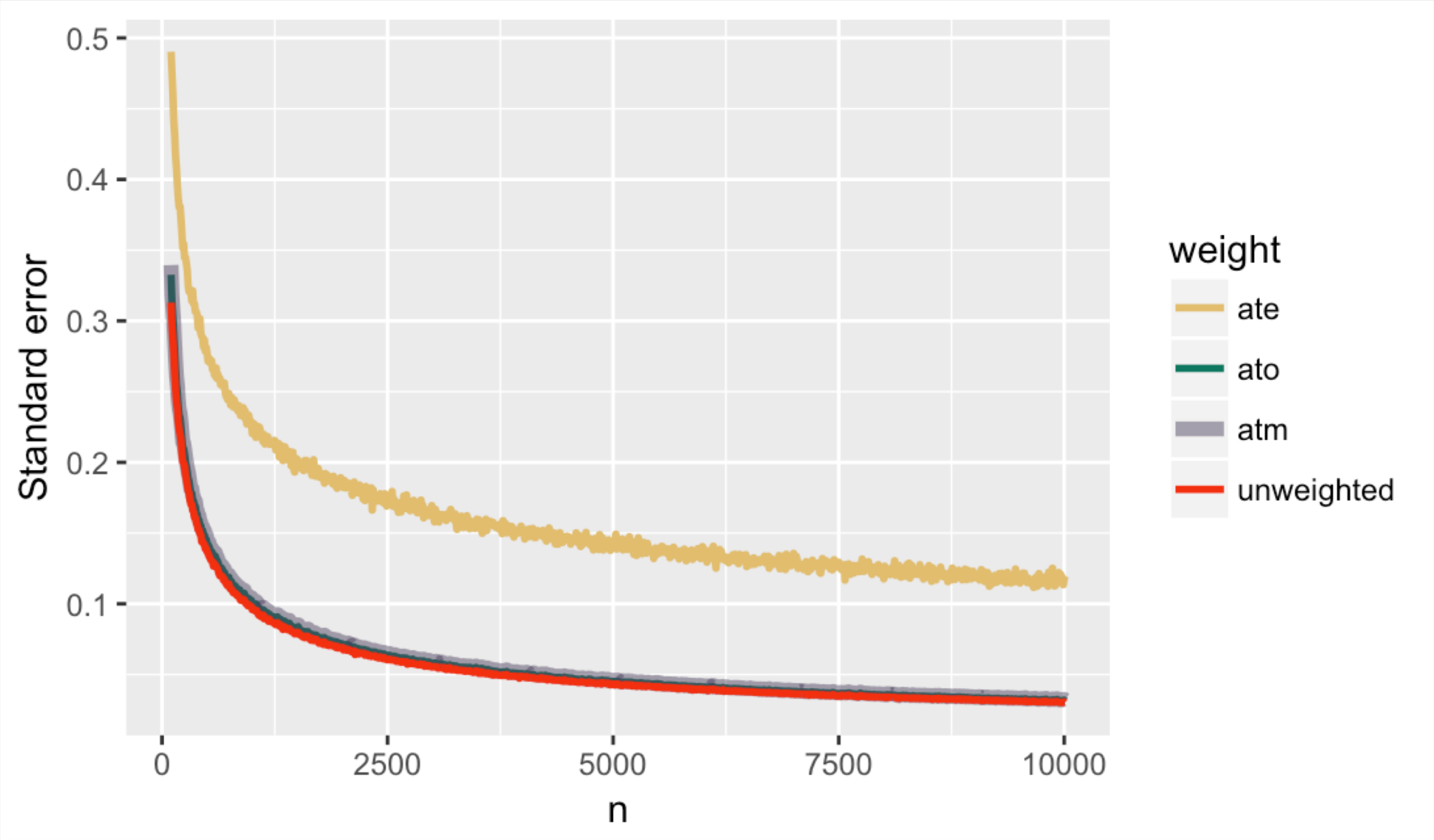
Standard errors

Outcome model





0.5
0.4
0.3
0.2
0.1
Standard error
0
2500
5000
7500
10000
n
weight
ate
unweighted



Standard error
0.5
0.4
0.3
0.2
0.1
0
2500
5000
7500
10000
n
weight
ate
ato
atm
unweighted

Recap

- **Replicated** the finite-sample bias seen by Freedman and Berk using the **ATE** weights
- **ATM** and **ATO** weights had **improved** finite-sample properties
- The **variance** for the **ATO** and **ATM** is **preferable** to that of the ATE

Unmeasured confounding
the problem

ed: 1 January 2017 | Accepted: 3 January 2017
10.1111/jep.12714
WILEY Journal of Evaluation
International Journal of Public Health

RIGINAL ARTICLE

Improving causal inference with a doubly rob
combines propensity score stratification and

Ariel Linden DrPH[1,2]

[1] President, Linden Consulting Group, LLC Ann Arbor, MI, USA
[2] Research Scientist, Division of General Medicine, Medical School, University of Michigan, Ann Arbor, MI, USA

Correspondence
Ariel Linden, DrPH, Linden Consulting Group, LLC, 1301 North Bay Drive, Ann Arbor, MI 48103, USA.
Email: alinden@lindenconsulting.org

Abstract

Rationale, aims and objectives Wh
researchers typically use observational
founding when estimating treatment
estimators based on a model for
weighted estimators based on th
and (3) "doubly robust" (DR) esti
the same framework. In this p
weighting through stratifica
may prove more accurat
to be more accurate tha
compare the perform
estimators.

Method Mont
adjustment, 2 w
assess perfor
sity score m

Results
compo
estimato
equally as well as

Conclusions Health re
strategy in observational studies,
class.

al Statistics, 2007

Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data[1]

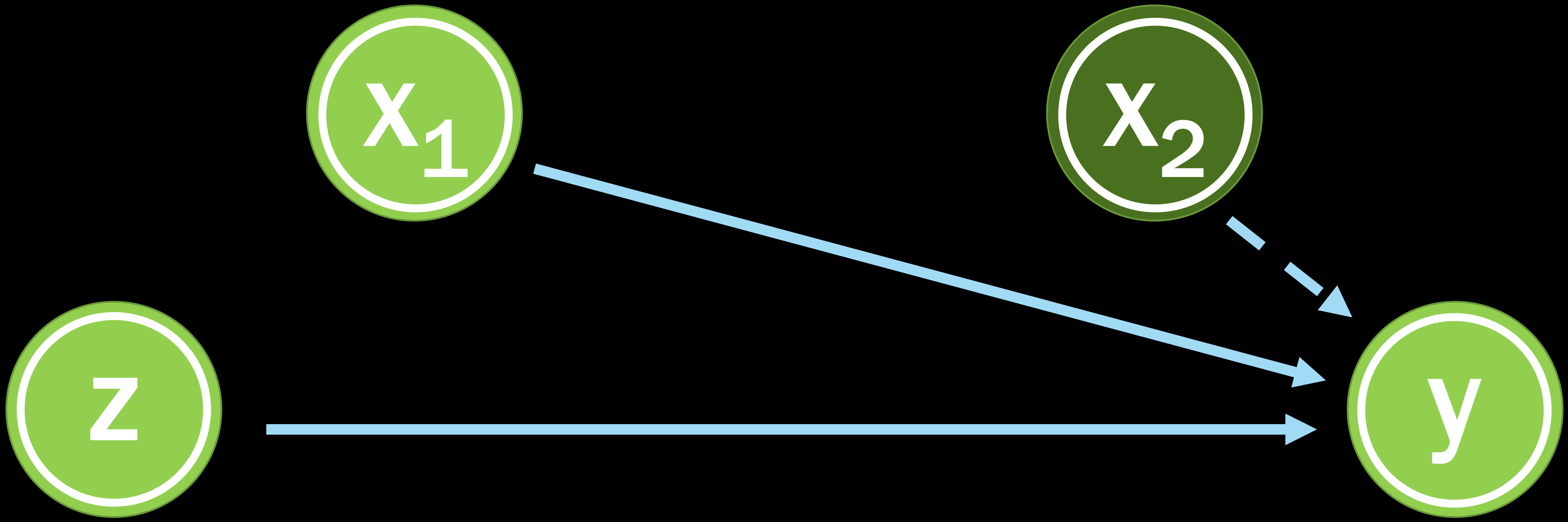
Joseph D. Y. Kang and Joseph L. Schafer

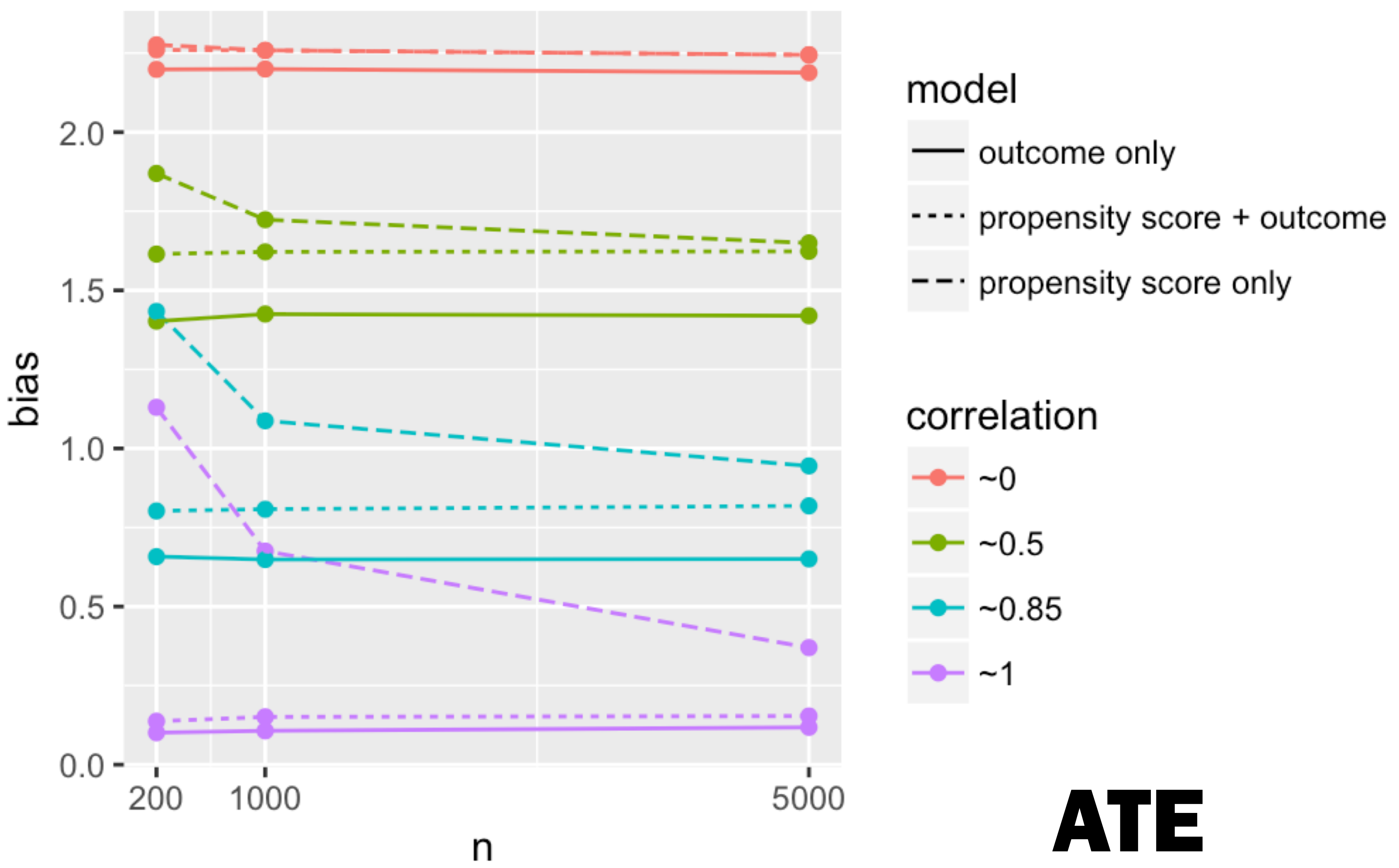
Abstract. When outcomes are missing for reasons beyond an investigator's control, there are two different ways to adjust a parameter estimate for covariates that may be related both to the outcome and to missingness. One approach is to model the relationships between the covariates and the outcome and use those relationships to predict the missing values. Another is to model the probabilities of missingness given the covariates and incorporate them into a weighted or stratified estimate. Doubly robust (DR) procedures apply both types of model simultaneously and produce a consistent estimate of the parameter if either of the two models has been correctly specified. In this article, we show that DR estimates can be constructed in many ways. We compare the performance of various DR and non-DR estimates of a population mean in a simulated example where both models are incorrect but neither is grossly misspecified. Methods that use inverse-probabilities as weights, whether they are DR or not, are sensitive to misspecification of the propensity model when some estimated propensities are small. Many DR methods perform better than simple inverse-probability weighting. None of the DR methods we tried, however, improved upon the performance of simple regression-based prediction of the missing values. This study does not represent every missing-data problem that will arise in practice. But it does demonstrate that, in at least some settings, two wrong models are not better than one.

Key words and phrases: Causal inference, missing data, propensity score, model-assisted survey estimation, weighted estimating equations.

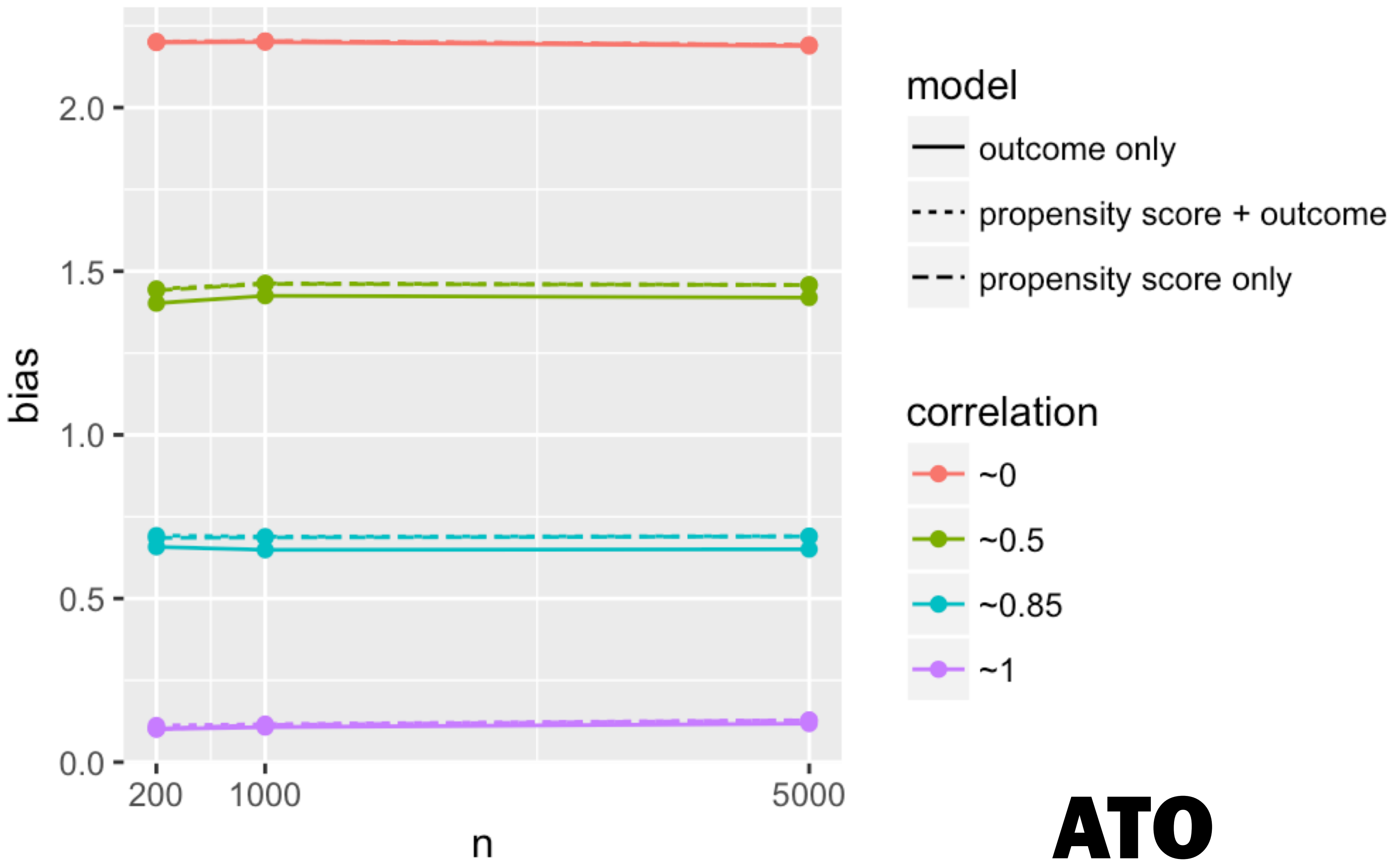
Propensity score model

Outcome model




model
outcome only
propensity score + outcome
propensity score only
correlation
~0
~0.5
~0.85
~1
bias
2.0
1.5
1.0
0.5
0.0
200
1000
5000
n

ATE



2.0
1.5
1.0
0.5
0.0
bias
200
1000
5000
n
model
outcome only
propensity score + outcome
propensity score only
correlation
~0
~0.5
~0.85
~1

ATO

Unmeasured confounding
a solution

E-value

$$\mathrm{E-value} = LB_{obs} + \sqrt{LB_{obs} \times (LB_{obs} - 1)}$$

VanderWeele and Ding (2017)

Adjusted E-value

$$\mathrm{E-value}_{adj} = \frac{LB_{obs}}{LB_{adj}} + \sqrt{\frac{LB_{obs}}{LB_{adj}} \times \left(\frac{LB_{obs}}{LB_{adj}} - 1\right)}$$

Adjusted E-value

$$\mathrm{E-value}_{adj} = \frac{LB_{obs}}{LB_{adj}} + \sqrt{\frac{LB_{obs}}{LB_{adj}} \times \left(\frac{LB_{obs}}{LB_{adj}} - 1\right)}$$

Right Heart Catheterization **Data**

Connors et al (1996)

ENAR 2018

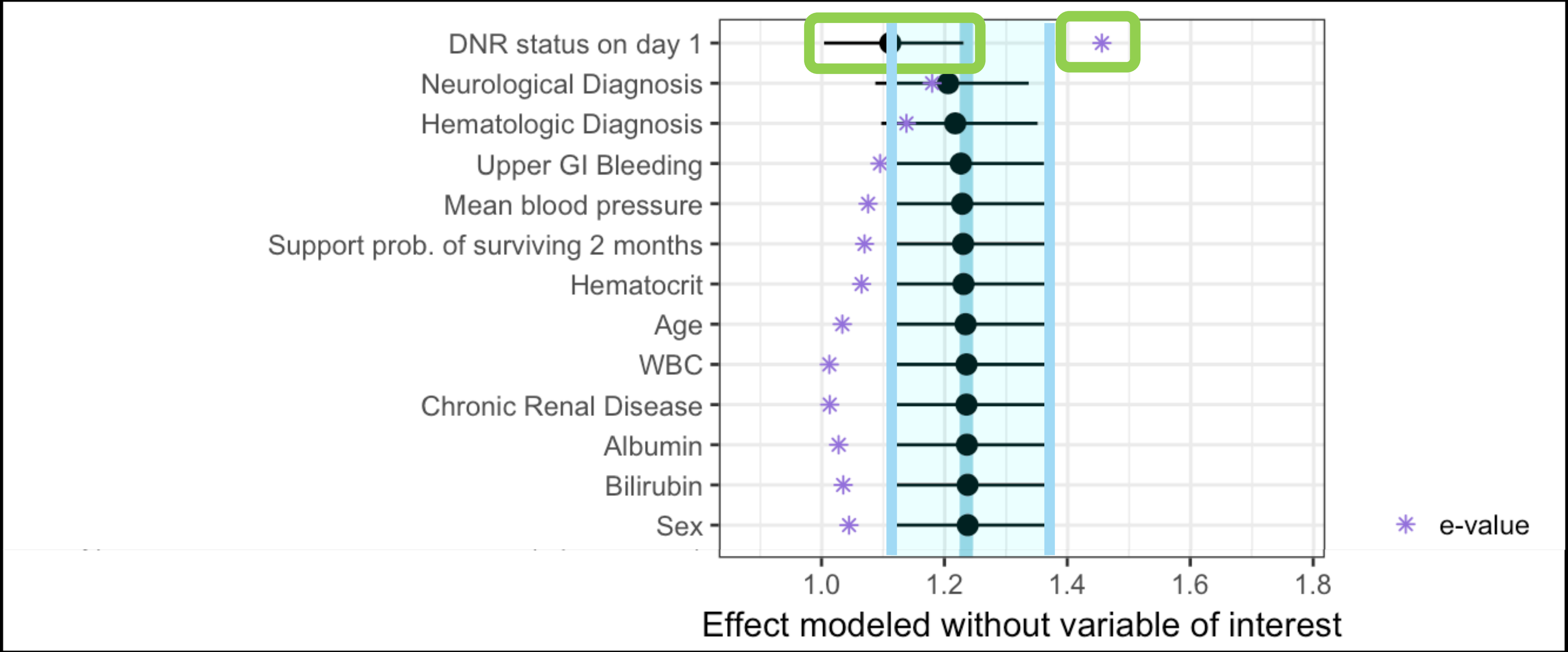
Right Heart Catheterization data

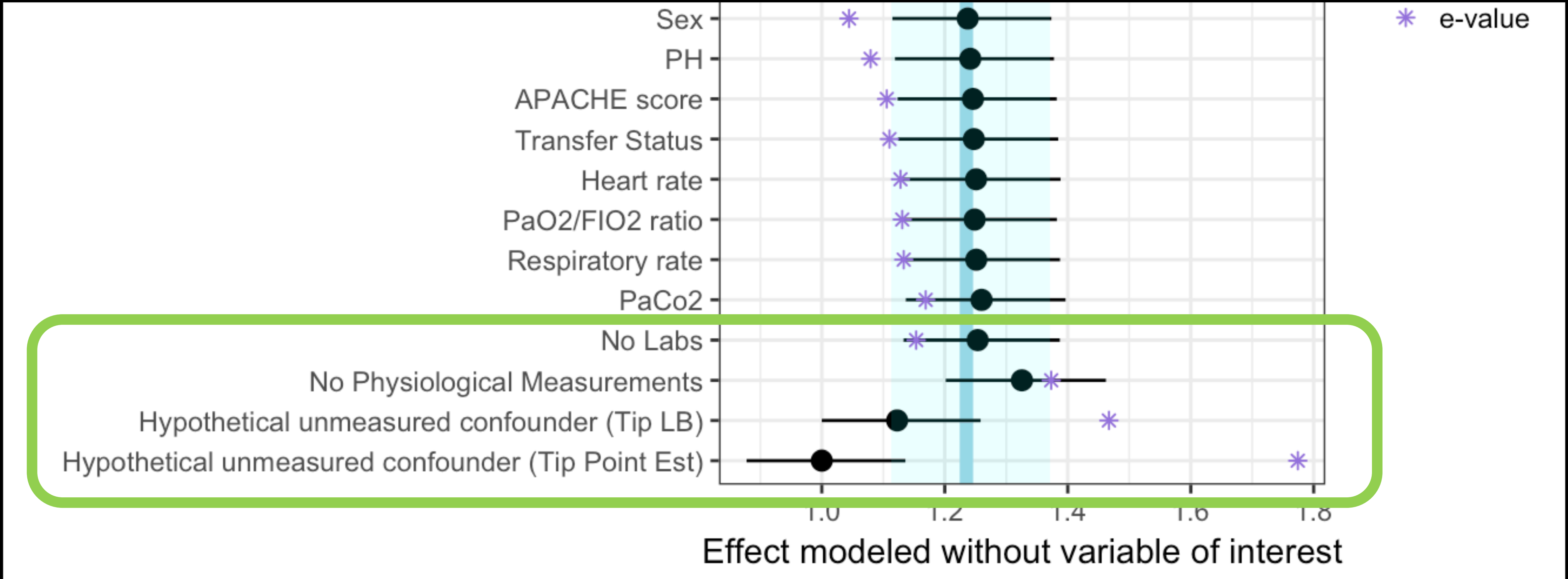
- We chose 20 covariates for demonstration purposes
  - demographics
  - comorbidities
  - physiological measurements
  - diagnosis categories
  - APACHE score
  - SUPPORT (probability of surviving 2 months)
  - DNR status on day 1

Connors et al (1996)

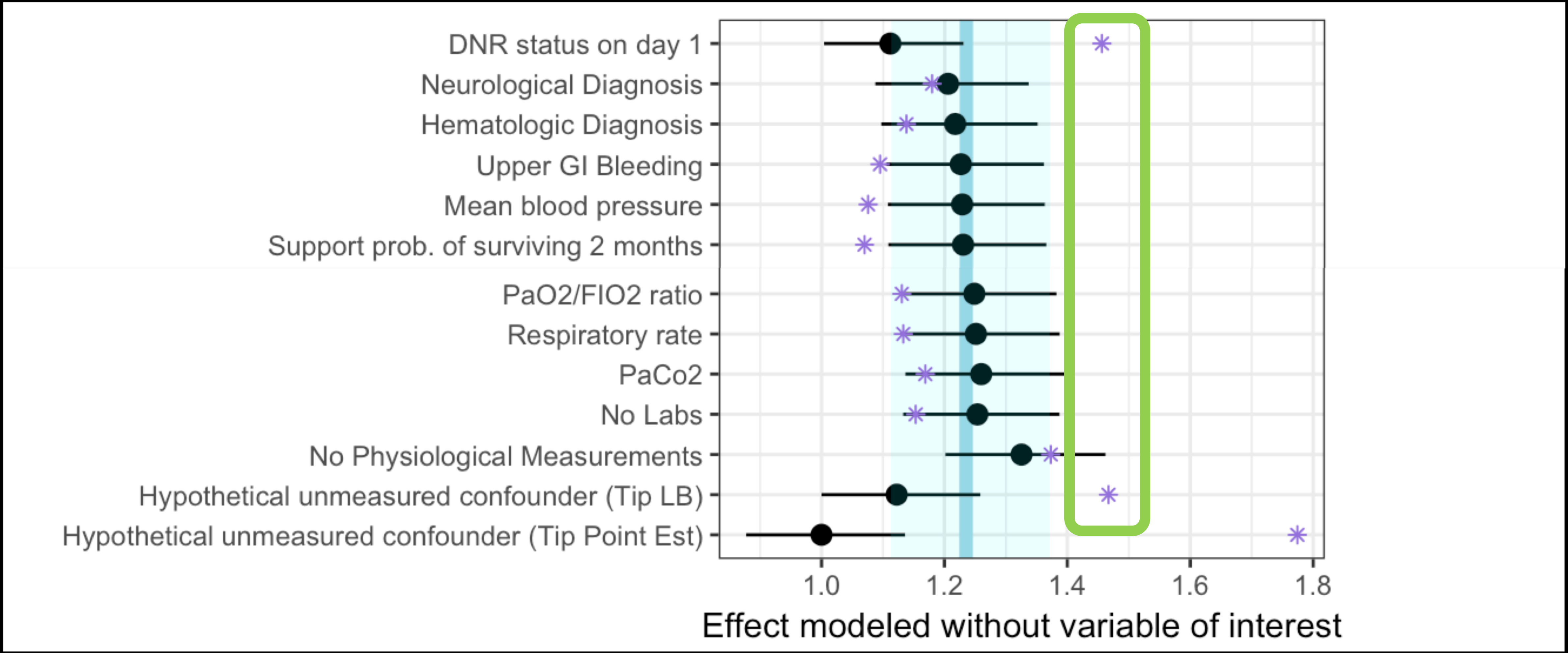
Right Heart Catheterization data

- Fit a propensity score model
- Use ATO weights
- Fit weighted cox model for 30 day survival


DNR status on day 1
Neurological Diagnosis
Hematologic Diagnosis
Upper GI Bleeding
Mean blood pressure
Support prob. of surviving 2 months
Hematocrit
Age
WBC
Chronic Renal Disease
Albumin
Bilirubin
Sex
1.0
1.2
1.4
1.6
1.8
Effect modeled without variable of interest
e-value



Sex
PH
APACHE score
Transfer Status
Heart rate
PaO2/FIO2 ratio
Respiratory rate
PaCo2
No Labs
No Physiological Measurements
Hypothetical unmeasured confounder (Tip LB)
Hypothetical unmeasured confounder (Tip Point Est)
1.0
1.2
1.4
1.6
1.8
Effect modeled without variable of interest
e-value



DNR status on day 1
Neurological Diagnosis
Hematologic Diagnosis
Upper GI Bleeding
Mean blood pressure
Support prob. of surviving 2 months
PaO2/FIO2 ratio
Respiratory rate
PaCo2
No Labs
No Physiological Measurements
Hypothetical unmeasured confounder (Tip LB)
Hypothetical unmeasured confounder (Tip Point Est)
1.0
1.2
1.4
1.6
1.8
Effect modeled without variable of interest

Thank you!

@LucyStats

http://bit.ly/LucyStatsENAR2018

lucymcgowan.com