Detecting critical treatment effect bias in small subgroups using randomized trials

October 28th 2024, ML Seminar, CMU

Fanny Yang

Statistical Machine Learning group, CS department, ETH Zurich



The hormone replacement therapy controversy

Wanted: effect of treatment on health for a patient → Decide: to give drug to patient or not

Treatment: hormone replacement therapy (HRT), Outcome Y: Coronary heart disease

Review > Ann Intern Med. 1992 Dec 15;117(12):1016-37. doi: 10.7326/0003-4819-117-12-1016.

Hormone therapy to prevent disease and prolong life in postmenopausal women

D Grady [1], S M Rubin, D B Petitti, C S Fox, D Black, B Ettinger, V L Ernster, S R Cummings

Based on observational studies – hospital patient data (no experimental interventions)

The hormone replacement therapy controversy

Wanted: effect of treatment on health for a patient → Decide: to give drug to patient or not

Treatment: hormone replacement therapy (HRT), Outcome Y: Coronary heart disease

Menopause drug scare hits women

True degree of therapy risk lost in the clamour of commentat

Expert panel backs HRT cancer warning

600,000 women warned to stop combined HRT medication

Hormone alert for canc

HRT linked to cancer and stroke: doctors demand drug restrictions

More needed to settle HRT scare

Randomized trials – experimental data collected in controlled environment

The hormone replacement therapy controversy

Wanted: effect of treatment on health for a patient → Decide: to give drug to patient or not

Treatment: hormone replacement therapy (HRT), Outcome Y: Coronary heart disease

Observational studies in 1999

suggest that HRT prevent heart disease

WHI randomized trial published in 2002

shows HRT increases risk of heart disease

So what's the effect of HRT on Y?

Plan for today

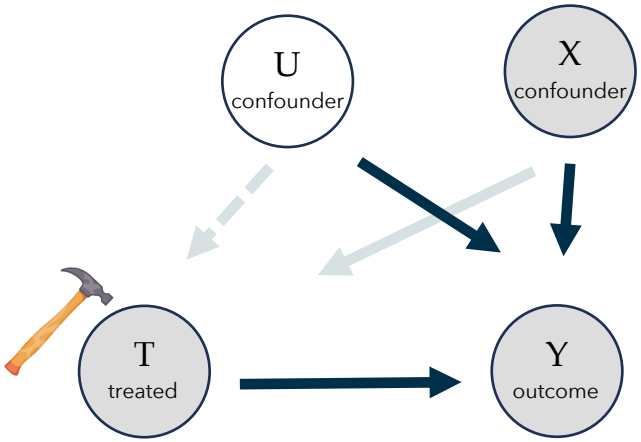
- Recap: Treatment effect estimation using RCT vs. observational studies
- Our credo: use observational studies when bias not too big in any subgroup
- Our two-stage "flagging" approach

tolerance granularity

- Empirically: Effects on tolerance and granularity on the "flagging" outcome

Recap: Causal effect

Wanted: effect of treatment on health for a patient → Decide: to give drug to patient or not



- Outcome Y : e.g. contract heart disease or not
- $T = 0$: do nothing; $T = 1$: treat with method
- X: measured patient features

Causal effect: Expected outcome Y

when $T = 1$ vs. when $T = 0$

$\mathbb{E}[Y|do(T = 1), X = x]$
$= \mathbb{E}[Y(1)|X = x]$

$\mathbb{E}[Y|do(T = 0), X = x]$
$= \mathbb{E}[Y(0)|X = x]$

→ we'd like to estimate $\tau^{rct}(x) = \mathbb{E}_{\mathbb{P}}[Y(1) - Y(0)|X = x]$ through data

Recap: Randomized trials and observational data

For x: Observed features of patient

Randomized trial

assign treatment

randomized

extract data

True effect $\tau^{rct}(x)$

Observational study

assign treatment

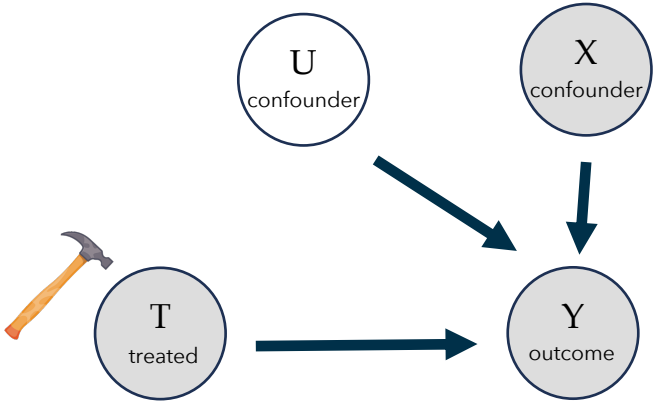
extract data

Estimable effect $\tau^{os}(x)$

Recap: Randomized trials and observational data

What's the effect of drug (for patient X) on disease risk? → whether to give drug to patient X

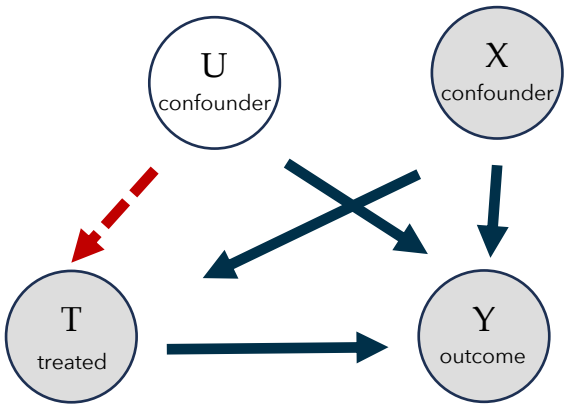
Randomized trial (intervention)



unbiased!

$$\mathbb{E}_{\mathbb{P}^{rct}}[Y|T=t,X=x] = \mathbb{E}[Y|do(T=t),X=x]$$

Observational data (no intervention)

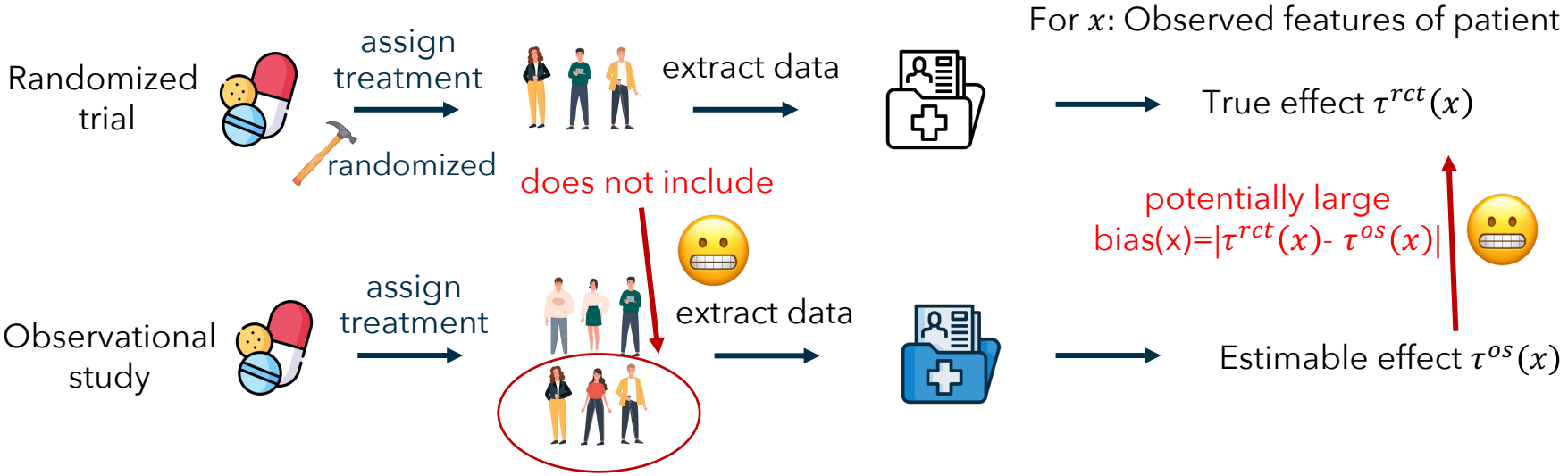


maybe biased!

$$\mathbb{E}_{\mathbb{P}^{os}}[Y|T=t,X=x] \neq \mathbb{E}[Y|do(T=t),X=x]$$

$$\Rightarrow \tau^{os}(x) = \mathbb{E}_{\mathbb{P}^{os}}[Y|T=1,X=x] - \mathbb{E}_{\mathbb{P}^{os}}[Y|T=0,X=x] \neq \mathbb{E}_{\mathbb{P}}[Y(1)-Y(0)|X=x] = \tau^{rct}(x)$$

Our goal: lower bounding confounding strength



Our question: Can we leverage best of both worlds when possible?

Next slide: Prior paradigms

Prior paradigms to mitigate bias – and their caveats

Randomized trial

assign treatment

randomized

extract data

Model may be wrong

Use parametric assumptions to debias $\tau^{os}(x)$

e.g. Kallus-Puli-Shalit '18, Yang-Gao-Zeng-Wang '23

Observational study

assign treatment

extract data

Test hypothesis: bias model holds **and** ATE is insignificant

Yadlowsky-Namkoong-Basu-Duchi-Tian-22, Jin-Ren-Candes-23

Often in epidemiology: Select features that capture all confounders of heuristic causal graph

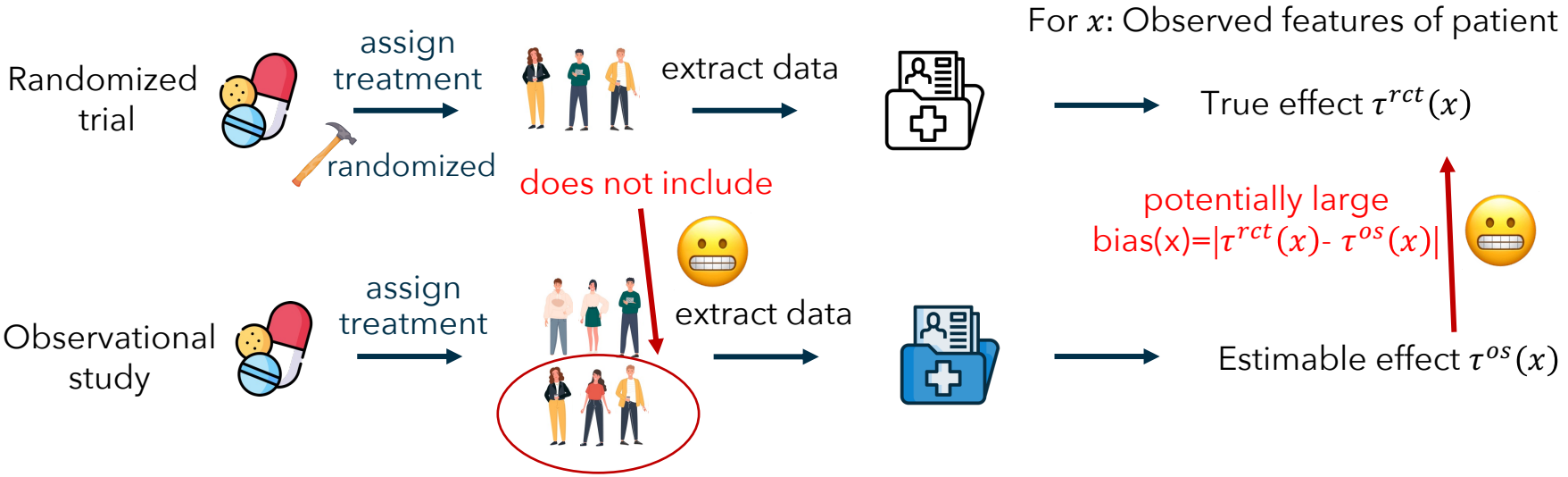
Model may be wrong

Compute critical value that would switch the sign of effect

e.g. vanderWeele-Ding-17, Jin-Ren-Candes-23

→ statement about true bias not possible!

Our paradigm: Testing whether bias is too high

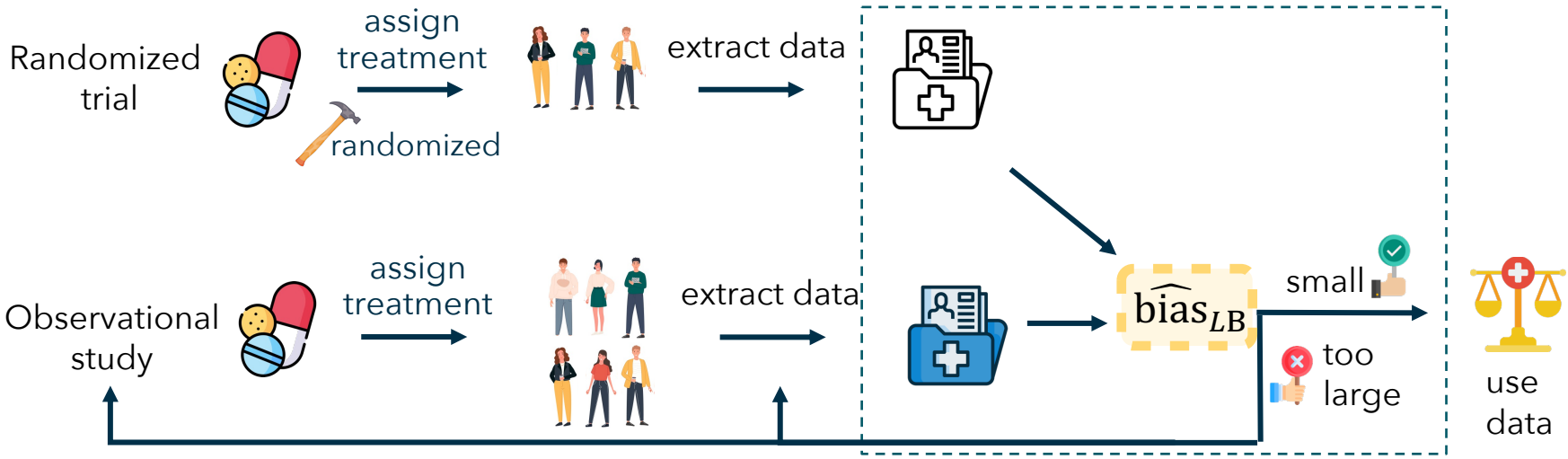


Our Goal: trust $\tau^{os}(x)$ if bias(x)=$|\tau^{rct}(x)- \tau^{os}(x)|$ is small using statistical tests

Practitioner problem with previous tests*: designed to reject when bias(x)≠0

*[Hussain et al. '22, '23, Demirel '24]

Our paradigm: Testing whether bias is too high



Goal: Compute a $\widehat{\text{bias}}_{LB}$ so that we're sure the true $bias(x) \geq \widehat{\text{bias}}_{LB}$ for some x & discard data only if $\widehat{\text{bias}}_{LB}$ is too large (compared with some critical value)

Plan for today

- Recap: Treatment effect estimation using RCT vs. observational studies
- Our credo: use observational studies when bias not too big in any subgroup
- Two-stage "flagging" approach
- Empirically: Effects on tolerance and granularity on the "flagging" outcome

Approach to finding a lower bound through testing

Our plug-and-play approach for desired significance α:

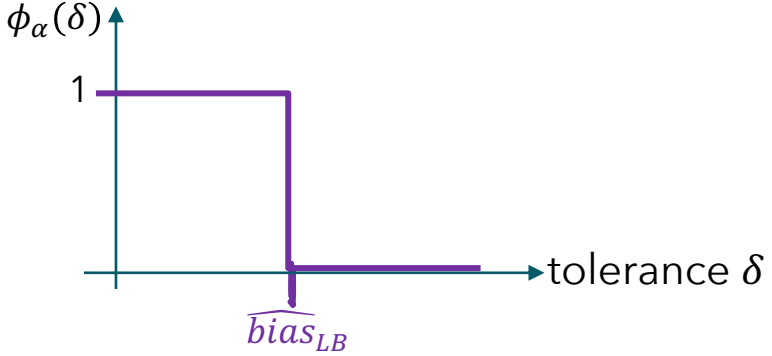
1. Test $\phi_\alpha(\delta)$ of the null $H_0(\delta)$: $bias(\mathrm{x}) \leq \delta$ for all x
2. Report $\widehat{bias}_{LB} = \inf\{\delta: \phi_\alpha(\delta)\ accepted\}$ and flag if $\widehat{bias}_{LB} > \delta_{\mathbf{thresh}}$

$\widehat{bias}_{LB}$: tests for bias smaller $\widehat{bias}_{LB}$ are rejected

has granularity

test tolerance

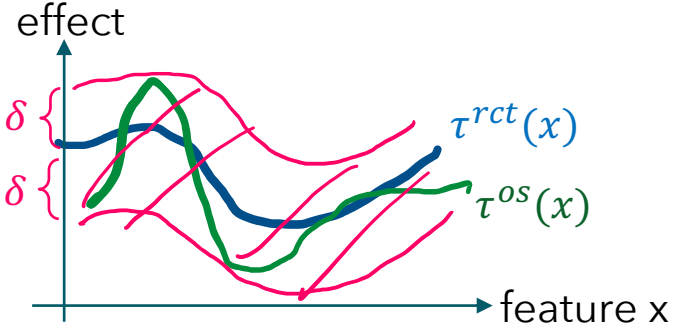
bias tolerance to neglect



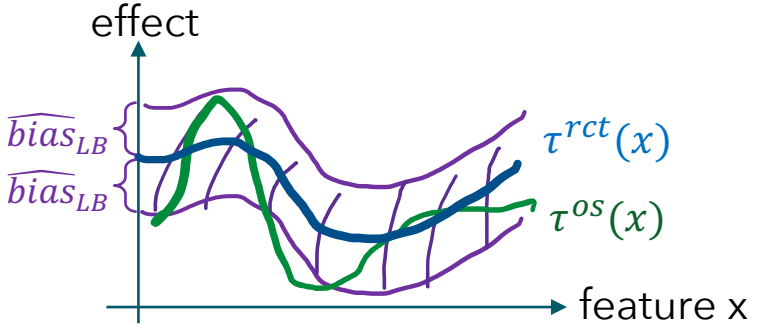
What we want $\phi_\alpha(\delta)$ and $\widehat{bias}_{LB}$ to satisfy

Our plug-and-play approach for desired significance α:

1. Test $\phi_\alpha(\delta)$ of the null $H_0(\delta)$: $bias(\mathrm{x}) \leq \delta$ for all x
2. Report $\widehat{bias}_{LB} = \inf\{\delta: \phi_\alpha(\delta)\ accepted\}$ and flag if $\widehat{bias}_{LB} > \delta_{\mathbf{thresh}}$



1. For $\phi_\alpha(\delta)$ testing null $H_0(\delta)$:
accepts if $\tau^{os}(x)$ is in δ-band around $\tau^{rct}(x)$



2. For lower bound $\widehat{bias}_{LB}$:
guaranteed $\mathbf{bias}(\mathrm{x}) \geq \widehat{bias}_{LB}$ for some x

[de Bartolomeis, Abad, Donhauser, Y. '24a, '24b]

$\text{bias}(x) := |\tau^{\text{os}}(x) - \tau^{\text{rct}}(x)|$

Guarantees for our lower bound

Our plug-and-play approach for desired significance α:

1. Test $\phi_\alpha(\delta)$ of the null $H_0(\delta)$: $bias(\text{x}) \leq \delta$ for all x
2. Report $\widehat{bias}_{LB} = \inf\{\delta: \phi_\alpha(\delta)\ accepted\}$ and flag if $\widehat{bias}_{LB} > \delta_{\text{thresh}}$

We propose test statistics for $\phi_\alpha(\delta)$, that is efficiently computable with data

Theorem (de Bartolomeis, Abad, Donhauser, Y. '24a, '24b)

If above assumptions hold, our test statistics yield tests $\phi_\alpha(\delta)$ that is asymptotically valid, and further $\text{P}\left(\max_{\text{x}} \text{bias}(x) \geq \widehat{\text{bias}}_{LB}\right) \geq 1 - \alpha + o(1)$ as sample size → infinity

$\text{bias}(x) := |\tau^{\text{os}}(x) - \tau^{\text{rct}}(x)|$

Guarantees for our lower bound

Assumptions besides internal validity of randomized trial

- Transportability: $\mathbb{E}_{\mathbb{P}^{rct}}[Y(1) - Y(0)|X = x] = \mathbb{E}_{\mathbb{P}^{os}}[Y(1) - Y(0)|X = x]$ for all $x \in \mathcal{X}$
- Support inclusion: $\mathbf{supp}(\mathbb{P}_X^{\text{rct}}) \subseteq \mathbf{supp}(\mathbb{P}_X^{\text{os}})$
- CATE can be estimated at rate $o(1/\sqrt{n_{os}})$ and $\lim_{n\to\infty} n_{rct}/n_{os} = 0$

We propose test statistics for $\phi_\alpha(\delta)$,
that is efficiently computable with data

Theorem (de Bartolomeis, Abad, Donhauser, Y. '24a, '24b)

If above assumptions hold, our test statistics yield tests $\phi_\alpha(\delta)$ that is asymptotically valid, and further $\mathrm{P}\left(\max_{\mathrm{x}} \text{bias}(x) \geq \widehat{\text{bias}}_{LB}\right) \geq 1 - \alpha + o(1)$ as sample size → infinity

Constructing a valid test $\phi_\alpha(\delta)$

Original null H_0: $\text{bias}(x) \leq \delta$ for all $x \in \mathcal{X} \subset \mathbb{R}^d$ ← how to test?

Step I: Find a null that (i) can be tested and (ii) is true if original H_0 is true

$\Leftrightarrow H_0$: There exists $g^\star: \mathcal{X} \to [0,1]$ $s.t.$ $\underbrace{\text{bias}(X) - \delta \cdot g^\star(X)}_{=: \psi_{g^\star}(x)} = 0$ $\mathbb{P}^{rct}$-almost surely

e.g. for RKHS $\mathcal{F}$

$\Rightarrow$ H_0 : There exists $g^\star: \mathcal{X} \to [0,1]$ $s.t.$ $\mathbb{E}_{\mathbb{P}^{\text{rct}}}\left[\psi_{g^\star}(X) f(X)\right] = 0$

1

for kernel k of $\mathcal{F}$

if corresponding $g^\star \in G$

$\Rightarrow$ H_0^G: There exists $g^\star: \mathcal{X} \to [0,1] \in G$ s. t. $\mathbb{E}_{\mathbb{P}^{\text{rct}}}\left[\psi_{g^\star}(X) k(X, X') \psi_{g^\star}(X')\right] = 0$

*each step is reduction of power but *not* of validity:*

prob H_0^G rejected < prob H_0 rejected

Constructing a valid test $\phi_\alpha(\delta)$

Step II: Find valid test for $H_0^G: \exists g^\star: X \to [0,1] \in G \; s.t. \; \mathbb{E}_{\mathbb{P}^{\text{rct}}}\left[\psi_{g^\star}(X) k(X, X') \psi_{g^\star}(X')\right] = 0$

- Use cross U-Statistic*

$$\hat{T}^2(g;\delta) \coloneqq \frac{1}{D_1^{rct}} \frac{1}{D_2^{rct}} \sum_{x \in D_1^{rct}} \sum_{x' \in D_2^{rct}} \psi_{g;\delta}(x) k(x, x') \psi_{g;\delta}(x')$$

- $\Rightarrow$ assuming bounded effects (i.e. $\left|\left|\psi_{g^\star}\right|\right|_\infty < \infty$) using result in*

$$\hat{T}_G^2(\delta) \coloneqq \min_{g \in G} \left| \frac{\sqrt{n}\, \hat{T}^2(g;\delta)}{\hat{\sigma}\left(\hat{T}^2(g;\delta)\right)} \right| \leq \left| \frac{\sqrt{n}\, \hat{T}^2(g^\star;\delta)}{\hat{\sigma}\left(\hat{T}^2(g^\star;\delta)\right)} \right| \to |N(0,1)|$$

- $\Rightarrow$ the test $\phi_\alpha(\delta) = \mathbb{I}(\hat{T}_G^2(\delta) > z_{1-\alpha})$ - where $z_{1-\alpha}$ is α-quantile of half-normal – is valid

[Kim-Ramdas '24]

Plan for today

- Recap: Treatment effect estimation using RCT vs. observational studies
- Our credo: use observational studies when bias not too big in any subgroup
- Two-stage "flagging" approach
- Empirically: Effects on tolerance and granularity on the "flagging" outcome

tolerance granularity

Empirical properties of our procedure

Our plug-and-play approach for desired significance α:

1. Compute for all δ test $\phi_\alpha(\delta) = \mathbb{I}\left(\min_{g \in G} \left|\frac{\sqrt{n}\,\hat{T}^2(g;\delta)}{\hat{\sigma}(\hat{T}^2(g;\delta))}\right| > z_{1-\alpha}\right)$ ← from now on fix α=0.05
2. Flag observational study if $\widehat{bias}_{LB} = \inf\{\delta\colon \phi_\alpha(\delta)\ accepted\} > \delta_{\text{thresh}}$

We next discuss how features of our approach affect experimental results:

- Effect of allowing tolerance on decisions compared to using $\phi_\alpha(\delta = 0)$ } real-world (HRT)
- Choice of function class G on power } semi-synthetic
- Effects of granularity on power

A family of tests of different granularity

Our test gives rise to a family of tests by varying the features that we condition on

- Remember bias(x)=$|\tau^{rct}(x) - \tau^{os}(x)|$

  $= \mathbb{E}_{\mathbb{P}}[Y(1) - Y(0)|X = x] - \mathbb{E}_{\mathbb{P}^{os}}[Y|T = 1, X = x] - \mathbb{E}_{\mathbb{P}^{os}}[Y|T = 0, X = x]$

- Most granular null hypothesis H_0: bias(x)$\leq \delta$ for all features x

  we call corresponding test* $\hat{\phi}^{CATE}(\delta)$

- Coarsest (non-granular) null hypothesis H_0: $\mathbb{E}_X$ bias(X) $\leq \delta$

  we call corresponding test* $\hat{\phi}^{ATE}(\delta)$

As x can pick any *subset* of all features to condition on (the more, the better it can pick up subgroup bias)

coarse test: [de Bartolomeis, Abad, Donhauser, Y. '24a], granular test: [de Bartolomeis, Abad, Donhauser, Y. '24b]

Back to hormone replacement therapy controversy

- Treatment: Hormone replacement therapy (HRT)
- Outcome Y: Coronary heart disease

Observational studies in 1999

suggest that HRT prevent heart disease

WHI randomized trial in 2002

shows HRT increases risk of heart disease

Final resolution (only in 2005)*:

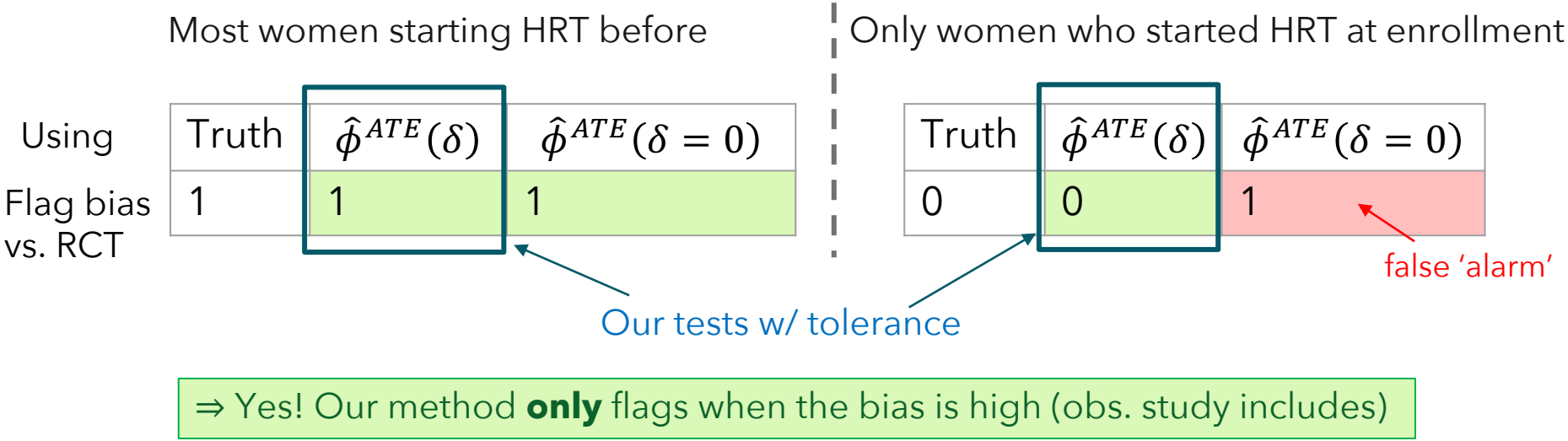
- Problem: most women in WHI obs. studies **took HRT earlier** and survived side effects (but this variable was unmeasured)
- In obs. study: among those starting HRT **when enrolling**, older women did have risk of getting heart disease

Could our method have flagged obs. studies except when controlled for HRT start time?

see [Cagnacci-Vernier '19] for a summary of controversy; *[Prentice et al. '05]

Effect of tolerance in real-world scenarios

- As threshold δ_{thresh} use e.g. estimated critical value $bias_{CT} = |\mathbb{E}_{\mathbb{P}^{OS}}[\tau^{OS}(X)]|$
- Our procedures flags 1 if $\widehat{bias}_{LB} > bias_{CT}$ with $\widehat{bias}_{LB} = \inf\{\delta : \hat{\phi}^{ATE}(\delta) = 0\}$



⇒ Yes! Our method **only** flags when the bias is high (obs. study includes)

[de Bartolomeis, Abad, Donhauser, Y. '24a]

Semi–synthetic experimental setups

Dataset: MineThatData Email

- X: customer data
- T: whether exposed to ads
- Y: dollars spent (synthetic)
- True bias: $\delta^{\star}(x) = \tau^{os}(x) - \tau^{rct}(x)$

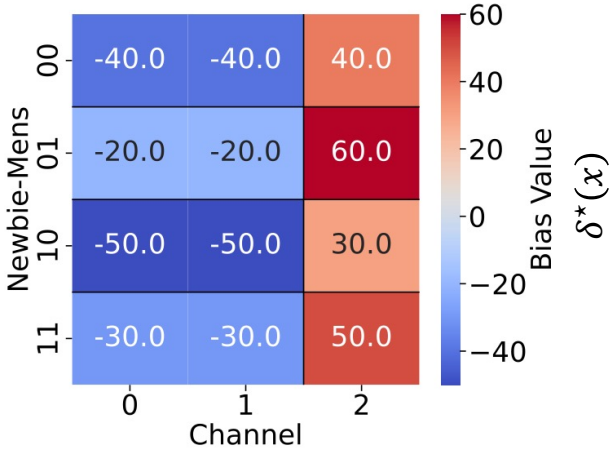
Experiment 1

One group of varying proportion

biased with $\delta^{\star} = 60$

Experiment 2

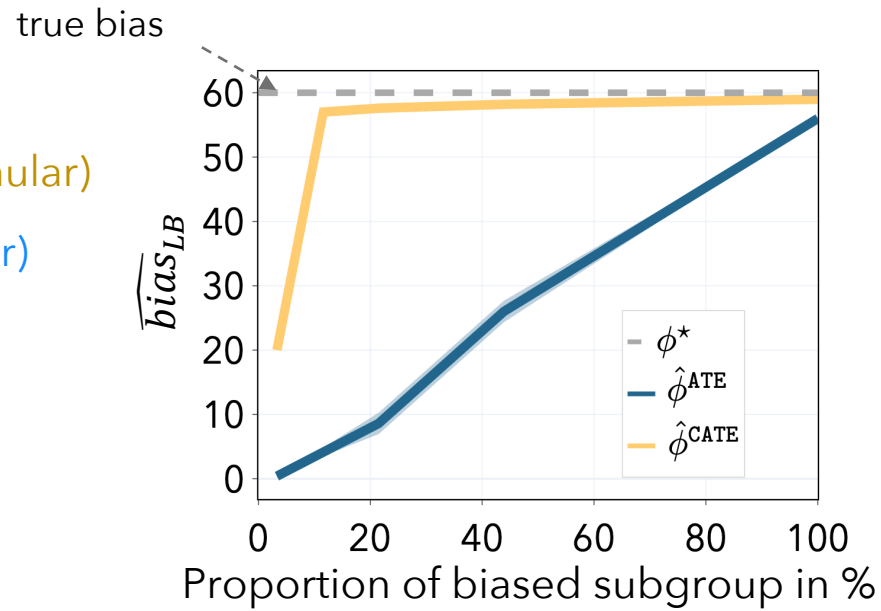
different bias values $\delta^{\star}$ for subgroups according to 3 features (newbie, men, channel)



Effect of granularity – **Experiment 1**

- $\hat{\phi}^{CATE}$ test hypothesis bias(x) $\leq \delta$ (granular)
- $\hat{\phi}^{ATE}$ tests $\mathbb{E}_x$ bias(x) $\leq \delta$ (non-granular)
- $\widehat{bias}_{LB} = \inf \{\delta : \hat{\phi}(\delta)\ accepted\}$

- Larger power corresponds to $\widehat{bias}_{LB}$ being closer to "true bias"

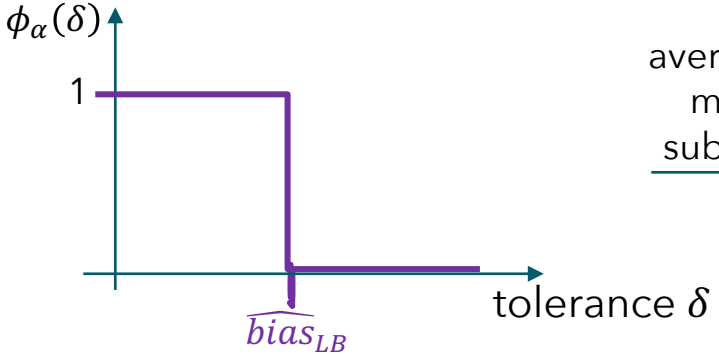


Effect of function class G on power – **Experiment 2**

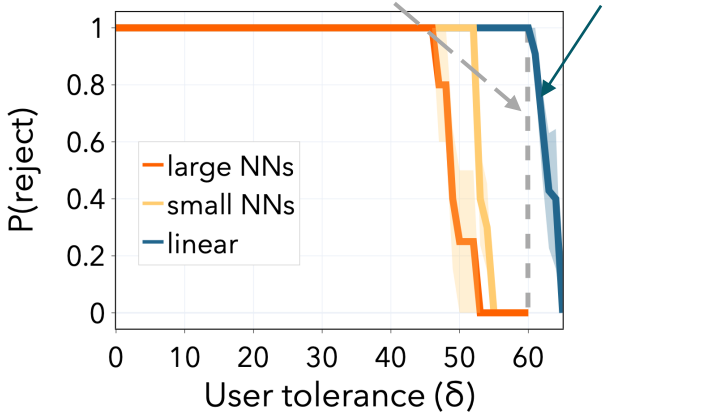
Our plug-and-play approach for desired significance α:

1. Compute for all δ test $\phi_\alpha(\delta) = \mathbb{I}\left(\min_{g \in G} \left|\frac{\sqrt{n}\,\hat{T}^2(g;\delta)}{\hat{\sigma}(\hat{T}^2(g;\delta))}\right| > z_{1-\alpha}\right)$
2. Flag observational study if $\widehat{bias}_{LB} = \inf\{\delta\colon \phi_\alpha(\delta)\ accepted\} > \delta_{\text{thresh}}$

Larger power corresponds to $\widehat{bias}_{LB}$ being closer to "true bias"



average over multiple subsamples



Take-aways for our approach

- Sometimes we can trust observational data over randomized trials!
- Solution: use a statistical test to detect bias in observational data
  - but… real-world data is messy: we need tolerance!
  - but… averaging hides the bias on small subgroups: we need granularity!
- **Our paradigm**: test if the (point-wise) bias is larger than a critical value!

sml.inf.ethz.ch

- “Hidden yet quantifiable: A lower bound for confounding strength using randomized trials” by Piersilvio De Bartolomeis*, Javier Abad*, Konstantin Donhauser, FY, AISTATS 2024a
- “Detecting critical treatment effect bias in small subgroups” by

  Piersilvio De Bartolomeis, Javier Abad, Konstantin Donhauser, FY, UAI, 2024b