

**DINFK**

# Detecting critical treatment effect bias in small subgroups using randomized trials

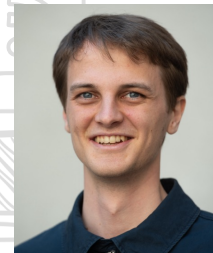
October 28<sup>th</sup> 2024, ML Seminar, CMU

Fanny Yang



Statistical Machine Learning group, CS department, ETH Zurich

**ETH** zürich



# 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](#) <sup>1</sup>, [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 commentators**

Experts claim that women have been alarmed, HRT users more than a small possibility of breast cancer, says John Kralovich.

Some women, however, have been alarmed by the possibility of breast cancer, heart disease, stroke and blood clots. The risk of breast cancer, heart disease, stroke and blood clots is small, but it is not zero. The risk of breast cancer is about 1 in 100 for women aged 50 to 60. The risk of heart disease is about 1 in 100 for women aged 50 to 60. The risk of stroke is about 1 in 100 for women aged 50 to 60. The risk of blood clots is about 1 in 100 for women aged 50 to 60.

### 600,000 women warned to stop combined HRT medication

## Hormone alert for cancer

600,000 women have been warned to stop taking combined hormone replacement therapy (HRT) medication because of a risk of breast cancer, heart disease, stroke and blood clots.

### More needed to settle HRT scare

The hormone replacement therapy scare inspired last month by US researchers is having predictable results. Australia's biggest supplier of the oestrogen-progestin combination has reported a 30 per cent decline in sales since American doctors cut short a long-term study of 16,000 HRT users to warn the world that the therapy increased the risk of breast cancer, heart disease, stroke and blood clots, particularly among women who took the therapy for five years or more.

## Expert panel backs HRT cancer warning

**Label guidelines**

- Limit HRT therapy to the lowest dose and shortest duration possible
- Review reason for the treatment of menopause
- Reassess on appropriate short-term basis

**'There can be risks with stopping medication suddenly without supervision'**

An expert panel has backed a warning that women taking hormone replacement therapy (HRT) should not stop the medication abruptly. The panel, which included doctors and scientists, said that stopping HRT suddenly could lead to a flare-up of symptoms, such as hot flashes and mood swings. The panel also recommended that women should consult their doctor before stopping HRT.

### HRT linked to cancer and stroke: doctors demand drug restrictions

Deborah Smith, Science writer

The NSW Cancer Council has called for a common term for hormone replacement therapy to be restricted to short-term use after a new study linked it to breast cancer.

**HORMONE THERAPY**

**THE RISKS**  
41% increase in stroke; 29% increase in heart attacks; doubling of venous blood clots; 26% increase in breast cancer.

**THE BENEFITS**  
37% cut in osteoporosis cases; one-third reduction in hip fractures; 26% reduction in all fractures.

For the defenders of HRT, the American report provoked understandable panic among its users. This might have been avoided, or at least lessened, had the researchers not highlighted their findings with a simplistic, misleading and, arguably, misleading set of statistics. The ensuing furore left little room, for instance, to counter arguments such as women being twice as likely to develop breast cancer if they took two alcoholic drinks a day, instead of HRT. The American report said an HRT user's breast cancer risk, for example, jumped 26 per cent (with similarly alarming rises in the risks of other side effects). To women who know little about statistical interpretation, this might (and probably did) suggest their odds of developing breast cancer would increase by 26 chances in 100. In fact, the odds grew by 0.08 per cent. In Australia, where 600,000 women used HRT pre-scare, this would mean 1200 extra cases a year of life-threatening heart attacks, stroke, breast cancer and colorectal cancers.

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

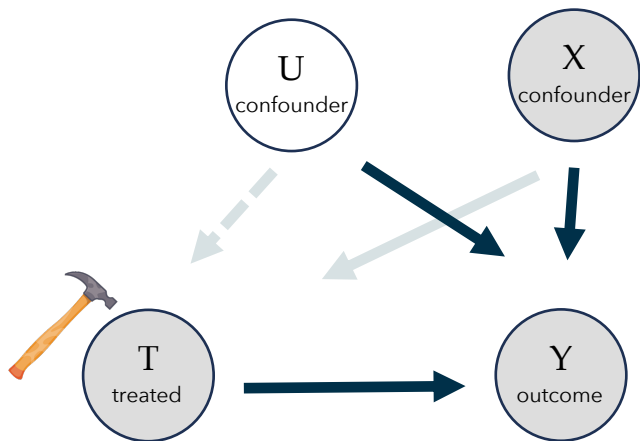
- Recap: Treatment effect estimation using RCT vs. observational studies
- Our credo: use observational studies when bias not too big in any subgroup
- Our two-stage “flagging” approach
- Empirically: Effects on tolerance and granularity on the “flagging” outcome

tolerance

granularity

# 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.

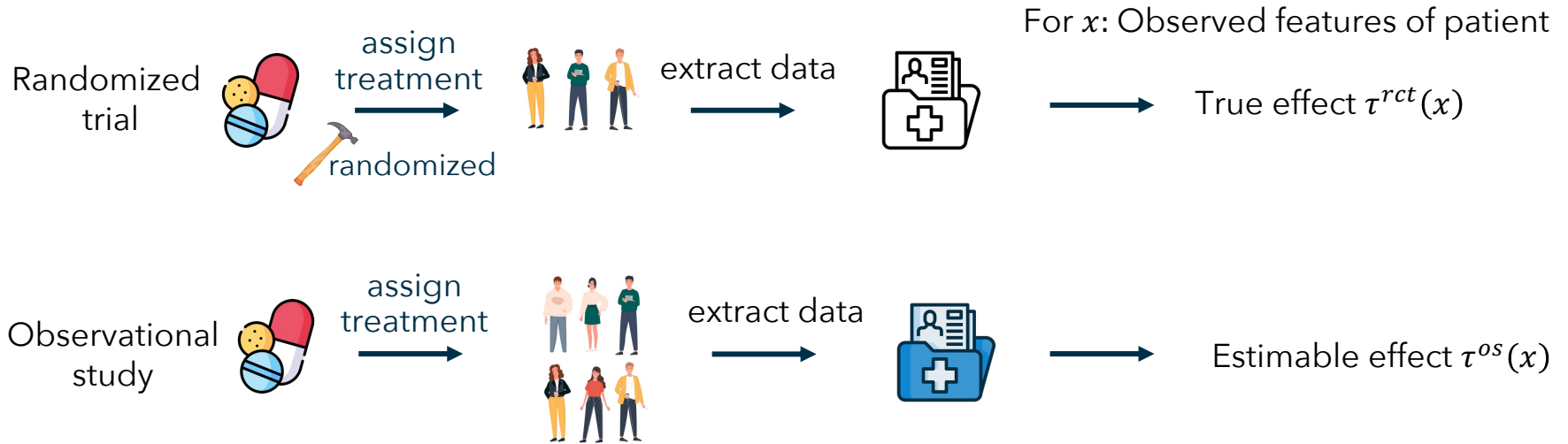
when  $T = 0$

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

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

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

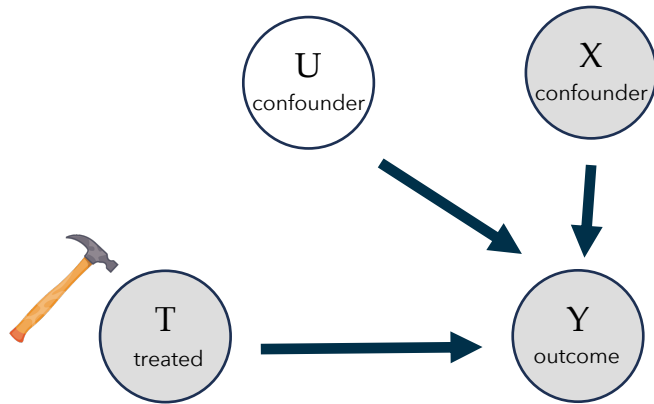
# Recap: Randomized trials and observational data



# 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

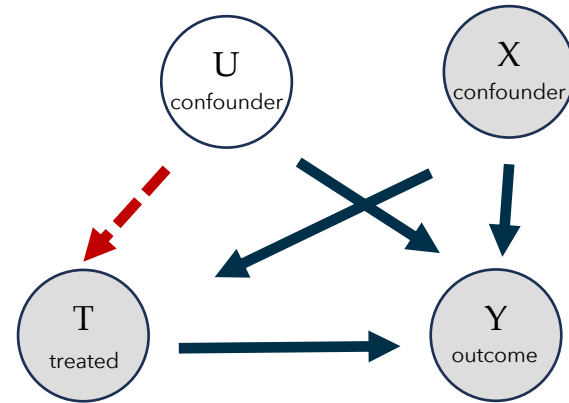
Randomized trial (intervention)



unbiased!

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

Observational data (no intervention)



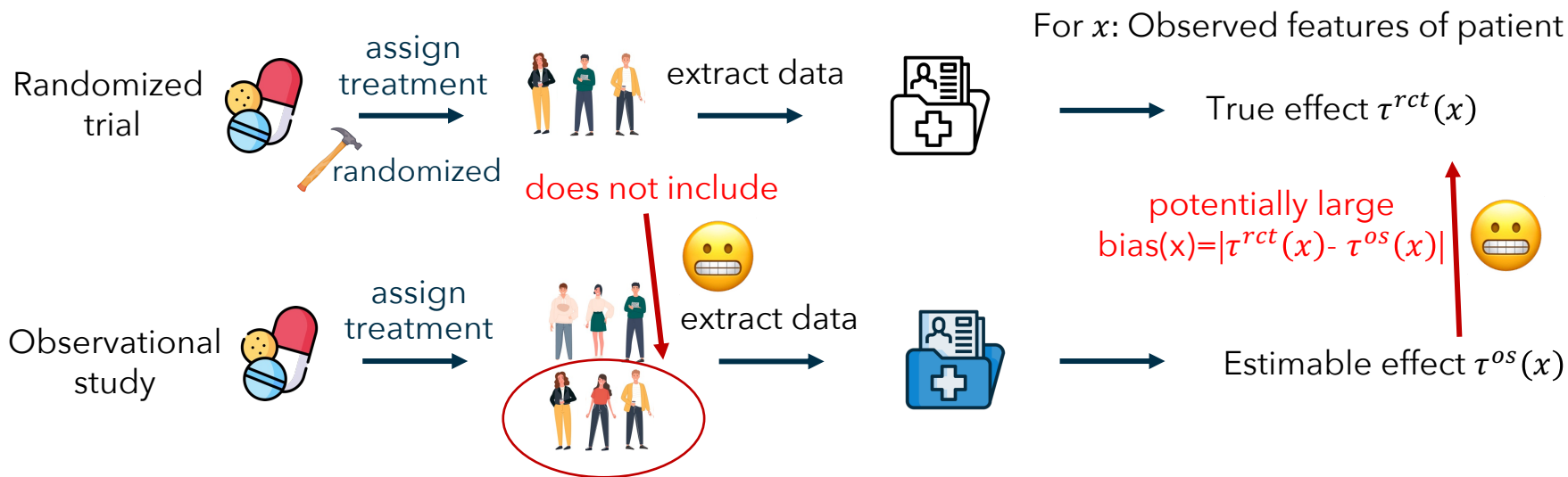
*maybe biased!*

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

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



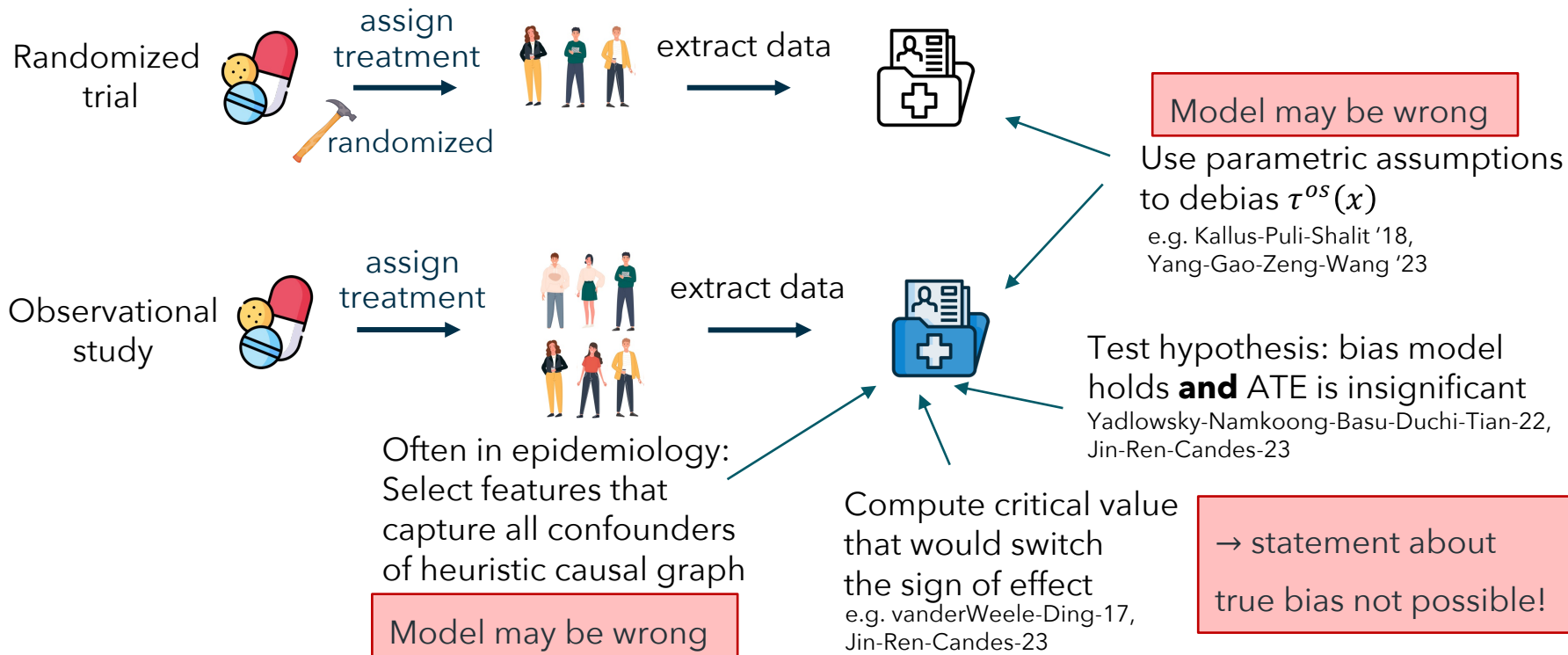
# Our goal: lower bounding confounding strength



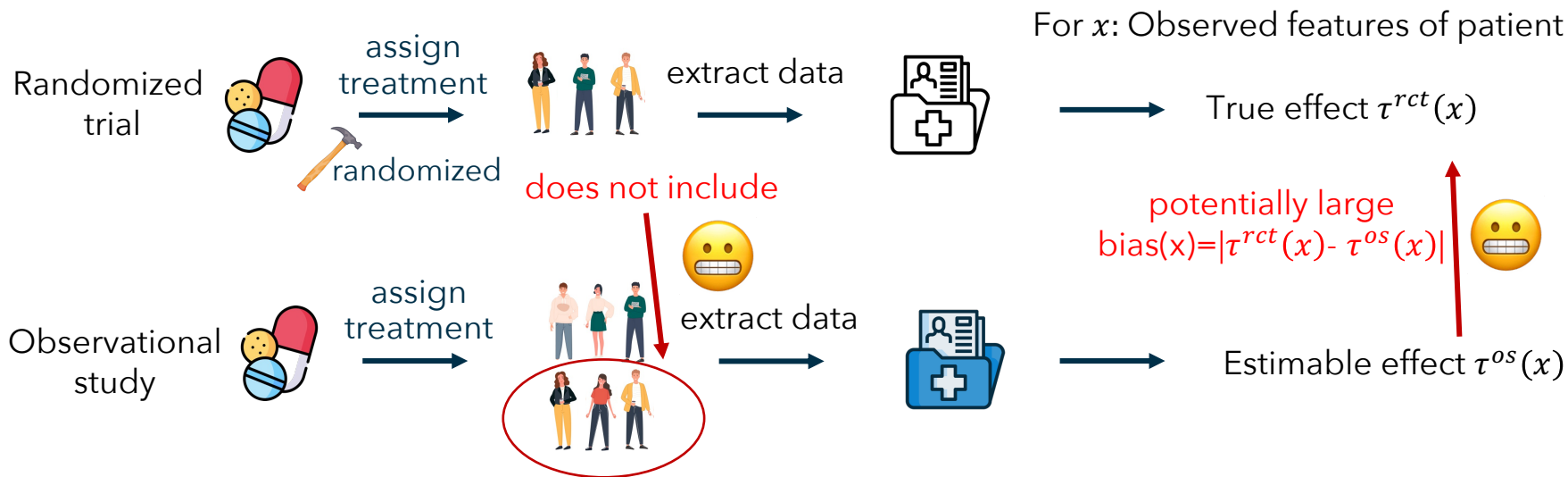
**Our question:** Can we leverage best of both worlds when possible?

Next slide: Prior paradigms

# Prior paradigms to mitigate bias - and their caveats



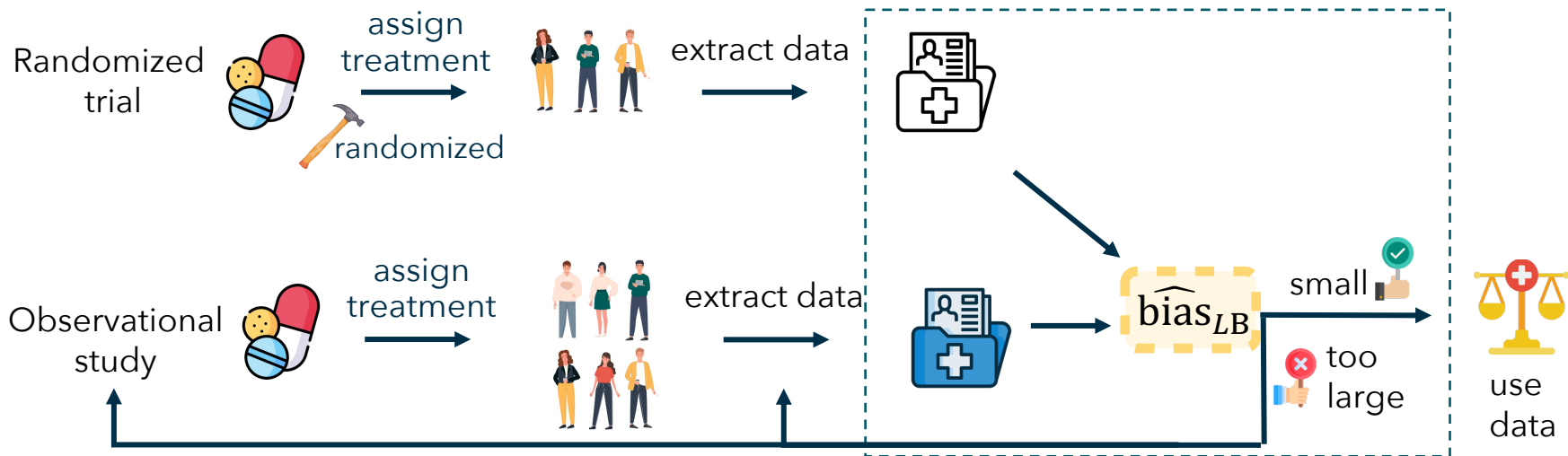
# Our paradigm: Testing whether bias is too high



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

Practitioner problem with previous tests\*: designed to reject when  $\text{bias}(x) \neq 0$

# Our paradigm: Testing whether bias is too high



**Goal:** Compute a  $\widehat{bias}_{LB}$  so that we're sure the true  $bias(x) \geq \widehat{bias}_{LB}$  for some  $x$  & discard data only if  $\widehat{bias}_{LB}$  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

tolerance

granularity

# Approach to finding a lower bound through testing

Our plug-and-play approach for desired significance  $\alpha$ :

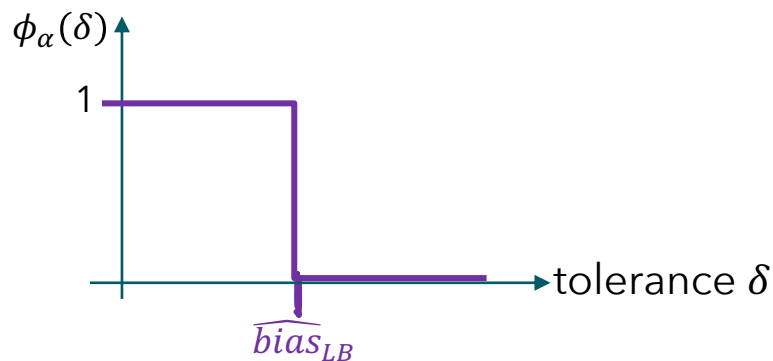
1. Test  $\phi_\alpha(\delta)$  of the null  $H_0(\delta): \text{bias}(x) \leq \delta$  for all  $x$
2. Report  $\widehat{\text{bias}}_{LB} = \inf \{ \delta: \phi_\alpha(\delta) \text{ accepted} \}$  and flag if  $\widehat{\text{bias}}_{LB} > \delta_{\text{thresh}}$

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

has granularity

test tolerance

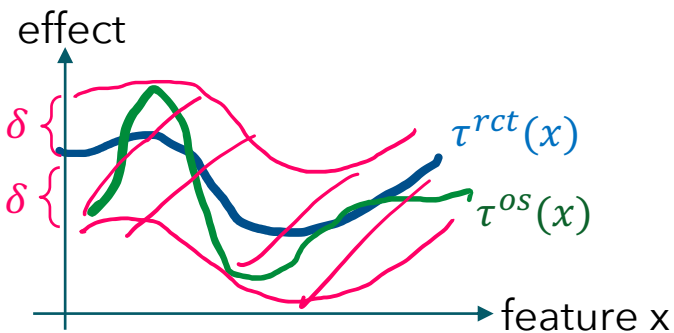
bias tolerance  
to neglect



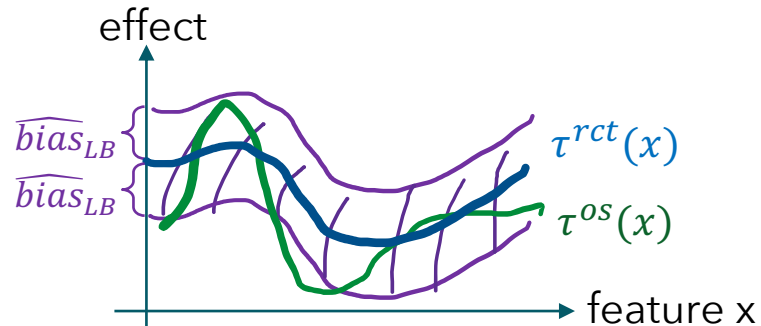
# What we want $\phi_\alpha(\delta)$ and $\widehat{bias}_{LB}$ to satisfy

Our plug-and-play approach for desired significance  $\alpha$ :

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



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



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

# Guarantees for our lower bound

Our plug-and-play approach for desired significance  $\alpha$ :

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



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  $P\left(\max_x \text{bias}(x) \geq \widehat{\text{bias}}_{LB}\right) \geq 1 - \alpha + o(1)$  as sample size  $\rightarrow$  infinity



# Guarantees for our lower bound

## Assumptions besides internal validity of randomized trial

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



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  $\mathbb{P}\left(\max_x \text{bias}(x) \geq \widehat{\text{bias}}_{\text{LB}}\right) \geq 1 - \alpha + o(1)$  as sample size  $\rightarrow$  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  $g^*: \mathcal{X} \rightarrow [0,1]$  s. t.  $\underbrace{\text{bias}(X) - \delta \cdot g^*(X)}_{=: \psi_{g^*}(x)} = 0$   $\mathbb{P}^{rct}$ -almost surely

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

$\Rightarrow H_0$  : There exists  $g^*: \mathcal{X} \rightarrow [0,1]$  s. t.

$$\mathbb{E}_{\mathbb{P}^{rct}}[\psi_{g^*}(X)f(X)] = 0$$

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

prob  $H_0^G$  rejected < prob  $H_0$  rejected

if corresponding  $g^* \in G$

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

for kernel  $k$  of  $\mathcal{F}$

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

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

- Use cross U-Statistic\*

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

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

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

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

# 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  $\alpha$ :

1. Compute for all  $\delta$  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  $\alpha=0.05$
2. Flag observational study if  $\widehat{bias}_{LB} = \inf \{ \delta: \phi_\alpha(\delta) \text{ accepted} \} > \delta_{\text{thresh}}$

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

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

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

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

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

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)

# 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?

# Effect of tolerance in real-world scenarios

- As threshold  $\delta_{thresh}$  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  $\widehat{bias}_{LB} = \inf \{\delta: \hat{\phi}^{ATE}(\delta) = 0\}$

Most women starting HRT before

Using  
Flag bias  
vs. RCT

Truth	$\hat{\phi}^{ATE}(\delta)$	$\hat{\phi}^{ATE}(\delta = 0)$
1	1	1

Our tests w/ tolerance

Only women who started HRT at enrollment

Truth	$\hat{\phi}^{ATE}(\delta)$	$\hat{\phi}^{ATE}(\delta = 0)$
0	0	1

false 'alarm'

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



# Semi-synthetic experimental setups

Dataset: MineThatData Email

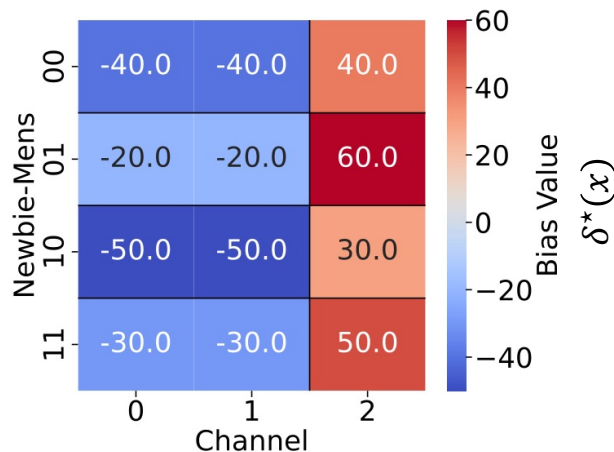
- X: customer data
- T: whether exposed to ads
- Y: dollars spent (synthetic)
- True bias:  $\delta^*(x) = \tau^{os}(x) - \tau^{rct}(x)$

## Experiment 1

One group of varying proportion  
biased with  $\delta^* = 60$

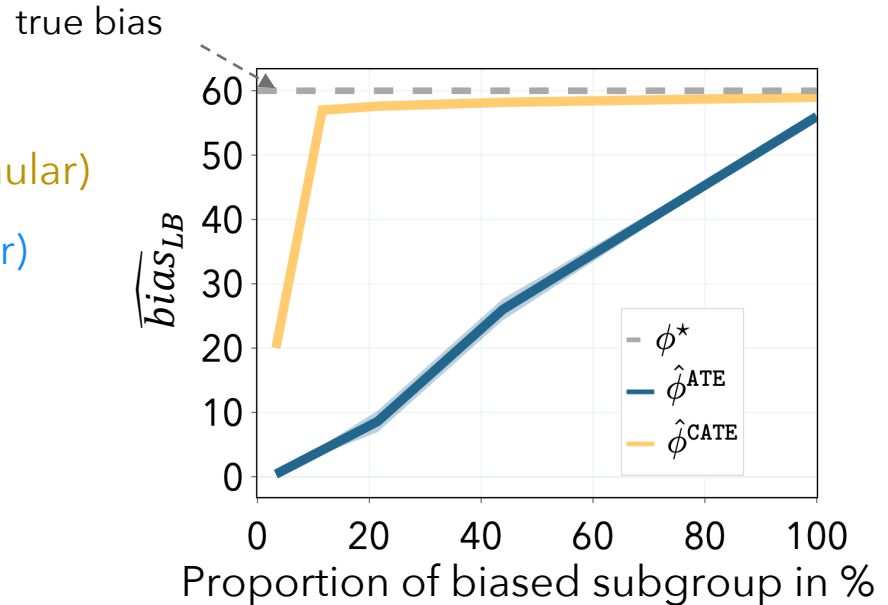
## Experiment 2

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



# Effect of granularity - **Experiment 1**

- $\hat{\phi}^{CATE}$  test hypothesis  $\text{bias}(x) \leq \delta$  (granular)
- $\hat{\phi}^{ATE}$  tests  $\mathbb{E}_x \text{bias}(x) \leq \delta$  (non-granular)
- $\widehat{\text{bias}}_{LB} = \inf \{ \delta : \hat{\phi}(\delta) \text{ accepted} \}$
- Larger power corresponds to  $\widehat{\text{bias}}_{LB}$  being closer to "true bias"

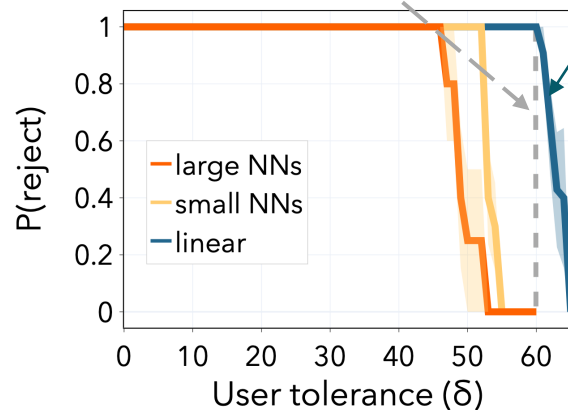
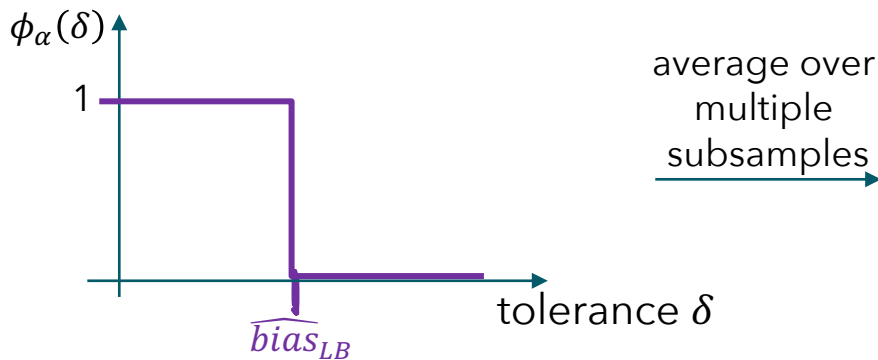


# Effect of function class G on power - **Experiment 2**

Our plug-and-play approach for desired significance  $\alpha$ :

1. Compute for all  $\delta$  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  $\widehat{bias}_{LB} = \inf \{ \delta : \phi_\alpha(\delta) \text{ accepted} \} > \delta_{\text{thresh}}$

Larger power corresponds to  $\widehat{bias}_{LB}$  being closer to "true bias" not valid



# 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!

Thanks!

 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