Combining Experimental and Observational Data for Causal Inference

Eric and Wendy Schmidt Center
Organisms – Clinical Trials Group
November 27, 2023

MIT EECS

ERIC AND WENDY
SCHMIDT CENTER
AT BROAD INSTITUTE

Overview

I will walk through some mechanics of causal inference and mention some of our related work along the lines

- 🎯 Causal inference
  - → objective
  - → assumptions
- 🎯 Data
  - → experimental and observational
  - → strengths and weaknesses
- 🎯 Benchmarking results
  - → are conclusions from two studies compatible?
- 🎯 Combining observational and experimental data
  - → how to leverage their complementary strengths?

Causal inference: objective

- X ›› baseline characteristics / covariates / features
- $A \in \{0, 1\}$ ›› treatment assignment (e.g. placebo vs treatment)
- Y ›› observed outcome
- Y^0, Y^1 ›› potential outcomes under treatment $A = 0$ and $A = 1$

→ only one of them observed for every patient

Causal inference: objective

- X ›› baseline characteristics / covariates / features
- $A \in \{0, 1\}$ ›› treatment assignment (e.g. placebo vs treatment)
- Y ›› observed outcome
- Y^0, Y^1 ›› potential outcomes under treatment $A = 0$ and $A = 1$

→ only one of them observed for every patient

What is the average treatment effect (ATE)?

$$\mathbf{E}\,[Y^1 - Y^0]$$

= ❓

$$\mathbf{E}\,[Y \mid A=1] - \mathbf{E}[Y \mid A=0]$$

Causal inference: objective

- X ›› baseline characteristics / covariates / features
- A ∈ {0, 1} ›› treatment assignment (e.g. placebo vs treatment)
- Y ›› observed outcome
- Y^0, Y^1 ›› potential outcomes under treatment A = 0 and A = 1

→ only one of them observed for every patient

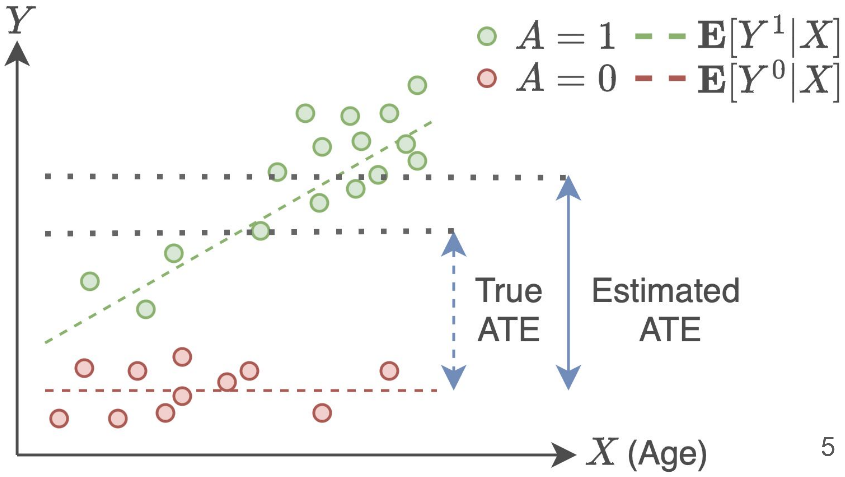
What is the average treatment effect (ATE)?

$\mathbf{E}\,[Y^1 - Y^0]$

= ❓

$\mathbf{E}\,[Y|A=1] - \mathbf{E}[Y|A=0]$

Not if there is ⇒ confounding



Causal inference: assumptions

- Assumptions for causal inference
  - → $\mathbb{E}[Y^a \mid X] = \mathbb{E}[Y^a \mid X, A=a]$ (ignorability/no unmeasured confounding)
    - › we can attribute differences in outcomes to treatment after fixing X
  - → $\mathbb{P}(A=a \mid X) > 0$ (positivity)
    - › for any baseline X, we have some data from both treatments
  - → If A=a, then $Y = Y^a$ (consistency)
    - › when we assign a treatment, we observe its outcome

Causal inference: assumptions

- Assumptions for causal inference
  - → $\mathbf{E}[Y^a \mid X] = \mathbf{E}[Y^a \mid X, A=a]$ (ignorability/no unmeasured confounding)
    - › we can attribute differences in outcomes to treatment after fixing X
  - → $\mathbf{P}(A=a \mid X) > 0$ (positivity)
    - › for any baseline X, we have some data from both treatments
  - → If A=a, then $Y = Y^a$ (consistency)
    - › when we assign a treatment, we observe its outcome

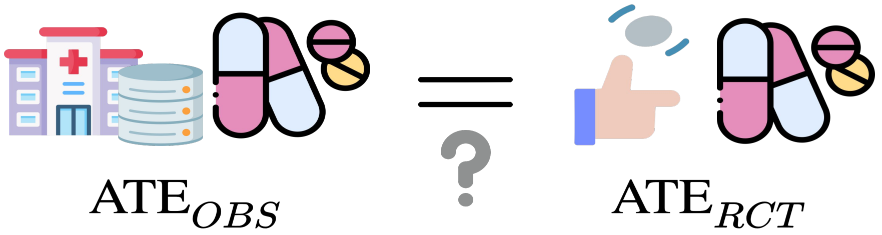
$$\begin{aligned}
\mathbf{E}[Y^a] &= \mathbf{E}_X[\mathbf{E}[Y^a|X]] && \text{Law of total expectation} \\
&= \mathbf{E}_X[\mathbf{E}[Y^a|X, A=a]] && \text{Ignorability + Positivity} \\
&= \mathbf{E}_X[\mathbf{E}[Y|X, A=a]] && \text{Consistency}
\end{aligned}$$

Causal inference: assumptions

- Assumptions for causal inference
  - → $\mathbf{E}[Y^a \mid X] = \mathbf{E}[Y^a \mid X, A=a]$ (ignorability/no unmeasured confounding)
  - → $\mathbf{P}(A=a \mid X) > 0$ (positivity)
  - → If $A=a$, then $Y = Y^a$ (consistency)
- When do these assumptions hold?
  - → Always for experimental data
    - › RCTs satisfy them by design via randomized treatment assignments
    - › strong internal validity
  - → Not always with observational data
    - › treatment assignments are not random
    - › do we control for all the confounders X?
    - › how is positivity affected by including more features in X?

Experimental data (RCT)

- RCTs are internally valid
  → great! we can make causal inference for the trial population
- Are they externally valid?
  → External validity of randomised controlled trials: “to **whom** do the **results** of this **trial apply**?”, PM Rothwell, *The Lancet*, 2005



$$\text{ATE}_{OBS} \overset{?}{=} \text{ATE}_{RCT}$$

→ strict exclusion criteria

› e.g. people with a history of cardiovascular disease

› what if those criteria also affects prognosis?

Causal inference with observational data

We use observational data to make causal inference when evidence from RCT is not available or limited

- $S \in \{0, 1\}$ ›› population index
  → $S = 1$ for OBS, $S = 0$ for RCT
- Last week: How to specify and emulate a target trial from the observational data?
  → an intricate and laborious process
  → define causal question of interest, follow-up, endpoints…

Causal inference with observational data

We use observational data to make causal inference when evidence from RCT is not available or limited

- $S \in \{0, 1\}$ ›› population index
  - → $S = 1$ for OBS, $S = 0$ for RCT
- Last week: How to specify and emulate a target trial from the observational data?
  - → an intricate and laborious process
  - → define causal question of interest, follow-up, endpoints…
- After specifying the target trial, we still have to assume:
  - → $\mathbf{E}[Y^a \mid X, S = 1] = \mathbf{E}[Y^a \mid X, A=a, S=1]$ (no unmeasured confounding)
  - → $\mathbf{P}(A = a \mid X, S = 1) > 0$ (positivity)
  - → If $A = a$ and $S = 1$, then $Y = Y^a$ (consistency)

Causal inference with observational data

We use observational data to make causal inference when evidence from RCT is not available or limited

- After specifying the target trial, we still have to assume:
  - → $\mathbf{E}[Y^a \mid X, S = 1] = \mathbf{E}[Y^a \mid X, A=a, S=1]$ (no unmeasured confounding)
  - → $\mathbf{P}(A = a \mid X, S = 1) > 0$ (positivity)
  - → If $A = a$ and $S = 1$, then $Y = Y^a$ (consistency)

- No unmeasured confounding assumption is unverifiable

Causal inference with observational data

We use observational data to make causal inference when evidence from RCT is not available or limited

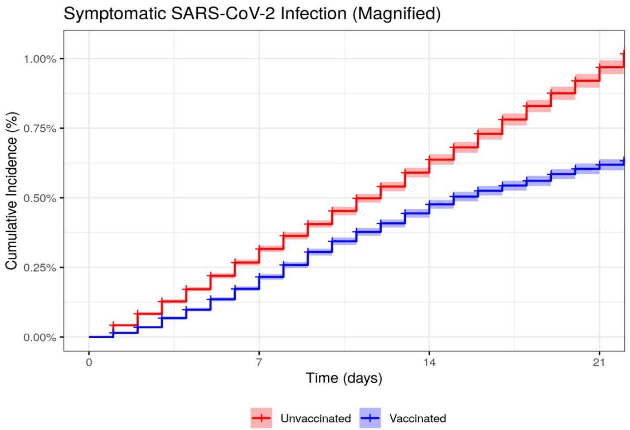
- After specifying the target trial, we still have to assume:
  - → $\mathbf{E}[Y^a \mid X, S = 1] = \mathbf{E}[Y^a \mid X, A=a, S=1]$ (no unmeasured confounding)
  - → $\mathbf{P}(A = a \mid X, S = 1) > 0$ (positivity)
  - → If $A = a$ and $S = 1$, then $Y = Y^a$ (consistency)

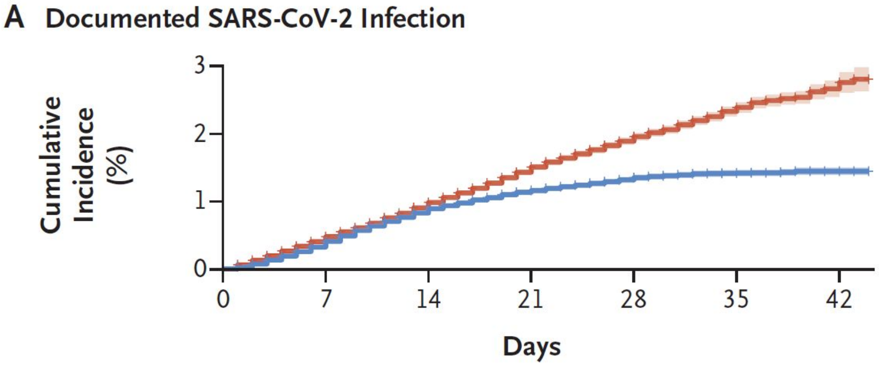
- No unmeasured confounding assumption is unverifiable
- High dim. X makes propensity scores $P(A|X)$ extreme
  - → high variance in estimations
  - → people use standardized weights and clip extreme values (e.g. >0.99)
  - → reduces variance but introduces bias
    - › rely heavily on “extrapolation”

Benchmarking observational data

We use observational data to make causal inference when evidence from RCT is not available or limited

- We would like to get a sense of whether the causal assumptions indeed hold for the observational study (recall: internal validity)

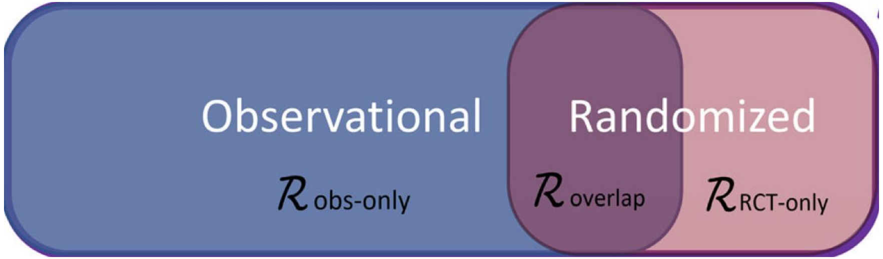




[1] Dagan+ BNT162b2 mRNA Covid-19 vaccine in a nationwide mass vaccination setting, *New England Journal of Medicine* (2021).
[2] @_MiguelHernan, URL: https://twitter.com/_MiguelHernan/status/1364700315044438023, Twitter Thread.

Benchmarking observational data

Benchmark an observational study to an RCT in “overlap region” before using it for “no RCT region”

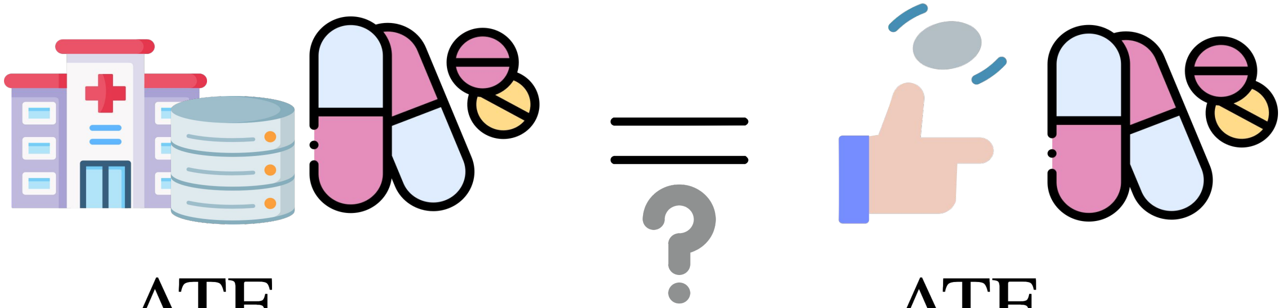

Observational
Randomized
$\mathcal{R}_{\text{obs-only}}$
$\mathcal{R}_{\text{overlap}}$
$\mathcal{R}_{\text{RCT-only}}$

Goal: Use experimental data as a form of validation — *falsify* assumptions of internal and external validity in observational studies

[3] (Figure on top right) Degtiar+, Conditional cross-design synthesis estimators for generalizability in Medicaid, *Biometrics* 2023.

Compare average effects

Goal: Use experimental data as a form of validation — *falsify* assumptions of internal and external validity in observational studies



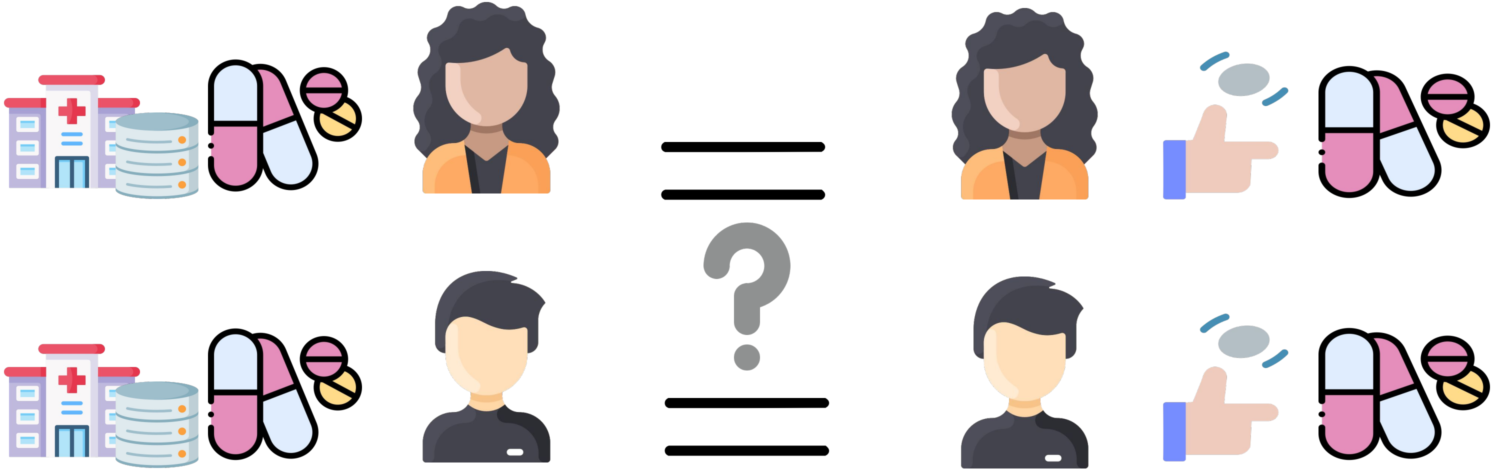
$$\mathrm{ATE}_{OBS} \overset{?}{=} \mathrm{ATE}_{RCT}$$

👎 No explanation for rejection

👎 Could lead to false negatives

Compare group-level average effects

Goal: Use experimental data as a form of validation — *falsify* assumptions of internal and external validity in observational studies



Need a-priori specification of the subgroups; would be nice to find these automatically

[4] Hussain+, Falsification before Extrapolation in Causal Effect Estimation, *NeurIPS* 2022.

Compare conditional average effects

Goal: Use experimental data as a form of validation — *falsify* assumptions of internal and external validity in observational studies

S	X_0	…	X□	A	Y	ψ_0	ψ_1	ψ
0	1		-.5	0	3	-20	0	20
0	1		.2	1	7	37	0	-37
⋮	⋮			⋮	⋮	⋮	⋮	
1	1		.3	1	9	0	46	46
1	1		-.4	0	11	0	-12	-12

CATE signal ψ_1 from OBS (S=1)

CATE signal ψ_0 from RCT (S=0)

[4] Hussain+, Falsification of Internal and External Validity in Observational Studies via Conditional Moment Restrictions, *AISTATS* 2023.

Compare conditional average effects

Goal: Use experimental data as a form of validation — *falsify* assumptions of internal and external validity in observational studies

S	X_0	…	X□	A	Y	ψ_0	ψ_1	ψ
0	1		-.5	0	3	-20	0	20
0	1		.2	1	7	37	0	-37
⋮	⋮			⋮	⋮	⋮	⋮	
1	1		.3	1	9	0	46	46
1	1		-.4	0	11	0	-12	-12

CATE signal ψ_1 from OBS (S=1)

(CATE in RCT)

$$\mathbf{E}\,[Y^1 - Y^0 \mid X, S=1]$$

CATE signal ψ_0 from RCT (S=0)

(CATE in OBS)

$$\mathbf{E}\,[Y^1 - Y^0 \mid X, S=0]$$

[4] Hussain+, Falsification of Internal and External Validity in Observational Studies via Conditional Moment Restrictions, *AISTATS* 2023.

Compare conditional average effects

Goal: Use experimental data as a form of validation — *falsify* assumptions of internal and external validity in observational studies

S	X_0	…	X□	A	Y	ψ_0	ψ_1	ψ
0	1		-.5	0	3	-20	0	20
0	1		.2	1	7	37	0	-37
⋮	⋮			⋮	⋮	⋮	⋮	
1	1		.3	1	9	0	46	46
1	1		-.4	0	11	0	-12	-12

(CATE in RCT)

$$\mathbf{E}\,[\psi_1 \mid X] = \mathbf{E}\,[Y^1 - Y^0 \mid X, S=1]$$

CATE signal ψ_1 from OBS (S=1)

Internal Validity of RCT and OBS

(CATE in OBS)

$$\mathbf{E}\,[\psi_0 \mid X] = \mathbf{E}\,[Y^1 - Y^0 \mid X, S=0]$$

CATE signal ψ_0 from RCT (S=0)

[4] Hussain+, Falsification of Internal and External Validity in Observational Studies via Conditional Moment Restrictions, *AISTATS* 2023.

Compare conditional average effects

Goal: Use experimental data as a form of validation — *falsify* assumptions of internal and external validity in observational studies

S	X_0	…	X□	A	Y	ψ_0	ψ_1	ψ
0	1		-.5	0	3	-20	0	20
0	1		.2	1	7	37	0	-37
⋮	⋮			⋮	⋮	⋮	⋮	
1	1		.3	1	9	0	46	46
1	1		-.4	0	11	0	-12	-12

(CATE in RCT)

$$\mathbf{E}\,[\psi_1 \mid X] = \mathbf{E}\,[Y^1 - Y^0 \mid X, S=1]$$

CATE signal ψ_1 from OBS (S=1)

Internal Validity of RCT and OBS

(CATE in OBS)

$$\mathbf{E}\,[\psi_0 \mid X] = \mathbf{E}\,[Y^1 - Y^0 \mid X, S=0]$$

CATE signal ψ_0 from RCT (S=0)

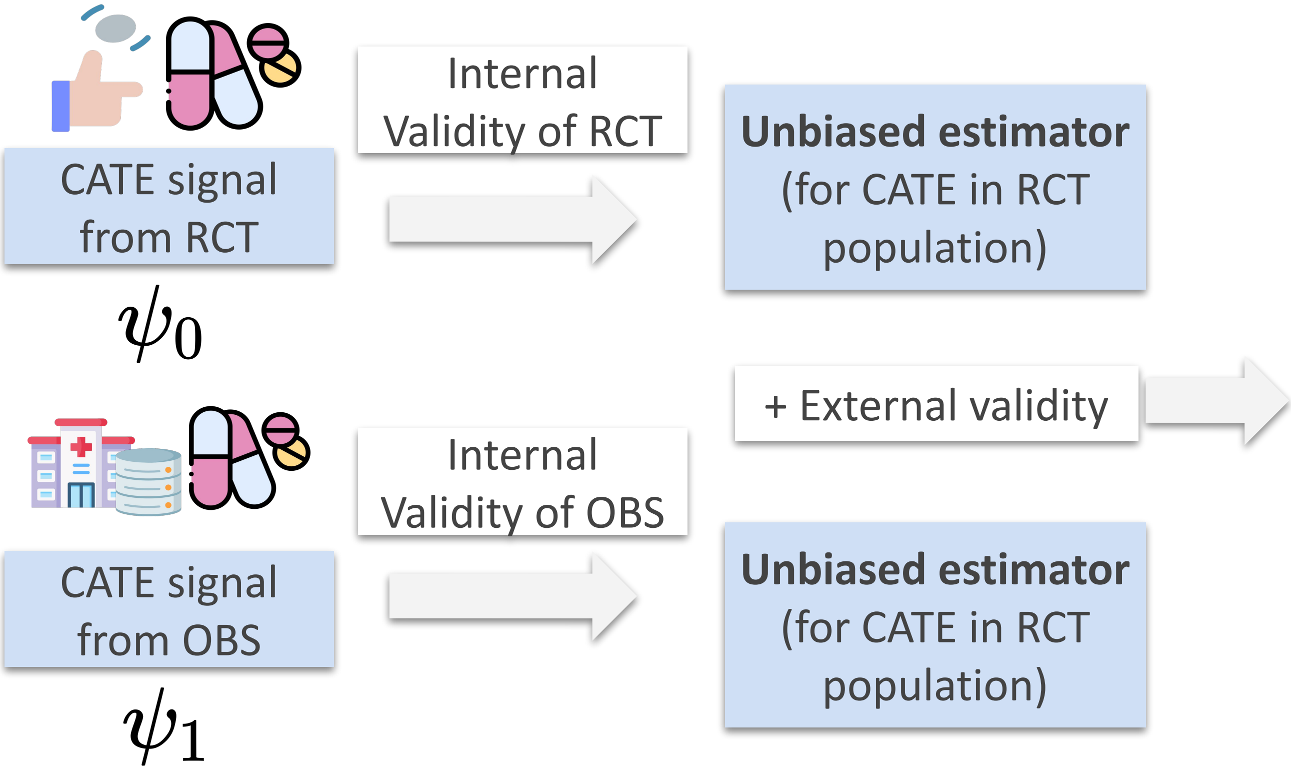
+ External validity

H_0: $\mathbf{E}\,[\psi \mid X] = 0, \;\; \psi := \psi_1 - \psi_0$

Our null hypothesis to test

[4] Hussain+, Falsification of Internal and External Validity in Observational Studies via Conditional Moment Restrictions, *AISTATS* 2023.

Compare conditional average effects



CATE signal from RCT

ψ_0

Internal Validity of RCT

Unbiased estimator (for CATE in RCT population)

CATE signal from OBS

ψ_1

Internal Validity of OBS

Unbiased estimator (for CATE in RCT population)

+ External validity

$$\psi := \psi_1 - \psi_0$$

$$H_0 : \mathbb{E}[\psi|X] = 0$$

$$P_X - \text{almost surely}$$

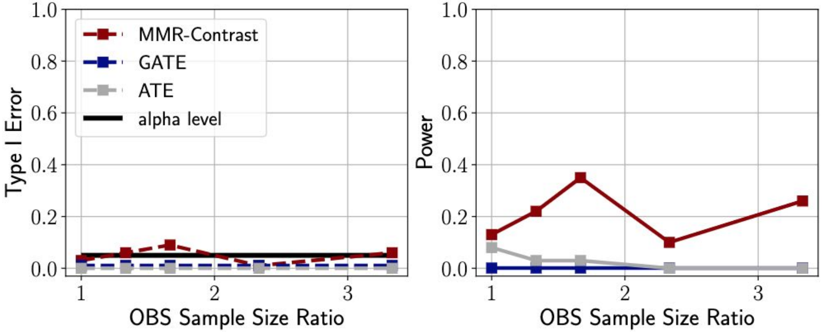
The null hypothesis is a set of conditional moment restrictions (**CMR**)

Muandet et al. shows how to test this hypothesis with a kernel-based method

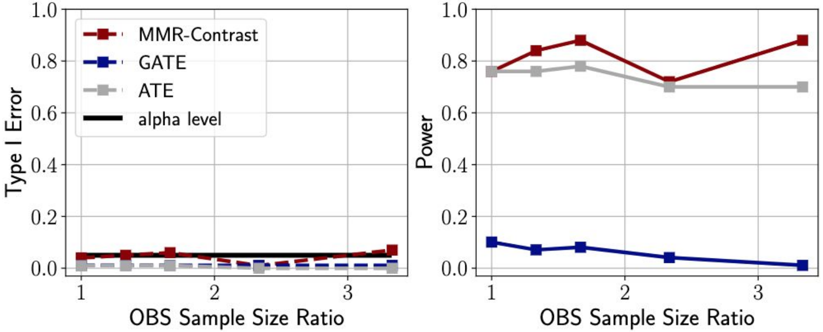
Also returns a “witness function” ***f(X)*** which gets larger with the discrepancy

[5] Muandet+, Kernel conditional moment test via maximum moment restriction, *UAI* 2020.

IHDP and WHI experiments



(a) Low confounder strength ($\max(\gamma) = 1.$). (left) no unobserved confounders; (right): one confounder concealed



(b) High confounder strength ($\max(\gamma) = 2.75$). (left) no unobserved confounders; (right): one confounder concealed

Selection Bias	**MMR-Contrast**	**ATE**	**GATE**
$p = 0$	0.29	0.32	**0.17**
$p = 0.05$	**0.67**	0.58	0.40
$p = 0.10$	**0.94**	0.88	0.67
$p = 0.15$	**1.0**	0.98	0.91

Table 1: Rejection rate when introducing different amounts of selection bias into the observational data in WHI study. p stands for the strength of selection introduced in the the data (refer to Section 5 for details).

[4] Hussain+, Falsification of Internal and External Validity in Observational Studies via Conditional Moment Restrictions, *AISTATS* 2023.

Compare conditional average effects

- Censored case?
  - → we do not observe the outcome in some patients
    - › there is censoring time
    - › loss-to-follow-up, end of study, drop-out…
  - → internal validity not enough for causal inference
    - › need new assumptions on censoring mechanism

Compare conditional average effects

- Censored case?
  - → we do not observe the outcome in some patients
    - › there is censoring time
    - › loss-to-follow-up, end of study, drop-out…
  - → internal validity not enough for causal inference
    - › need new assumptions on censoring mechanism
- Independent censoring
  - → censoring time (C) is independent of time-to-event outcome (Y)
- Global censoring
  - → they are dependent, but the same way in RCT and OBS
    - › e.g., similar drop-out due to adverse side effects

Compare conditional average effects

- Naively handling censored data does not work

  → easily bias the results
- We develop new signals ψ_0 and ψ_1
- Under independent censoring we can recover internal validity

  → $\mathbf{E}[\psi_1|X]=\mathbf{E}[Y^1 - Y^0|X, S=1]$ (can estimate CATE despite censoring)
  - business as usual: test **H_0:** $\mathbf{E}\,[\psi \mid X] = 0$
- Global censoring is more interesting, cannot estimate CATE

  → but still, internal + external validity implies: $\mathbf{E}\,[\psi \mid X] = 0$

Compare conditional average effects

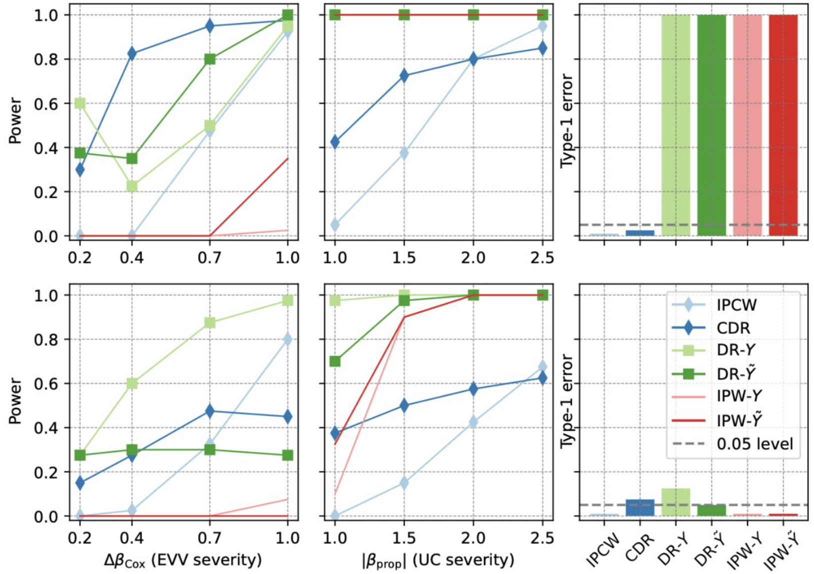
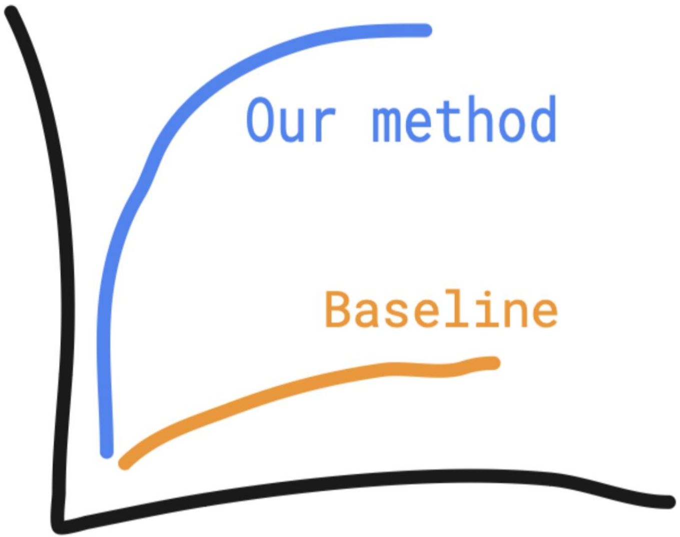
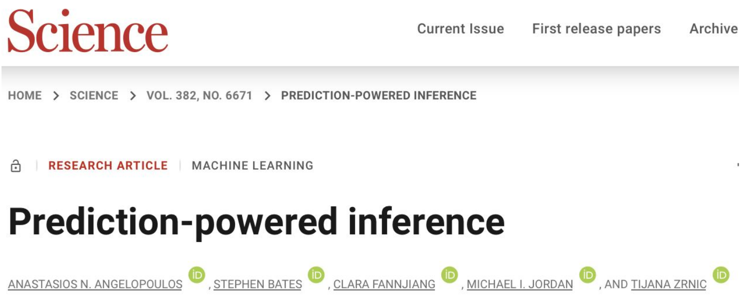


Figure 2: *Top row:* Conditionally independent censoring results. *Bottom row:* Global censoring results. EVV = External validity violation (A2). UC = Unobserved confounding (A1). OS size is $n_1 = 2955$.

(Under review)



Combining RCT + Observational Data



Science

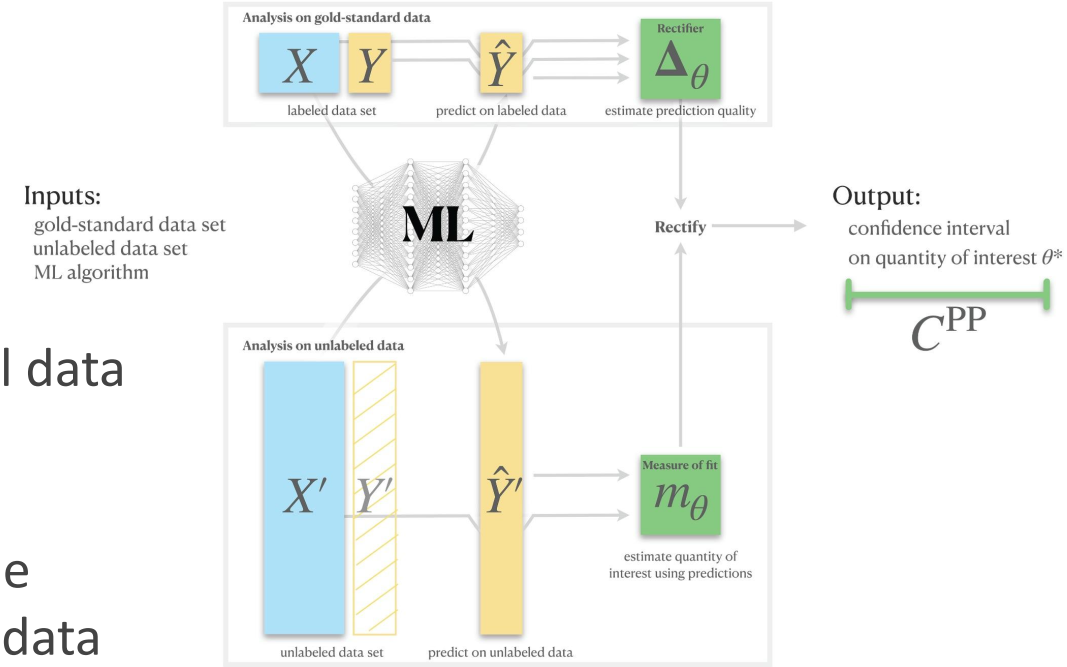
Current Issue First release papers Archive

HOME > SCIENCE > VOL. 382, NO. 6671 > PREDICTION-POWERED INFERENCE

RESEARCH ARTICLE | MACHINE LEARNING

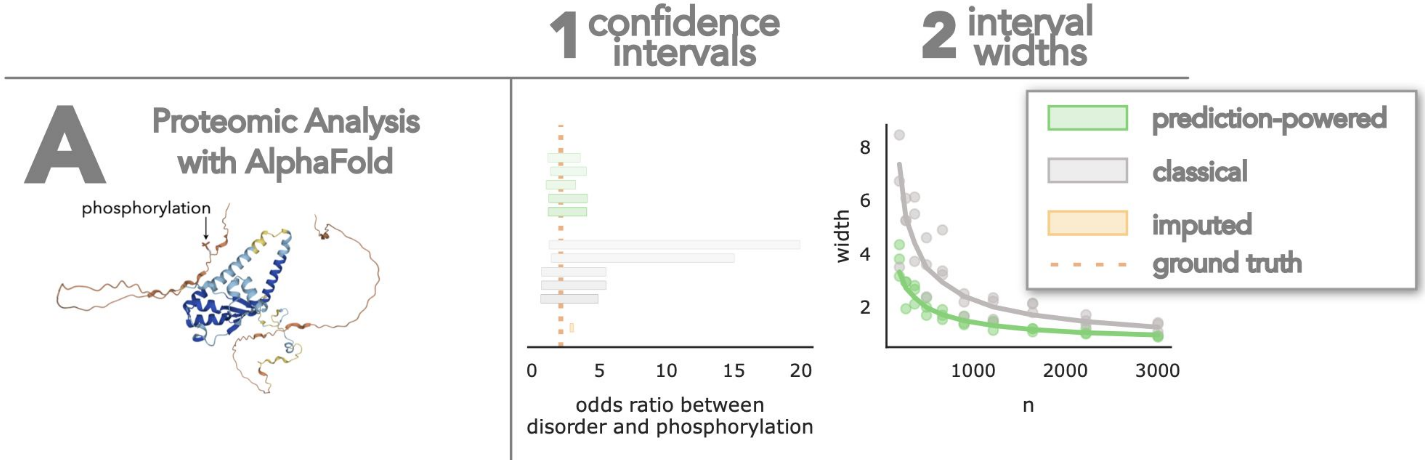
Prediction-powered inference

ANASTASIOS N. ANGELOPOULOS, STEPHEN BATES, CLARA FANNJIANG, MICHAEL I. JORDAN, AND TIJANA ZRNIC



- Gold-standard (labeled) small data
- Unlabeled huge data
- A predictive ML model f
  → cannot use it at face value
- Measure its error on labeled data
- Impute the unlabeled data
  → correct for the error learned
  → interval imputations rather than point imputations

Combining RCT + Observational Data



- Odds ratio (OR) between post-translational modifications (PTM) and intrinsically disordered regions (IDR)
- Use AlphaFold to predict protein structures
- Quantify its error on a small labeled (IDR) dataset (for OR)
- Correct for the error in the imputations for the unlabeled proteins

Combining RCT + Observational Data

(Work in progress)

- Recall the problem with RCTs: limited external validity
  - → consider trial (S = 0) and target (S = 2) populations
  - → $\mathbf{E}\,[Y^1 \mid S = 0] \neq \mathbf{E}\,[Y^1 \mid S = 2]$
  - → confounding! trial participants are systematically different than our target population in prognostic factors

$$\begin{aligned}\mathbf{E}[Y^1|S=2] &= \mathbf{E}_{P_{X|S=2}}[\mathbf{E}[Y^1|X,S=2]] \\ &= \mathbf{E}_{P_{X|S=2}}[\mathbf{E}[Y^1|X,S=0]] \\ &= \mathbf{E}_{P_{X|S=2}}[\mathbf{E}[Y|X,S=0,A=1]]\end{aligned}$$