Estimating long-term treatment effects: a semiparametric proximal DiD approach

extending randomized evidence using real-world data under unmeasured confounding, heterogeneity, and informative dropout

Shu Yang
North Carolina State University
syang24@ncsu.edu

IMSI Workshop
“New Horizons on Model Transportability and Data Integration”
06/22/2026

NC STATE UNIVERSITY

Disclaimer

Funding: This project is supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award, U01FD007934, totaling $2,556,429 over three years, with 100 percent of the funding provided by FDA/HHS.

Disclaimer: The contents are those of the speaker and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government.

Collaboration & Context

This work was developed in collaboration with industry partners to address practical challenges in:

- post-approval evidence generation, and
- long-term effectiveness assessment.

Sihyung Park, North Carolina State University
Mingyang Shan, Eli Lilly and Company
Wenyu Ye , Eli Lilly and Company
Ilya Lipkovich, Eli Lilly and Company

NC STATE UNIVERSITY

Roadmap

- Emerging RWD/E: policy as a catalyst
- The interface of RCT, RWD/E, and CI: combining convention & new
- Proximal DiD Framework for robust learning of long-term treatment effect
- Summary and path forward

NC STATE UNIVERSITY

The 21st century cures act
FDA's RWE framework

The 21st Century Cures Act (Cures Act), signed into law on December 13, 2016, is designed to help accelerate medical product development and bring new innovations and advances to patients who need them faster and more efficiently.

The US Food and Drug Administration (FDA) has developed a RWE framework (FDA, 2018) for evaluating the potential use of real-world evidence (RWE)

NC STATE UNIVERSITY

Regularity initiatives

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

Digital Health Technologies for Remote Data Acquisition in Clinical Investigations

Guidance for Industry, Investigators

Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry

Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 90 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document, contact (CDER) Ansalan Stewart, 240-402-6631, or (CBER) Office of Communication, Outreach and Development, 800-835-4709 or 240-402-8010.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Oncology Center of Excellence (OCE)

November 2021
Real World Data/Real World Evidence (RWD/RWE)

GOV.UK

Home > Clinical trials and investigations > MHRA guidance on the use of real-world data in clinical studies to support regulatory decisions

Medicines & Healthcare products Regulatory Agency

Guidance

MHRA guidance on the use of real-world data in clinical studies to support regulatory decisions

Published 16 December 2021

Contents
1. Scope
2. Introduction to MHRA RWD guidelines
3. Data Quality
4. Advice

Print this page

1. Scope

This document provides an introduction to the MHRA's real-world data (RWD) guideline series, and points to consider when evaluating whether a RWD source is of sufficient quality for the intended use.

Sponsors interested in the use of RWD in their development programmes are encouraged to engage with the MHRA for further advice on specific proposals.

2. Introduction to MHRA RWD guidelines

There are vast amounts of data being collected on patients, for example, in electronic health records (EHR), and disease and patient registries. Such data are commonly called RWD, reflecting that they are collected while patients go about their regular lives, as opposed to being specifically collected in a clinical study. When such data are analysed, the information produced may be referred to as real-world evidence (RWE).

For the purposes of this series of guidelines, RWD are defined as data relating to patient health status or delivery of health care collected outside of a clinical study. Sources of RWD include electronic healthcare records (EHR) defined as structured, digital collections of patient level medical data, primary and secondary care records, disease

国家药品监督管理局药品审评中心
CENTER FOR DRUG EVALUATION, NMPA

关于《药物真实世界研究设计与方案框架指导原则（征求意见稿）》公开征求意见的通知

发布日期：20220707

为了指导申办者科学合理的设计真实世界研究，明确真实世界研究方案撰写的技术要求，我中心组织起草了《药物真实世界研究设计与方案框架指导原则（征求意见稿）》，现公开征求意见。欢迎各界提出宝贵意见和建议，并请及时反馈给我们。

征求意见时限为自发布之日起2个月。

请将您的反馈意见发到邮箱：赵骏 zhaojun@cde.org.cn

感谢您的参与和大力支持。

国家药品监督管理局药品审评中心

2022年7月7日

相关附件	
序号	附件名称
1	《药物真实世界研究设计与方案框架指导原则（征求意见稿）》.pdf
2	《药物真实世界研究设计与方案框架指导原则（征求意见稿）》起草说明.pdf

リアルワールドって何だ？と思ったらここ

RWE

Home | いまさら聞けない？疫学の基礎 | おすすめ書籍 | お問い合わせ | このサイトについて | サイトマップ | プライバシーポリシー | 最新記事

HOME > エビデンス全般 >

エビデンス全般 | ビッグデータ | 規制

Points to Consider for Ensuring the Reliability in Utilization of Registry Data for Applications

2021年5月23日

目次

1. Purpose of the points to consider
2. Concepts of ensuring the reliability
3. Concepts of ensuring the reliability in utilization of registry data as Application data/documents for marketing approval
   (1) Compliance matters for applicants utilizing registry data
      1) Confirmation of data quality management implemented by registry holders
      2) Contracts with registry holders
      3) Statistical analysis
      4) Preparation of application data/documents
      5) Storage of records
   (2) Points to consider for the registry utilized as application data/documents for marketing approval
      1) Governance by registry holders
      2) Computerized system
      3) Quality Management of registry data
      4) Quality Assurance for Registry
      5) Data Extraction and Datasets Preparation
4. Consideration for protection of personal information
5. Concept of ensuring the reliability when registry data are utilized as application data/documents for reexamination or use-results evaluation, etc.
6. Others
7. Definitions of terms

Reference

NC STATE UNIVERSITY

Convention versus new
RCT vs RWD

- Conventional gold-standard: randomized clinical trial (RCT)
- Increasingly trending: real-world data (RWD)

	RCT	RWD
Randomization	✓	X**
Enrollment criteria	restrictive	broad
Generalizability	less*	more
Patient diversity	less	rich
Sample size	small	large
Follow up	short	long
Cost	high	low

* e.g., only 8% of cancer patients participated in oncology trials

** e.g., unmeasured confounding

Research Questions

NC STATE UNIVERSITY

that can be addressed by integrative analysis

Causal inference methods for combining RCT & RWD

Yang and Wang (2022) ASA Biop Report RWE

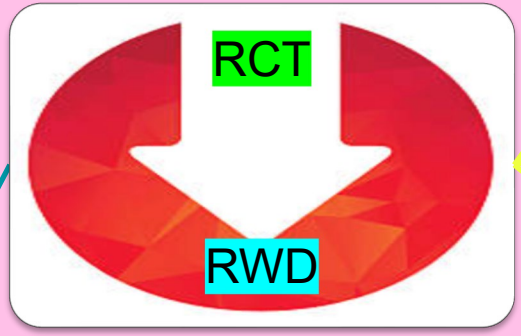
Colnet et al (2023) Statistical Science a review

Zhu et al. (2026) arxiv A scoping review of HCTs

Github repo IntegrativeStats

Leveraging RWD to improve the generalizability of RCT findings

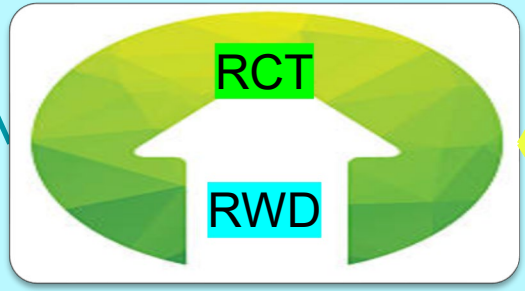
ATE: average treatment effect; ITR: individual treatment regimes



- ATE → Lee et al (2021, 2022, 2024) Biom, JCI, JBS
- ITR → Wu & Yang (2022) JCGS; Chu et al (2023) ICML; Zhao et al (2025) JRML
- ITR with summary stat → Chu et al (2023) Biometrika

Leveraging RWD to improve RCT analysis efficiency

ATE: average treatment effect; HTE: heterogeneity of treatment effect



- ATE: control var → Yang and Ding (2020) JASA
- HTE: Test-then-pool → Yang et al (2023) JRSSB; Gao&Yang (2023) EJS
- ATE/HTE: Bias modeling → Wu & Yang (2022) CLEAR; Cheng et al (2023) UAI; Yang et al (2025) Bernoulli; Mao et al (2025) Biom
- ATE: Selective borrowing → Gao et al. (2024, 2025) Biometrika, ICML; Zhu et al (2025) ICML; Liu et al (2025) arxiv

NC STATE UNIVERSITY

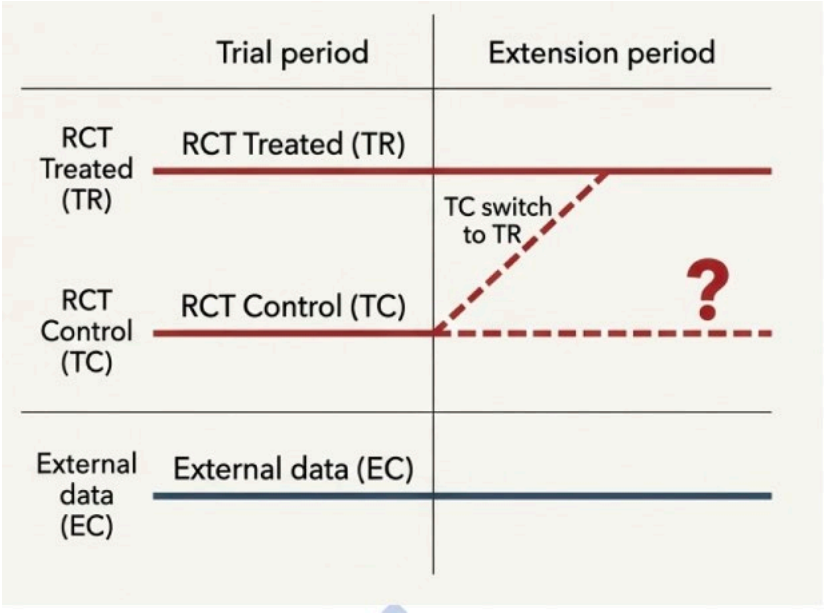
Key objectives **Integrative evidence synthesis**

Integrate evidence from randomized clinical trials and real-world data to

1. improve the generalizability and transportability of RCT findings,
2. increase the efficiency and statistical power of treatment effect estimation,
3. **enable long-term safety and effectiveness monitoring and estimation**,
4. and many others

Long-term treatment effects are critical, but RCTs provide only short-term answers.

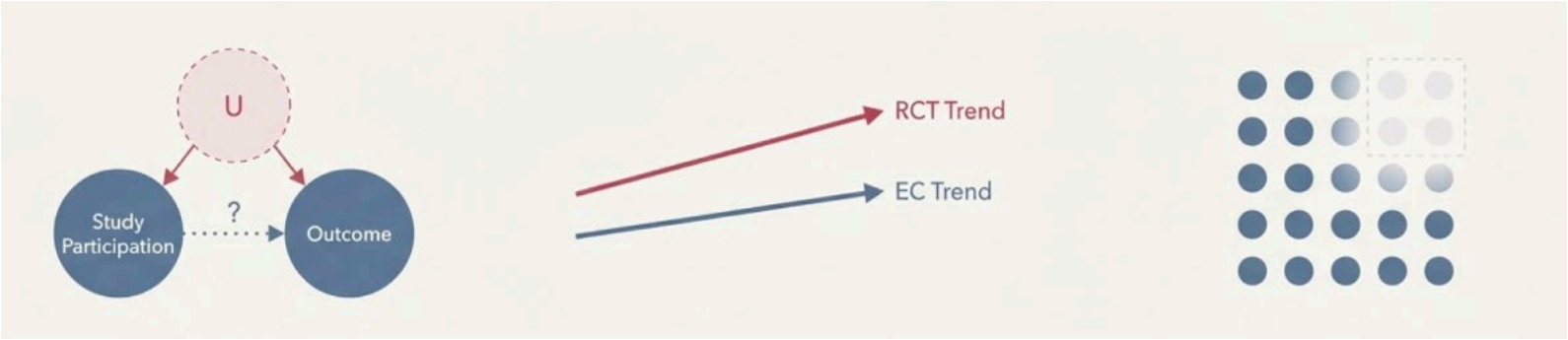
- Assessing long-term treatment effects is crucial for understanding sustained efficacy, adverse effects, drug tolerance, and for cost-effectiveness analysis required by health technology assessment bodies.
- However, long-term RCTs are often infeasible due to prohibitive costs and ethical issues like patient crossover.
- This creates a critical evidence gap. The central question is: What would have happened to the RCT control group over the long term if they hadn't crossed over to the treatment arm? This is the 'missing counterfactual' we need to estimate.



NC STATE UNIVERSITY

Simply augmenting RCTs with external data **is fraught with hidden biases.**

Combining data sources is not straightforward due to fundamental differences. We face three core challenges:



Unmeasured Confounding (U)
Key patient characteristics that can influence both study participation and health outcomes (e.g., health consciousness, underlying disease severity) are not recorded, creating spurious correlations.

Study Heterogeneity
The experience of being in a trial may be fundamentally different from receiving routine care (e.g., placebo effects, Hawthorne effect). This means outcome trends can diverge between RCT and external cohorts, even for otherwise identical patients.

Informative Missingness
Patient dropout is rarely random. Sicker patients are more likely to discontinue or be lost to follow-up. The observed data represent a selected, healthier, biased subset of the original population.

Ignoring any one of these can materially bias long-term effect estimates.

NC STATE UNIVERSITY

Existing methods
fall short of jointly addressing all challenges

No existing method jointly addresses unmeasured confounding, study heterogeneity, and informative missingness.

Method	Key Assumption	Why It Fails in This Context
Standard Data Integration	Exchangeability between data sources, conditional on confounders.	Unlikely in practice due to study-specific effects (e.g., placebo, Hawthorne effect), creating inherent data heterogeneity.
Synthetic Control /Matrix Completion	Linear factor models, cannot allow for systematic differences due to U	Cannot handle data heterogeneity or informative missingness that are common in real-world data integration scenarios.
Difference-in-Differences	Confounding effects are constant over time	Violated by common data structures like linear factor models where a confounder (U)'s impact can evolve over time.
Proximal Inference	Assume data homogeneity and fully observed outcomes	Cannot handle data heterogeneity or informative missingness that are common in real-world data integration scenarios.

A gap exists for a method that can handle all these challenges simultaneously.

NC STATE UNIVERSITY

Notation

Data structure and potential outcomes

Data structure

- $S = 1$ if the subject is from RCT, 0 if from EC
- X: baseline covariates
- $t \in \{pre, 0\}$: time periods
- (A_{pre}, A_0): treatment sequence
- (Y_{pre}, Y_0): outcome sequence
- (R_{pre}, R_0): observation sequence
- abbreviate (A_0, Y_0, R_0) as (A, Y, R)
- U: latent confounder $(S \leftarrow U \rightarrow Y)$

Potential outcomes

- $Y_t^{s,(a_{pre},a_0)}$: potential outcome had subject been assigned to study s and treatment (a_{pre}, a_0)

Stable Unit Treatment Value Assumption (SUTVA)

- Future treatment does not affect past potential outcomes
- e.g. Y_t in EC is $Y_t^{0,(0,0)}$ and $Y_t^{(0,0)}$ is $Y_t^{S,(0,0)}$

Defining the statistical challenge: Estimating the counterfactual

The Objective

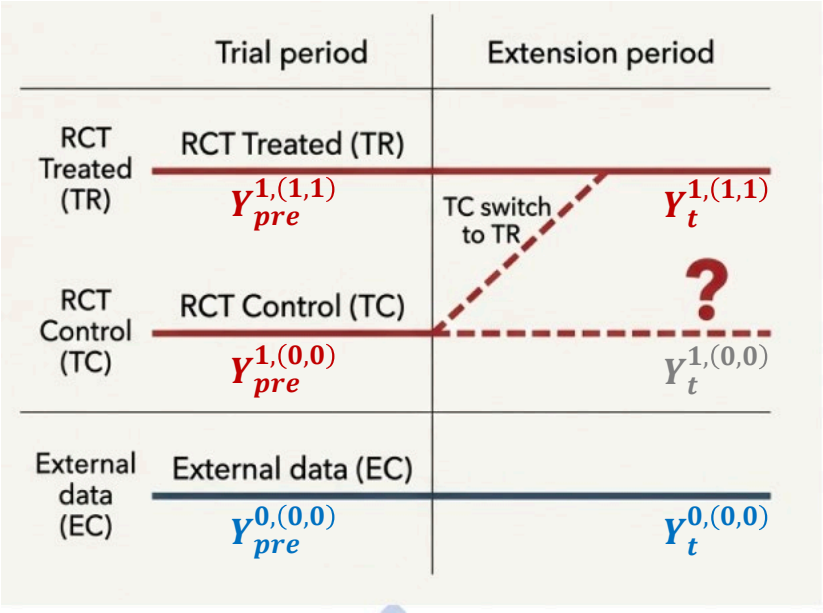
- Estimate the long-term average treatment effect (LTATE) for RCT subjects at time t in the extension period:

$$LTATE_t = \mathbb{E}\left[Y_t^{1,(1,1)} - Y_t^{1,(0,0)} \mid S = 1 \right].$$

The core problem

- The treated outcome $\mathbb{E}\left[Y_t^{1,(1,1)} \mid S = 1 \right]$ is observed. The challenge is estimating the counterfactual control outcome

$$\tau_t = \mathbb{E}\left[Y_t^{1,(0,0)} \mid S = 1 \right]$$



NC STATE UNIVERSITY

A Motivating Example
Solanezumab trial

- **RCT**: Solanezumab trial data that examine the effectiveness of Solanezumab on slowing the progression of Alzheimer's disease (Siemers et al., 2016) + **EC**: Geras observational data (Wimo et al., 2013) to showcase estimators for LTATE in clinical settings

- 19 covariates forming X(including BMI, age, duration of diagnosis, and cognitive and functional scores such as ADL, NPI, EQ-5D, ADAS-13), baseline Mini-Mental State Examination (MMSE) score forming U

- Instrumental Activities of Daily Living (iADL) scores measured across four time points (t = base, 28, 52, 80) were selected as outcomes Y_t.

- The trial period ends at week 28. **Our objective** is to estimate the RCT control trend at week 52.

A Motivating Example
Solanezumab trial

We conducted a plasmode simulation:

- The study inclusion probability

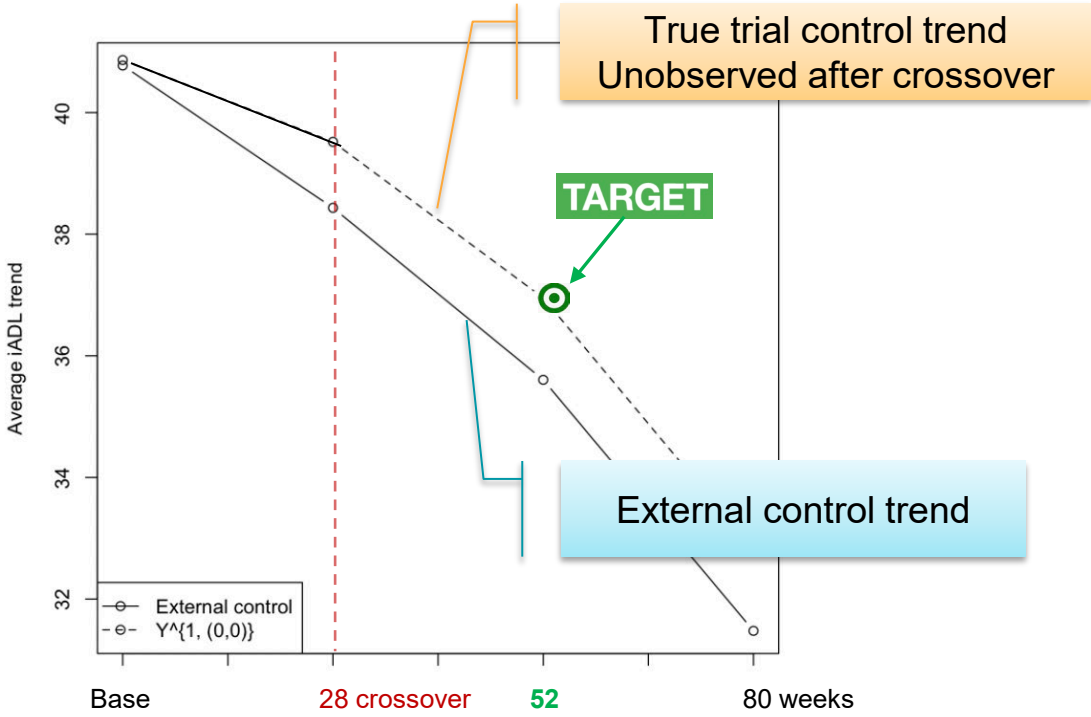
$$P(S = 1 \mid U, X) = \frac{1}{1 + \exp(X\,\eta_x + U\,\eta_u)\}}$$

- Potential outcomes

$$Y^{s,(0,0)} = XB_0 + U\Gamma_0 + s \cdot (XB_1 + U\Gamma_1) + E,$$

where $B_s \in \mathbb{R}^{p\times T}, \Gamma_s \in \mathbb{R}^{m\times T}$

- $Y_{base} \perp S \mid (X, U)$ by generating it from the model without $s \cdot (X\,B_1 + U\,\Gamma_1)$



Existing literature on borrowing external controls for long-term treatment effect estimation

Standing on the shoulder of the giant

NC STATE UNIVERSITY

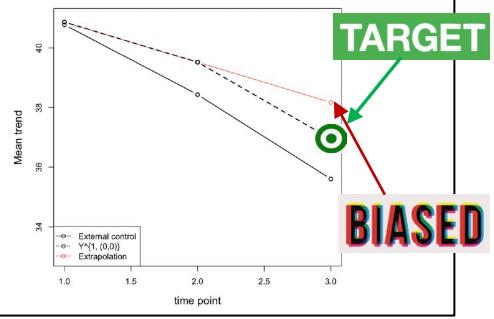
Standard data integration

Extrapolation, prediction and matching approaches

- These methods rely on **Structural Models or Ignorability**, assuming all relevant confounders are measured and accounted for.

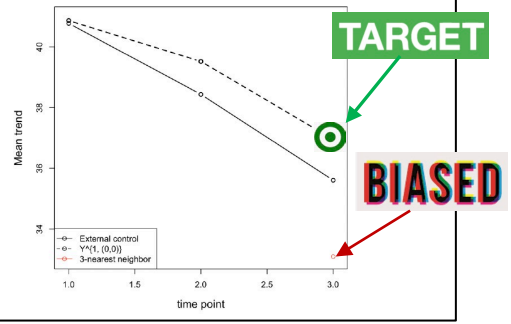
Extrapolation

Fit a parametric model (e.g., linear) to the RCT control arm's outcomes during the trial period and extend the curve into the future.



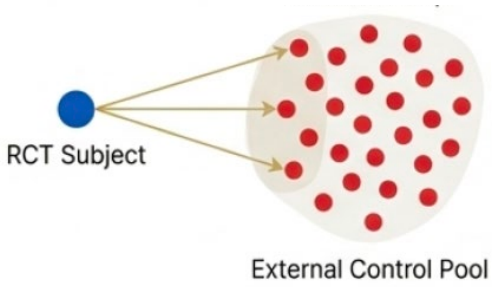
Predictive Model

Train a model on EC data to predict future outcomes based on covariates. Apply this model to RCT control subjects.



Matching

For each RCT control subject, find matching subjects in the EC pool with similar baseline characteristics to impute the missing long-term counterfactual control outcome.



These methods are straightforward but are sensitive to model misspecification and unobserved differences between the RCT and external control populations.

NC STATE UNIVERSITY

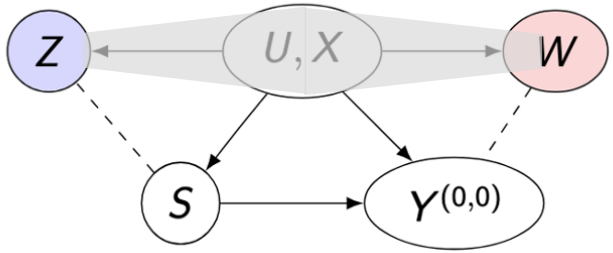
Proximal Causal Inference

Taming unmeasured confounding with observable proxi

How do we operationalize assumptions about an unobserved U? Through **Proximal Causal Inference**.

The key is to find **observed proxies** that are informative about U but satisfy specific **conditional independence properties**.

Mataphor: Proxies are like shadows cast by the hidden confounder U. By observing the shadows, we can infer/correct the impacts of U.



- $W \perp\!\!\!\perp (Z, S) | (X, U)$.
- $Z \perp\!\!\!\perp (Y^{(0,0)}, W) | (X, U, S)$.

NC STATE UNIVERSITY

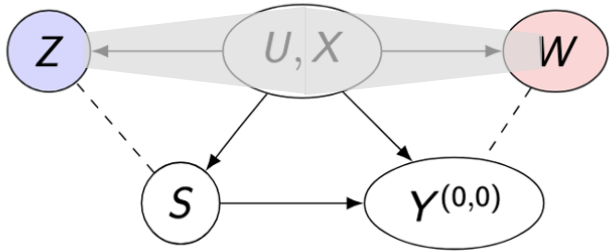
Proximal Causal Inference

Taming unmeasured confounding with observable proxi

In our context, we use two types of observable proxies that are related to U:

- **Outcome proxies** W: variables related to the outcome, but not trial participation, given U, X.
  Example: baseline functional scores or comorbidities not used in trial eligibility criteria.
- **Source proxies** Z: variables related to study participation, but not the outcome, given U, X.
  Examples: distance to study site or caregiver support.

Mechanism: Using both types of proxies allows us to mathematically "bridge" between observed data and the latent confounder U.



- $W \perp\!\!\!\perp (Z, S) | (X, U)$.
- $Z \perp\!\!\!\perp (Y^{(0,0)}, W) | (X, U, S)$.

NC STATE UNIVERSITY

Proximal Causal Inference

Taming unmeasured confounding with observable proxi

ID	Covariates X	Proxy W	Treatment	Short-term outcome	Long-term outcome
001	X_1	W_1	treated	$Y_{pre,1}$	$Y_{t,1}$
				$Y_{pre,2}$	$Y_{t,2}$
003	X_3	W_3	treated	$Y_{pre,3}$	$Y_{t,3}$
004	X_4	W_4	control	$Y_{pre,4}$	?
005	X_5	W_5	control	$Y_{pre,5}$	?
006	X_6	W_6	control	$Y_{pre,6}$	?

Outcome bridge function $h(W, X)$

- **The goal:** predict what Y_t would be in the RCT control group, accounting for U.
- Define the outcome confounding bridge function
$$\mathbb{E}[\,Y \mid X, Z, S = 0\,] = \mathbb{E}[\,h(W, X) \mid X, Z, S = 0\,]$$
- Under suitable completeness conditions
$$\mathbb{E}\big[\,Y^{(0,0)} \mid X, U, S = 0\,\big] = \mathbb{E}[\,h(W, X) \mid X, U, S = 0\,]$$
- Under study homogeneity and ignorability
$$\tau = \mathbb{E}[\,h(W, X) \mid S = 1\,]$$

Interpretation:

- Use EC to learn $h(W, X)$
- Use Z as a lever to identify the relationship.
- Once learned, $h(W, X)$ acts as a "proxy outcome" for $Y^{(0,0)}$ in the RCT.

NC STATE UNIVERSITY

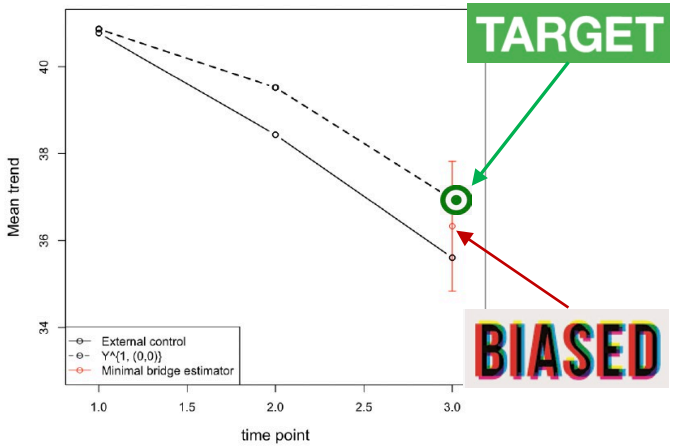
Proximal Causal Inference

Taming unmeasured confounding with observable proxi

ID	Covariates X	Proxy W	Treatment	Short-term outcome	Long-term outcome
001	X_1	W_1	treated	$Y_{pre,1}$	$Y_{t,1}$
				$Y_{pre,2}$	$Y_{t,2}$
003	X_3	W_3	treated	$Y_{pre,3}$	$Y_{t,3}$
004	X_4	W_4	control	$Y_{pre,4}$	?
005	X_5	W_5	control	$Y_{pre,5}$	?
006	X_6	W_6	control	$Y_{pre,6}$	?

Outcome bridge function
$h(W, X)$

- However, under study heterogeneity, proximal methods alone become biased:



NC STATE UNIVERSITY

DiDaddresses study heterogeneity

Assuming parallel trends (PT)

- Classical parallel trend assumption

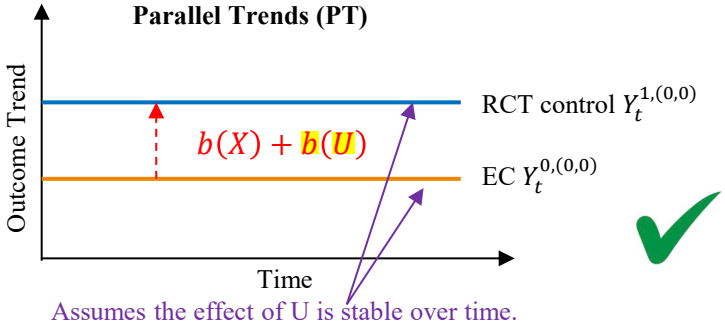
$$\mathbb{E}\left[Y^{s,(0,0)} - Y_{pre}^{s,(0,0)} \mid U, X \right] = \mathbb{E}\left[Y^{s,(0,0)} - Y_{pre}^{s,(0,0)} \mid X \right]$$
$$\mathbb{E}\left[Y^{1,(0,0)} - Y_{pre}^{1,(0,0)} \mid X \right] = \mathbb{E}\left[Y^{0,(0,0)} - Y_{pre}^{0,(0,0)} \mid X \right]$$

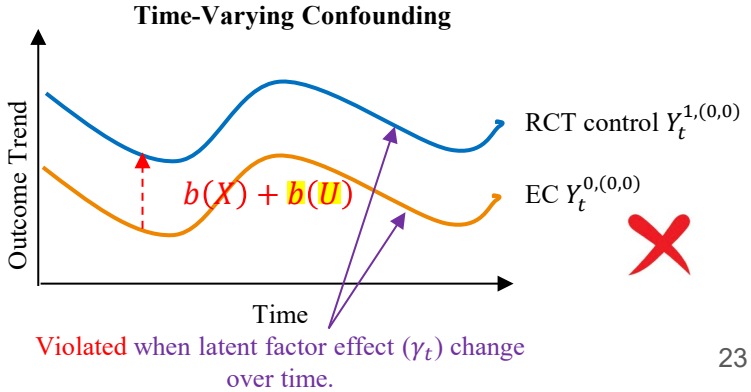
- The effect of U on $Y_t^{s,(0,0)}$ should be **additive and stable** across time:

$$Y_t^{s,(0,0)} = \nu_t(X) + \nu(U) + s \cdot \{b(X) + b(U)\} + \epsilon_{s,t}$$

- Limitation: fails for common models, e.g., linear factor models with time-varying effects of U

- In many biomedical settings, the impact of latent disease severity changes over time.





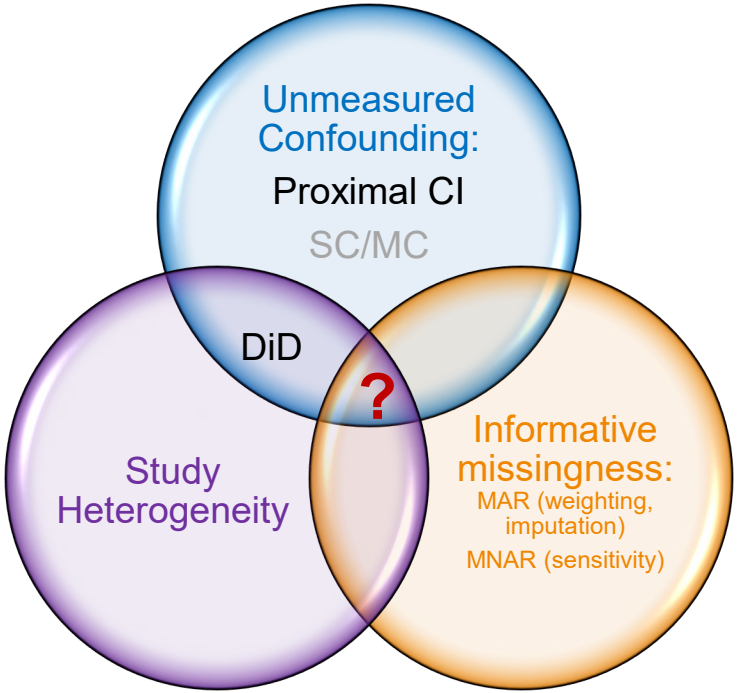
NC STATE UNIVERSITY

Summary

Existing method landscape

Existing methods address one or two challenges in the context;

Is there a unified framework which can address all these challenges?

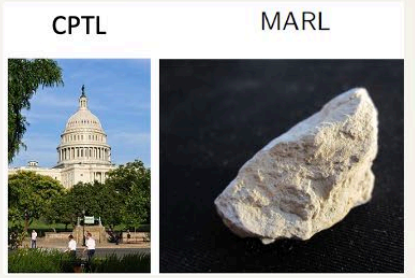


We introduce a unified framework to overcome **confounding, heterogeneity, and missingness**

- Our approach integrates the power of Pillar 1 Proximal Causal Inference with Pillar 2 flexible Difference-in-Differences structure.

This is the first semiparametric methodology to systematically address these challenges in concert by introducing:

1. A more realistic parallel trends assumption (**CPTL**).
2. A principled approach to informative missingness (**MARL**).
3. A doubly robust and efficient estimator (**PDR**).



NC STATE UNIVERSITY

Pillar 2
Generalizing parallel trends with CPTL

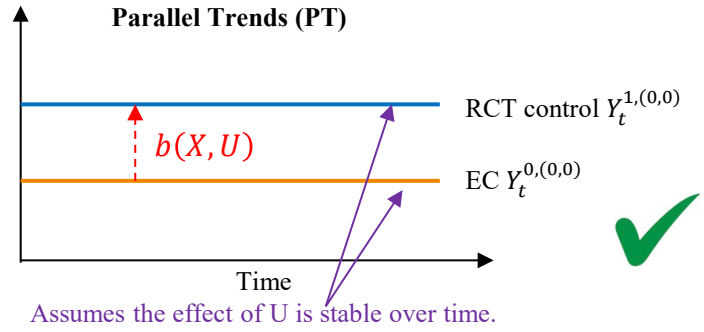
- **Assumption 1 (Conditional Parallel Trend given Latent factors or CPTL)**

$$\mathbb{E}\left[Y_t^{1,(0,0)} - Y_t^{0,(0,0)} \mid X, U \right] = b(X,U)$$

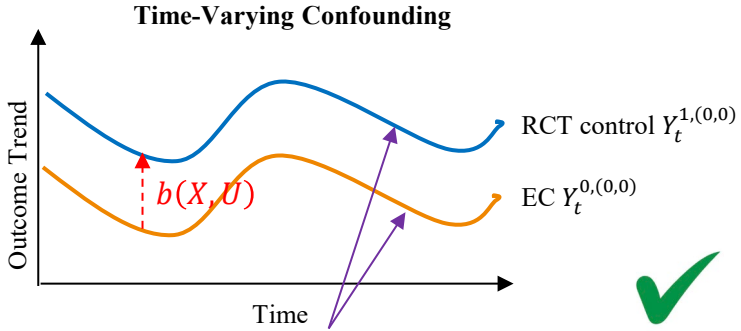
- Similar to PT, gap is stable over time!

- But this directly allows for time-varying confounder effects. CPTL is compatible with flexible outcome models like

$$Y_t^{s,(0,0)} = \mu_t^{EC}(X,U) + s \cdot b(X,U) + \epsilon_{t,s}$$



Assumes the effect of U is stable over time.

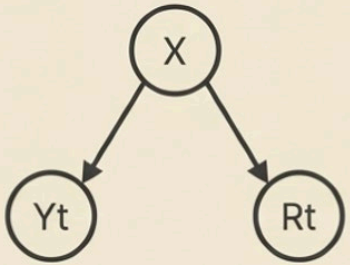


Violated when latent factor effect (γ_t) change over time.

NC STATE UNIVERSITY

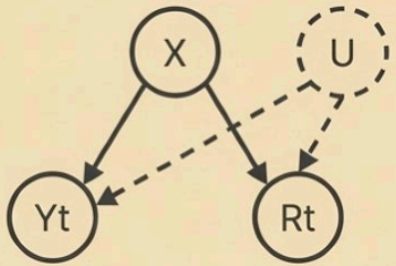
Pillar 3
Informative missingness (MARL)

MAR (Missing At Random)



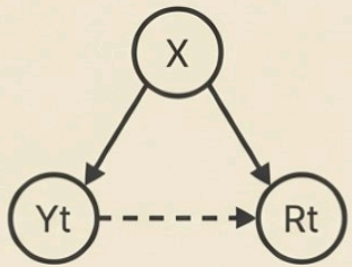
Assumes missingness depends only on observed data. This is often an **overly optimistic** assumption.

MARL (Our Approach)



A **powerful middle ground** where missingness depends on the **unmeasured confounder U**. This is realistic (e.g., sicker patients with a high value of U drop out) and, crucially, **identifiable** using our **proximal** framework.

MNAR (Missing Not At Random)



Assumes missingness depends on generally unobserved outcome values. This is **generally intractable**.

NC STATE UNIVERSITY

ProximaDiD Framework

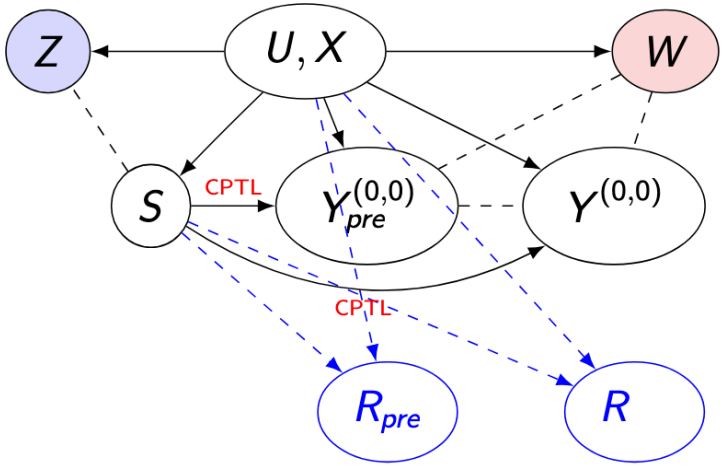
CPTL+MARL

Integrated handling of bias

Unmeasured confounding
→ handled by proxies

Study heterogeneity
→ handled by CPTL

Informative missingness
→ handled by MARL



A causal diagram illustrates the proximal DiD setup under proxies, CPTL, and MARL

NC STATE UNIVERSITY

Identification

Multiple paths

$$Y_t^{s,(0,0)} = \mu_t^{EC}(X,U) + s \cdot b(X,U) + \epsilon_{s,t}$$

$$\tau = \mathbb{E}[Y^{1,(0,0)} \mid S = 1] = \mathbb{E}[\mu_0^{EC}(X,U) + b(X,U) \mid S = 1]$$

Path 1: The Outcome Bridge
Imputation

ID	Covariates X	Proxy W	Treatment	Short-term outcome	Long-term outcome
001	X_1	W_1	treated	$Y_{pre,1}$	$Y_{t,1}$
			d	$Y_{pre,2}$	$Y_{t,2}$
			d	$Y_{pre,3}$	$Y_{t,3}$
004	X_4	W_4	control	$Y_{pre,4}$	?
005	X_5	W_5	control	$Y_{pre,5}$	?
006	X_6	W_6	control	$Y_{pre,6}$	?

Bias-corrected outcome bridge function $h(W,X) + h_{pre}(W,X,1) - h_{pre}(W,X,0)$

The goal: predict what Y_t would be in the RCT control group, accounting for U and study heterogeneity.

Define the outcome bridge functions for pre- and post-extension times

$$\mathbb{E}[\, Y - h(W,X) \mid Z, X, S = 0, R = 1 \,] = 0$$
$$\mathbb{E}\big[\, Y_{pre} - h_{pre}(W,X,S) \mid Z, X, S, A_{pre} = 0, R_{pre} = 1 \,\big] = 0$$

Interpretation:

- Under CPTL, in oracle, impute ? by $\mu_0^{EC}(X,U) + b(X,U)$
- $h(W,X)$ connects to EC trend $\mu_0^{EC}(X,U)$
- $h_{pre}(W,X,1) - h_{pre}(W,X,0)$ connects to study difference $b(X,U)$

Thus, $h(W,X) + h_{pre}(W,X,1) - h_{pre}(W,X,0)$ acts as a "proxy outcome" for long-term RCT control counterfactual.

NC STATE UNIVERSITY

Identification

Multiple paths

$$Y_t^{s,(0,0)} = \mu_t^{EC}(X,U) + s \cdot b(X,U) + \epsilon_{s,t}$$

$$\tau = \mathbb{E}[Y^{1,(0,0)} \mid S=1] = \mathbb{E}[\mu_0^{EC}(X,U) + b(X,U) \mid S=1]$$

Path 1

Outcome bridge identification

$$\tau = \mathbb{E}\big[\, h(W,X) + h_{pre}(W,X,1) - h_{pre}(W,X,0) \mid S=1 \,\big]$$

Path 2

Path 3

Path 2: The Source Bridge

Reweighting

EC observed $(1-S)RY$ is biased for τ due to
i) selection bias and missingness and ii) study heterogeneity

The goal: Rewright the external control so its confounder distribution looks like the RCT, and correct the bias due to study heterogeneity.

Define **the source bridge functions** for pre- and post- extension period

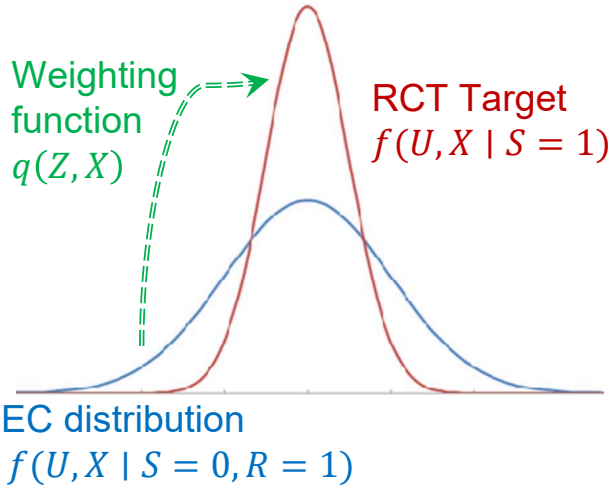
$$\mathbb{E}[\,(1-S)Rq(Z,X) - S \mid W,X\,] = 0$$
$$\mathbb{E}\big[\,1\{S=s\}R_{pre}q_{pre}(Z,X,s) - S \mid W,X\,\big] = 0$$

Interpretation:

- (q_{pre}, q) recover ratios of selection and observation probabilities

$$\mathbb{E}[\,q(Z,X) \mid U,X,S=0,R=1\,] = \frac{P(S=1 \mid U,X)}{P(S=0,R=1 \mid U,X)}$$
$$\mathbb{E}\big[\,q_{pre}(Z,X,s) \mid U,X,S=s,R_{pre}=1\,\big] = \frac{P(S=1 \mid U,X)}{P(S=s,R_{pre}=1 \mid U,X)}$$

- Reweight and correct: $(1-S)Rq(Z,X)Y + (2S-1)q_{pre}(Z,X,S)R_{pre}Y_{pre}$



NC STATE UNIVERSITY

Identification

Multiple paths

$$Y_t^{s,(0,0)} = \mu_t^{EC}(X,U) + s \cdot b(X,U) + \epsilon_{s,t}$$
$$\tau = \mathbb{E}[Y^{1,(0,0)} \mid S=1] = \mathbb{E}[\mu_0^{EC}(X,U) + b(X,U) \mid S=1]$$

Path 1

Outcome bridge identification

$$\tau = \mathbb{E}\big[\, h(W,X) + h_{pre}(W,X,1) - h_{pre}(W,X,0) \mid S=1 \,\big]$$

Path 2

Source bridge identification

$$\tau = \mathbb{E}\big[(1-S)Rq(Z,X)Y + (2S-1)q_{pre}(Z,X,S)R_{pre}Y_{pre}\big]/P(S=1)$$

Path 3

NC STATE UNIVERSITY

Identification

Multiple paths

$$Y_t^{s,(0,0)} = \mu_t^{EC}(X,U) + s \cdot b(X,U) + \epsilon_{s,t}$$
$$\tau = \mathbb{E}[Y^{1,(0,0)} \mid S=1] = \mathbb{E}[\mu_0^{EC}(X,U) + b(X,U) \mid S=1]$$

Path 1

Outcome bridge identification

$$\tau = \mathbb{E}\big[\, h(W,X) + h_{pre}(W,X,1) - h_{pre}(W,X,0) \mid S=1 \,\big]$$

Path 2

Source bridge identification

$$\tau = \mathbb{E}\big[(1-S)Rq(Z,X)Y + (2S-1)q_{pre}(Z,X,S)R_{pre}Y_{pre}\big]/P(S=1)$$

Path 3

Combined identification

$$\tau = \mathbb{E}\big[\, h(W,X) + h_{pre}(W,X,1) - h_{pre}(W,X,0) \mid S=1 \,\big] +$$
$$\mathbb{E}\big[(1-S)Rq(Z,X)e + (2S-1)q_{pre}(Z,X,S)R_{pre}e_{pre}\big]/P(S=1)$$

- $(e = Y - h,\ e_{pre} = Y_{pre} - h_{pre})$

NC STATE UNIVERSITY

Three estimators deliver Increasing levels of robustness

- Based on these bridge functions $H = (h, h_{pre})$ and $Q = (q, q_{pre})$, we derive a class of semiparametric estimators for the counterfactual mean τ.

Estimator	Name	Relies on correct specification of …	Key Property
$\hat{\tau}_{POR}$	Proximal outcome regression	Outcome model (H)	Simple to implement, relies on correctly modeling outcome trends.
$\hat{\tau}_{PIPW}$	Proximal inverse probability weighting	Source model (Q)	Uses weighting to balance populations, relies on correctly modeling study participation and missingness.
$\hat{\tau}_{PDR}$	Proximal doubly robust	Outcome model (H) **OR** Source model (Q)	Consistent if either model is correct. Efficient if both are correct.

- PDR is the preferred choice, offering protection against model misspecification and achieving the semiparametric efficiency bound.

Nuisance parameter estimation
Bridge functions

$$\mathbb{E}[\{Y - h(W,X)\}g(Z,X) \mid S = 0, R = 1] = 0$$
$$\mathbb{E}\left[\{Y_{pre} - h_{pre}(W,X,S)\}g(Z,X,S) \mid A_{pre} = 0, R_{pre} = 1\right] = 0$$
$$\mathbb{E}[\{(1-S)Rq(Z,X) - S\}g(W,X)] = 0$$
$$\mathbb{E}[\{1\{S = s\}R_{pre}q_{pre}(Z,X,s) - S\}g(W,X)] = 0$$

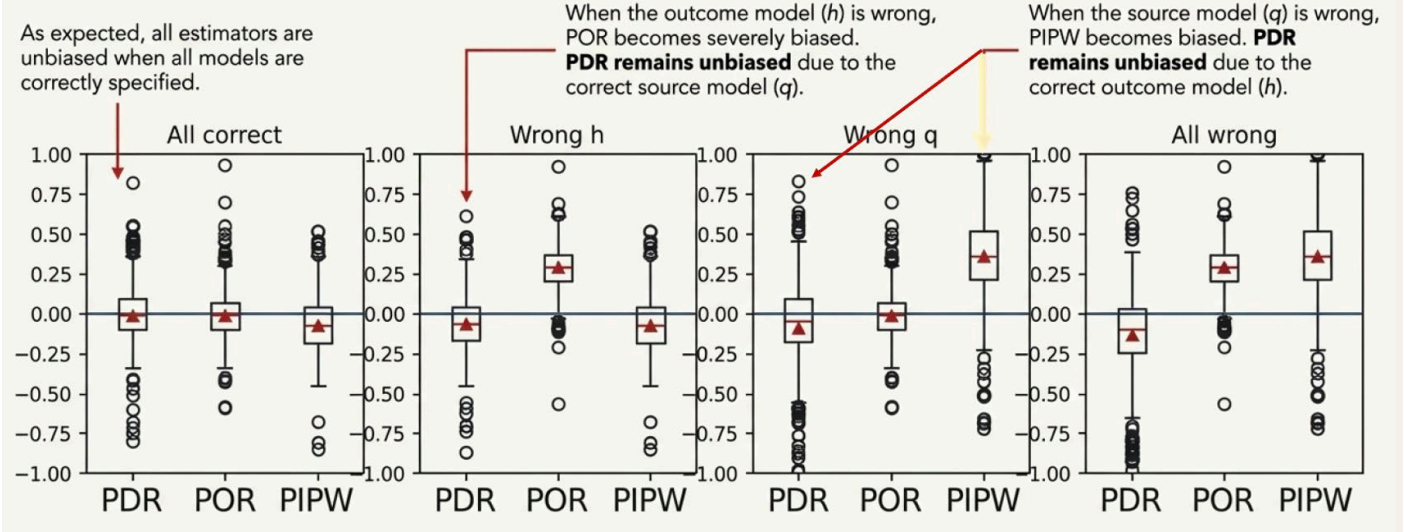
for any reasonable $g(\cdot)$

Implications: The bridge functions can be solved using
- Generalized method of moments (Imbens et al., 2021)
- Minimax approach with kernels (Ghassami et al., 2022; Kallus et al., 2021)
- Neural networks (Zhang et al., 2025)

Simulation

Confirm the PDR estimator's robustness.

We conducted a simulation study to evaluate the properties of our estimators under different scenarios of model misspecification.



Conclusion: the double robustness property provides powerful protection against model misspecification in practice.

NC STATE UNIVERSITY

The clinical challenge

Estimating Alzheimer's progression after crossover

Data Sources:

RCT: Phase 3 trial data for **Solanezumab**, a treatment for Alzheimer's disease.

EC: **GERAS** observational study data.

The Problem: In the Solanezumab trial, control patients crossed over to active treatment, leaving the **long-term (52-week) control trajectory unobserved**.

The Goal: Estimate the mean 52-week iADL score for the RCT control group, a quantity essential for understanding the drug's sustained effect.

Proxy Selection: We used the longitudinal outcomes themselves as proxies, consistent with our framework:

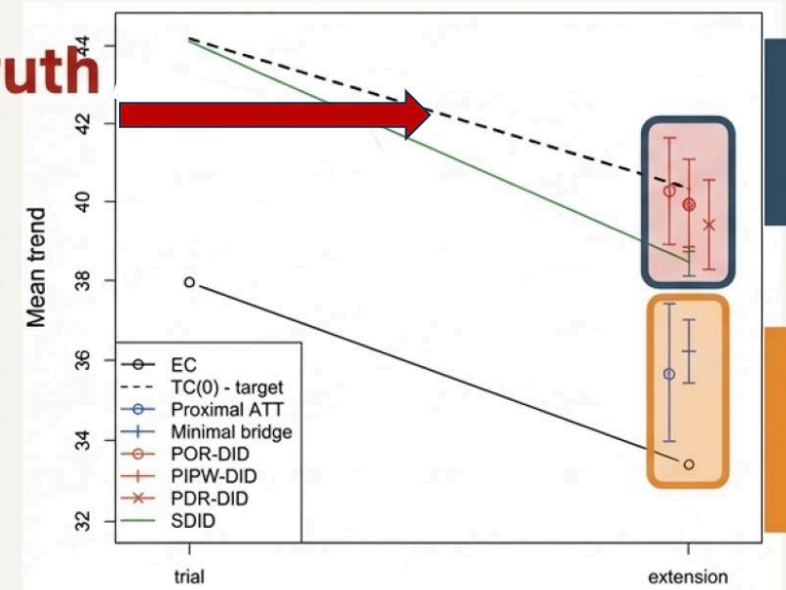
$\boldsymbol{W}$ (Outcome-Proxy) = Y_{base} (Baseline iADL score)

$\boldsymbol{Z}$ (Source-Proxy) = Y_{80} (Week 80 iADL score from EC data)

NC STATE UNIVERSITY

Our method uniquely recovers the true trend, outperforming state-of-the-art alternatives.

The Ground Truth



Proposed methods successfully estimate the true long-term outcome.

Existing methods show significant bias, failing to account for the complex confounding.

NC STATE UNIVERSITY

A flexible and robust solution for a critical evidence gap.

A Novel Unified Framework	•The first to systematically address unmeasured confounding, study heterogeneity, and informative missingness when integrating external data
Weaker, More Realistic Assumptions	•Introduced CPTL and MARL, relaxing the restrictive assumptions that limit the validity of existing methods
A High-Performance Estimator	•Developed PDR, doubly robust and semiparametrically efficient
Demonstrated Real-World Value	•Proved the method's superiority in a challenging and clinically relevant Alzheimer's disease simulation

How to "design" ECs? → How to select and utilize proxy variables? → General outcomes such as time-2-event

NC STATE UNIVERSITY

Thank You!
Any Questions?

syang24@ncsu.edu
https://shuyang.wordpress.ncsu.edu/

NC STATE UNIVERSITY

Special Case
No missingness

- When outcomes are fully observed:
  - Simplified identification
  - Reduced data requirements
  - Longitudinal outcomes act as proxies

NC STATE UNIVERSITY

Special Case
No missingness

$$R_{pre} = R = 1, q_{pre}(Z, X, 1) = 1, q_{pre}(Z, X, 0) = q(Z, X)$$

$$\tau = \mathbb{E}\big[\, d(W, X) + Y_{pre} \mid S = 1 \,\big] + \frac{\mathbb{E}[(1 - S)q(\boldsymbol{Z}, X)\delta]}{P(S = 1)}$$

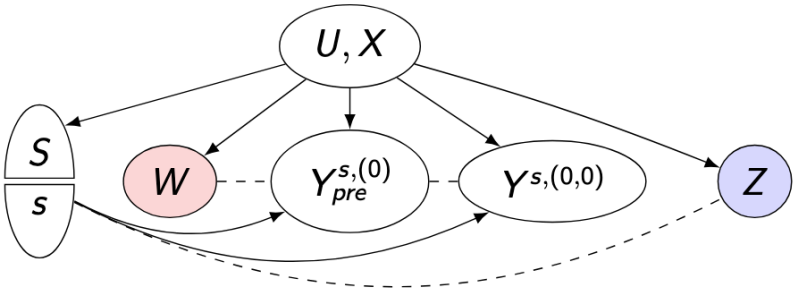
- $d(W, X)$ is a bridge function for $Y - Y_{pre}$ in RCT controls
- $\delta = Y - Y_{pre} - d(W, X)$
- $\boldsymbol{Z}$ **is not necessary from RCT data!**

NC STATE UNIVERSITY

The practical advantage

Longitudinal outcomes can serve as their own proxie

- Use observed outcomes themselves as proxies. This makes the method highly practical for longitudinal studies.
- Baseline outcome (Y_{base}) as 'W': if it is "externally valid" $Y_{base} \perp S \mid (X, U)$
- Post-baseline outcome (Y_{post}) as 'Z': a later outcome observed in the EC can serve as the source-proxy 'Z'



- Assuming factor model $(W, Y_{pre}^{s,(0)}, Y^{s,(0,0)}, Z) \sim X + U$
- W: outcome prior to Y_{pre}.
- Z: outcome after Y.