

Robust Causal Inference in Real-world Evidence Studies (RWE) with Double Machine Learning (DML)



Jing Dai

*Director Biostatistics
Jazz Pharmaceuticals*



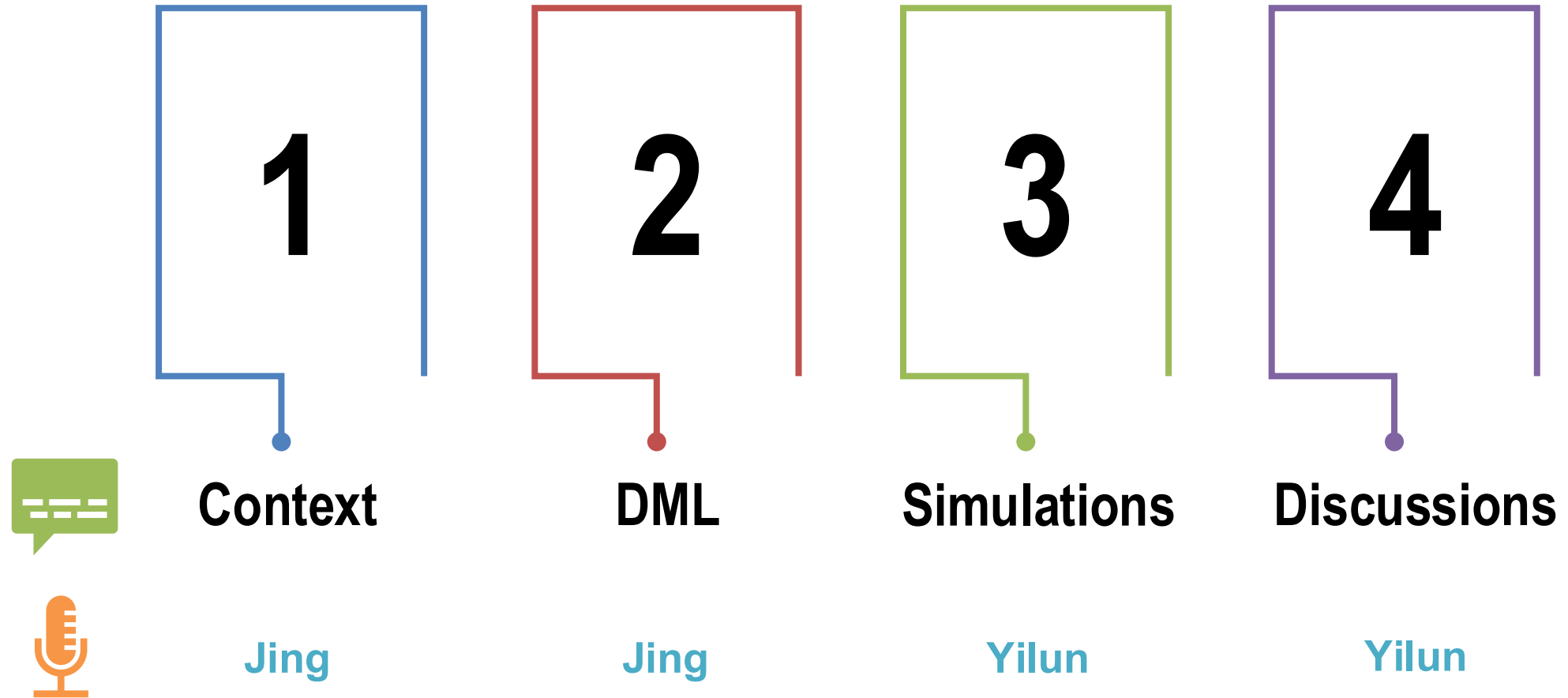
Yilun Huang

*Principal Biostatistician
Novartis*

Disclaimer

- ❖ The views and opinions expressed in this presentation are those of the speakers and do not necessarily reflect the positions of any entities they represent.
- ❖ Select visual elements in this presentation were created using generative AI tools.

Agenda





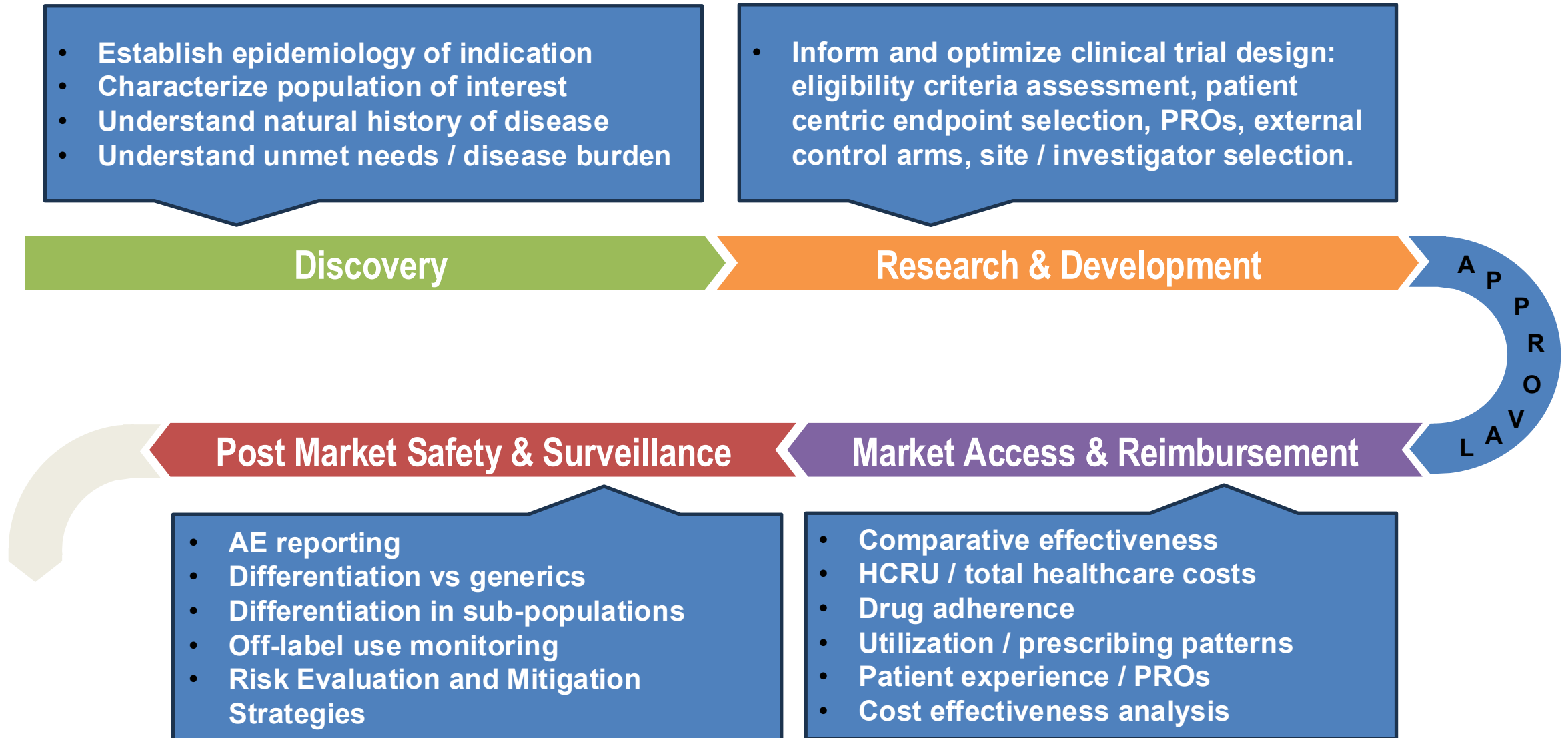
CONTEXT

RWE | CAUSAL INFERENCE

The Rise of Real-world Evidence (RWE)

Clinical Data	Claims Data	Patient Generated Data	Registries / Cohort Studies / Public Data	Other
• Electronic Health Records (EHRs) • Case Report Forms (CRFs)	• Medical claims • Pharmacy claims	• Survey data • Patient Reported Outcomes (PROs)	• Registry studies • Longitudinal cohort studies • Government / nonprofit networks	• Literature • Social media • Device / mobile health • Molecular profiling

RWE Application throughout Drug Development Lifecycle



Deloitte RWE Benchmark Report 2025

KEY STATISTICS:

96%

of biopharma companies believe RWD/E is very important or critical to their organizational strategy.

40%

of C-suite leaders consistently discuss RWD/E within their organization.

96%

of organizations will continue to invest in RWD/E over the next two to three years, in part to prepare for agentic AI.

80%

of participants agree or strongly agree that GenAI will have a transformative impact on RWE generation in the next 12 months.

SURVEYED IMPACTS:

56%

Reducing costs and timelines

44%

Increasing efficiency and productivity

SURVEY RESPONSES WERE DIVERSIFIED ACROSS:

52%

Post-market surveillance

44%

Health technology assessments and market access

40%

Payer and provider

40%

Clinical development

THE MOST PROMISING GENAI USE CASES EMERGING FROM THE SURVEY INCLUDE:

76% believe GenAI will accelerate the analysis of RWD

48% anticipate they will use GenAI to automate the summary of real-world findings

44% stated they will use GenAI to support literature summarization

The Warming Climate: Unlocking the De-identified Data

FDA NEWS RELEASE

FDA Eliminates Major Barrier to Using Real-World Evidence in Drug and Device Application Reviews

[More Press Announcements \(\)](#)

For Immediate Release:

December 15, 2025

The U.S. Food and Drug Administration (FDA) today removed a key limitation on the use of real-world evidence (RWE) used in drug and device applications reviews. In new guidance for certain types of medical device submissions, the agency states it will accept RWE without requiring that identifiable individual patient data collected from real-world data sources always be submitted in a marketing submission. The FDA similarly intends to consider updating its guidance for drugs and biologics.

The Promise and Challenge of RWE

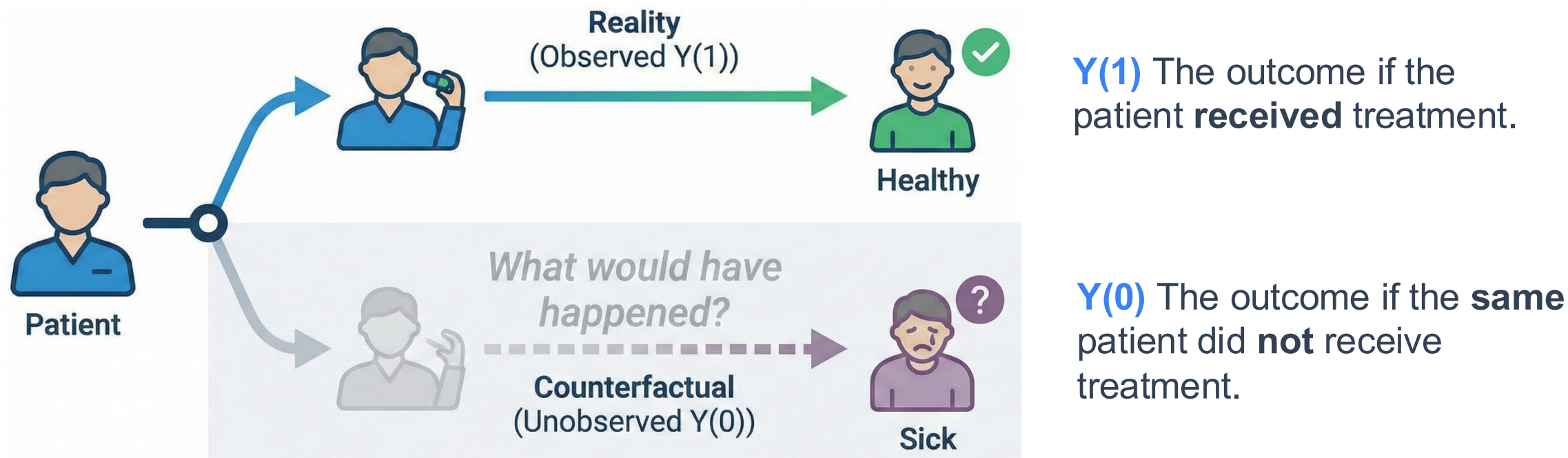


- Widespread adoption
- Market maturity
- Regulatory acceleration
- Investment surge
- Growing regulatory & payer impact
- Unlocking high value with AI



- Uncontrolled confounding
- Selection bias
- Measurement bias
- Missingness
- Etc.

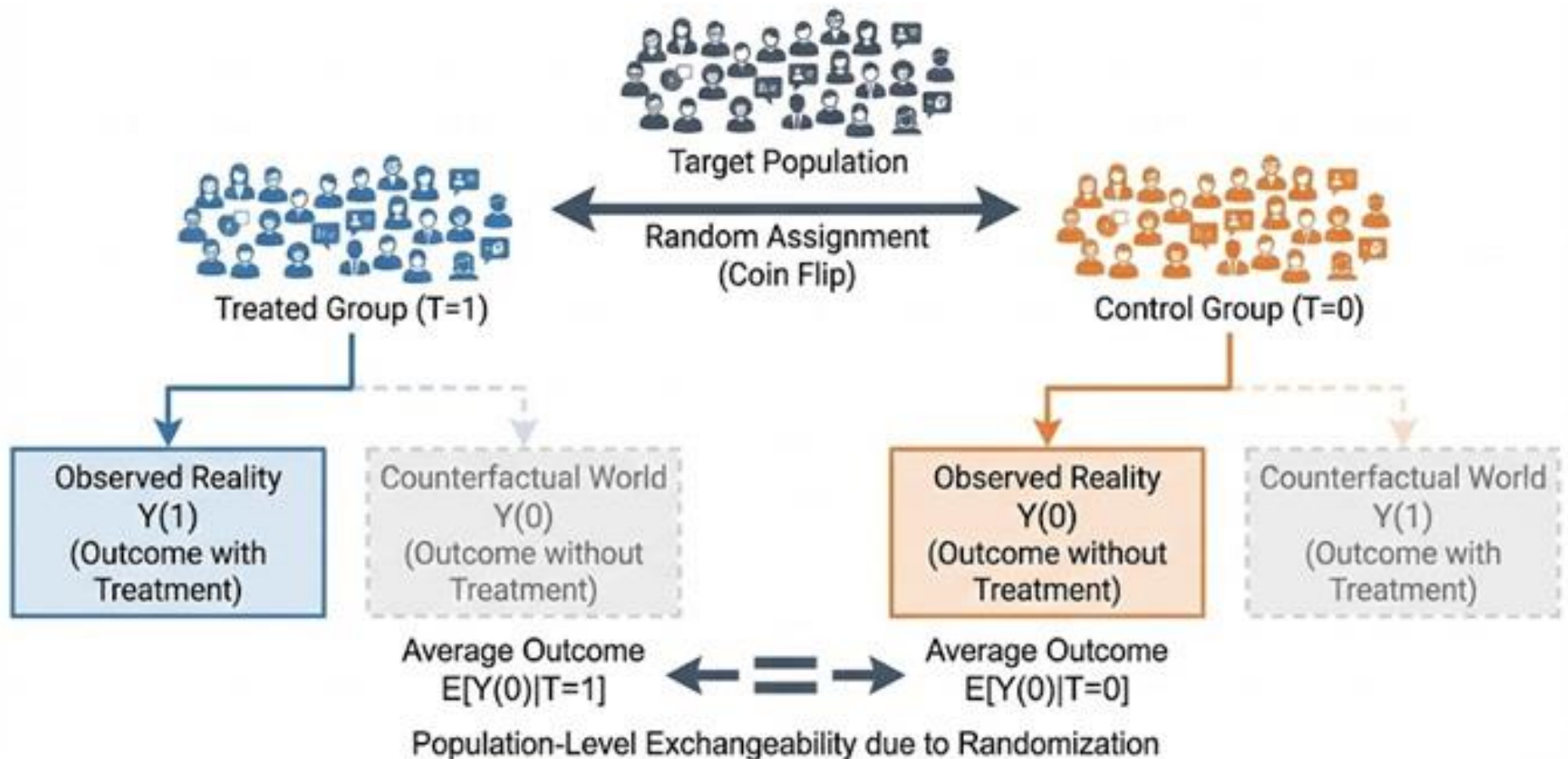
Potential Outcome and the Counterfactual



$$\text{Causal Effect ATE} = E[Y(1) - Y(0)]$$

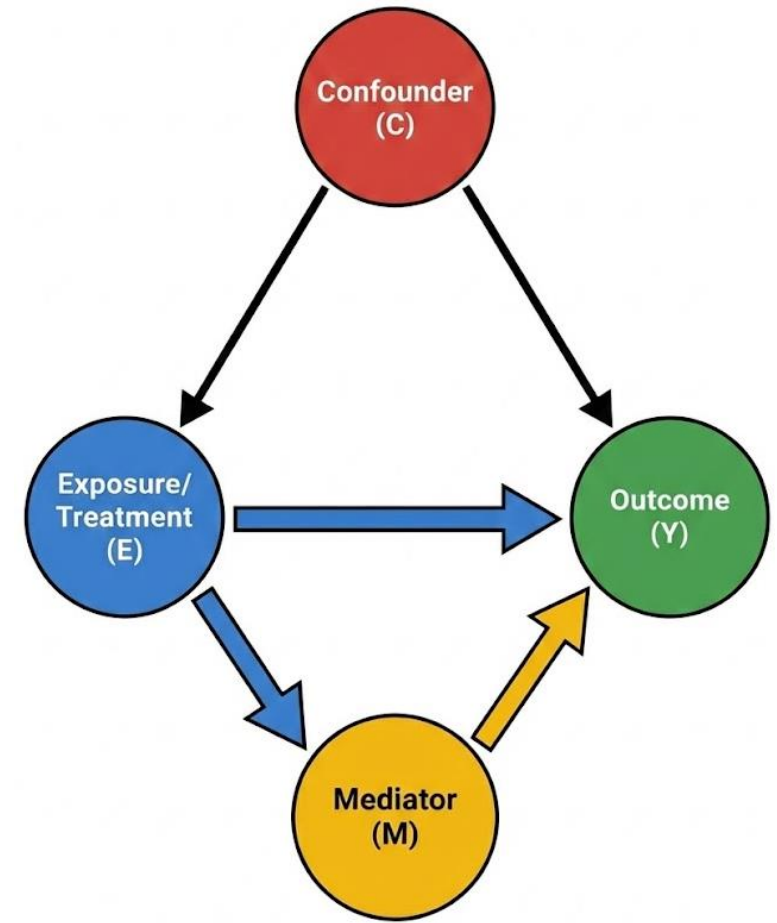
The Fundamental Problem: We want to know the difference between two parallel universes for the same patient, but we only get to observe one outcome per patient.

Randomized Controlled Trial and Target Trial Emulation



Traditional Adjustment Methods in RWE

Method	Brief Methodology	What It's Really Doing
Multiple Regression	Models the outcome using treatment and all confounders as independent variables.	'Comparing Apples to Apples' : Levels the playing field by mathematically 'adjusting' everyone to a common baseline.
Propensity Score Matching (PSM)	Pairs treated patients with untreated patients who have the same probability (propensity score) of receiving treatment.	'Finding a Twin' : For every treated patient, it finds a 'statistical twin' in the untreated group.
Inverse Probability of Treatment Weighting (IPTW)	Applies weights to patients based on their treatment probability to balance groups.	'Synthetic RCT' : Re-weights the population to create a balanced 'pseudo-group' that mimics a randomized trial.



The High Dimensionality in Rare Disease Setting

Large P (Thousands of High-Dimensional Features)

Small N
(N=5
Patients)

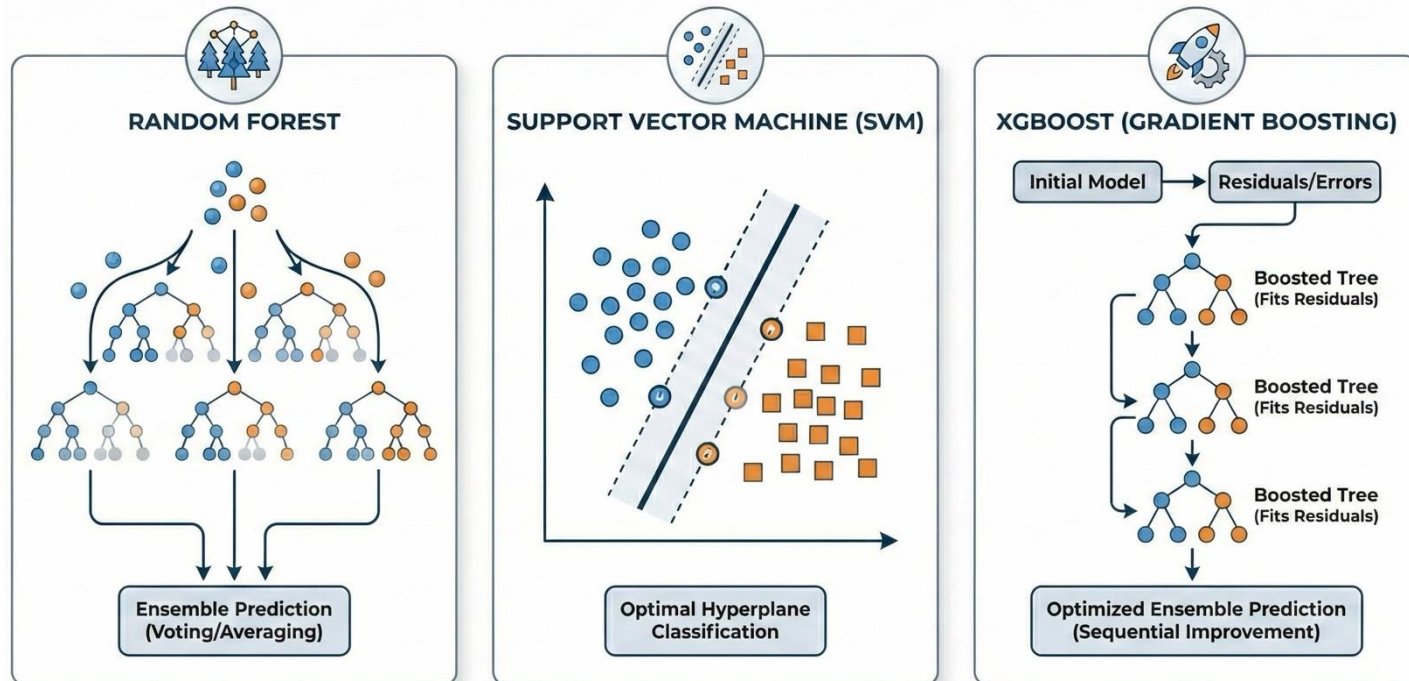
Patient ID	Biomarker A	Gene_X1	Lab_Val_2	Clinical_Score_B		Gene_Y99	Proteomic_Z5	Outcome (Y)
Patient 1	0.45	Positive	120	High	...	...	...	High
Patient 2	0.45	Positive	120	High		...	...	High
Patient 3	0.45	Positive	120	High		...	...	High
Patient 4	0.45	Positive	120	High		...	...	High
Patient 5	...	...	...	High		...	...	High

Hidden complex nonlinearity

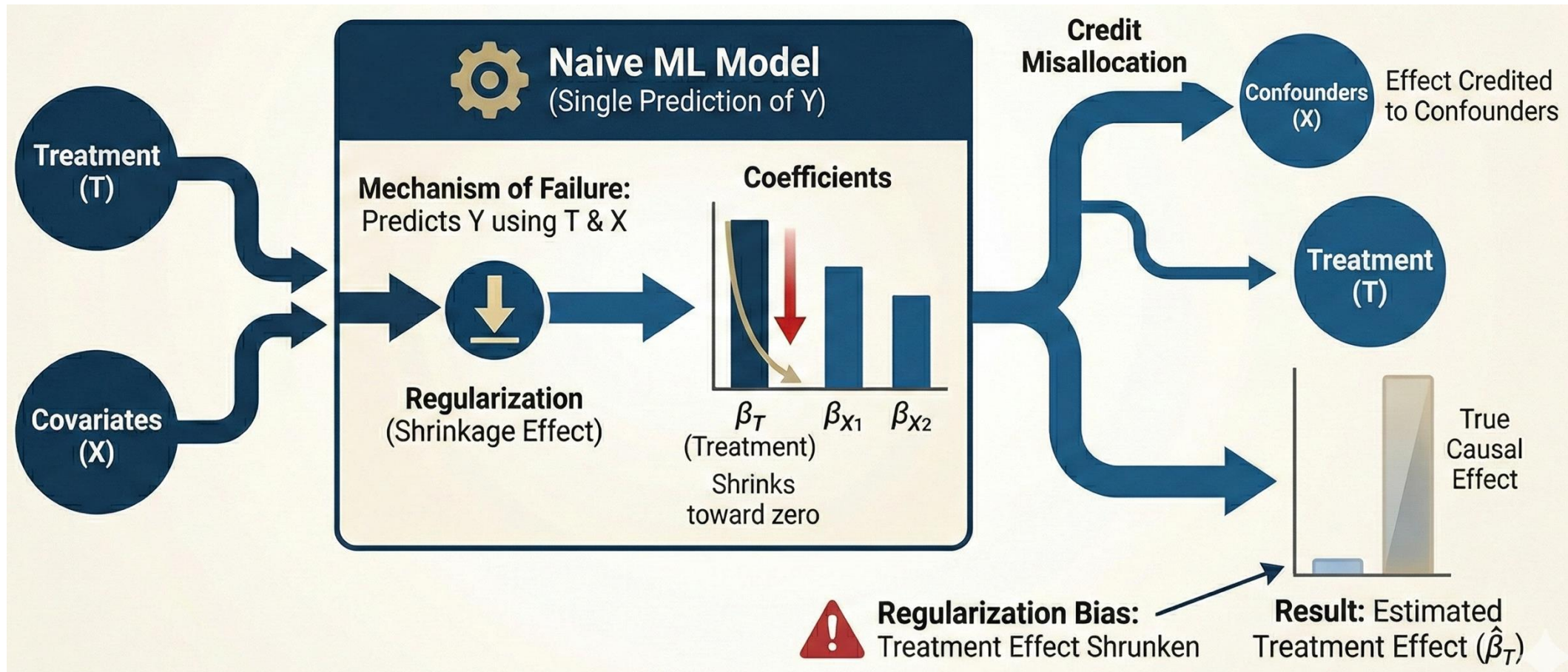
Can we get an *unbiased causal effect estimate* without overfitting noise or relying on fragile parametric assumptions?

Naive Application of ML Models

- Handles high dimensionality and captures nonlinearity
- ML is designed for prediction, not inference. To prevent overfitting, ML models apply regularization. Naive application can lead to regularization bias which can distort or even eliminate causal effect.



Regularization Bias of Naive ML Models



Naive ML models don't know your 'treatment' variable is SPECIAL

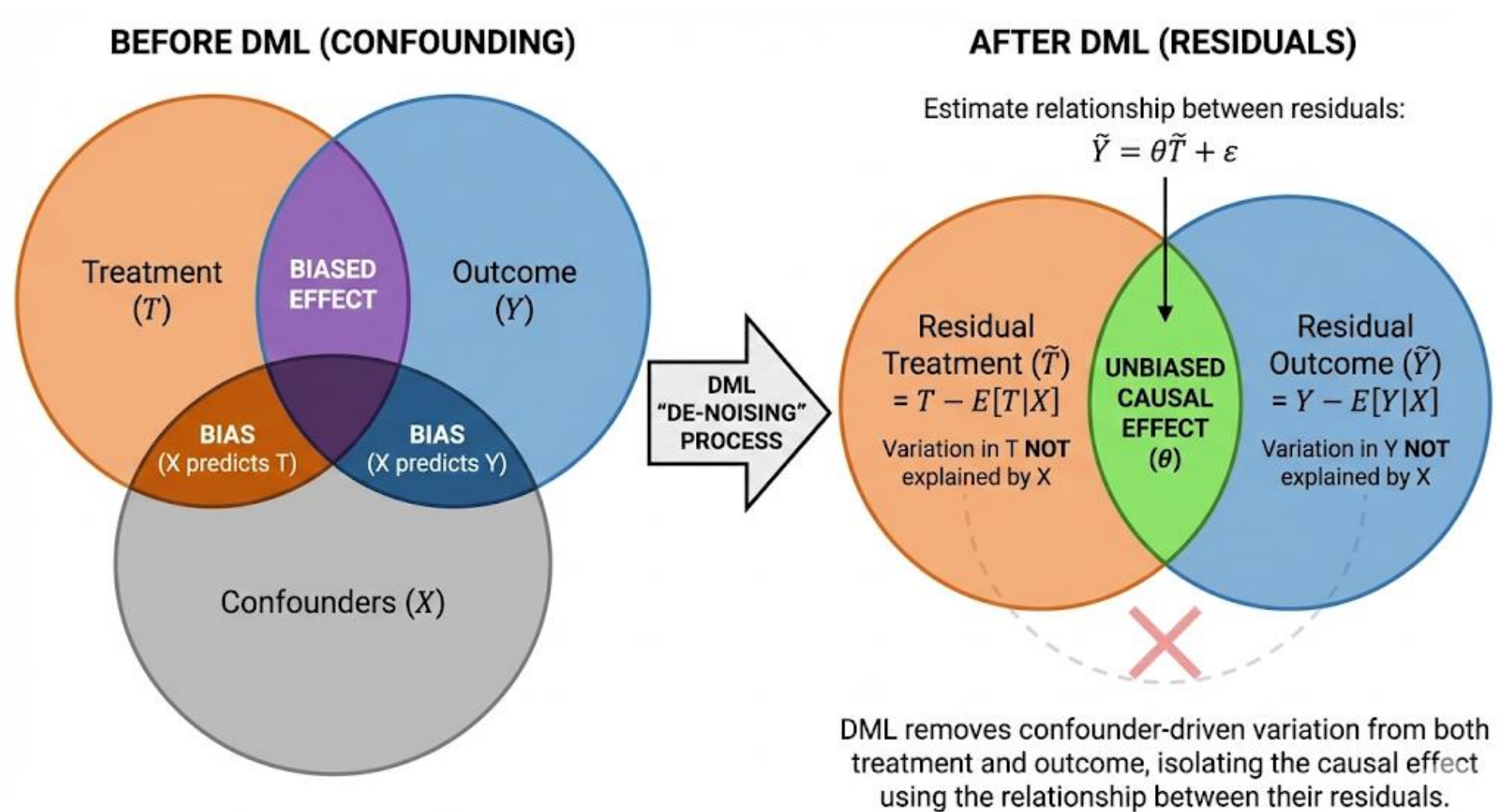


2

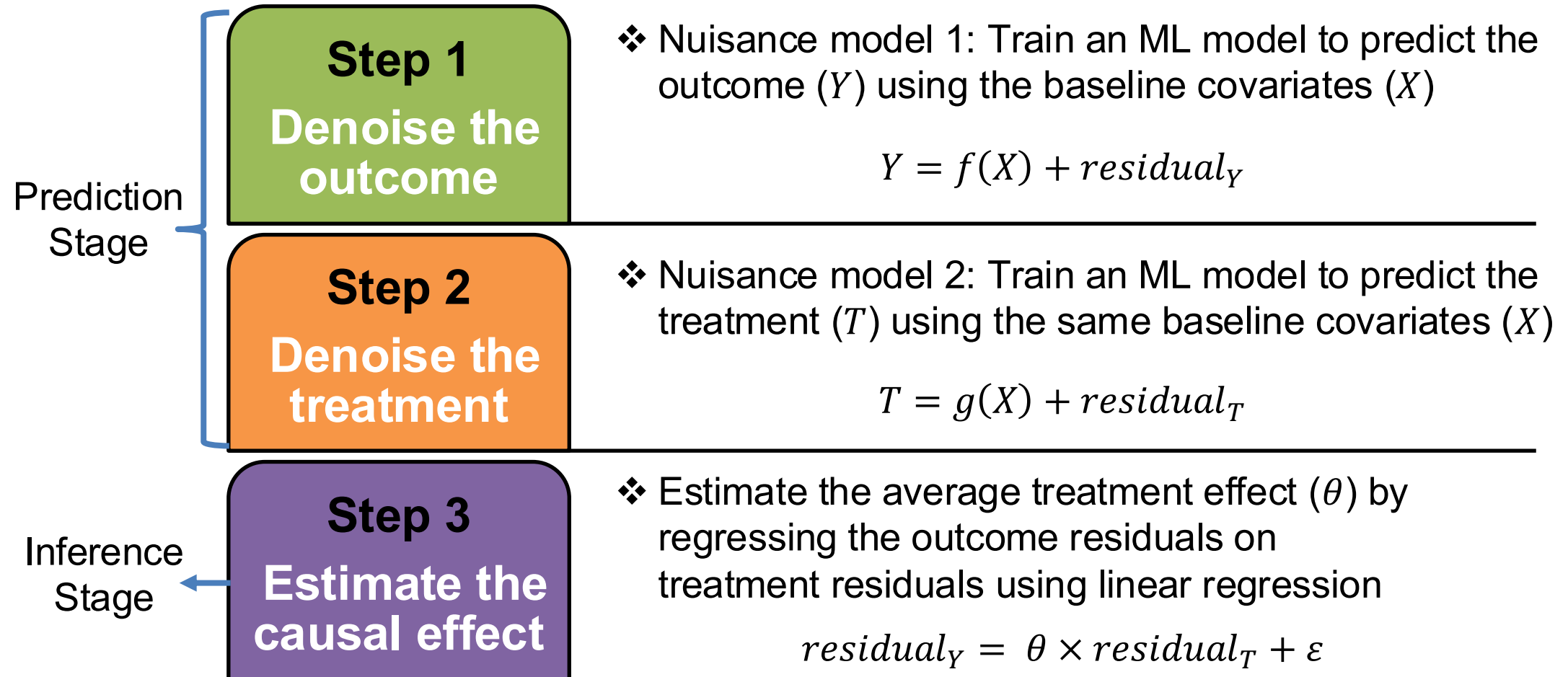
DML

METHODOLOGY

DML – The ‘Denoising’ Operation

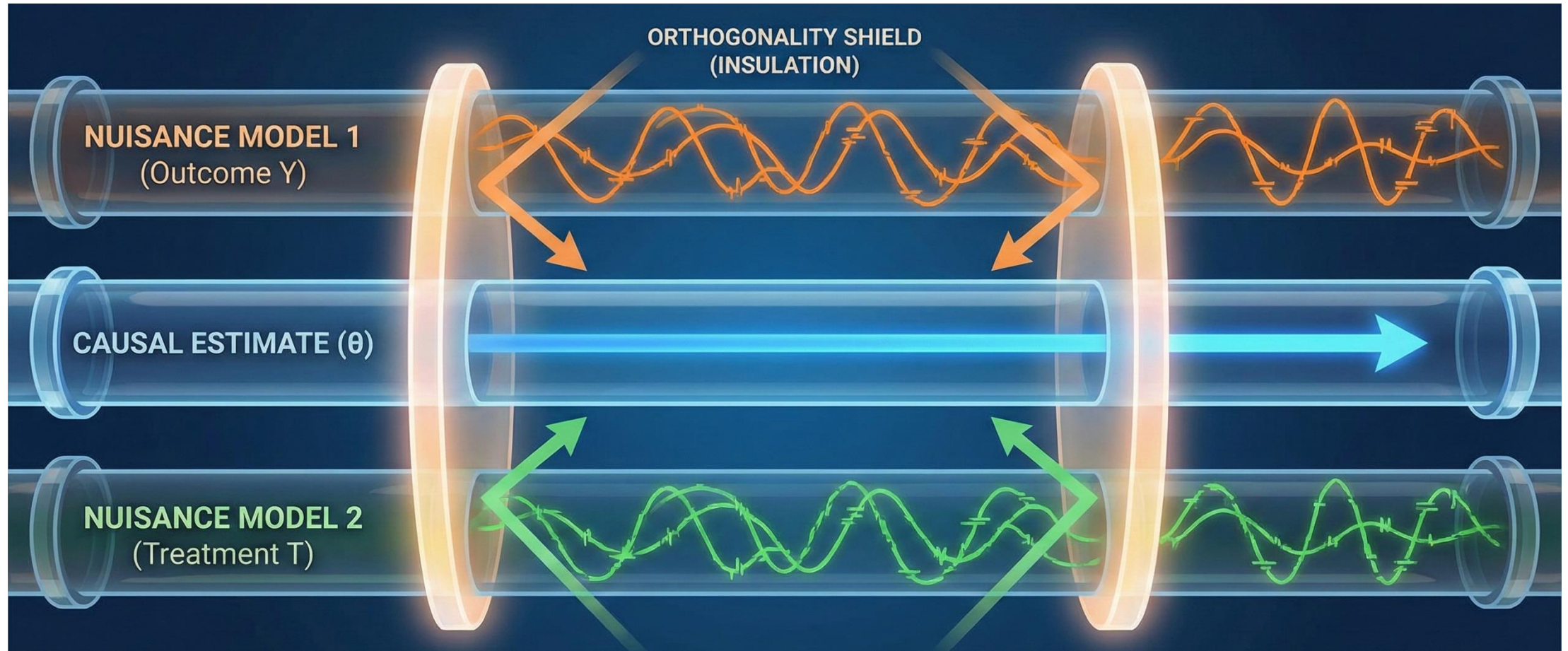


The DML Framework



ε -- The 'clean' noise that is uncorrelated with treatment, after **orthogonalization** from Steps 1&2

Orthogonality in DML – The ‘Insulation’ Shield



How DML Works: Orthogonality and Robustness

❖ Traditional methods: First-order sensitivity

- Bias is **linear** to estimation error.
- These models rely on perfect specification. The estimation error transfers linearly to the final result.

❖ DML: Neyman orthogonality and second-order robustness

- Bias scales with **squared** estimation error.
- DML employs Neyman Orthogonality, a property that immunizes the causal estimator against noise in the nuisance parameters. This allows for valid causal inference even when the underlying ML models are 'noisy' or imperfect.

Practical Advantages: Stability and Flexibility

Both PSM and IPTW relies on the accurate estimation of propensity scores.

- In high dimensionality and nonlinearity, the PS based methods becomes unreliable because the model is only as accurate as the propensity scores estimated by logistic regressions.
- **IPTW:** Leads to massive weights for some influential noise points
- **PSM:** Leads to poor matches (matching on noise variables).
- **DML** avoids these issues by focusing on the **residuals**, rather than the PS or the weights. DML also offers flexible ML modeling approach for nuisance models.

While DML is robust to model misspecification, it is **not a cure for unobserved confounding**. Neither DML nor traditional methods (PSM/IPTW) can account for **unobserved confounders**. DML optimizes adjustment for the observed data, but it cannot remove bias from factors it cannot see.



SIMULATIONS

SETUP | FINDINGS

Simulation Objective

To evaluate the performance of DML vs traditional methods in a real-world setting, we conducted a simulation study designed to mimic a rare disease cohort. This design reflects common real-world scenarios involving both continuous and binary outcomes.

Continuous Outcome

Treatment: Low adherence (relative to high adherence)

Outcome: Total healthcare costs

True Effect: \$1,000 increase annually

Binary Outcome

Treatment: Combo therapy (relative to monotherapy)

Outcome: Adverse event risk

True Effect: ~15% Absolute risk increase

Simulation Design

- **Sample Size (N):** 500 subjects.
- **Covariates (P):** 100 baseline covariates.
- **Sparsity:** Only 5 variables are 'true' confounders affecting both the treatment and the outcome. The remaining 95 are noise variables.
- **Non-linearity:** The 5 'true' confounders affect outcomes via interactions and quadratics
- **Repetitions:** 1,000 Monte Carlo simulation runs for both continuous and binary outcome scenarios.
- **Results presentation:** The estimated effect represents the mean and SD across 1,000 runs, with 95% CI defined by the 2.5th and 97.5th percentiles of the simulation results.

Methods Compared

- **Unadjusted:** Simple regression on treatment alone (ignoring confounders).
- **Regression adjustment:** Multivariable linear/logistic regression including all baseline covariates (full adjustment).
- **IPTW:** Inverse probability weighting using logistic regression for propensity scores.
- **PSM:** 1:1 Nearest Neighbor Matching using logistic regression for propensity scores.
- **DML:** Gradient Boosting for nuisance models with 3-fold cross-fitting.

Implementation Ecosystem

DML is accessible via open-source Python and R packages.

EconML

The package used in this paper was published by Microsoft Research. It's an excellent support for flexible DML models and CATE estimation. This package is only available in Python.

<https://github.com/py-why/EconML>

DoubleML

This package is available in both Python and R. The package provides a direct implementation of the DML framework proposed by Chernozhukov et al. (2018).

<https://github.com/DoubleML>

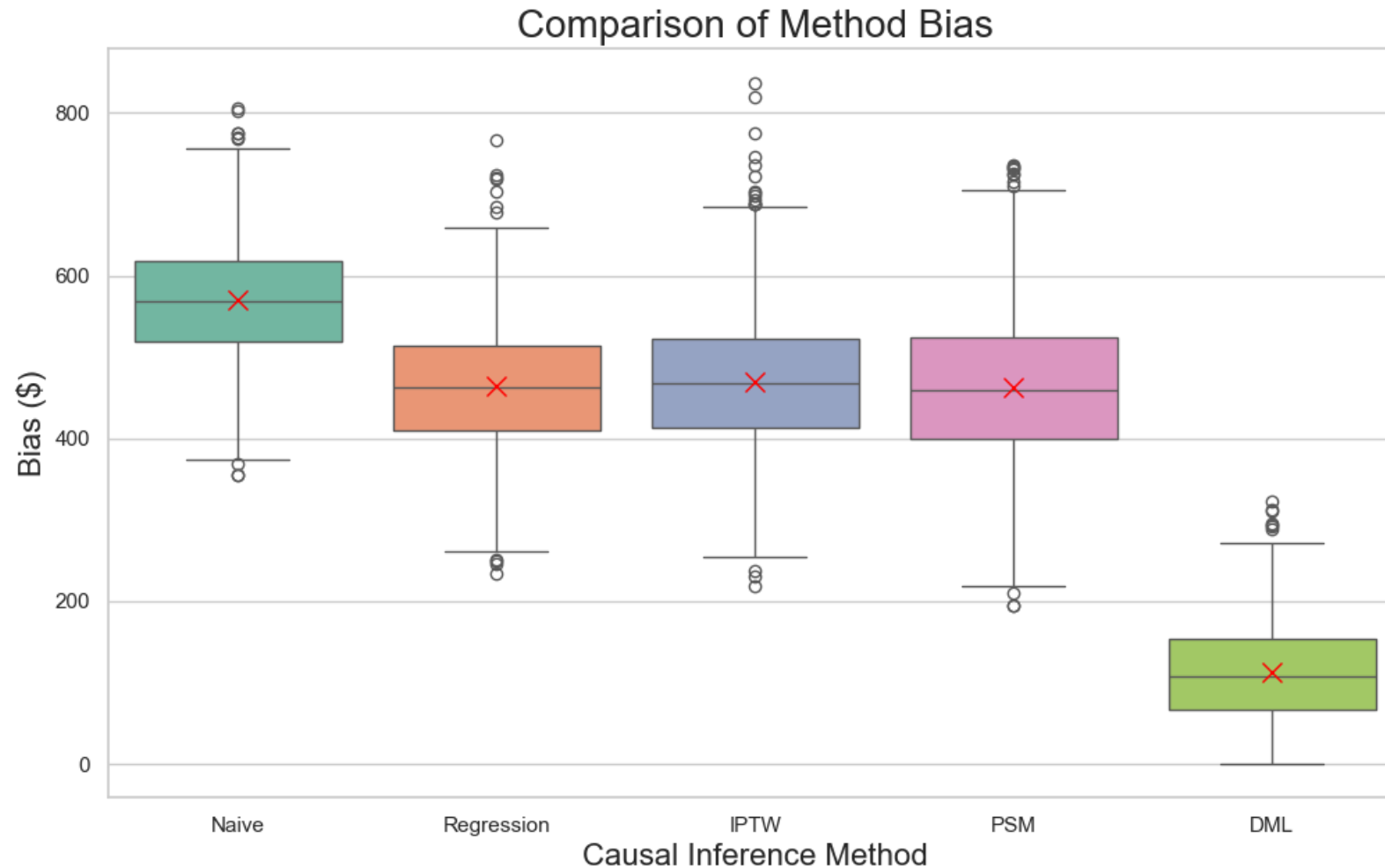
Simulation Results: Continuous Outcome

True Effect: \$1,000

Estimator	Estimated Effect (SD)	Bias (95% CI)
Unadjusted	\$1,569.7 (\$73.44)	\$569.7 (\$429.20, \$709.11)
Regression* adjusted	\$1,463.4 (\$78.34)	\$463.4 (\$309.28, \$618.87)
IPTW	\$1,469.7 (\$86.49)	\$469.7 (\$300.96, \$642.66)
PSM	\$1,463.1 (\$92.09)	\$463.1 (\$256.17, \$692.05)
DML	\$1,109.6 (\$62.84)	\$112.4 (\$6.38, \$241.34)

*Linear regression

Bias Comparison: Continuous Outcome



DML reduced bias by ~**80%** compared to naive linear regression

DML reduced bias by ~**75%** compared to PSM

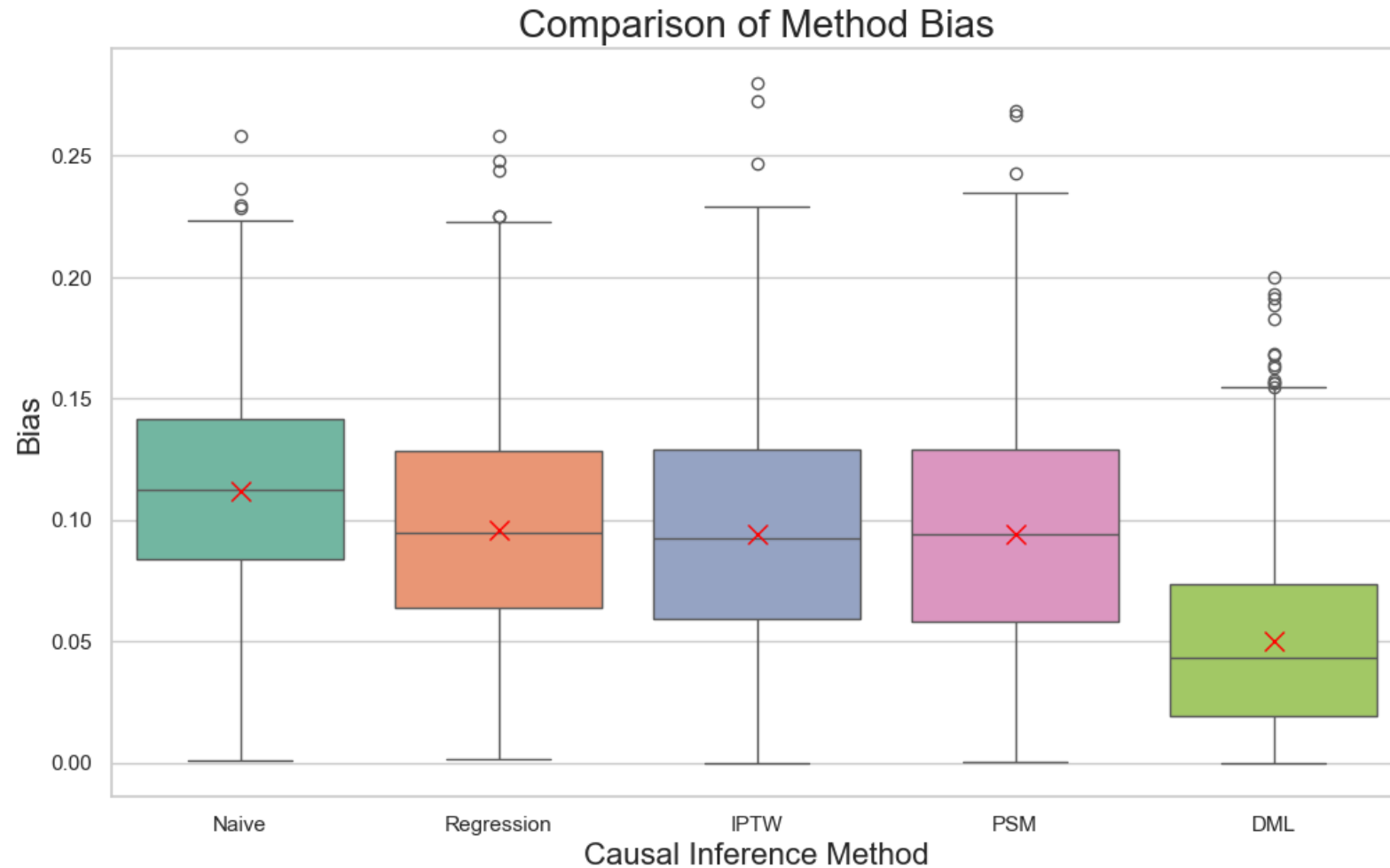
Simulation Results: Binary Outcome (Risk Difference)

True Effect: ~0.15 (15%)

Estimator	Estimated Effect (SD)	Bias (95% CI)
Unadjusted	0.27 (0.041)	0.11 (0.032 - 0.191)
Regression* adjusted	0.25 (0.048)	0.10 (0.010 - 0.192)
IPTW	0.25 (0.052)	0.09 (0.005 - 0.200)
PSM	0.25 (0.054)	0.09 (0.006 - 0.198)
DML	0.19 (0.052)	0.05 (0.002 - 0.137)

*Logistic regression

Bias Comparison: Binary Outcome



DML reduced bias by ~**55%** compared to naive logistic regression

DML reduced bias by ~**45%** compared to PSM



4

DISCUSSIONS

Main Findings

- DML consistently yielded estimates closest to the truth effect, outperforming traditional adjustment methods in both continuous and binary settings. The framework demonstrated a clear advantage in minimizing bias within complex, data-scarce environments.
- In the binary outcome setting, logistic regression can fail to converge in extreme conditions where we specified more variables than patients. The high dimensionality relative to sample size likely led to issues of perfect separation, causing coefficient estimates to become unstable or undefined, highlighting a practical limitation of traditional approaches.
- Most methods consistently overestimated the true effect (positive bias). This reflects the underlying 'positive confounding' structure of the simulations, where specific comorbidities were generated to simultaneously increase the likelihood of treatment and the magnitude of adverse outcomes (higher costs/higher risk).

Comparative Assessment of Causal Estimators

Linear regression adjustment

- **High-dimensional failure:** Prone to severe overfitting when the number of predictors is large relative to sample size
- **Linearity assumption:** Fails to capture complex relationships, such as interactions or non-linear structures, unless explicitly manually specified.

Logistic regression adjustment

- **Performance:** Can be robust in low-dimensional settings, especially when penalized (e.g., Ridge/Lasso) to handle collinearity
- **Selection challenge:** Struggles to automatically distinguish true confounders from noise variables without advanced regularization, often leading to omitted variable bias.

Comparative Assessment of Causal Estimators

IPTW

- **Weight instability:** Highly sensitive to extreme weights (propensity scores near 0 or 1), which can inflate variance and destabilize estimates.
- **Misspecification risk:** Standard propensity models (logistic) often fail to capture complex non-linear confounding, leading to residual bias.

PSM

- **Dependency on overlap:** Effective only when excellent balance is achievable; poor PS overlap results in data loss.
- **Estimand drift:** Discarding unmatched subjects alters the target population.
- **Misspecification risk:** Like IPTW, relies on correct specification of the propensity model and struggles with unknown non-linearities.

Comparative Assessment of Causal Estimators

DML

- **Computational cost:** Requires significantly higher computational resources and training time due to cross-fitting and fine-tuning steps.
- **Non-linear robustness:** Effectively captures complex, high-dimensional, and non-linear confounding structures without manual specification.
- **Robust inference:** Provides mathematically consistent estimates and robust inference, even when the underlying ML models are imperfect.
- **Specification requirement:** Requires clear definition of the treatment and outcome variables; it is a causal estimation tool, not a causal discovery tool.

Model Selection Strategy: DML vs PS Methods

Feature	IPTW / PSM	DML
Primary Goal	Covariate balance across groups.	Bias reduction in high-dimensional and nonlinear settings.
Model Focus	Mostly treatment assignment.	Both treatment and outcome models.
Complexity	Low; easy to explain to clinicians.	High; requires careful explanation.
Best Use Case	Low-dimensional, well-understood data.	High-dimensional data with complex interactions.

Acknowledgements and Contact

Communication author:

Jing Dai (Jazz Pharma and Wayne State University)

Email: jdai@jazzpharma.com

Co-authors:

Yilun Huang (Novartis; former Jazz intern)

Sarah Markt (Jazz Pharma)

Adeniyi Togun (Jazz Pharma)