Augmented Targeted Trial Emulation

Huiyuan Wang Ph.D.

University of Pennsylvania

JSM 2025, Nashville

August 4, 2025

Joint work with

Jingyue Huang (Postdoc, Upenn)

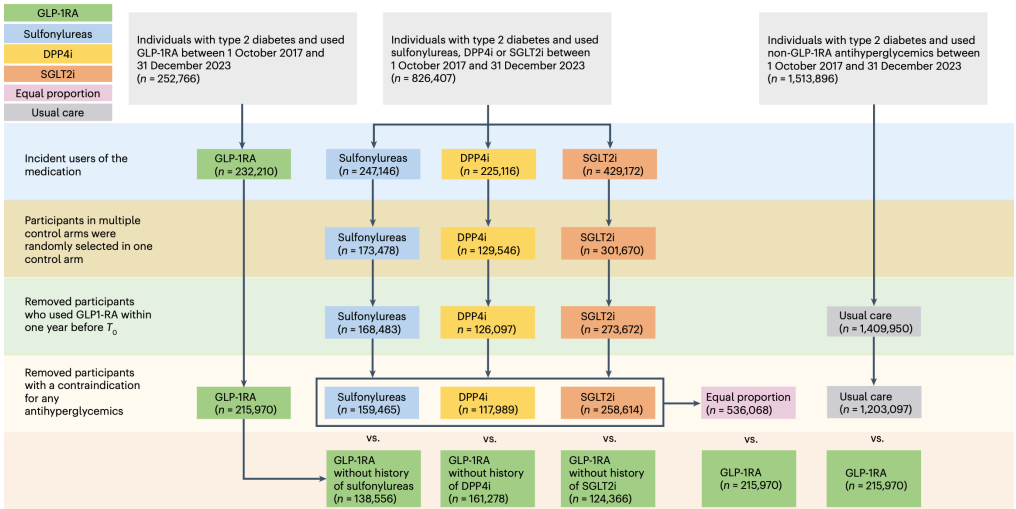
Yong Chen (Professor, Upenn)

Background

- Drawing clinical evidence from real-world data follows a structured process, typically consisting of the following steps (Hernán et al. 2025)
  - Step 1 Formulate the clinical question and define the estimand
  - Step 2 Identify the target population and select the study sample based on eligibility criteria
  - Step 3 Perform statistical analysis to estimate the parameter of interest
  - Step 4 Interpret and report the results

Hernán, Miguel A., Issa J. Dahabreh, Barbra A. Dickerman, and Sonja A. Swanson. "The Target Trial Framework for Causal Inference From Observational Data: Why and When Is It Helpful?." *Annals of Internal Medicine* (2025).

Example Adepted from Xie et al. (2025)



Xie, Yan, Taeyoung Choi, and Ziyad Al-Aly. "Mapping the effectiveness and risks of GLP-1 receptor agonists." *Nature Medicine* (2025): 1–12.

Potential Problems in Step 2

- Step 2: Identify the target population and selecting the study sample based on eligibility criteria
  - Applying eligibility criteria can reduce the sample size by more than 30%;
    e.g., 124,366/252,766≈49.2% in the above example (SGLT2i vs. GLP-1RA)
  - Efficiency matters, and a smaller sample size may lead to wider confidence intervals
  - The width of a 95% confidence interval is approximately $2 \times 1.96 \frac{\widehat{\mathsf{Var}}(\varphi)}{\sqrt{n}}$
- Question: *Can we use the excluded samples to obtain a more efficient estimator?*

Notation

- A: The treatment assignment; e.g., $A = 1$ for GLP-1RA and $A = 0$ for usual care
- $Y = AY(1) + (1 - A)Y(0)$ with Y denoting the observed outcome of interest and $Y(a)$ the potential outcome for $A = a$
- $X = (X^{(1)}, X^{(2)}, X^{(3)})$: baseline covariates, which is categorized by three subtypes:
  - $X^{(1)} \in \mathbb{R}^{p_1}$ includes covariates that do not directly affect the eligibility
  - $X^{(2)} \in \mathbb{R}^{p_2}$ affects only eligibility but not treatment or outcome
  - $X^{(3)} \in \mathbb{R}^{p_3}$ affects both eligibility and serves as a confounder

Notation

- Z: The eligible indicator variable; e.g., $Z = 1$ indicates that the following conditions are satisfied:
  - GLP-1RA, if used, would be used for the first time during 1 October 2017 and 31 December 2023 $X^{(2)}$
  - The individual was diagonosed with type-2 diabetes $X^{(3)}$
  - The individual was safe to take any blood sugar–lowering medications $X^{(3)}$
- The target estimand is $\theta = \mathbb{E}\{Y(1) - Y(0) \mid Z = 1\}$, i.e., the average treatment effect of GLP-1RA against the usual care on the eligible subpopulation

Illustration of the Data Structure

Subject	Y	A	X	Z	
1	Y_1	A_1	X_1	$Z_1 = 1$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	Eligible Sample ($\mathcal{E}$)
n	Y_n	A_n	X_n	$Z_n = 1$	
$n+1$	Y_{n+1}	A_{n+1}	X_{n+1}	$Z_{n+1} = 0$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	Ineligible Sample ($\mathcal{I}$)
N	Y_N	A_N	X_N	$Z_N = 0$	

Inverse Propensity Score Weighting Estimator of θ

1. **Propensity score estimation:** We fit a logistic regression model for the probability of treatment given observed covariates,

$$\mathbb{P}(A = 1 \mid X) = \text{expit}\left(\alpha_0 + \alpha_1^\top X^{(1)} + \alpha_3^\top X^{(3)}\right),$$

using only individuals in the eligible cohort

- The resultant estimator is denoted by $(\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\alpha}_3)$
- The estimated propensity scores $\widehat{e}(X) = \text{expit}(\widehat{\alpha}_0 + \widehat{\alpha}_1^\top X^{(1)} + \widehat{\alpha}_3^\top X^{(3)})$ are obtained for each individual

2. **Weight calculation:** Compute inverse probability weights

$$w_i = \frac{A_i}{\widehat{e}(X_i)} + \frac{1 - A_i}{1 - \widehat{e}(X_i)}$$

for each individual $i : Z_i = 1$

Inverse Propensity Score Weighting Estimator of θ

3. **Weighted outcome regression:** We fit a weighted linear regression model

$$Y = c + \theta A + \epsilon,$$

using w_i as sampling weights, and the resultant coefficient $\widehat{\theta}$ is the estimated average treatment effect

Note: Under some conditions, the IPW estimator is regular and asymptotic linear in the sense that

$$\widehat{\theta} - \theta = n^{-1} \sum_{i=1}^{n} \psi(A_i, X_i, Y_i) + o_p\big(n^{-1/2}\big),$$

where ψ denotes its influence function.

4. **Variance estimation and inference:** If a consistent estimator of ψ, denoted by $\widehat{\psi}$, can be derived, then the asymptotic variance can be estimated by

$$\widehat{\mathsf{Var}}(\widehat{\theta}) = n^{-1} \sum_{i=1}^{n} \big\{\widehat{\psi}(A_i, X_i, Y_i)\big\}^2$$

How to Construct More Efficient Estimator?

- The asymptotic variance of $\widehat{\theta}$ is the variance of ψ
- If we can construct an influence function of a smaller variance, we can obtain a more efficient estimator

Motivation: Control-variate Method for Variance Reduction

- Consider estimating the integral $\int_\Omega h(X)\, dP(X)$ using Monte Carlo with n i.i.d. samples $X_1, \ldots, X_n \sim P$:
  - The standard Monte Carlo estimator is $n^{-1} \sum_{i=1}^n h(X_i)$, with variance $\mathsf{Var}(\widehat{\mu}_h) = n^{-1}\mathsf{Var}_P\{h(X)\}$
  - Suppose we have a control function $f(X)$ with known mean $\mu = \mathbb{E}_P\{f(X)\}$
  - Then the control variate estimator (Nelson 1990) is

$$\widehat{\mu}_{\mathsf{cv}} = n^{-1} \sum_{i=1}^n \Big[h(X_i) - \mathsf{Cov}_P(h, f)\big\{\mathsf{Var}_P(f)\big\}^{-1}\big\{f(X_i) - \mu\big\}\Big],$$

  which has reduced variance:

$$\mathsf{Var}(\widehat{\mu}_{\mathsf{cv}}) = n^{-1}\Big[\mathsf{Var}_P(h) - \mathsf{Cov}_P(h, f)\big\{\mathsf{Var}_P(f)\big\}^{-1}\mathsf{Cov}_P(f, h)\Big]$$

Nelson, B. L. (1990). Control variate remedies. *Operations Research*, 38(6):974–992.

Motivation: Control-variate Method for Variance Reduction

- In the case of parameter estimation:
  - If we would know the mean $\mu = \mathbb{E}\{f(X, A, Y) \mid Z = 1\}$ for some function f, then the estimator

$$\widehat{\theta} - \mathsf{Cov}_P(\psi, f)\{\mathsf{Var}(f)\}^{-1}\left\{n^{-1}\sum_{i=1}^{n} f(A_i, X_i, Y_i) - \mu\right\}$$

    would have the asymptotic expansion

$$\theta + n^{-1}\sum_{i=1}^{n}\left[\psi(A_i, X_i, Y_i) - \mathsf{Cov}_P(\psi, f \mid Z = 1)\{\mathsf{Var}(f \mid Z = 1)\}^{-1}\left\{f(A_i, X_i, Y_i) - \mu\right\}\right] + o_p(n^{-1/2})$$

    with a reduced asymptotic variance

- However, we usually do not know the expectation $\mathbb{E}(f \mid Z = 1)$

Augmented Target Trial Emulation

- In target trial emulation, Z is solely determined by $X^{(2)}, X^{(3)}$, and thus
$$X^{(1)} \perp\!\!\!\perp Z \mid X^{(2)}, X^{(3)}$$
- Suppose that $\gamma \in \mathbb{R}^d$ can correctly specify the distribution $P(X^{(1)} \mid X^{(2)}, X^{(3)})$, i.e.,
$$P(X^{(1)} \mid X^{(2)}, X^{(3)}) = P(X^{(1)} \mid X^{(2)}, X^{(3)}; \gamma = \gamma_0) \quad \text{for some } \gamma_0 \in \mathbb{R}^d$$
- Let $\widehat{\gamma}_1$ denote the MLE of γ_0 derived using the eligible sample
- Let $\widehat{\gamma}_0$ denote the MLE of γ_0 derived using the ineligible sample

Augmented Target Trial Emulation

- $X^{(1)} \perp\!\!\!\perp Z \mid X^{(2)}, X^{(3)}$, together with some regularity conditions, implies that

$$\widehat{\gamma}_1 = \gamma_1 + n^{-1} \sum_{i=1}^{n} \phi_1(X_i) + o_p(n^{-1/2})$$

$$\widehat{\gamma}_0 = \gamma_1 + (N-n)^{-1} \sum_{i=n+1}^{N} \phi_0(X_i) + o_p(N^{-1/2})$$

- Thus, $\widehat{\gamma}_1 - \widehat{\gamma}_0$ can serve as a proxy of control-variate, with

$$f = \phi_1 \quad \text{and} \quad \mu = (N-n)^{-1} \sum_{i=n+1}^{N} \phi_0(X_i)$$

- The augmented estimator is

$$\widehat{\theta}^{\mathsf{Aug}} = \widehat{\theta} - \beta(\widehat{\gamma}_1 - \widehat{\gamma}_0),$$

where β is a weighting coefficient to be determined such that the asymptotic variance of $\widehat{\theta}^{\mathsf{Aug}}$ is minimized

Augmented Target Trial Emulation

- The asymptotic variance of $\widehat{\theta}^{\mathsf{Aug}}$ consists of
  - The variance of $\widehat{\theta}$
  - The variance of $\widehat{\gamma}_1$ and the covariance between $\widehat{\theta}$ and $\widehat{\gamma}_1$
  - The variance of $\widehat{\gamma}_0$
- Proposition. The optimal β is

$$\beta^* = \mathsf{Cov}(\psi, \phi_1 \mid Z = 1)\{\mathsf{Var}(\phi_1 \mid Z = 1) + \rho \mathsf{Var}(\phi_0 \mid Z = 0)\}^{-1},$$

where $\rho = \lim n/(N - n)$

Augmented Target Trial Emulation

- Theorem. If $\widehat{\beta}$ is a consistent estimator for β^*, then under some regularity conditions,

$$\sqrt{n}\big(\widehat{\theta}_1^{\mathsf{Aug}} - \theta\big) \to_d N(0, \Sigma^{\mathsf{Aug}}),$$

and $\Sigma^{\mathsf{Aug}} = \mathsf{Var}(\psi \mid Z = 1) - \beta^{*\top}\big\{\mathsf{Var}(\phi_1 \mid Z = 1) + \rho\,\mathsf{Var}(\phi_0 \mid Z = 0)\big\}\beta^*$ which can be consistently estimated by

$$\widehat{\mathsf{Var}}(\widehat{\psi} \mid Z = 1) - \widehat{\beta}^{\top}\big\{\widehat{\mathsf{Var}}(\widehat{\phi}_1 \mid Z = 1) + \rho\,\widehat{\mathsf{Var}}(\phi_0 \mid Z = 0)\big\}\widehat{\beta}$$

Some Nice Properties

- *"Do-no-harm" property:* The augmented estimator is at least as efficient as the eligibility-restricted estimator
- *Procedure-agnostic property:* The proposed augmented estimation can improve the efficiency of a wide spectrum of estimation procedures, e.g., propensity score matching, stratification, weighting, or doubly robust, as long as the resultant estimator is regular and asymptotic linear
- *Modular property:* The proposed augmented estimation process can be seamlessly integrated into conventional target trial emulation pipeline

Data Generation Overview

1. **Covariates:**
   - $X_i^{(2)} \sim \mathcal{N}(0, I_{p_2}), \quad X_i^{(3)} \sim \mathcal{N}(0, I_{p_3})$
   - $X_i^{(1)} = \Gamma_2 X_i^{(2)} + \Gamma_3 X_i^{(3)} + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma_X^2 I_{p_1})$
2. **Treatment Assignment:**

$$\mathbb{P}(A_i = 1 \mid X_i) = \text{expit}\left(\alpha_0 + \alpha_1^\top X_i^{(1)} + \alpha_3^\top X_i^{(3)}\right)$$

3. **Potential Outcomes:**

$$Y_i(0) = \beta_0 + \beta_1^\top X_i^{(1)} + \beta_3^\top X_i^{(3)} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$
$$Y_i(1) = Y_i(0) + \beta_A + \beta_M^\top X_i^{(1)}, \quad Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$$

4. **Eligibility Rule:** Use Section 1 rules to compute Z_i, then determine eligibility ($Z_i = 1$ or 0).

Data Storage and Summary

- Two subsets stored:
  - Eligible: $\mathcal{E} = \{i : Z_i = 1\}$, size n
  - Ineligible: $\mathcal{I} = \{i : Z_i = 0\}$, size $N - n$

- Parameter settings:
  - $N = 10{,}000$, $p_1 = 3$, $p_2 = 2$, $p_3 = 1$, $q = 3$
  - $\gamma_2 = 0.3$, $\gamma_3 = 0.3$, $\sigma_X^2 = 1$
  - $\alpha_0 = -1$, $\alpha_1 = 0.2$, $\alpha_3 = -0.3$
  - $\beta_0 = 0$, $\beta_1 = 1$, $\beta_3 = 1$, $\beta_A = 1$, $\beta_M = 3$
  - $\sigma^2 = 1$, Threshold $c = 0$

Simulation Results: Table Summary

Table 1: Performance of the baseline and proposed estimators with RB: Relative Bias; AV: Average Variance; CP: Coverage Probability; AL: Average CI Length

Estimator	RB ($\times 100$)	Var ($\times 100$)	AV ($\times 100$)	CP (%)	AL ($\times 100$)
Baseline	0.12	3.58	3.60	95.4	74.30
Proposed	0.54	2.27	2.21	94.2	58.21

The variance of the proposed estimator is 63.4% of that of the baseline estimator

Take-away Message

- The proposed method has three key advantages:
  - *Do-no-harm*
  - *Procedure-agnostic*
  - *Modular*
- **Limitation:** The method relies on correct model specification
- Future work may explore robustness checks or model relaxation strategies

Thanks!

Any Questions?