Robust Estimation and Inference in Hybrid Controlled Trials

Ke Zhu

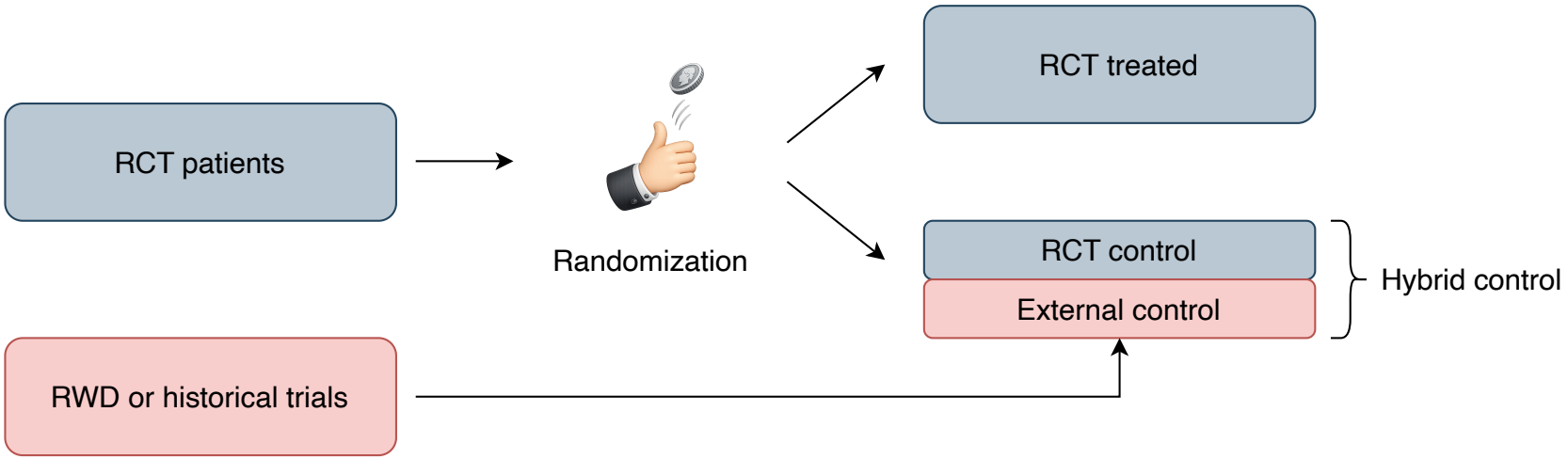
Department of Statistics, North Carolina State University
Department of Biostatistics and Bioinformatics, Duke University

Joint work with Shu Yang (NCSU) and Xiaofei Wang (Duke)

Presented at IMSI Workshop New Horizons on Model Transportability and Data Integration
June 23, 2026

Disclaimer
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, funded by the FDA/HHS. This work is also supported by R01AG066883, funded by the NIH/HHS. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS, NIH/HHS, or the U.S.

Hybrid Controlled Trials

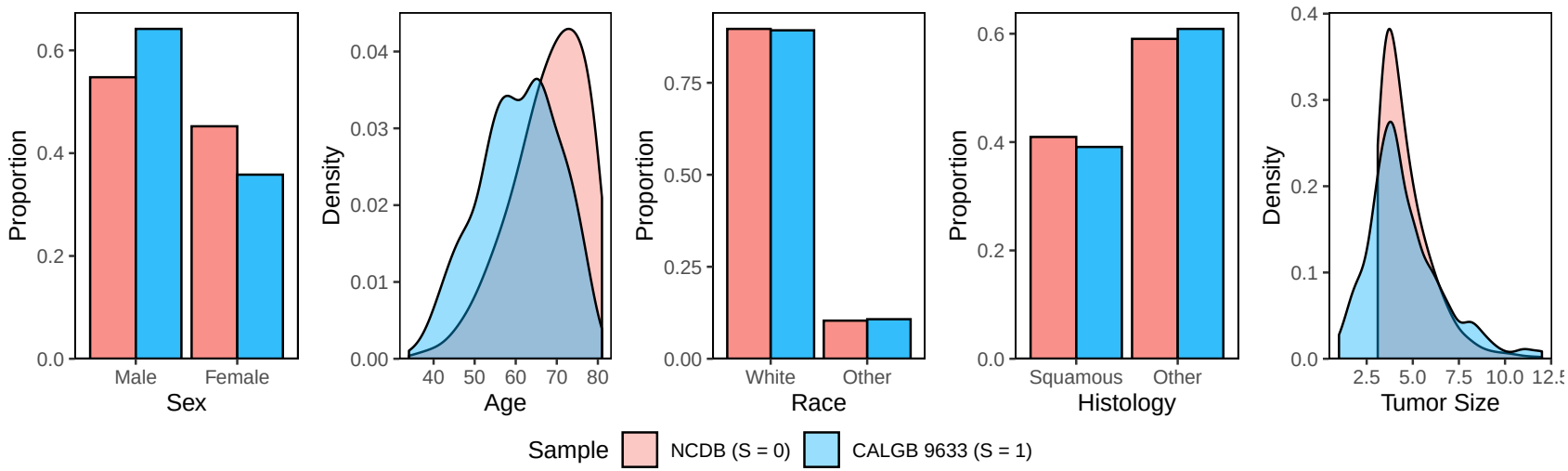


(i) Borrowing ECs can improve efficiency while targeting the **same estimand** as the RCT.

(ii) RCT controls help address both **covariate shift** and **outcome drift**.

(iii) Randomization enables **randomization inference**.

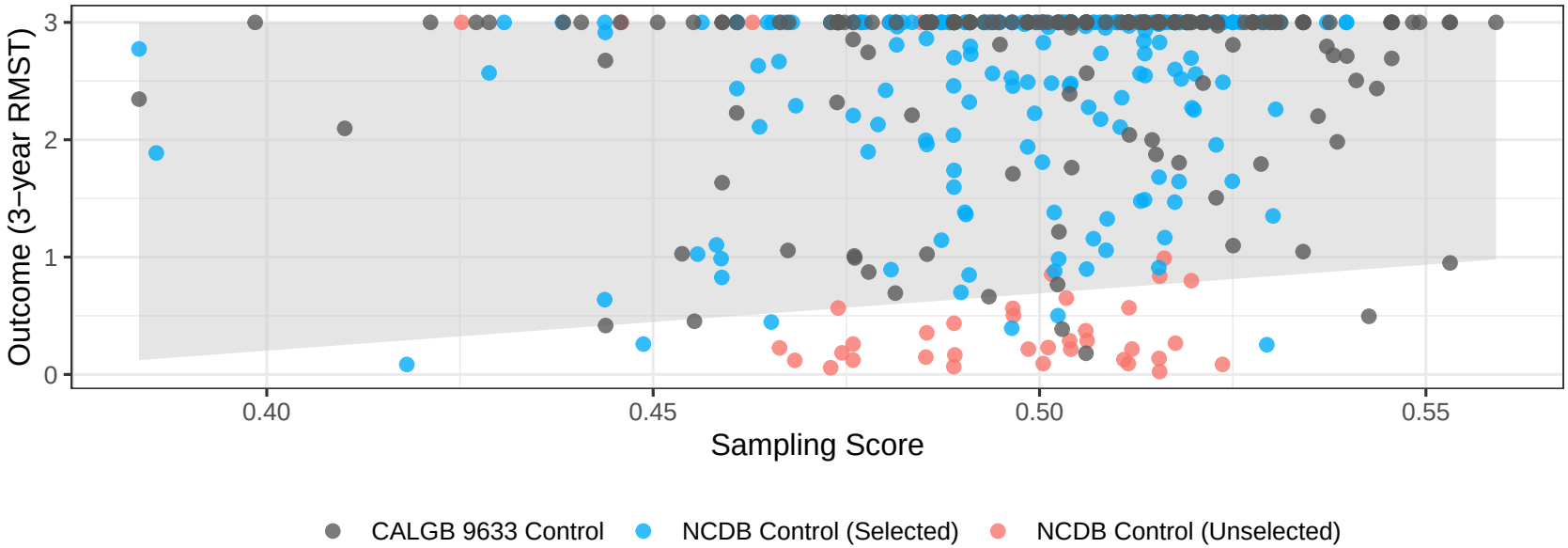
Addressing Covariate Shift



Full Borrowing AIPW

- Doubly robust and semiparametrically efficient (Li et al., 2023).
- Biased w/o Conditional Mean Exchangeability Assumption.

Addressing Outcome Drift



Conformal Selective Borrowing (CSB) AIPW

- Assess exchangeability for each EC using **conformal *p*-values** (Vovk et al., 2005): general, flexible, and model-agnostic.
- Borrow a subset of ECs using a **targeted-tuned threshold** to minimize the MSE of the AIPW estimator (Rothenhäusler, 2024).

Two Targets of Inference

ATE in the RCT population is zero, $\tau = 0$

- Wald inference with sandwich variance estimators
- Asymptotic normality, double robustness, selection consistency

No individual treatment effect in RCT sample, $Y_i(0) = Y_i(1), \forall i \in \mathcal{R}$

- Fisher Randomization Test (FRT)
- Finite-sample exact, model-free, and post-selection valid

These two inferential targets play complementary roles in different clinical and regulatory settings (Berger, 2000; Carter et al., 2024).

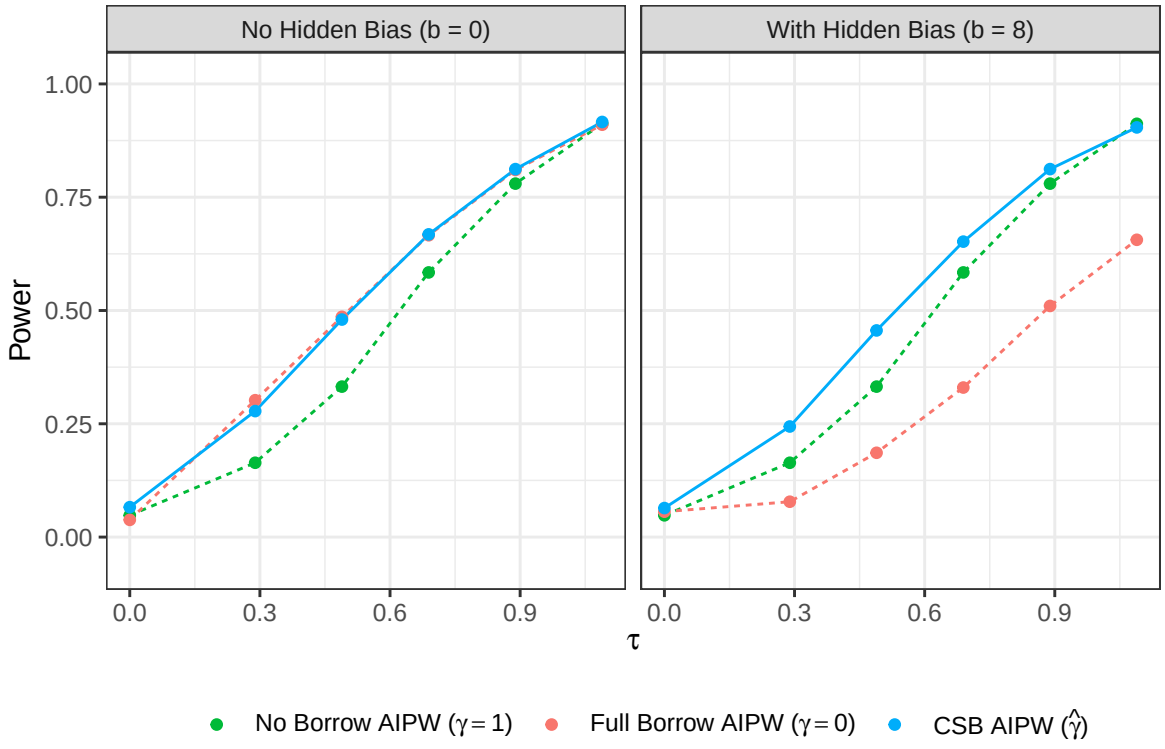
Fisher Randomization Tests in Hybrid Controlled Trials

Fisher Randomization Test (Fisher, 1935)

1. **Test Statistic**: Compute $T(\boldsymbol{A}^{\text{obs}})$ for actual assignment $\boldsymbol{A}^{\text{obs}}$.
2. **Analyze as you randomize**:
   - Generate A_i^b for RCT patients per the actual randomization procedure.
   - Keep $A_i^b \equiv 0$ for ECs, as they remain fixed during randomization in RCT.
3. **Reconstruct the entire HCT analysis** under $\boldsymbol{A}^b$, including EC selection (if applicable) and estimation, and compute the corresponding test statistic $T(\boldsymbol{A}^b)$.
4. **Compute p value**: Repeat for B iterations and compute:

$$\hat{p}^{\text{FRT}} = \frac{\sum_{b=1}^{B} \mathbb{I}\{T(\boldsymbol{A}^b) \geq T(\boldsymbol{A}^{\text{obs}})\} + 1}{B+1}.$$

Simulation: Power Curves of FRTs



- **FRTs** control Type I error rate ($\tau = 0$) for any test statistic.
- **FB-AIPW** loses power when outcome drift is present.
- **CSB-AIPW** achieves the highest power.

Papers and Software

Ke Zhu, Shu Yang, and Xiaofei Wang (2025). Enhancing Statistical Validity and Power in Hybrid Controlled Trials: A Randomization Inference Approach with Conformal Selective Borrowing. *ICML*.

Case Study: Jiajun Liu, Ke Zhu, Shu Yang, and Xiaofei Wang (2026). Robust Estimation and Inference in Hybrid Controlled Trials for Binary Outcomes: A Case Study on Non-Small Cell Lung Cancer. *arXiv:2505.00217*.

Tutorial: Ke Zhu, Hairong Huang, Shu Yang, and Xiaofei Wang (2026). Robust Estimation and Inference with Selective Borrowing in Hybrid Controlled Trials: A Tutorial with SelectiveIntegrative and intFRT.

R package: *intFRT* available at *github.com/IntegrativeStats/intFRT*.

Review: Ke Zhu, Rima Izem, Peng Yang, Ying Yuan, Herbert Pang, Mark van der Laan, Lei Nie, Birol Emir, Pallavi Mishra-Kalyani, Hana Lee, and Shu Yang (2026). Externally Controlled Trials: A Review of Design and Borrowing Through a Causal Lens. *arXiv:2605.03282*.

Transfer Learning Through Conditional Quantile Matching

Yikun Zhang[1]

Joint work with *Steven Wilkins-Reeves*[2], *Wesley Lee*[2], and *Aude Hofleitner*[2]

[1]Department of Statistics, University of Washington
[2]Central Applied Science, Meta

New Horizons on Model Transportability and Data Integration
June 23, 2026

UNIVERSITY of WASHINGTON

Central Applied Science

Meta

Business Question: How do Meta's family of apps perform compared to an app developed by a competing company?

Business Question: How do Meta's family of apps perform compared to an app developed by a competing company?

- **Metric:** Country-level monthly active user (MAU) statistics for an app.
- **Limitation:** Complete MAU data for Meta's apps but limited MAU data for the target app.

Business Question: How do Meta's family of apps perform compared to an app developed by a competing company?

- **Metric:** Country-level monthly active user (MAU) statistics for an app.
- **Limitation:** Complete MAU data for Meta's apps but limited MAU data for the target app.

Objective: Predict the country-level MAU of a target app using data from Meta's family of apps.

Regression Task: Predict the country-level MAU $Y^{(0)} \in \mathcal{Y}$ for a target app from country-specific covariates $X^{(0)} \in \mathcal{X}$.

1. **Source-Domain Data:** $\mathcal{D}_S^{(k)} = \left\{ \left(X_i^{(k)}, Y_i^{(k)} \right) \right\}_{i=1}^{n_k} \sim P^{(k)}$ for $k = 1, \ldots, K$.
2. **Target-Domain Data:** $\mathcal{D}_T = \left\{ \left(X_i^{(0)}, Y_i^{(0)} \right) \right\}_{i=1}^{n_0} \sim P^{(0)}$, where $n_0 \ll n_k$ for $k = 1, \ldots, K$.

Regression Task: Predict the country-level MAU $Y^{(0)} \in \mathcal{Y}$ for a target app from country-specific covariates $X^{(0)} \in \mathcal{X}$.

1. **Source-Domain Data:** $\mathcal{D}_S^{(k)} = \left\{ \left(X_i^{(k)}, Y_i^{(k)} \right) \right\}_{i=1}^{n_k} \sim P^{(k)}$ for $k = 1, \ldots, K$.
2. **Target-Domain Data:** $\mathcal{D}_T = \left\{ \left(X_i^{(0)}, Y_i^{(0)} \right) \right\}_{i=1}^{n_0} \sim P^{(0)}$, where $n_0 \ll n_k$ for $k = 1, \ldots, K$.

Covariate Shift (Shimodaira, 2000):
$P^{(k)}(Y|X) = P^{(0)}(Y|X)$ but $P^{(k)}(X) \neq P^{(0)}(X)$.

Label Shift (Saerens et al., 2002):
$P^{(k)}(X|Y) = P^{(0)}(X|Y)$ but $P^{(k)}(Y) \neq P^{(0)}(Y)$.

General Target Shift:

$$P^{(k)}(X) \neq P^{(0)}(X) \quad \text{and} \quad P^{(k)}(Y|X) \neq P^{(0)}(Y|X) \quad \text{for } k = 1, \ldots, K.$$

Proposed Transfer Learning Framework (TLCQM)

Key Idea: Jointly calibrate source distributions for data augmentation in the target domain (Zhang et al., 2026).

Key Idea: Jointly calibrate source distributions for data augmentation in the target domain (Zhang et al., 2026).

1 Learn source distributions $\widehat{P}^{(k)}(\cdot|X)$ for each $k = 1, \ldots, K$. $\longrightarrow$

1. Learn a generative model $\widehat{P}^{(k)}(\cdot|x)$ using $\left\{ \left(X_i^{(k)}, Y_i^{(k)} \right) \right\}_{i=1}^{n_k}$ for each source domain k.

Key Idea: Jointly calibrate source distributions for data augmentation in the target domain (Zhang et al., 2026).

1 Learn source distributions $\widehat{P}^{(k)}(\cdot|X)$ for each $k = 1, \ldots, K$. $\longrightarrow$

1. Learn a generative model $\widehat{P}^{(k)}(\cdot|x)$ using $\left\{ \left(X_i^{(k)}, Y_i^{(k)} \right) \right\}_{i=1}^{n_k}$ for each source domain k.
   - Model $P^{(k)}(\cdot|x) = g^{(k)}(x, \cdot)_{\#} P_\eta$ and estimate $\widehat{g}^{(k)}$ via engression (Shen and Meinshausen, 2025):

$$\widehat{g}^{(k)} \in \underset{g \in \mathcal{G}}{\arg\min} \frac{1}{n_k} \sum_{i=1}^{n_k} \left[\frac{1}{m} \sum_{j=1}^{m} \left| Y_i^{(k)} - g(X_i^{(k)}, \eta_{ij}) \right| - \frac{1}{2m(m-1)} \sum_{j=1}^{m} \sum_{j'=1}^{m} \left| g(X_i^{(k)}, \eta_{ij}) - g(X_i^{(k)}, \eta_{ij'}) \right| \right],$$

   where P_η is a prespecified noise distribution and $\mathcal{G}$ is a neural network class.

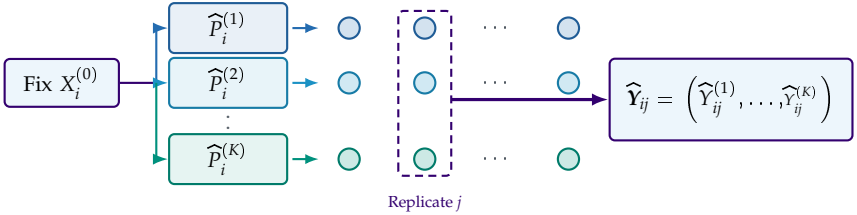
1 Learn source distributions
$\widehat{P}^{(k)}(\cdot|X)$ for each $k = 1, \ldots, K$.

→

2 Generate at target covariates
$\widehat{Y}_{ij}^{(k)} \sim \widehat{P}^{(k)}(\cdot|X_i^{(0)})$ for $j = 1, \ldots, M$.

→

2. For each target covariate $X_i^{(0)}$, independently draw M synthetic response vectors:



$$\left\{\left(X_i^{(0)}, \widehat{\mathbf{Y}}_{ij}\right) : i = 1, \ldots, n_0,\ j = 1, \ldots, M\right\}, \quad \widehat{\mathbf{Y}}_{ij} = \left(\widehat{Y}_{ij}^{(1)}, \ldots, \widehat{Y}_{ij}^{(K)}\right), \quad \widehat{Y}_{ij}^{(k)} \sim \widehat{P}^{(k)}\left(\cdot|X_i^{(0)}\right).$$

1 Learn source distributions
$\widehat{P}^{(k)}(\cdot|X)$ for each $k = 1, \ldots, K$.

→

2 Generate at target covariates
$\widehat{Y}_{ij}^{(k)} \sim \widehat{P}^{(k)}(\cdot|X_i^{(0)})$ for $j = 1, \ldots, M$.

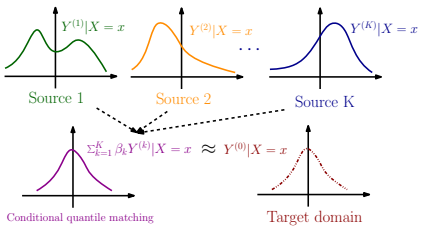
→

3 Conditional quantile calibration
$\widehat{\beta}_0 + \sum_{k=1}^{K} \widehat{\beta}_k \widehat{Y}^{(k)}$ matches the target distribution.

3. Compute the quantile matching estimator (Sgouropoulos et al., 2015):

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^{K+1}}{\arg\min} \sum_{i=1}^{n_0} \sum_{j=1}^{M} \left[Y_{(i)}^{(0)} - \left(\boldsymbol{\beta}^T \widehat{\boldsymbol{V}} \right)_{((i-1)M+j)} \right]^2.$$

- $\widehat{\boldsymbol{V}}_{ij} = \left(1, \widehat{\boldsymbol{Y}}_{ij}\right)^T$, while $Y_{(1)}^{(0)} \leq \cdots \leq Y_{(n_0)}^{(0)}$ and $\left(\boldsymbol{\beta}^T \widehat{\boldsymbol{V}} \right)_{(1)} \leq \cdots \leq \left(\boldsymbol{\beta}^T \widehat{\boldsymbol{V}} \right)_{(n_0 M)}$ are order statistics, respectively.



Proposed Transfer Learning Framework (TLCQM)

1 Learn source distributions
$\widehat{P}^{(k)}(\cdot|X)$ for each $k = 1, \ldots, K$.

→

2 Generate at target covariates
$\widehat{Y}_{ij}^{(k)} \sim \widehat{P}^{(k)}(\cdot|X_i^{(0)})$ for $j = 1, \ldots, M$.

→

3 Conditional quantile calibration
$\widehat{\beta}_0 + \sum_{k=1}^{K} \widehat{\beta}_k \widehat{Y}^{(k)}$ matches the target distribution.

→

4 Augment and fit on $\mathcal{D}_T \cup \mathcal{D}_A$
XGBoost, kernel ridge regression (KRR), neural network (NN),...

4. Augment the target domain with synthetic, target-calibrated labels:

$$\mathcal{D}_A = \bigcup_{k=1}^{K} \left\{ \left(X_i^{(k)}, \widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{V}}_i^{(k)} \right) \right\}_{i=1}^{n_k} \quad \text{with} \quad \widehat{\boldsymbol{V}}_i^{(k)} = \left(1, \widehat{Y}_i^{(1,k)}, \ldots, \widehat{Y}_i^{(K,k)}\right)^T \in \mathbb{R}^{K+1},$$

where $\widehat{Y}_i^{(j,k)}$ is a prediction of $Y^{(j)}$ at $X_i^{(k)}$, such as $\widehat{\mathbb{E}}(Y^{(j)}|X_i^{(k)})$.

5. (Optional) Estimate $w_k(x) = \frac{dP_X^{(0)}(x)}{dP_X^{(k)}(x)}$ for $k = 1, \ldots, K$ to correct for covariate shift.

Theory: Excess Risk Decomposition

- $R(f) := \mathbb{E}_{P^{(0)}}\left[\ell\left(Y^{(0)}, f(X^{(0)})\right)\right]$ and $f^{(0)} \in \arg\min_{f\in\mathcal{F}} R(f)$ under a loss function ℓ.
- $\mathrm{Rad}_n(\mathcal{F})$ is the Rademacher complexity of $\mathcal{F}$.

Target-Only: $\widehat{f}^{(0)} = \arg\min_{f\in\mathcal{F}} \frac{1}{n_0}\sum_{i=1}^{n_0} \ell\left(Y_i^{(0)}, f(X_i^{(0)})\right)$ satisfies

$$R(\widehat{f}^{(0)}) - R(f^{(0)}) \lesssim \mathrm{Rad}_{n_0}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{n_0}} \quad \text{w. p. } \geq 1-\delta.$$

- $R(f) := \mathbb{E}_{P^{(0)}}\left[\ell\left(Y^{(0)}, f(X^{(0)})\right)\right]$ and $f^{(0)} \in \arg\min_{f\in\mathcal{F}} R(f)$ under a loss function ℓ.
- $\mathrm{Rad}_n(\mathcal{F})$ is the Rademacher complexity of $\mathcal{F}$.

Target-Only: $\widehat{f}^{(0)} = \arg\min_{f\in\mathcal{F}} \frac{1}{n_0}\sum_{i=1}^{n_0} \ell\left(Y_i^{(0)}, f(X_i^{(0)})\right)$ satisfies

$$R(\widehat{f}^{(0)}) - R(f^{(0)}) \lesssim \mathrm{Rad}_{n_0}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{n_0}} \quad \text{w. p. } \geq 1-\delta.$$

TLCQM: $\widehat{f}^{(0,tl)} = \arg\min_{f\in\mathcal{F}} \frac{1}{N}\left\{\sum_{i=1}^{n_0} \ell\left(Y_i^{(0)}, f(X_i^{(0)})\right) + \sum_{k=1}^{K}\sum_{i=1}^{n_k} \widehat{w}_k(X_i^{(k)}) \cdot \ell\left(\widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{V}}_i^{(k)}, f(X_i^{(k)})\right)\right\}$ satisfies

$$R(\widehat{f}^{(0,tl)}) - R(f^{(0)}) \lesssim \mathrm{Rad}_N(\mathcal{F}) + \sqrt{\frac{K\log(1/\delta)}{N}} + \frac{1}{N}\sum_{k=1}^{K} \|\widehat{w}_k - w_k\|_1 + \inf_{\boldsymbol{\beta}_*\in\mathcal{B}} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\|_1$$
$$+ \sum_{k=1}^{K} \|\widehat{g}^{(k)} - g^{(k)}\|_\infty + \inf_{\boldsymbol{\beta}_*\in\mathcal{B}} \left(\int_0^1 \left[Q_{Y^{(0)}}(\alpha) - Q_{\boldsymbol{\beta}_*^T \boldsymbol{V}}(\alpha)\right]^2 d\alpha\right)^{1/2} \quad \text{w. p. } \geq 1-\delta,$$

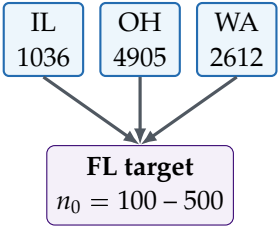
where $N = \sum_{k=0}^{K} n_k$ and $\mathcal{B} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{K+1}} \int_0^1 \left[Q_{Y^{(0)}}(\alpha) - Q_{\boldsymbol{\beta}^T \boldsymbol{V}}(\alpha)\right]^2 d\alpha$.

$$R(\widehat{f}^{(0,tl)}) - R(f^{(0)}) \lesssim \underbrace{\mathrm{Rad}_N(\mathcal{F}) + \sqrt{\frac{K\log(1/\delta)}{N}}}_{\text{Standard generalization error}} + \underbrace{\frac{1}{N}\sum_{k=1}^{K}\|\widehat{w}_k - w_k\|_1}_{\text{Importance weight error}} + \underbrace{\sum_{k=1}^{K}\|\widehat{g}^{(k)} - g^{(k)}\|_\infty}_{\text{Distributional learning error}}$$

$$+ \underbrace{\inf_{\boldsymbol{\beta}_* \in \mathcal{B}} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\|_1}_{\text{Quantile matching error}} + \underbrace{\inf_{\boldsymbol{\beta}_* \in \mathcal{B}} \left(\int_0^1 [Q_{Y^{(0)}}(\alpha) - Q_{\boldsymbol{\beta}_*^T \boldsymbol{V}}(\alpha)]^2 \, d\alpha\right)^{1/2}}_{\text{Transfer bias}} \quad \text{w. p. } \geq 1-\delta.$$

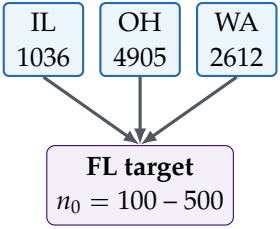
1. $N = \sum_{k=0}^{K} n_k \gg n_0$ reduces the generalization/complexity error.
2. $\inf_{\boldsymbol{\beta}_* \in \mathcal{B}} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\|_1 = O\left(\sqrt{K\sum_{k=1}^{K}\|\widehat{g}^{(k)} - g^{(k)}\|_\infty}\right) + O_P\left(\sqrt{\frac{K\log\log n_0}{n_0}} + \sqrt{K}\left[\frac{\log\log n_0}{n_0} \cdot \inf_{\boldsymbol{\beta}_* \in \mathcal{B}} \int_0^1 \{Q_{Y^{(0)}}(\alpha) - Q_{\boldsymbol{\beta}_*^T \boldsymbol{V}}(\alpha)\}^2 \, d\alpha\right]^{1/4}\right)$.
3. Transfer bias vanishes when $P^{(0)}(Y|X)$ lies in the convex hull of $\{P^{(k)}(Y|X)\}_{k=1}^{K}$.

Goal: Predict apartment rental prices in FL using data from IL, OH, and WA.



500 Monte Carlo replications.

Goal: Predict apartment rental prices in FL using data from IL, OH, and WA.

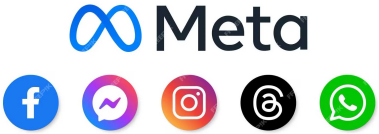


500 Monte Carlo replications.

MSE reduction versus target-only training

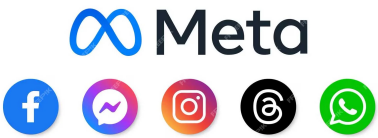
n_0	100	200	300	500
XGBoost	7.2%	1.6%	−2.3%	−6.8%
KRR	**94.2%**	**91.4%**	**88.4%**	**82.5%**
NN	**99.4%**	**99.4%**	**96.9%**	**88.0%**

Result: MSE reductions are larger for more flexible learners (NN and KRR) and generally decrease as n_0 increases.



Objective: Predict country-level MAU for a held-out app at Meta.

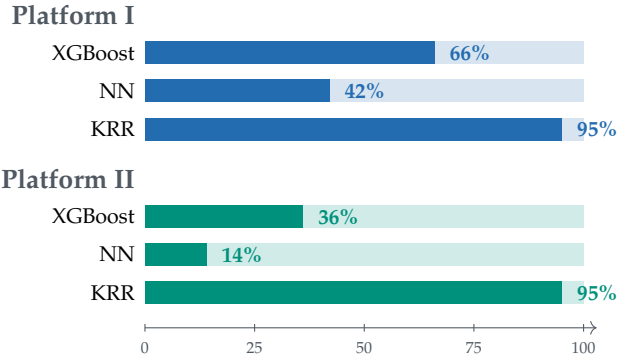
- Four source apps with ≈ 230 observations per app across two device platforms.



Objective: Predict country-level MAU for a held-out app at Meta.

- Four source apps with ≈ 230 observations per app across two device platforms.

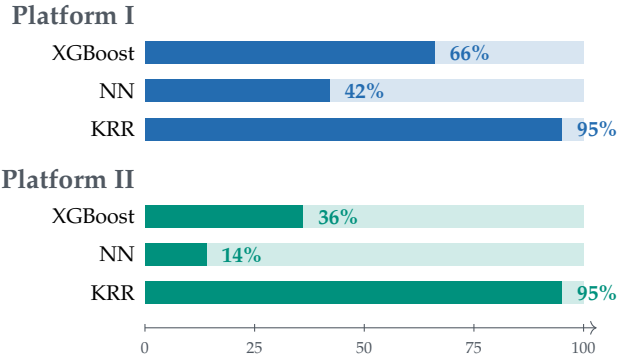
MSE reduction from target-only to TLCQM



Objective: Predict country-level MAU for a held-out app at Meta.

- Four source apps with ≈ 230 observations per app across two device platforms.

MSE reduction from target-only to TLCQM



Average test MSE (TLCQM standard errors in parentheses)

	TLCQM			TKRR	CDAR	DARC
	XGBoost	NN	KRR			
Platform I	**1.545 (0.198)**	**1.689 (0.146)**	**1.556 (0.207)**	2.204	3.183	36.380
Platform II	**1.364 (0.034)**	**1.841 (0.015)**	**1.377 (0.034)**	2.348	1.867	10.884

Main Takeaways of TLCQM

1. **Generate distributions, not point predictions,** from each source at target covariates.

2. **Match target quantiles** to correct discrepancies between $P^{(0)}(Y|X)$ and $P^{(k)}(Y|X)$, $k = 1, \ldots, K$.

3. **Train a standard predictor** on a target-aligned augmented dataset.

Main Takeaways of TLCQM

1. **Generate distributions, not point predictions,** from each source at target covariates.

2. **Match target quantiles** to correct discrepancies between $P^{(0)}(Y|X)$ and $P^{(k)}(Y|X)$, $k = 1, \ldots, K$.

3. **Train a standard predictor** on a target-aligned augmented dataset.

Thank you!

More details are in

https://arxiv.org/abs/2602.02358.

Bottom Line: Conditional quantile matching helps avoid negative transfer, while transfer bias governs downstream performance.

References

M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41, 2002.

N. Sgouropoulos, Q. Yao, and C. Yastremiz. Matching a distribution by matching quantiles estimation. *Journal of the American Statistical Association*, 110(510):742–759, 2015.

X. Shen and N. Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):653–677, 2025.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

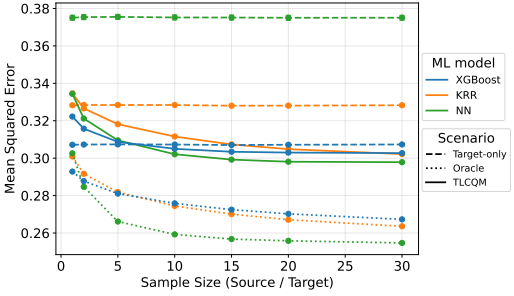
Y. Zhang, S. Wilkins-Reeves, W. Lee, and A. Hofleitner. Transfer learning through conditional quantile matching. *arXiv preprint arXiv:2602.02358*, 2026.

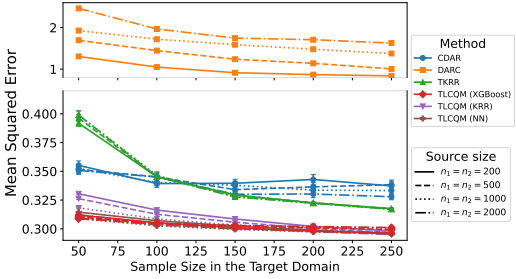
- *Two Source Domains:*

$$Y^{(1)} = \sin\left(3\theta^T X^{(1)}\right) + 1 + \epsilon, \quad Y^{(2)} = 2\cos\left(3\theta^T X^{(2)}\right) + 1 + \epsilon,$$

where $\theta = \left(1, \frac{1}{2}, \ldots, \frac{1}{6}\right)^T \in \mathbb{R}^6$ and $X^{(1)}, X^{(2)} \sim \mathcal{N}(\mathbf{1}_6, \boldsymbol{I}_6)$, $\epsilon \sim \mathcal{N}\left(0, \frac{1}{4}\right)$.

- *Target Domain:* $Y^{(0)} = \frac{1}{3}\sin\left(3\theta^T X^{(0)}\right) - 3 + \epsilon$ with $X^{(0)} \sim \mathcal{N}(\mathbf{0}_6, 0.25 \cdot \boldsymbol{I}_6)$ and $\epsilon \sim \mathcal{N}(0, 0.25)$.





A Note on High-Dimensional Sparse Sequential Change-Point Detection

One Affected Stream among Many

Presenter: Xinyuan Zhang (Georgia Institute of Technology)

Joint work with Dr. Benjamin Yakir (Hebrew University of Jerusalem) and Dr. Yajun Mei (New York University)

June 2, 2026

Sparse Change in High-dimensional System

- **Change-Point Model:** A large-scale system of data streams $\{X_{k,t}\}_{t\geq 1}$ for $k = 1, \ldots, K$
  - When the system is in-control, $X_{k,t} \sim f$
  - At some time point ν, **exactly one** data stream (with index j) is affected: $X_{j,t} \sim g$ for $t \geq \nu$
- **Statistical Procedure:** An integer-valued stopping rule T with respect to the observations
  - A goal is to design T that optimizes

$$\bar{D}(T) := \sup_{\nu \geq 1} \mathbb{E}_\nu\big(T - \nu + 1 \mid T \geq \nu\big), \quad \text{subject to} \quad \mathbb{E}_\infty(T) \geq \gamma \tag{1}$$

  where $\gamma > 0$ is a pre-specified constant.
- How does the optimal detection delay depend jointly on the false-alarm constraint γ and the dimension K?

Asymptotic Regimes

- **Prior Work Across Asymptotic Regimes** [1]
  - $\gamma \to \infty$ with fixed $K < \infty$, e.g., Tartakovsky and Veeravalli (2008)
  - $\log \gamma \gg K$, e.g., Sum-CUSUM in Mei (2010)
  - $\log \gamma \asymp K$, e.g., Mixture-Likelihood-Ratio procedure in Xie and Siegmund (2013)
  - $\log \gamma \ll K$, e.g., Chan (2017) considers $\log \gamma \sim K^{\zeta}$ and $s \sim K^{1-\beta}$ for some $\zeta \in (0,1)$ and $\beta \in (0,1)$, where s is the number of affected data streams
- **Our Focus**
  - The extremely sparse case $s = 1$ under the asymptotic regime

$$K^{c} \ll \gamma \ll e^{K^{\zeta}} \quad \text{for some } c > 0 \text{ and every } \zeta > 0. \tag{2}$$

    i.e., γ grows faster than some **polynomial power** of K but slower than every stretched exponential in K (the latter implies $\log \gamma \ll K^{\zeta}$)
  - Setting γ to be a polynomial power of K is more practical (e.g., 100^2 vs e^{100})

[1] K is the number of data streams in the system and γ is the constraint on ARL to false alarm.

Results

- **An Information Lower Bound:**
  - Suppose $\gamma \gg K^c$ for some constant $c > 1$ and some other assumptions, for any T such that $\mathbb{E}_\infty(T) \geq \gamma$,

$$\liminf_{\gamma, K \to \infty} \left\{ \sup_{1 \leq \nu < \infty} \bar{D}(T) - \frac{\log \gamma + \log K}{I(g, f)} \right\} \geq - \frac{\mathbb{E}[\log(1 + R_1^* + R_2^*)]}{I(g, f)}, \tag{3}$$

  where R_1^*, R_2^* are some r.v.s, $I(g, f)$ is the KL divergence between g and f.
- **A First-Order Optimal Procedure with Federated Monitoring Structure:**
  - The Sum Shiryaev-Roberts-Pollak (Sum-SRP) Procedure $N(A) := \inf\{n \geq 1 : R(n) \geq A\}$ with $R(n) = K^{-1} \sum_{j=1}^{K} R(n; k)$, the summation of local SRP statistics.
  - Under some assumptions, with $A = \lambda\gamma$

$$\lim_{\gamma, K \to \infty} \mathbb{E}_\infty \big[N(A)/\gamma \big] = 1. \tag{4}$$

  and

$$\limsup_{\gamma, K \to \infty} \left\{ \bar{D}[N(A)] - \frac{\log \gamma + \log K}{I(g, f)} \right\} \leq \frac{1}{I(g, f)} \{ \rho + \log \lambda - \mathbb{E}[\log(1 + R_2^*)] \}, \tag{5}$$

  for some $\lambda, \rho > 0$.

Improving Trial-Based CATE Estimation with Observational Data Borrowing

Borrowing Evidence from Real-World Data Without Introducing Bias

Samhita Pal · Jared Huling · Amir Asiaee

Department of Biostatistics, Vanderbilt University Medical Center

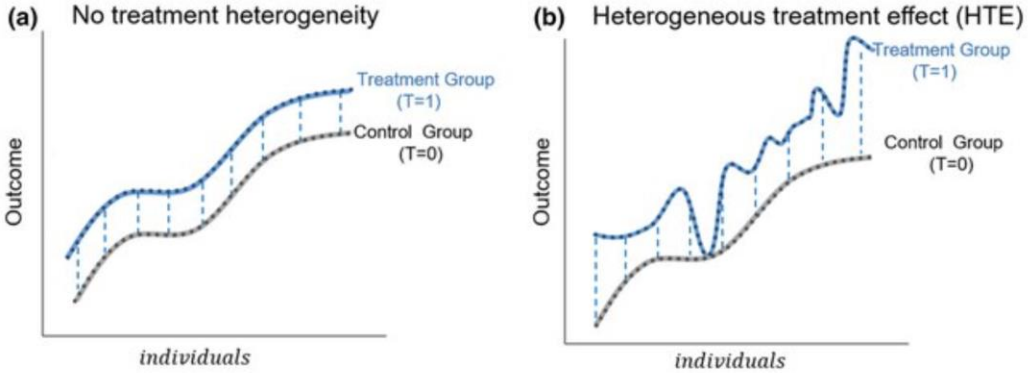
IMSI New Horizons on Model Transportability and Data Integration **2026**

One drug — but who actually benefits?

The same treatment can
help some patients and
harm others.

Precision medicine asks: for this specific patient, how large is the benefit?

This patient-level variation is called a Heterogeneous Treatment Effect (HTE). Estimating HTE precisely lets us match the right treatment to the right patient — and avoid giving a harmful therapy to someone who won't benefit.



Goal: Efficient and consistent estimation of the conditional average treatment effect (CATE) in the trial population,

$$\tau^r(\boldsymbol{x}) = \mathbb{E}\{Y(1) - Y(-1) \mid \boldsymbol{X} = \boldsymbol{x},\ S = r\},$$

where $Y(a)$ is the potential outcome under treatment $a \in \{-1, 1\}$ and $S = r$ indicates the RCT.

The gold standard: Randomized Controlled Trials (RCTs)

Why RCTs are trusted

→ Treatment assigned randomly

→ Groups balanced on known and unknown factors

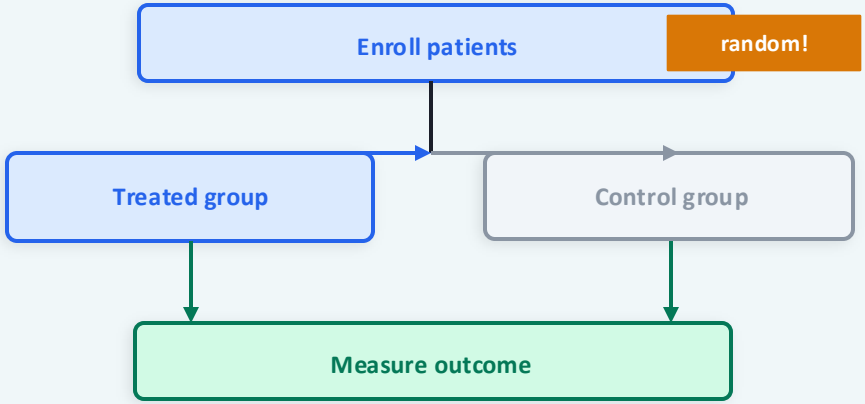
→ Causal conclusions are valid

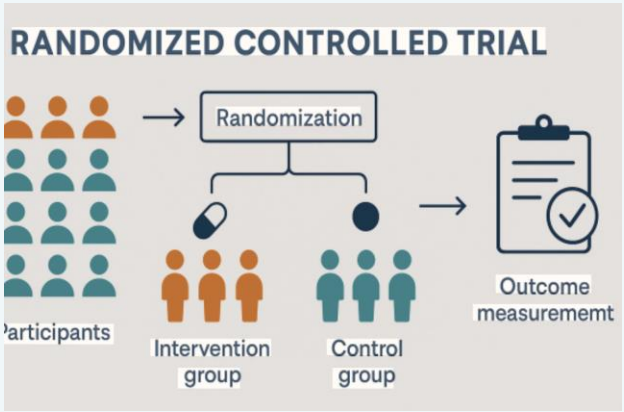
The catch

→ Typically, only a few hundred patients

→ Powered for average effects — not for who benefits most

→ Too small to find fine-grained HTE





Observational studies: big but messy

Electronic health records, registries, claims data…

Strengths

→ Thousands of patients

→ Rich, detailed covariates

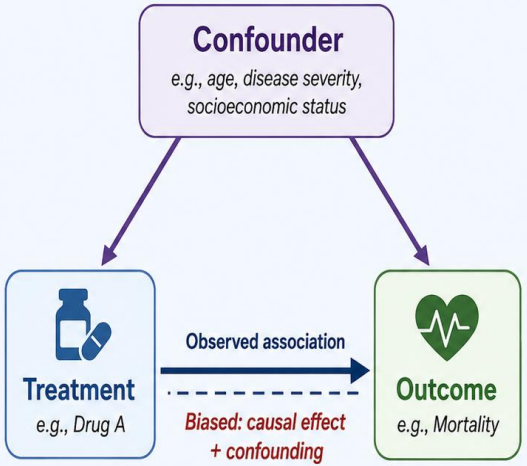
→ Mirrors the real world

The fundamental problem: Confounding

→ Sicker patients more likely to receive treatment

→ Cannot directly conclude causation

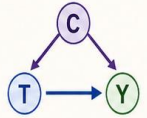
Observational study: confounding bias



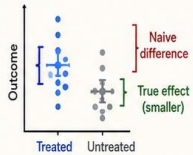
Why bias occurs

1 Patients are not randomly assigned. ≠

2 Baseline differences affect both treatment choice and outcome.



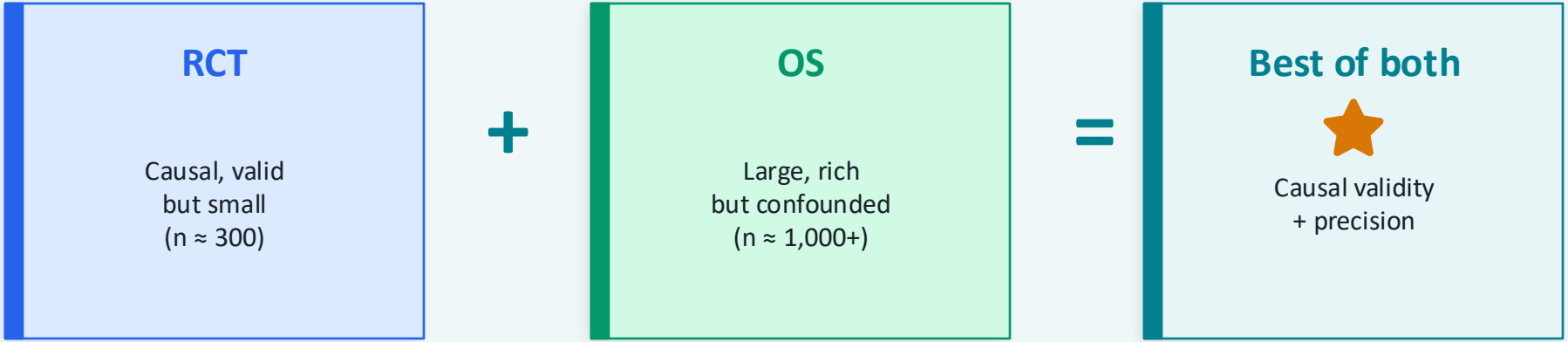
3 Naive comparisons can overstate or understate the true treatment effect.



Confounding creates spurious treatment–outcome associations in observational data unless adjusted for.

Can we combine both data sources?

What if the big observational data helps us sharpen treatment-effect estimates from the small RCT?



Key challenge: The OS may measure different variables than the trial, and its outcome predictions carry confounding bias. We need a method that borrows carefully — without corrupting the causal estimate.

How R-OSCAR borrows from the OS — without importing its bias

CATE as pseudo-outcome regression

Choose augmentation $m(X)$ → Build pseudo-outcome → Regress pseudo-outcome on X in the RCT

$$\tau_m(X, A, Y) = \frac{A\{Y - m(X)\}}{\pi_A(X)}$$

$$\hat{\tau}^r \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n_r} \sum_{i:S_i=r} \{\tau_m(X_i^r, A_i, Y_i) - f(X_i^r)\}^2$$

Best choice: $m(X) = \mu^r(X)$, the RCT counterfactual mean outcome (CMO).

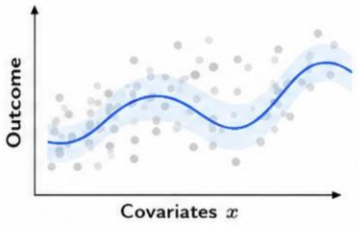
Counterfactual mean outcome (CMO)

$$\hat{\mu}^r(x) = \sum_{a \in \{-1,1\}} \pi^r_{-a}(x)\, \hat{\mu}^r_a(x)$$

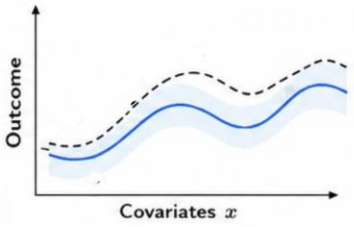
where $\pi^r_a(x) = \Pr(A = a \mid X = x, S = r)$ are randomization (propensity) probabilities in the RCT.

1 Learn from OS → **2 RCT outcomes vs OS** → **3 Learn penalized discrepancy** → **4 Calibrated outcome model**

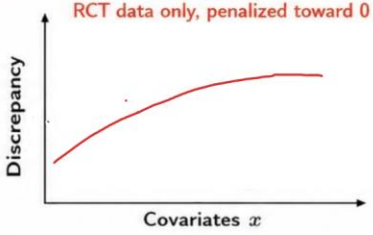
OS outcome model $\hat{\mu}^o_a(x)$



RCT data inform discrepancy

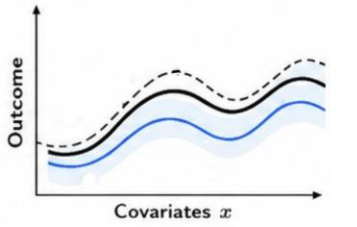


Discrepancy $\hat{\delta}_a(x)$ estimated on RCT data only, penalized toward 0



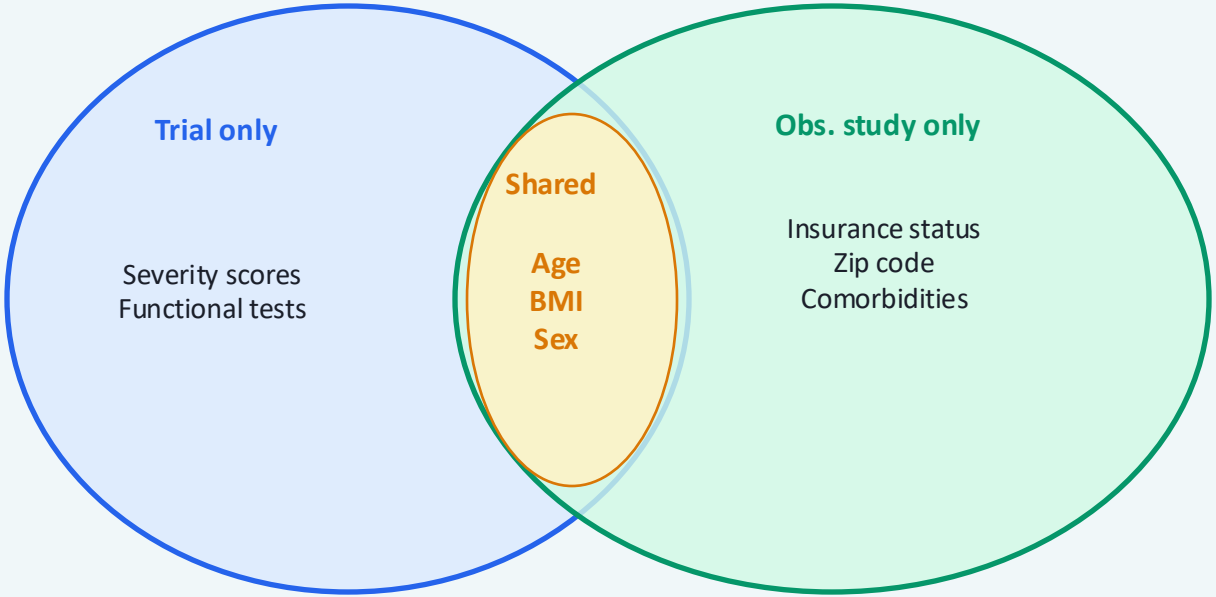
Penalization shrinks toward 0:
- prevents overfitting
- preserves borrowing from OS

$\hat{\mu}^{cal}_a(x) = \hat{\mu}^o_a(x) + \hat{\delta}_a(x)$



But what if the two datasets don’t even share the same variables?

R-OSCAR assumes shared covariates. In practice, the OS often records variables the trial never collected — this is covariate mismatch, and it requires a further extension: MR-OSCAR.



Problem

Covariate mismatch

A model trained on OS data cannot be applied to trial patients — those variables are missing. R-OSCAR’s next extension (MR-OSCAR) addresses this.

This is called covariate mismatch. R-OSCAR handles shared variables; MR-OSCAR extends the framework to impute and borrow across mismatched covariate sets — the subject of ongoing work.

Thank you!

Questions welcome.

samhita.pal@vumc.org

ArXiv preprint: R-OSCAR and MR-OSCAR

Joint work with Dr. Amir Asiaee and Dr. Jared Huling · Vanderbilt University Medical Center

Federated Learning of Robust Individualized Decision Rules

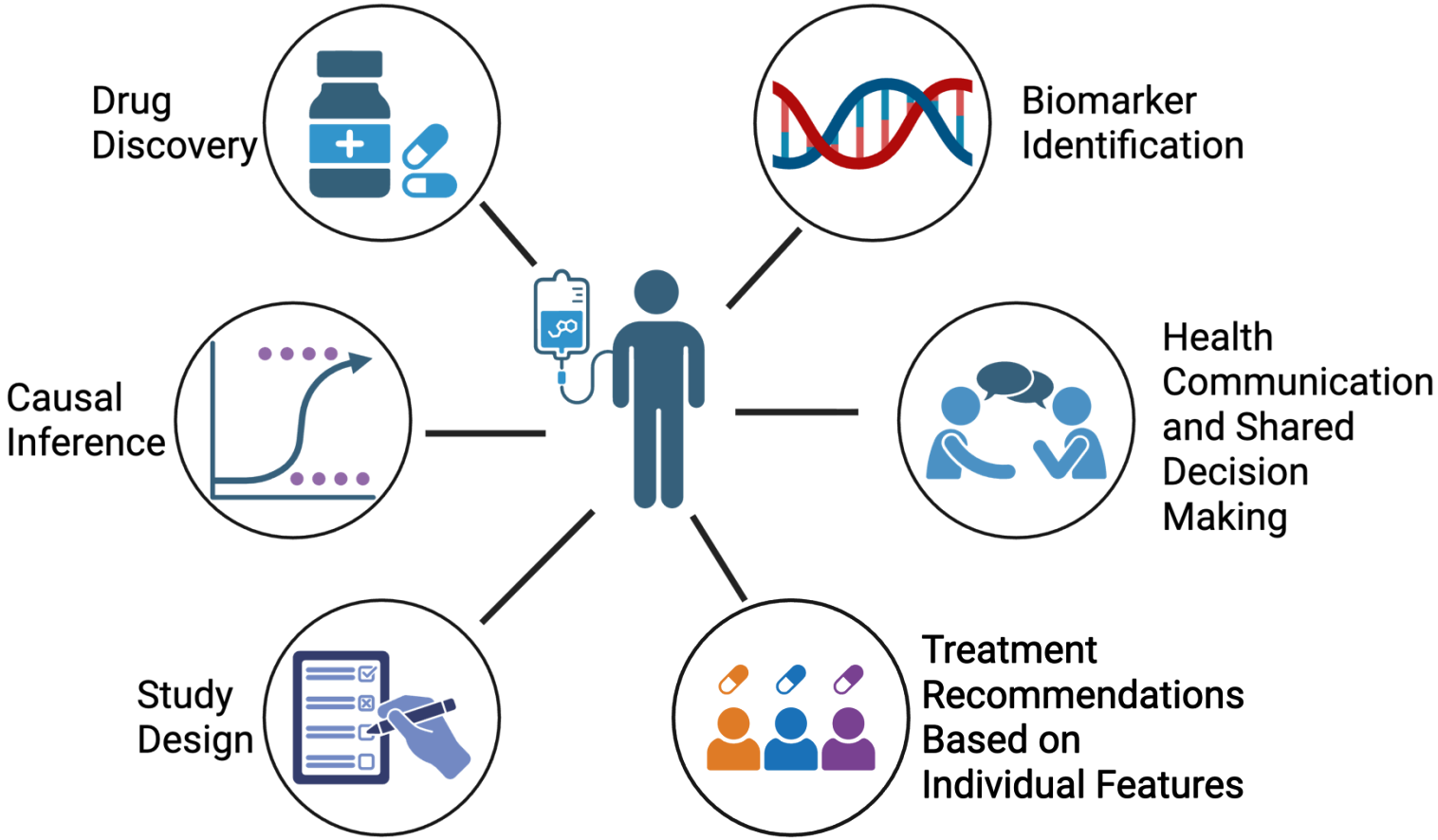
Xinlei Chen

University of Pittsburgh

Department of Biostatistics and Health Data Science

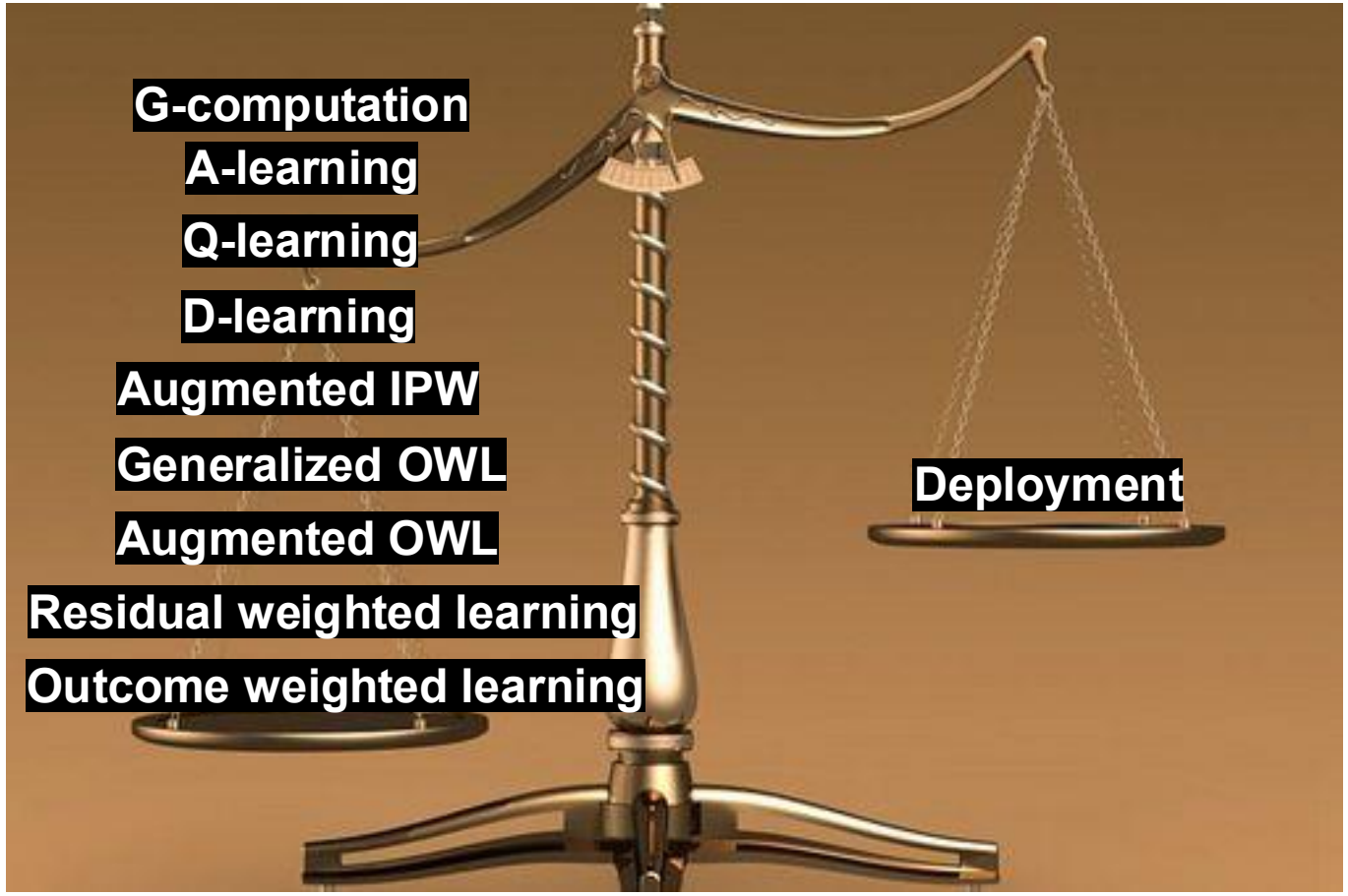
June 23, 2026

Treatment Recommendation Systems



Created with BioRender.com

Challenges In Clinical Deployment



Challenge 1: Infrastructure Barrier

- Not all hospitals have the data infrastructure required to collect the data and resources to build and maintain the treatment recommendation model.
- **Generalizability:** A model trained on one patient population is not expected to have the same accuracy when applied to another.

Patients in hospitals used for training

Generalizable?

Patient from unknown population

Challenge 2: Trustworthiness

- Data-driven models suffer from biases in data and model building and consequently may lead to treatment solutions that are against the principle of clinic practices.

What would an ideal treatment recommendation system look like?

What Would An Ideal System Look Like?

Recommends treatment that is safe to patients

Robust to population uncertainties

Feasible under real-world infrastructure constraints

Recommends treatment that improves patient outcome

What if, there is a system that can do everything at the same time…

University of Pittsburgh®

Thank You For Listening!

xic120@pitt.edu

Assembling a Synthetic Life Course Cohort to Study the Incidence of Non-AIDS-Defining Conditions among People with Perinatally Acquired HIV in the United States:

A DATA FUSION STUDY

Jason Haw, PhD (nhaw1@jh.edu)

Department of Epidemiology, Johns Hopkins University

on behalf of co-authors Keri Althoff, PhD; Allison Agwu, MD ScM; Catherine Lesko, PhD

Lightning Talk · June 23, 2026 · New Horizons on Model Transportability and Data Integration

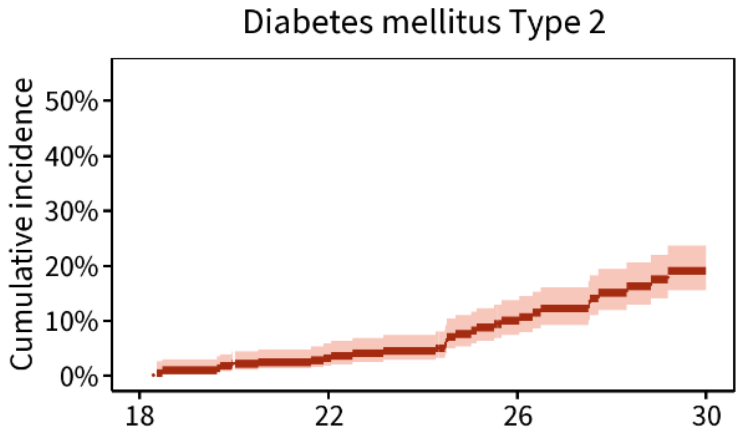
Adults with perinatal HIV have been living with HIV for decades longer than adults with non-perinatal HIV

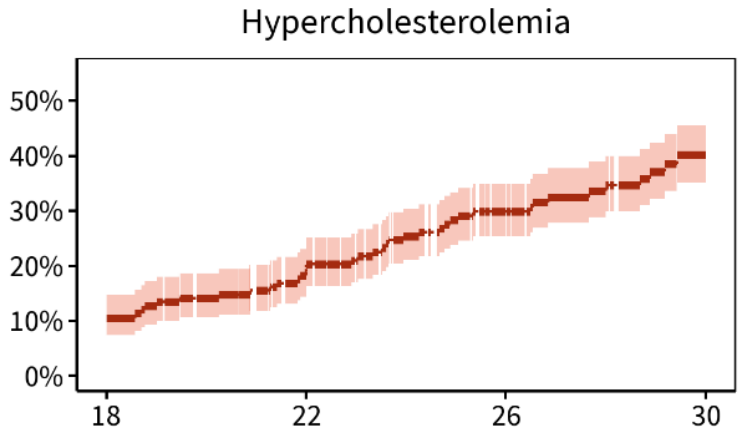
Perinatal HIV acquisition ("vertical transmission"):
Passed on during pregnancy, childbirth, or breastfeeding

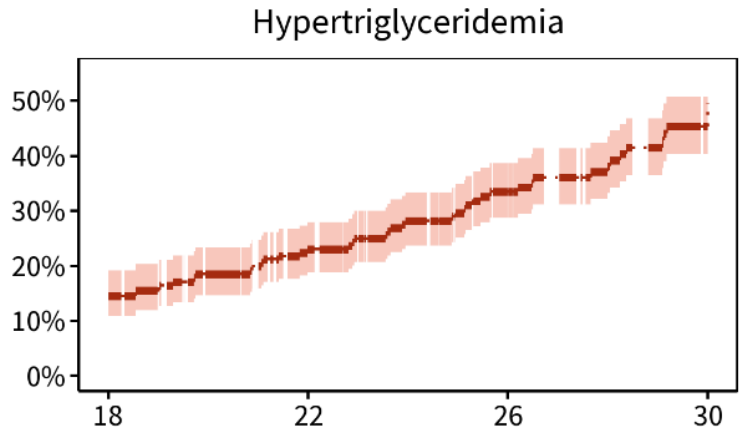
Consider a 30 year old…

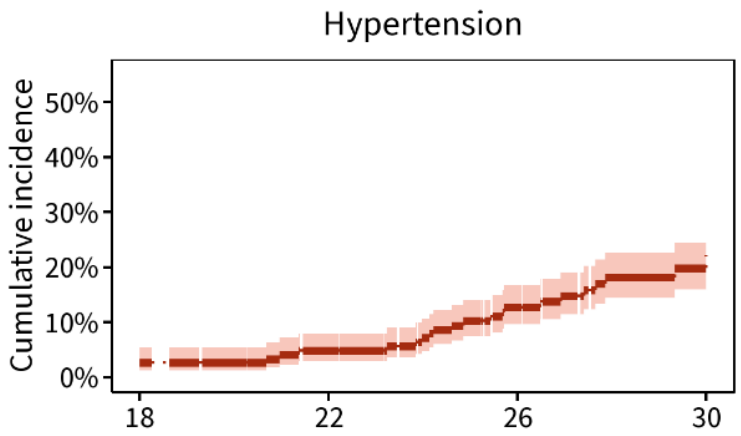
with **perinatal HIV**:	Living with HIV for	**30 years**
with **non-perinatal HIV**:	On average, living with HIV for	**5-10 years**

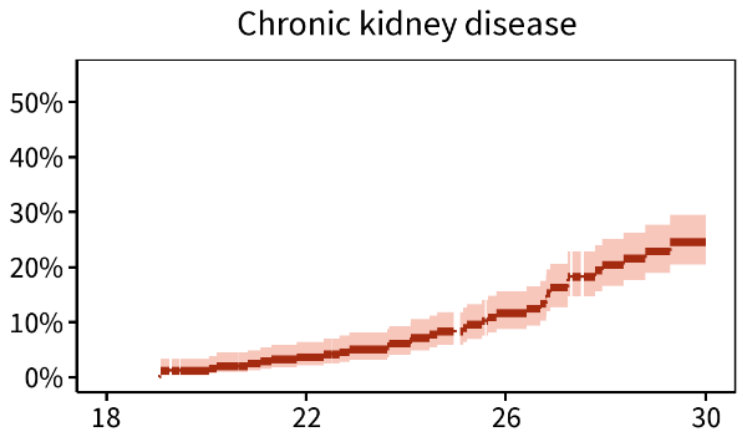
Adults with perinatal HIV may experience “accelerated” aging, with high incidence of non-AIDS defining conditions by age 30











Cumulative incidence of non-AIDS defining conditions by age 30 among young adults with perinatally acquired HIV in North America 2000 to 2019

Early exposure to suboptimal antiretroviral therapy may be a factor, but no single cohort can answer this question

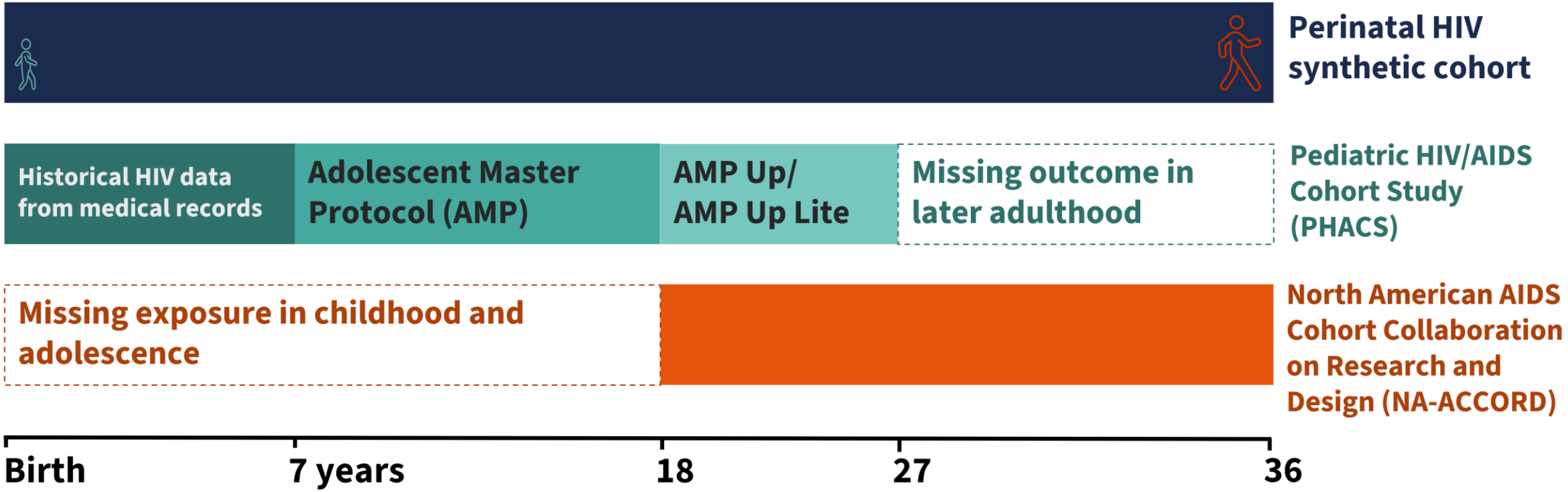
Pediatric cohorts lack adult trajectories

- Lost to follow-up due to aging out of pediatric HIV care
- Prospective adult data are decades away

Adult cohorts lack early life histories

- Limitations in linkages to pediatric HIV providers
- Restrictions with institutional review board approvals

Assembling a synthetic cohort of perinatal HIV using two long-standing HIV cohorts



1 NA-ACCORD participant is MATCHED with replacement to 1 PHACS participant based on shared age region (18-27 years)

Variables	PHACS (Birth – 27 years)	NA-ACCORD (18 – 36 years)
Sociodemographic characteristics (sex, birth year, race/ethnicity, region of site, insurance)	✓ baseline study visit interview	✓ first eligible clinical care visit
HIV-related characteristics (antiretroviral therapy, CD4 cell count, HIV viral load, AIDS history)	✓ study interval visits, medical record abstraction	✓ medical record abstraction, historical data available
Non-AIDS defining conditions (biomarkers, medications, and clinical diagnoses)	✓ medical record abstraction	✓ medical record abstraction

Preliminary data: 302 PHACS, 676 NA-ACCORD

	NA-ACCORD (as of 2024)
Birth year, median (IQR)	1992 (1988, 1996)
Age (years), median (IQR)	at entry: 21.1 (18.7, 25.3); at exit: 31.7 (27.3, 36.3)
Year, median (IQR)	at entry: 2014 (2011, 2019); at exit: 2022 (2015, 2024)
Female, *n* (%)	382 (56%)
Race/ethnicity, *n* (%)	Non-Hispanic White: 73 (11%); Non-Hispanic Black: 309 (46%) Hispanic: 126 (19%); Other/Multi-racial: 168 (25%)

IQR: interquartile range, study exit defined as last activity date

ACKNOWLEDGMENTS

PHACS cohort data access made possible through a public use data set available on the Harvard Dataverse and/or the NICHD BRICS repository

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was supported by National Institutes of Health grants U01AI069918, F31AI124794, F31DA037788, G12MD007583, K01AI093197, K01AI131895, K23EY013707, K24AI065298, K24AI118591, K24DA000432, KL2TR000421, N01CP01004, N02CP055504, N02CP91027, P30AI027757, P30AI027763, P30AI027767, P30AI036219, P30AI050409, P30AI050410, P30AI094189, P30AI110527, P30MH62246, R01AA016893, R01DA011602, R01DA012568, R01 AG053100, R24AI067039, U01AA013566, U01AA020790, U01AI038855, U01AI038858, U01AI068634, U01AI068636, U01AI069432, U01AI069434, U01DA03629, U01DA036935, U10EY008057, U10EY008052, U10EY008067, U01HL146192, U01HL146193, U01HL146194, U01HL146201, U01HL146202, U01HL146203, U01HL146204, U01HL146205, U01HL146208, U01HL146240, U01HL146241, U01HL146242, U01HL146245, U01HL146333, U24AA020794,U54MD007587, UL1RR024131, UL1TR000004, UL1TR000083, Z01CP010214 and Z01CP010176; contracts CDC-200-2006-18797 and CDC-200-2015-63931 from the Centers for Disease Control and Prevention, USA; contract 90047713 from the Agency for Healthcare Research and Quality, USA; contract 90051652 from the Health Resources and Services Administration, USA; grants CBR-86906, CBR-94036, HCP-97105 and TGF-96118 from the Canadian Institutes of Health Research, Canada; Ontario Ministry of Health and Long Term Care; and the Government of Alberta, Canada. Additional support was provided by the National Institute Of Allergy And Infectious Diseases (NIAID), National Cancer Institute (NCI), National Heart, Lung, and Blood Institute (NHLBI), Eunice Kennedy Shriver National Institute Of Child Health & Human Development (NICHD), National Human Genome Research Institute (NHGRI), National Institute for Mental Health (NIMH) and National Institute on Drug Abuse (NIDA), National Institute On Aging (NIA), National Institute Of Dental & Craniofacial Research (NIDCR), National Institute Of Neurological Disorders And Stroke (NINDS), National Institute Of Nursing Research (NINR), National Institute on Alcohol Abuse and Alcoholism (NIAAA), National Institute on Deafness and Other Communication Disorders (NIDCD), and National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

This research was funded in part by a 2025 developmental grant from the Johns Hopkins University CFAR, an NIH-funded program (1P30AI094189) supported by: NIAID, NCI, NICHD, NHLBI, NIDA, NIA, NIGMS, NIDDK, NIMHD.

The authors would also like to thank PHACS and NA-ACCORD participants.

CascadeS2F:

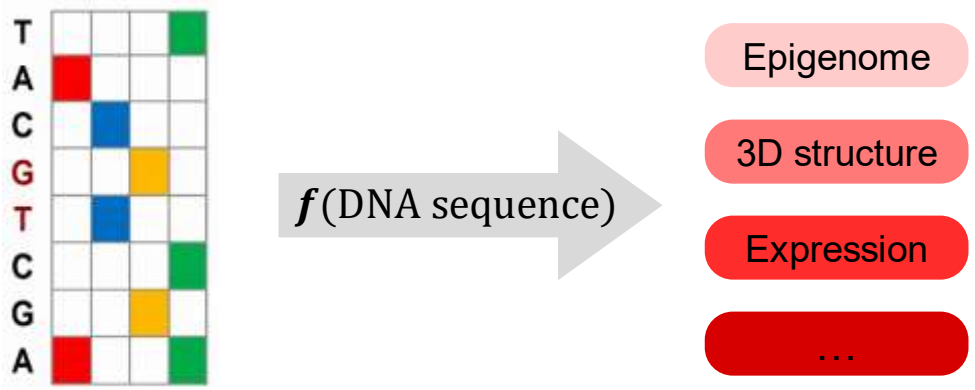
A Cascaded, Causally-Aware Sequence-to-Function Model Across Molecular Layers

Shan Leng | Keleş Lab

Department of Biostatistics and Medical Informatics
University of Wisconsin–Madison

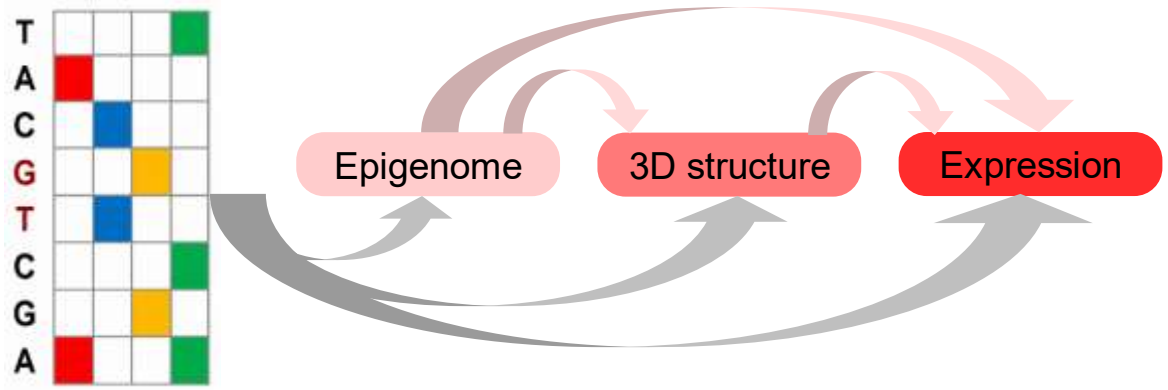
Motivation: From S2F to CascadeS2F

Sequence-to-function (S2F) models



Variant effect: $f(\mathrm{ALT}) - f(\mathrm{REF})$

CascadeS2F



Variant effect:

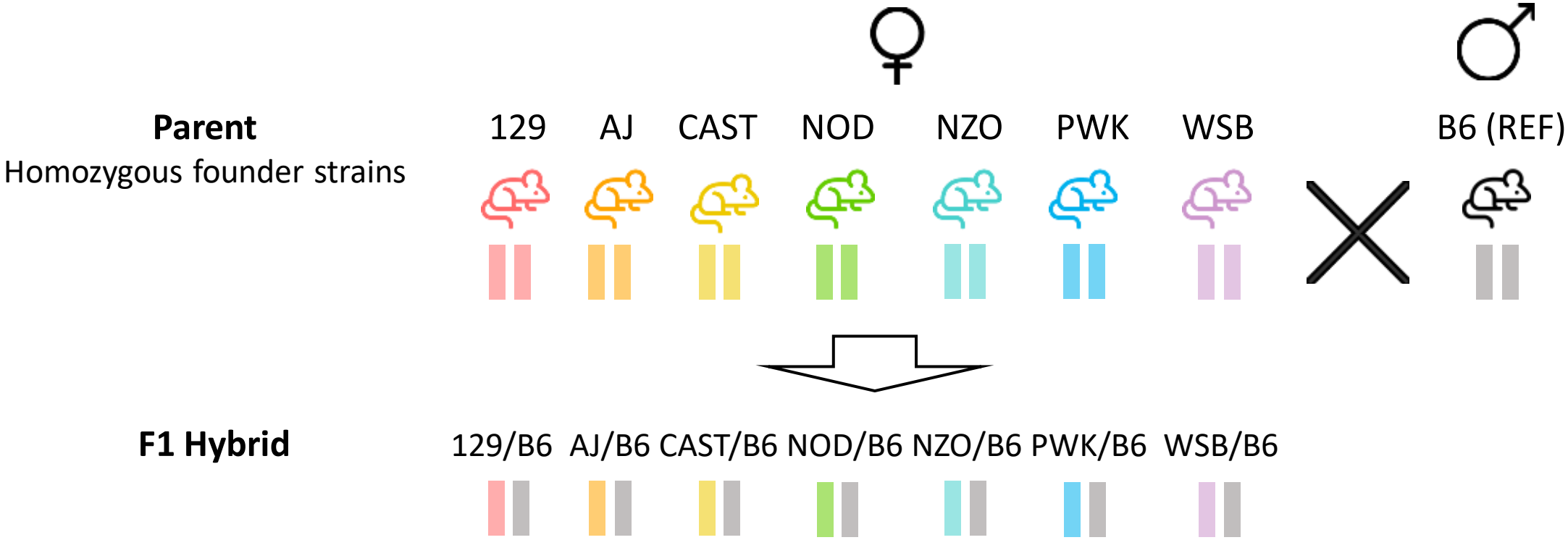
SNP-specific latent contribution that can

① directly affect each molecular layer ( )

② propagate through regulatory hierarchy ( )

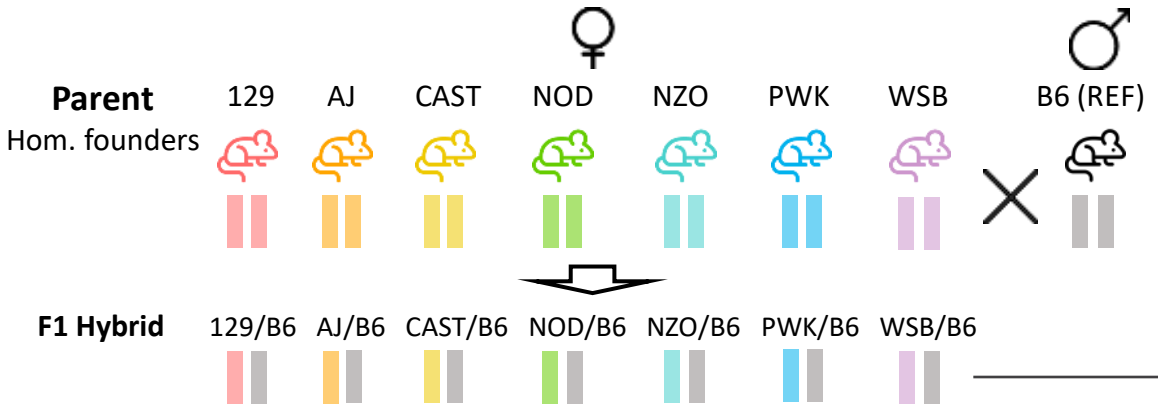
Experimental System

Large, trackable genetic variation in F1 crosses:



~40 million SNPs relative to the B6 reference genome,
heterozygous at every locus where the two parental strains differ.

Experimental System



Islets

β-cells → ATAC peaks and gene expression

β-cells → *in situ* Hi-C → chromatin loops

CTCF CUT&RUN → CTCF binding sites

F1 pseudobulk allele-specific（AS）data

AS ATAC peaks

AS CTCF binding sites

AS chromatin loops

AS gene expression

Maternal

Paternal

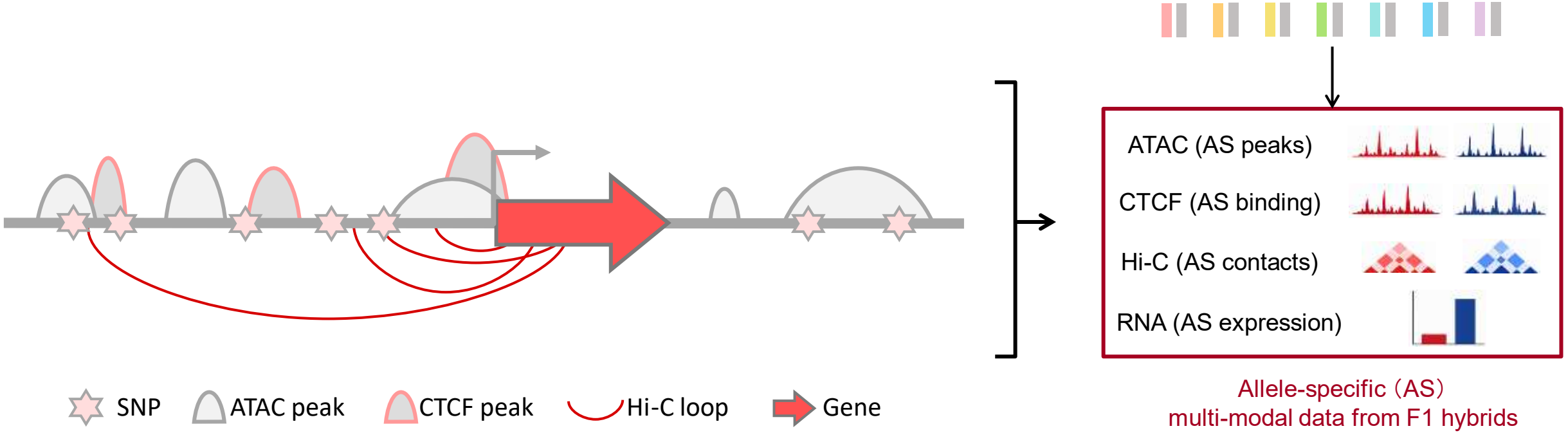
Undefined

SNP info

Total reads

CascadeS2F: Setup

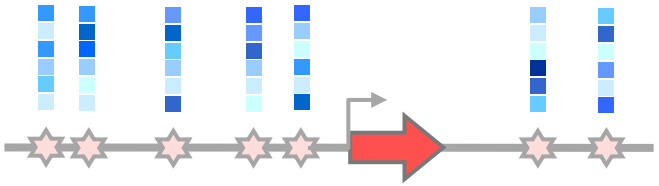
cis window surrounding a gene

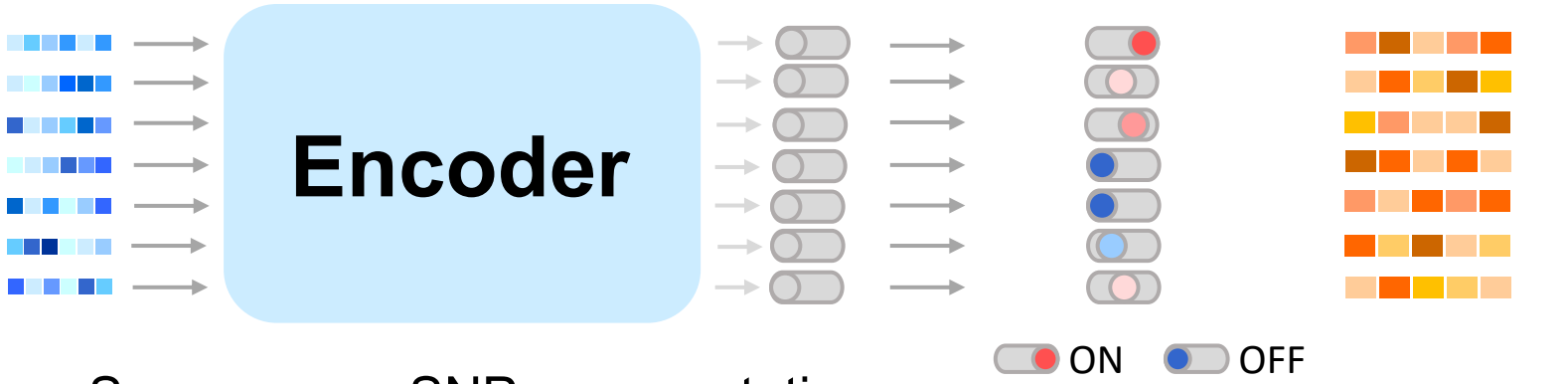


Allele-specific (AS)
multi-modal data from F1 hybrids

CascadeS2F: Model

Per-SNP input features:
motif disruption, distance to gene, conservation

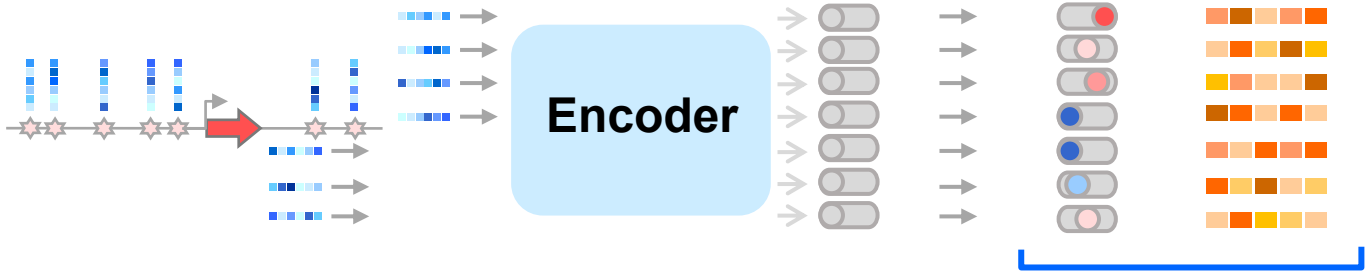




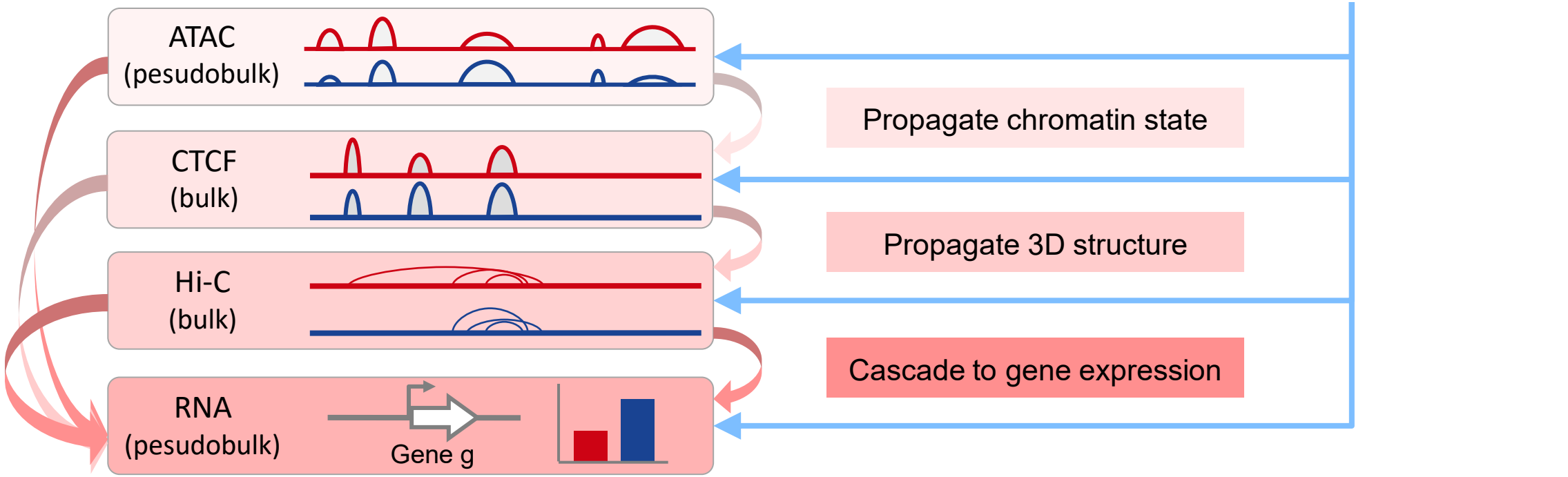
Sequence → SNP representation

Encoder uses SNP features in the *cis* window to learn:
which variants matter for the gene and **how they perturb regulation**.

CascadeS2F: Model

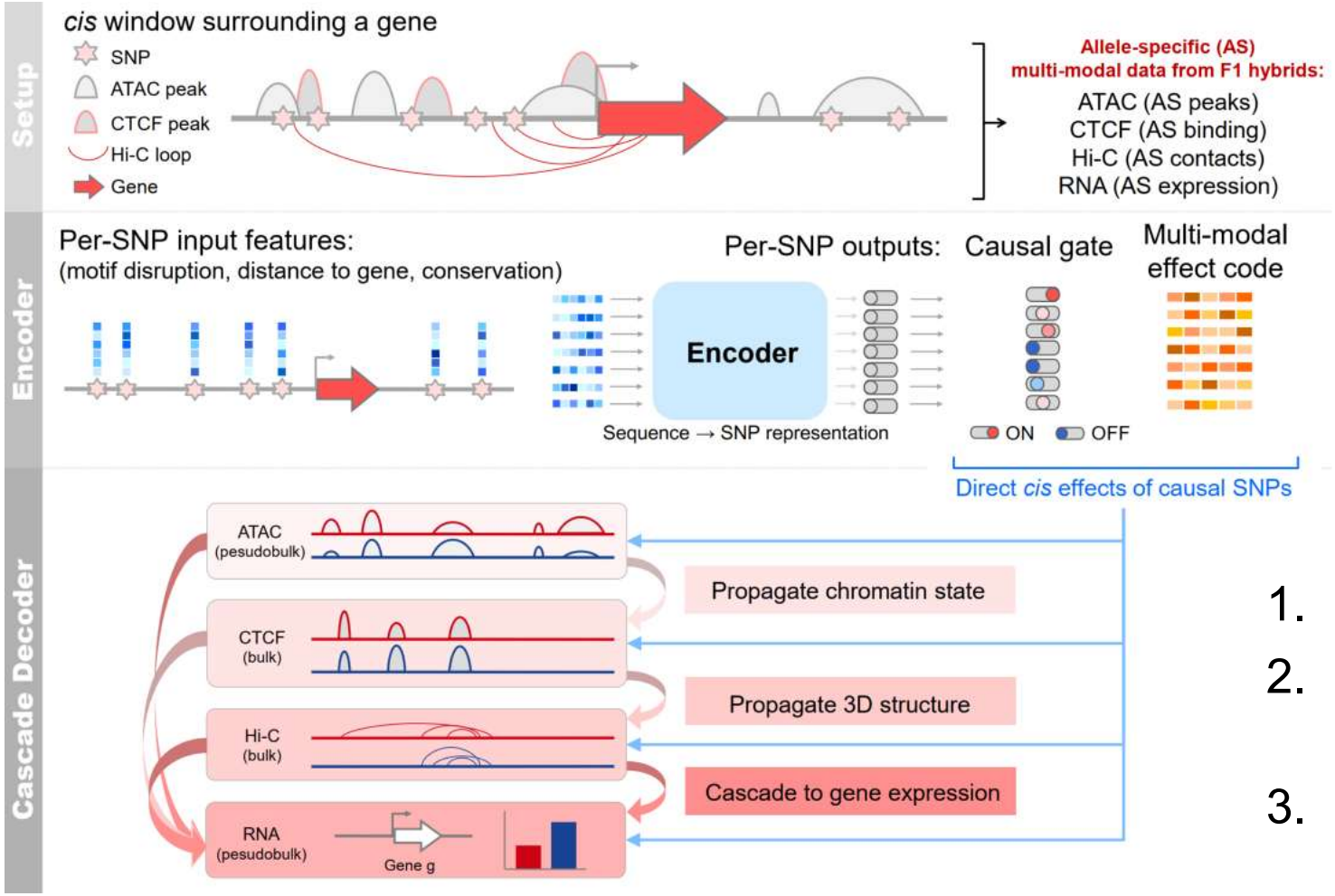


Cascade Decoder:



Variant effects selected by the **encoder** first alter local chromatin (ATAC), **then** propagate through CTCF binding and 3D contacts (Hi-C), ultimately shaping allele-specific RNA expression of gene g.

CascadeS2F: Summary



1. Shared SNP encoder across genes.
2. Gene-specific cascade decoder across ATAC → CTCF → Hi-C → RNA.
3. Jointly explains allelic imbalance in all modalities and all F1 crosses.

History-Aware Conformal Prediction Sets for Censored Time-to-Event Outcomes

Yuyao Wang

Department of Public Health and Health Sciences, Northeastern University

Joint work with:

Alexander W. Levis, University of Pennsylvania

Shu Yang, North Carolina State University

Larry Han, Northeastern University

Conformal prediction for time-to-event outcomes

- Early developments focus on lower prediction bounds (LPB) under type-I censoring (Candes et al., 2023; Gui et al, 2024)
- Under the general right censoring setting:
  - censoring imputation (Sesia and Svetnik, 2025); weighting (Davidov et al., 2025)
  - EIF-based methods(Farina et al., 2025, Si & Qiu, 2025)
- Recent work on mixed-type prediction sets based on censoring status prediction (Holmes & Marandon, 2024), and two-sided prediction intervals (Yi et al., 2025; Qin et al., 2025).
- Existing conformal survival methods are all **static**
- Limitation: **not informative**

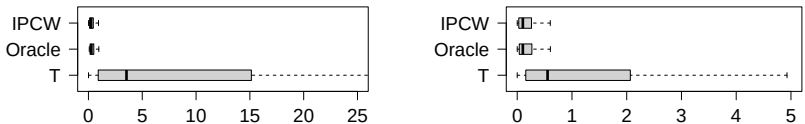


Figure: Boxplots of 90% LPBs under (left) simulation setup 1 in Gui et al. (2024), and (right) simulation setup 1 in Farina et al. (2025).

Notation and prediction target

- T: event time; C: censoring time.
- Z_t: covariates measured at t; $\bar{Z}_t = \{Z_s : 0 \le s \le t\}$.
- Observed data: an i.i.d. sample $\mathcal{D} = \{O_i\}_{i=1}^n$, where $O = (X, \Delta, \bar{Z}_X)$, $X = \min(T, C)$, $\Delta = \mathbb{1}(T < C)$.

Prediction target

- For a fixed prediction time $\tau \ge 0$, construct a prediction set $\hat{\mathcal{C}}(\bar{Z}_{\tau,n+1})$, satisfying *survivor conditional coverage*:

$$\mathbb{P}\Big\{ T_{n+1} \in \hat{\mathcal{C}}(\bar{Z}_{\tau,n+1}) \mid T_{n+1} > \tau \Big\} \ge 1 - \alpha.$$

- *Probably Asymptotically Approximately Correct (PAAC)* coverage: with prob. $\ge 1 - \epsilon$ over $\mathcal{D}$,

$$\mathbb{P}\Big\{ T_{n+1} \in \hat{\mathcal{C}}(\bar{Z}_{\tau,n+1}) \mid T_{n+1} > \tau, \mathcal{D} \Big\} \ge 1 - \alpha + o_p(1)$$

Assumptions and calibration condition

Assumptions:

- **Conditional independent censoring given covariate history**: $\lambda_C(t|\bar{Z}_T, T) = \lambda_C(t|\bar{Z}_t)$
- **Positivity**: (i) $\mathbb{P}(T > \tau) > 0$; (ii) there exists $\eta > 0$ s.t. $\mathbb{P}(C > T \mid T, \bar{Z}_T) > \eta$ a.s..

Calibration Condition:

- **Nested candidate prediction sets:** e.g., $\mathcal{C}_{\tau,\theta}(\bar{z}_\tau; \mathcal{A}) = (\hat{q}_{1-\theta}(\bar{z}_\tau), \hat{q}_\theta(\bar{z}_\tau))$, $\theta \in [0.5, 1]$.
- **Censoring-free calibration condition:**

$$\mathbb{E}\left[\mathbb{1}(T > \tau)\left\{\mathbb{1}\left(T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\right) - (1-\alpha)\right\}\right] \geq 0.$$

- **With censored data, apply IPCW** (Robins & Rotnitzky, 1992; Rotnitzky & Robins, 2005)**:**

$$U(\theta; G, \mathcal{A}) = \frac{\Delta \mathbb{1}(X > \tau)}{G(X|\bar{Z}_X)}\left[\mathbb{1}\{X \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} - (1-\alpha)\right].$$

$\lambda_C(t|\cdot) = \lim_{h \to 0^+} \mathbb{P}(t \leq C < t+h \mid \cdot,\ C \geq t,\ T \geq t)/h$; $G(t|\bar{Z}_t) = \exp\{-\int_0^t \lambda_C(u|\bar{Z}_u)\, du\} = \mathbb{P}(C > t \mid \bar{Z}_t)$.

History-aware conformal prediction sets (HAPS)

Split-conformal algorithm:

1. Randomly split $\mathcal{D}$ (50:50) into a training set $\mathcal{D}_{\text{tr}}$ and a calibration set $\mathcal{D}_{\text{cal}}$.
2. Fit the prediction model $\hat{\mathcal{A}}$ on $\mathcal{D}_{\text{tr}}$, and construct nested candidate sets $\{\mathcal{C}_{\tau,\theta}\}_{\theta\in\Theta}$.
3. On $\mathcal{D}_{\text{tr}}$, estimate the nuisance parameter G, and denote the estimate as $\hat{G}$.
4. On $\mathcal{D}_{\text{cal}}$, compute $\hat{\theta} = \inf\left\{\theta \in \Theta : \sum_{i\in\mathcal{D}_{\text{cal}}} U_i(\theta; \hat{G}, \hat{\mathcal{A}}) \geq 0\right\}$.
5. **Output:** prediction set $\hat{\mathcal{C}}(\bar{Z}_{\tau,n+1}) = \mathcal{C}_{\tau,\hat{\theta}}(\bar{Z}_{\tau,n+1}; \hat{\mathcal{A}})$.

PAAC-type survivor conditional coverage

For any $\epsilon \in (0,1)$, there exists $K > 0$ such that, with prob. $\geq 1-\epsilon$ over $\mathcal{D}$,

$$\mathbb{P}(T_{n+1} \in \hat{\mathcal{C}}(\bar{Z}_{\tau,n+1}) \mid T_{n+1} > \tau, \mathcal{D})$$
$$\geq 1 - \alpha - s_\tau^{-1}\eta^{-2}\|\hat{G} - G\|_2 - s_\tau^{-1}\eta^{-1}\left(\sqrt{\frac{1}{2}\log\frac{1}{\epsilon}} + K\right)\frac{1}{\sqrt{|\mathcal{D}_{\text{cal}}|}}.$$

Doubly robust extensions

HAPS-DR: doubly robust post-processing

$$\hat{\mathcal{C}}_{\mathrm{DR}}(\bar{Z}_\tau) = \hat{\mathcal{C}}(\bar{Z}_\tau) \cup \left(\hat{q}_{\alpha/2}(\bar{Z}_\tau), \hat{q}_{1-\alpha/2}(\bar{Z}_\tau)\right).$$

- Achieves PAAC coverage if either $\hat{G}$ or $(\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2})$ is consistent.

HAPS-A: use augmented IPCW (AIPCW) estimating function,

$$U_{\mathrm{AIPCW}} = U(\theta; G, \mathcal{A}) + \int_0^X h_\tau(u, \bar{Z}_u; \theta, \mathcal{A}) \frac{dM_C(u; G)}{G(u \mid \bar{Z}_u)}.$$

where $h_\tau(u, \bar{Z}_u; \theta, \mathcal{A}) = \mathbb{E}\left(\mathbb{1}(T > \tau)\left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} - (1-\alpha)\right] \middle| \bar{Z}_u, T \geq u\right)$,
$N_C(t) = \mathbb{1}(X \leq t, \Delta = 0)$, $M_C(t; G) = N_C(t) - \int_0^t \mathbb{1}(X \geq u) dG(u|\bar{Z}_u)/G(u|\bar{Z}_u)$,

- Achieves PAAC coverage if either $\hat{G}$ or $\hat{h}_\tau$ is consistent.
- <u>Rate double robustness</u>: nuisance estimation errors affect the coverage gap only by the product of the estimation errors between $\hat{G}$ and $\hat{h}_\tau$.

Simulation

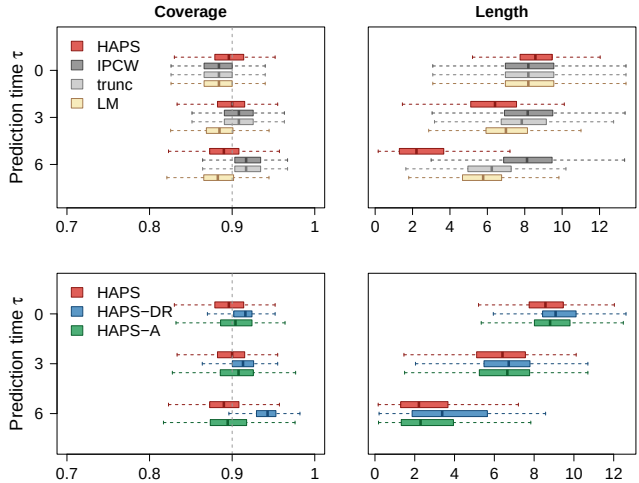


Figure: 90% prediction intervals under 200 Monte Carlo replications, with Dynamic-DeepHit prediction model and RSF/XGB-XGB nuisance models. Sample size $n = 1000$; test data with size 500 and no censoring. Relative to baseline methods, the median interval length is reduced by **5%–22%** at $\tau = 3$ and **34%–68%** at $\tau = 6$.

Applications: colon cancer data

- 929 patients; censoring rate: 51.3%.
- Baseline covariates: age (years), sex (male/female), treatment group (observation, Levamisole, or Levamisole+5-FU), number of positive lymph nodes (more than 4 vs. not)
- Time-varying covariate: recurrence (a 0–1 process).

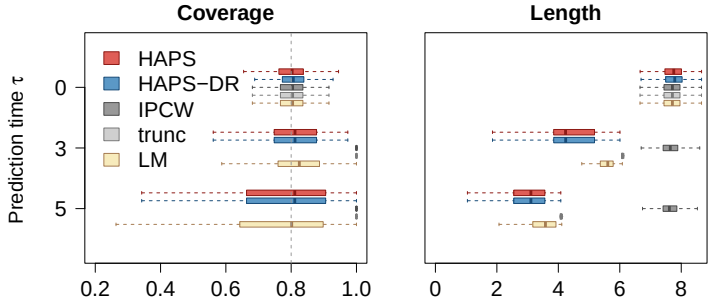


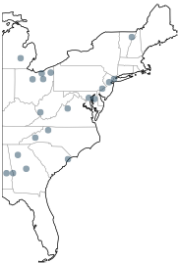
Figure: 80% prediction intervals with Dynamic-DeepHit prediction model across 200 random splits. Relative to baseline methods, the median interval length is reduced by **22%–44%** at $\tau = 3$ years and **13%–60%** at $\tau = 5$ years.

When Representative Samples Produce Worse Outcomes

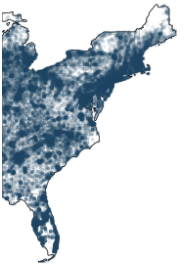
Scale-up Decisions and Testing in Small-Budget RCTs

Hannah Li Hongseok Namkoong Isaac Scheinfeld

Convenience or Enriched

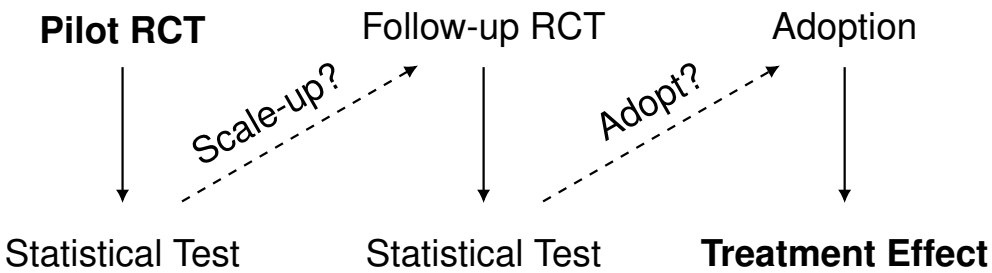


Representative Sample

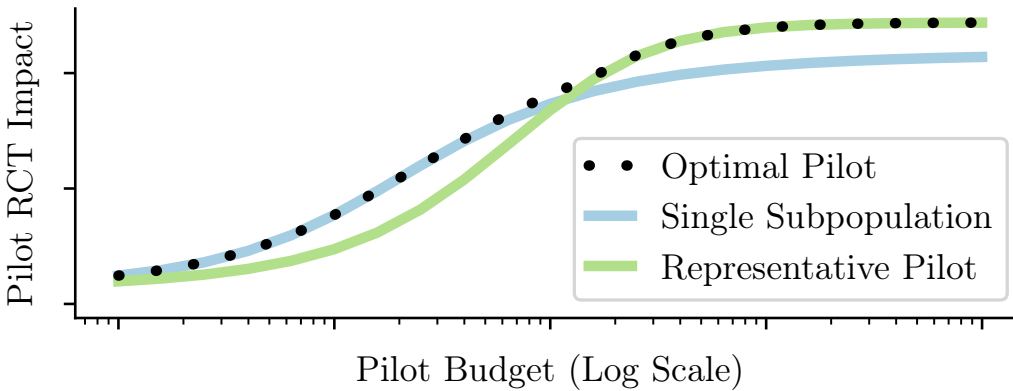


Target Population

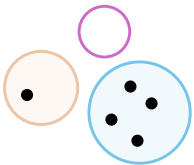
Status-Quo – Significance Testing



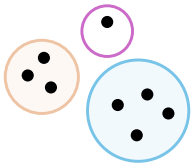
We Find – Optimal Depends on Budget



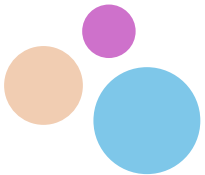
Sample Variables



$\mathbf{Sample}_{\text{pilot}}$,



$\mathbf{Sample}_{\text{follow-up}}$,



$\mathbf{Sample}_{\text{target}} \in \mathbb{R}_{\geq 0}^{\text{Groups}}$

Probabilistic Effect Model

$$\mathbf{CATE} \sim \mathcal{P}$$

prior on group effects

$$\widehat{\mathrm{ATE}}_t \sim \mathcal{N}\left(\mathrm{ATE}_t, \frac{1}{\mathbf{1}^\top \mathbf{Sample}_t}\right)$$

$$\mathrm{ATE}_t = \frac{\mathbf{Sample}_t^\top \mathbf{CATE}}{\mathbf{1}^\top \mathbf{Sample}_t}$$

$$\mathrm{Test}_t = \mathbb{I}\left\{\widehat{\mathrm{ATE}}_t > \frac{z_{1-\alpha_t}}{\sqrt{\mathbf{1}^\top \mathbf{Sample}_t}}\right\}$$

Impact Objective

$$\underset{\mathbf{Sample}_{\text{pilot}}}{\text{maximize}} \quad \mathbb{E}\left[\text{Test}_{\text{pilot}} \cdot \text{Test}_{\text{follow-up}} \cdot \text{ATE}_{\text{target}}\right]$$

$$\text{subject to} \quad \mathbf{Cost}^{\top}\mathbf{Sample}_{\text{pilot}} \leq \text{Budget}$$

Main Result – Optimal Small-Budget Design

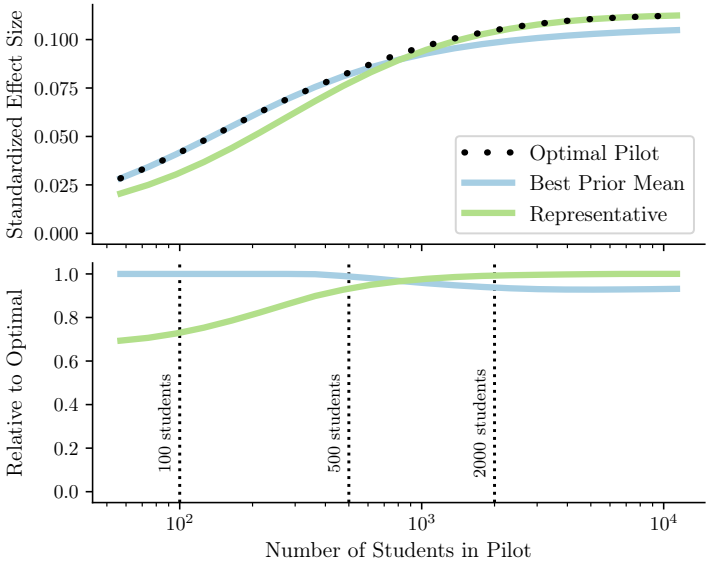
Theorem. Suppose prior $\mathcal{P}$ has finite fourth moments. Assume a unique group $g^\star$ maximizes the *small-budget index*

$$\frac{\mathbb{E}\left[\mathrm{CATE}_g \cdot \mathrm{Test}_{\text{follow-up}} \cdot \mathrm{ATE}_{\text{target}}\right]}{\sqrt{\mathrm{Cost}_g}}.$$

Then for all sufficiently small budgets, a unique optimal pilot exists and invests the entire budget in group $g^\star$.

Corollary. If the prior $\mathcal{P}$ is elliptical, the small-budget index equals

$$\frac{W_1 \cdot \mathbb{E}[\mathrm{CATE}_g] + W_2 \cdot \mathrm{Cov}\left(\mathrm{CATE}_g, \mathrm{ATE}_{\text{follow-up}}\right) + W_3 \cdot \mathrm{Cov}(\mathrm{CATE}_g, \mathrm{ATE}_{\text{target}})}{\sqrt{\mathrm{Cost}_g}}.$$



Calibrated Case Study

6 Groups
Equal Costs
Different Prior Means
Medium Correlation

Yeager et al. (2019).
A national experiment reveals where a growth mindset improves achievement.
Nature, 573(7774), 364–369.

Generalized Entropy Calibration for Inference with Partially Observed Data: A Unified Framework

Mst Moushumi Pervin[1], Hengfang Wang[2], and Jae Kwang Kim[3]

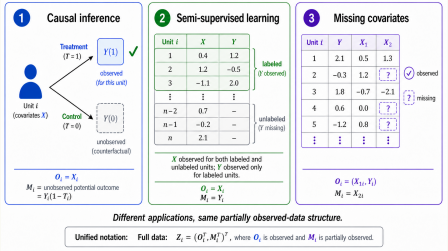
[1, 3]Department of Statistics, Iowa State University

[2]School of Mathematics and Statistics, Fujian Normal University

IMSI Presentation, 2026

June 23, 2026

Motivating Example



Research Question in Semi-supervised setting:

How can we efficiently integrate labeled and unlabeled data while correcting for the selection bias?

Research Question in Causal inference setting:

How can we infer causality and reduce the selection bias in the presence of imbalanced covariates?

Objective and Limitations of the Study

- Objective of the study: Develop a single flexible calibration weighting framework which can handle a broad class of incomplete-data problems- semi-supervised learning, regression with missing covariates, and causal inference.

Limitations of Existing Approaches:

- Both Outcome regression (OR) and propensity score (PS) models are vulnerable to model misspecification.
- Doubly robust (DR) methods, such as augmented inverse probability weighting (AIPW; Robins et al. (1994)), combine OR and PS models but can be unstable due to extreme weights and efficiency depends on how well the correction term is modeled.

Basic Set Up

- **Goal:** estimate a target parameter θ defined by $\mathbb{E}\{U(\theta; Z)\} = 0$, where $Z = (O^\top, M^\top)^\top$ and only O is fully observed. Here, $U(\theta; Z)$ is a user-specified estimating function.
- In the absence of missing data, the parameter vector θ can be obtained by solving:

$$\hat{U}_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} U(\theta; z_i) = 0, \tag{1}$$

- A class of estimators are obtained by solving:

$$\hat{U}_\omega(\theta) \equiv \sum_{i=1}^{n} \delta_i \, \hat{\omega}_i \, U(\theta; z_i) = 0, \tag{2}$$

Proposed Method

The calibrated weights $\hat{\omega}_i$ are obtained by solving the following GEC program.

Generalized entropy calibration (GEC) program: For a strictly convex, differentiable function G with $g = G'$,

$$\min_{\omega_1,\ldots,\omega_n} \sum_{i=1}^{n} \delta_i \, G(\omega_i)$$

$$\text{subject to} \quad \underbrace{\sum_{i=1}^{n} \delta_i \, \omega_i \, b(\theta; O_i) = \sum_{i=1}^{n} b(\theta; O_i)}_{\textbf{Balancing: uses auxiliary info}}$$

$$\underbrace{\sum_{i=1}^{n} \delta_i \, \omega_i \, g(\hat{\pi}_i^{-1}) = \sum_{i=1}^{n} g(\hat{\pi}_i^{-1})}_{\textbf{Debiasing: uses PS info (proposed by Kwon et al. (2025))}}$$

Choice of Entropy and Balancing function

Choice of entropy G:

$G(\omega)$	Name	Domain
$\omega \log \omega - \omega$	Exponential tilting (**ET**)	$(0, \infty)$
$-\sqrt{\omega}$	Hellinger distance (**HD**)	$(0, \infty)$

- **The calibration function $b(\theta; O_i)$:** The optimal choice is the conditional mean

$$b^*(\theta; O_i) = \mathbb{E}\{U(\theta; Z_i) \mid O_i\}.$$

- In practice, we approximate b^* by predicting the missing component M_i from O_i and plugging into U:

$$b^*(\theta; O_i) \approx U(\theta; O_i, \hat{M}_i),$$

where $\hat{M}_i$ is obtained via cross-fitted machine learning Angelopoulos et al. (2023).

Main Theoretical Result: Same Robustness, Higher Efficiency

- **Same robustness:** GEC keeps the doubly robust property of AIPW. It is consistent if either the OR model or the PS model is correct, and it is locally efficient when both models are correct.
-

Main efficiency result under correct PS:

$$V_{\mathrm{GEC}} \leq V_{\mathrm{AIPW}}$$

GEC can achieve smaller variance than AIPW.

Example Application: Average Treatment Effect

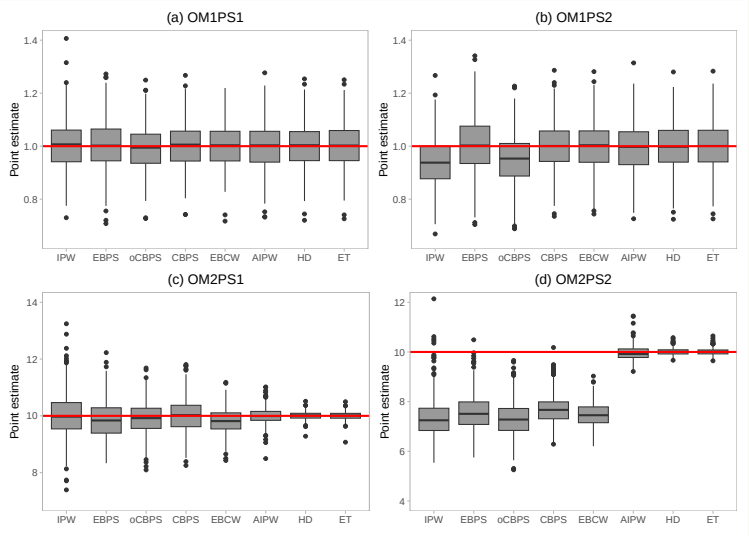
Mapping to the GEC framework:

$O_i = X_i$	observed covariates
$M_i = Y_i(t)$	potential outcome under treatment arm t
$\delta_i = \mathbf{1}(T_i = t)$	indicator that $Y_i(t)$ is observed
$U_t(\theta_t; X_i, Y_i(t)) = Y_i(t) - \theta_t$	estimating function for mean outcome

GEC estimator: For each treatment arm $t \in \{0, 1\}$, apply GEC within that arm to obtain weights $\hat{\omega}_{ti}$:

$$\hat{\theta}_t = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(T_i = t)\, \hat{\omega}_{ti}\, Y_i, \qquad \widehat{\mathrm{ATE}} = \hat{\theta}_1 - \hat{\theta}_0.$$

Results: Estimation of ATE



Estimation of Average Treatment Effects (ATE)

- OM1(2): outcome model correctly specified (misspecified).
- PS1(2): propensity score model correctly specified (misspecified).
- Proposed ET and HD estimators outperform existing methods, especially under OM2PS1 and OM2PS2.

Real-Data Illustration: LaLonde Job Training Study

Goal: Use observational controls (PSID, CPS) to recover the NSW experimental treatment-effect benchmark.

Evaluation Bias (EB) = estimate − NSW benchmark. Smaller |EB| means closer agreement with the randomized experiment.

Method	**PSID EB**	**CPS EB**
Unweighted	−17000	−10292
IPW	−320	−730
AIPW	−511	−503
ET	**−139**	**−179**

Key Takeaway

- **Calibration can do more than balance covariates.**
- GEC carefully chooses weights so the observed data better represent the full data.
- This gives one unified framework for several incomplete-data problems.
- **Compared with AIPW, GEC can be more efficient while preserving double robustness.**

QR code: Full Paper

References I

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023). Prediction-powered inference. *Science*, 382(6671):669–674.

Kwon, Y., Kim, J. K., and Qiu, Y. (2025). Debiased calibration estimation using generalized entropy in survey sampling. *Journal of the American Statistical Association*, pages 1–12.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

TransPFN

Transportability-Inspired Causal Meta-Learning for OOD-Robust Causal Effects

Yonghan Jung
University of Illinois Urbana-Champaign

IMSI New Horizons on Model Transportability and Data Integration

Fixed-prior causal meta-learners are OOD-fragile

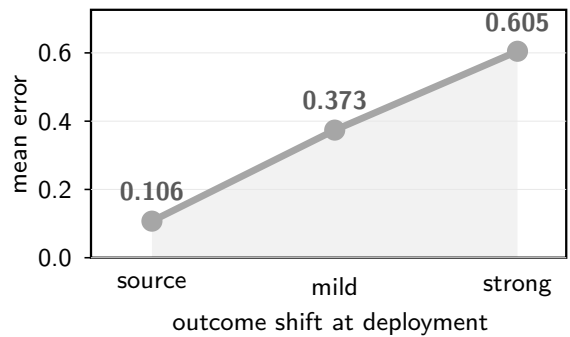
Causal Meta-Learner

$$(D \sim P, x, a) \;\mapsto\; \theta_a := \mathbb{E}_P[Y \mid \mathrm{do}(a), x].$$

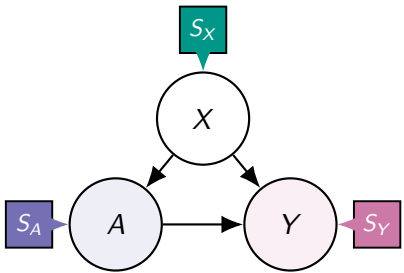
A fixed DGP generator π produces many (D, x, a, θ_a) episodes, and the learner fits this map.
Under ignorability and positivity: $\theta_a = \mathbb{E}_P[Y \mid A = a, X = x]$.

Fixed-prior learner breaks under OOD.

OOD = out of support of the DGP generator π.



OOD can be explained with selection nodes



$$P(X, A, Y) = P(X)\, P(A \mid X)\, P(Y \mid X, A).$$

S_X: covariate	S_A: treatment	S_Y: outcome
$P_\pi(X) \neq P^\star(X)$	$P_\pi(A \mid X) \neq P^\star(A \mid X)$	$P_\pi(Y \mid X, A) \neq P^\star(Y \mid X, A)$

$$\left\| \widehat{\theta}_a^\pi - \theta_a^\star \right\| \leq \underbrace{E_{\text{real}}}_{\text{finite learner}} + \underbrace{E_{\text{tr}}}_{\text{outcome mechanism}} + \underbrace{E_{\text{cov}}}_{\text{uncovered support}} .$$

finite: limited model and finite pretraining

outcome: S_Y affects coverage and transportability bias

uncovered: S_X and S_A affect coverage

TransPFN: keep the baseline, route the residual

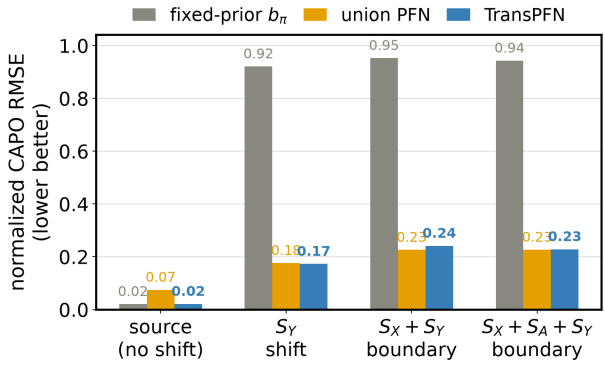
$$\widehat{\theta}(D, x, a) = b_\pi(D, x, a) + \sum_{s \neq s_0} q_S(s \mid D)\, r_s^\star(D, x, a).$$

b_π	$q_S(s \mid D)$	$r_s^\star$
off-the-shelf source baseline	selection router	small residual learner

$$r_s^\star \in \arg\min_r \left\| \theta_a^\star - b_\pi(D, x, a) - r(D, x, a) \right\|^2.$$

Lean guarantees: no-shift protection ($\widehat{\theta} \approx b_\pi$); in-library improvement when route/residual error is below the fixed-prior floor.

Semi-synthetic IHDP: source-safe OOD repair



IHDP covariates with constructed selection shifts; metric: normalized CAPO RMSE, lower is better.

- Union PFN: a PFN trained on the larger domain, without selection route q_S.
- Fixed-prior breaks under IHDP selection shift; TransPFN keeps source close and repairs OOD.