Soft-Constrained Bregman Calibration

Min-Yi Chen[1]
Joint work with Jae-Kwang Kim[1] and Yumou Qiu[2]

[1]Department of Statistics, Iowa State University
[2]School of Mathematical Sciences, Peking University

June 25, 2026

Motivation: Why Not Exact Calibration?

Calibration weighting is designed to correct imbalance using known auxiliary information.

Classical calibration enforces selected moments exactly:

$$\frac{1}{N}\sum_{i=1}^{N}\omega_i U_i(\theta) = 0.$$

This treats every selected moment as equally important. In practice, this can be too restrictive when the auxiliary information is high-dimensional or only partially reliable.

- Exact balance can create extreme weights and inflate variance.
- Exact balance on many moments may be infeasible or unstable.
- Some useful moments may be better controlled approximately.

Soft Calibration Principle

The key idea is to separate moments into two groups:

$$U_1 = \text{hard moments}, \qquad U_2 = \text{soft moments}.$$

Hard moments are imposed as exact calibration constraints:

$$\frac{1}{N}\sum_{i=1}^{N} \omega_i U_{1i}(\theta) = 0.$$

Soft moments are controlled through a penalty:

$$\frac{1}{2\rho}\left\|\frac{1}{N}\sum_{i=1}^{N} \omega_i U_{2i}\right\|^2.$$

The tuning parameter ρ controls the strength of the soft balance:

$$\rho \downarrow 0 \Rightarrow \text{nearly hard balance}, \qquad \rho \uparrow \infty \Rightarrow \text{soft moments are ignored}.$$

Setting

Assume we observe $(X_i, Y_i), \quad i = 1, \ldots, N$, where

$$Y_i \in \mathbb{R}^p, \qquad X_i = \begin{bmatrix} X_i^{(1)} \\ X_i^{(2)} \end{bmatrix}.$$

The target is $\theta_0 = \mathbb{E}(Y)$.
Known auxiliary means:

$$\mu_X^{(1)} = \mathbb{E}(X^{(1)}), \qquad \mu_X^{(2)} = \mathbb{E}(X^{(2)}).$$

Define

$$U_1(\theta; X, Y) = \begin{bmatrix} Y - \theta \\ X^{(1)} - \mu_X^{(1)} \end{bmatrix}, \qquad U_2(X) = X^{(2)} - \mu_X^{(2)}.$$

Estimator:

$$\hat{\theta} = \frac{\sum_{i=1}^N \hat{\omega}_i Y_i}{\sum_{i=1}^N \hat{\omega}_i}.$$

Soft-Constrained Bregman Calibration

We consider Bregman Calibration[1] with soft calibration.

$$\hat{\theta} = \arg\min_{\theta} \arg\min_{\omega} \left[\frac{1}{N} \sum_{i=1}^{N} D_G(\omega_i \| 1) + \frac{1}{2\rho} \left\| \frac{1}{N} \sum_{i=1}^{N} \omega_i U_{2i} \right\|^2 \right], \qquad (1)$$

subject to

$$\frac{1}{N} \sum_{i=1}^{N} \omega_i U_{1i}(\theta) = 0.$$

with Bregman divergence

$$D_G(\omega_i \| 1) = G(\omega_i) - G(1) - G'(1)(\omega_i - 1).$$

[1] Jae Kwang Kim, Yonghyun Kwon, and Yumou Qiu (2026). "Bregman projection for calibration estimation". In: *arXiv preprint arXiv:2603.20780*.

Dual Representation

Theorem (Dual of the soft Bregman calibration problem)

For every $\rho \in (0, \infty)$, the Fenchel dual of the fixed-θ soft-constrained Bregman calibration problem (1) is $\sup_{\lambda_1, \lambda_2} Q_{N,\rho}(\theta, \lambda)$, where

$$Q_{N,\rho}(\theta, \lambda) = -\frac{1}{N} \sum_{i=1}^{N} F\left(\lambda_1^\top U_{1i}(\theta) + \lambda_2^\top U_{2i}\right) - \frac{\rho}{2}\|\lambda_2\|^2,$$

and $F = G^*$ is the convex conjugate of G.

Our estimator can written as

$$\hat{\theta} = \arg\min_{\theta} \sup_{\lambda_1, \lambda_2} Q_{N,\rho}(\theta, \lambda)$$

Simulation Illustration

We use 1000 Monte Carlo replications with $N = 300$:

$$X_i \sim N(0,1), \qquad \varepsilon_i \sim N(0,1),$$

$$Y_i = X_i + X_i^2 + \varepsilon_i, \qquad \theta_0 = \mathbb{E}(Y) = 1.$$

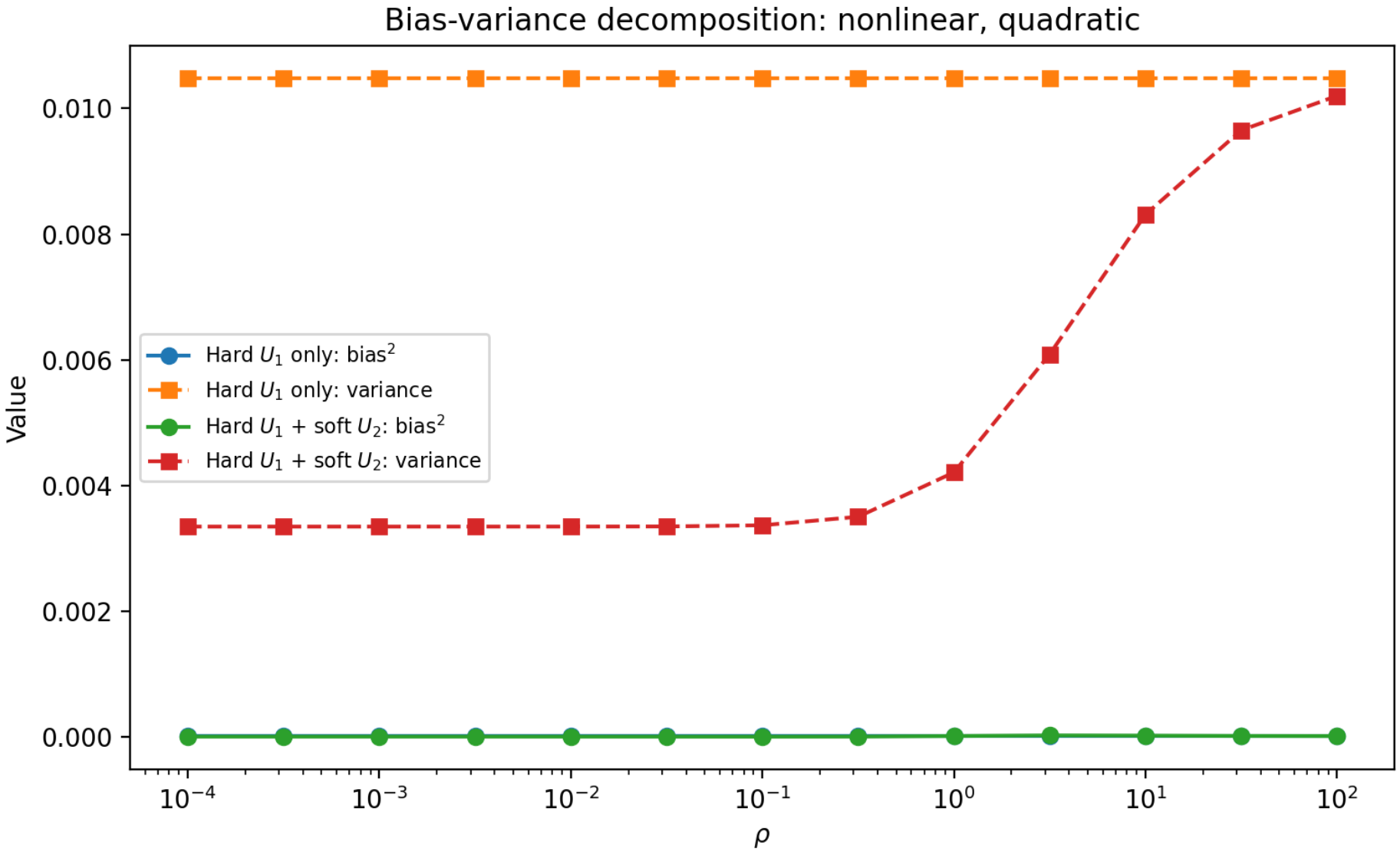
Hard moments:

$$U_{1i}(\theta) = \begin{bmatrix} Y_i - \theta \\ X_i \end{bmatrix}.$$

Soft moment:

$$U_{2i} = X_i^2 - 1.$$

- Small ρ: close to hard calibration on U_1 and U_2.
- Large ρ: close to hard-only calibration on U_1.
- Useful U_2 can reduce variance and MSE.


Bias-variance decomposition: nonlinear, quadratic
Value
0.010
0.008
0.006
0.004
0.002
0.000
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
ρ
Hard U_1 only: bias^2
Hard U_1 only: variance
Hard U_1 + soft U_2: bias^2
Hard U_1 + soft U_2: variance

References I

Kim, Jae Kwang, Yonghyun Kwon, and Yumou Qiu (2026). "Bregman projection for calibration estimation". In: *arXiv preprint arXiv:2603.20780*.

Integrating multi-omics data through deep learning for prioritizing genes associated with autism

Shengtong Han, PhD

Marquette University School of Dentistry

June 25th , 2026, IMSI workshop

Joint work with Dr. Di Wang, Dr. Wenhui Sheng, and Nathanael Chu

ASD is highly inheritable

JAMA Network Click to view article >

JAMA. 2017 Sep 26;318(12):1182–1184. doi: 10.1001/jama.2017.12141

The Heritability of Autism Spectrum Disorder

Sven Sandin [1,✉], Paul Lichtenstein [2], Ralf Kuja-Halkola [2], Christina Hultman [2], Henrik Larsson [3], Abraham Reichenberg [1]

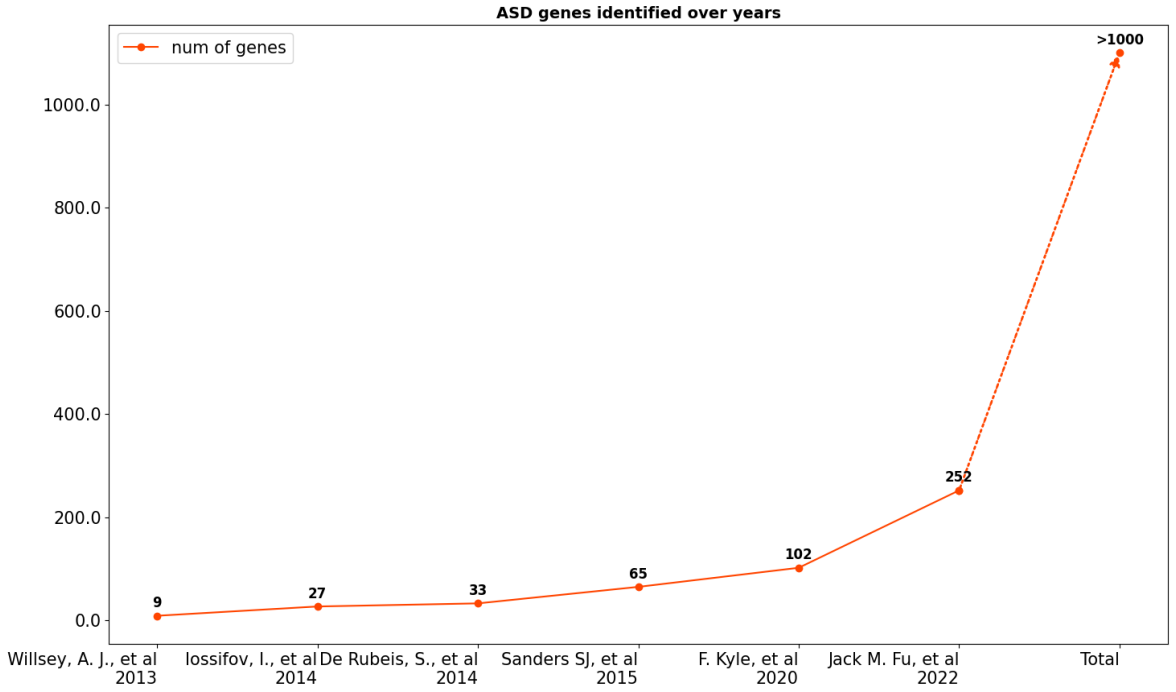
Author information · Article notes · Copyright and License information

PMCID: PMC5818813 PMID: 28973605

Abstract

This study reanalyzes Swedish cohort data to assess the stability under alternative assumptions and models of a previous estimate of the heritability of autism spectrum disorder.

Studies have found that autism spectrum disorder (ASD) aggregates in families, and twin studies estimate the proportion of the phenotype variance due to genetic factors (heritability) to be about 90%.



Current integrative methods for ASD

PLOS Genetics

Browse Publish About

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Integrated Model of *De Novo* and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes

Xin He, Stephan J. Sanders, Li Liu, Silvia De Rubeis, Elaine T. Lim, James S. Sutcliffe, Gerard D. Schellenberg, Richard A. Gibbs, Mark J. Daly, Joseph D. Buxbaum, Matthew W. State, Bernie Devlin, Kathryn Roeder

Published: August 15, 2013 • https://doi.org/10.1371/journal.pgen.1003671

Article | Authors | Metrics | Comments | Media Coverage

Molecular Autism BMC

Mol Autism. 2014 Mar 6;5:22. doi: 10.1186/2040-2392-5-22

DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics

Li Liu [1], Jing Lei [1], Stephan J Sanders [2,3], Arthur Jeremy Willsey [2,3], Yan Kou [4,5], Abdullah Ercument Cicek [6], Lambertus Klei [7], Cong Lu [1], Xin He [6], Mingfeng Li [8,9], Rebecca A Muhle [3,8,10], Avi Ma'ayan [5], James P Noonan [3,8], Nenad Šestan [8,9], Kathryn A McFadden [11], Matthew W State [2,3,9,12,13], Joseph D Buxbaum [4,14], Bernie Devlin [7], Kathryn Roeder [1,6,✉]

Author information • Article notes • Copyright and License information

PMCID: PMC4016412 PMID: 24602502

scientific reports

Explore content About the journal Publish with us

nature > scientific reports > articles > article

Article | Open access | Published: 12 March 2020

Forecasting risk gene discovery in autism with machine learning and genome-scale data

Leo Brueggeman, Tanner Koomar & Jacob J. Michaelson ✉

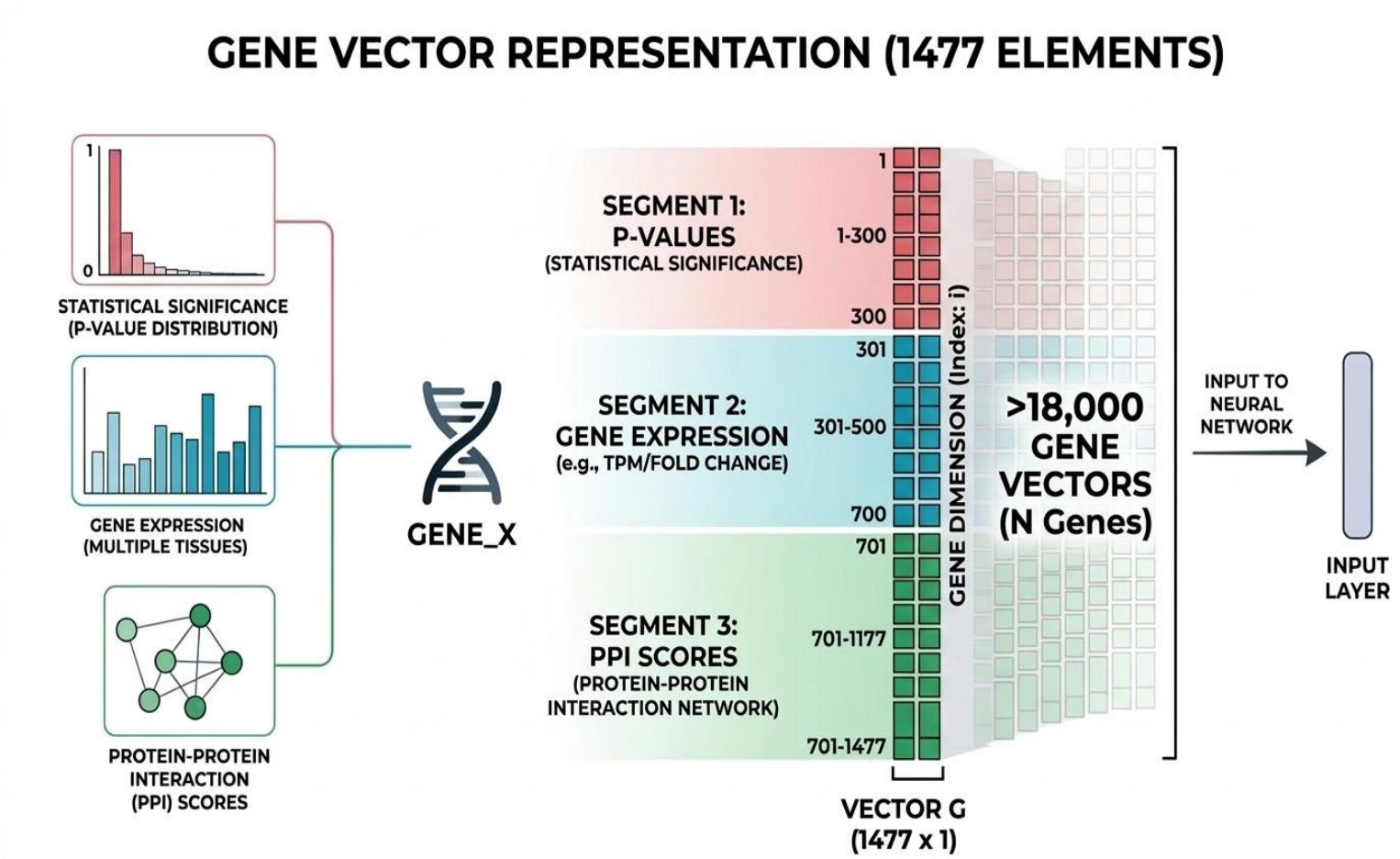
Scientific Reports **10**, Article number: 4569 (2020) | Cite this article

Challenges

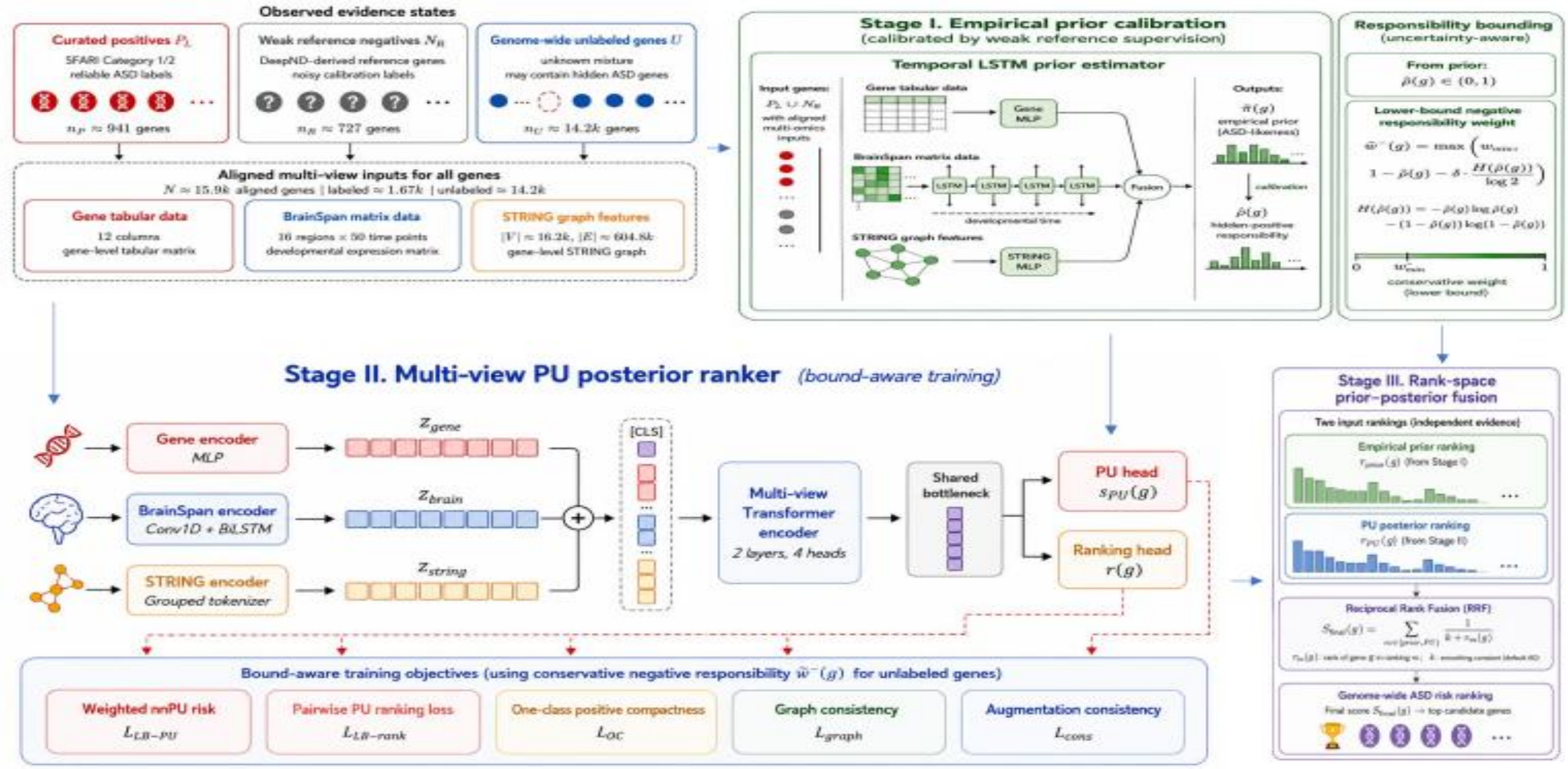
Challenges

1. Input data are measured at different biological level, i.e. variant, gene, protein

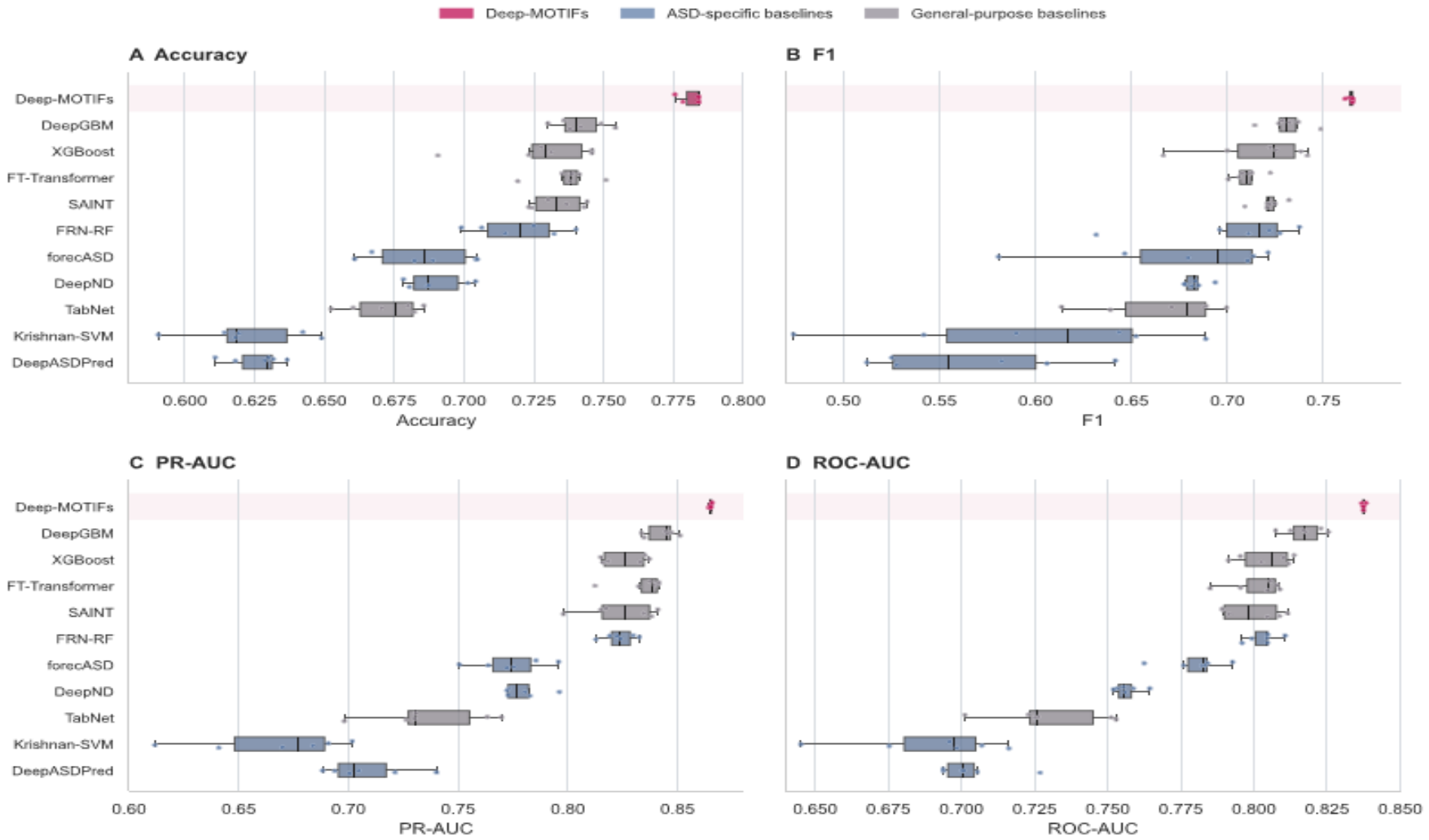
2. Each gene (>18,000 genes) was characterized by a vector of 1477 elements with both numerical and categorical values-> high dimensional tabular data



Deep learning framework for Multi-Omics inTegratIon For autiSm (Deep-MOTIFs)



Deep-MOTIFs outperforms many other methods



Summary

- A deep learning framework was proposed to integrate multi-omics data, not only for autism, but also for other human traits and diseases
- More autism novel genes are identified, supported by autism related features and pathways

Future research:

- Incorporate epigenetic evidence to further boost the power
- Validate functional effect of top novel genes by knockdown in mouse models

Why Representative Learning (RepL)?

Distributed Learning: The Reality

- Data distributed across nodes: $\mathcal{D}_k = \{(x_i, y_i)_{i \in I_k}\}$
- Goal: minimize global loss over decentralized data
- Core challenges:
  - Data/model heterogeneity
  - Communication latency

Federated Learning

Communicates models parameters

- Coupled to model structure
- Sensitive to heterogeneity
- Model drift & slow convergence

Key Idea: Representative Learning

Replace parameter sharing with data-centric communication

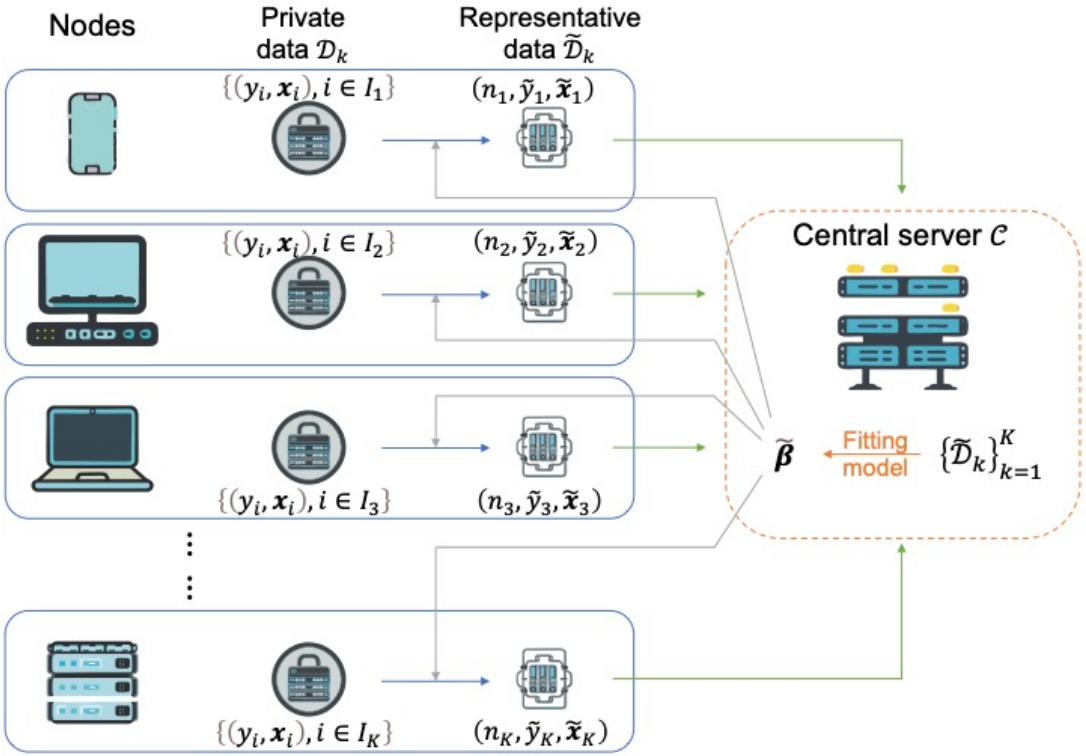
- Each node constructs representative

$$\mathcal{D}_k \Rightarrow \widetilde{\mathcal{D}}_k = (n_k, \tilde{x}_k, \tilde{y}_k)$$

  - Same format as original data → plug into any model
  - Compact → low communication cost
  - Structured → interpretable & traceable
  - Naturally adapts to heterogeneity

Keren Li
University of Alabama at Birmingham

Algorithmic Framework & Methods



Communication unit = data summaries, not parameters

General RepL Workflow

1. Partition local data
2. Construct representatives (pseudo-data)
3. Share representatives across nodes
4. Train/update model on combined representatives/local data

Core Representative Constructions

- MR/TMR
  - Matches variable means
  - Approximates full-data estimator with controlled error
- SMR/Anchored-SMR
  - Matches local score function (gradient)
  - Preserves likelihood structure
  - Geometric convergence under convexity
  - Stable with bounded representative error

Ongoing

Beyond Smooth Objectives

- Develop clouded representatives $\left(n_k, \tilde{y}_k, \tilde{\boldsymbol{x}}_{\boldsymbol{k}}, \tilde{\Sigma}_k\right)$
- Distributed SVM (in preparation)
- Robust learning

Functional Data Analysis

- Representatives summarize infinite-dimensional objects
  - curves, time series, spatial processes

Adaptive & Collaborative Learning

- Across nodes with heterogeneity awareness
- Links to collaborative transfer learning paradigms

Deep Learning & Neural Networks

…

Future Directions

Distributed Unsupervised Learning

Toward a General Distributed AI Framework

External Validity for Social Inquiry

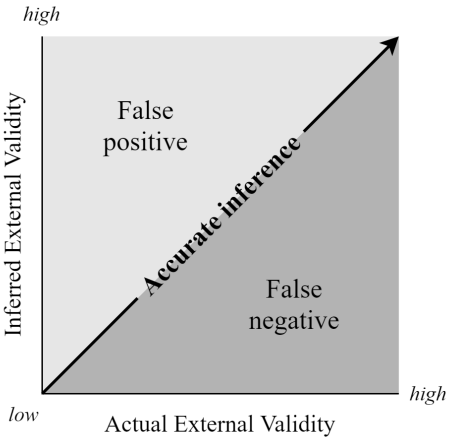
Mike Denly (Texas A&M)
joint with Michael G. Findley (UT Austin) & Kyosuke Kikuta (IDE-JETRO)

Institute for Mathematical and Statistical Innovation (IMSI)

New Horizons on Model Transportability & Data Integration Workshop

mdenly@tamu.edu

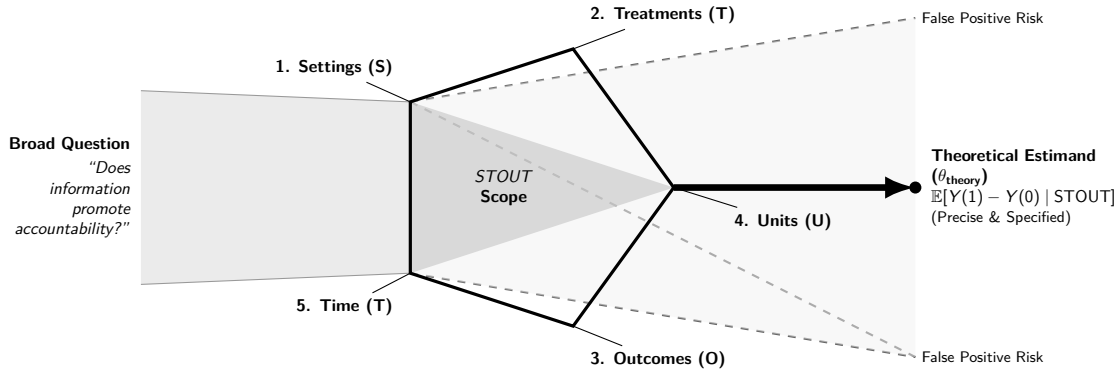
The Problem: External Validity (EV) False Positives *and* False Negatives



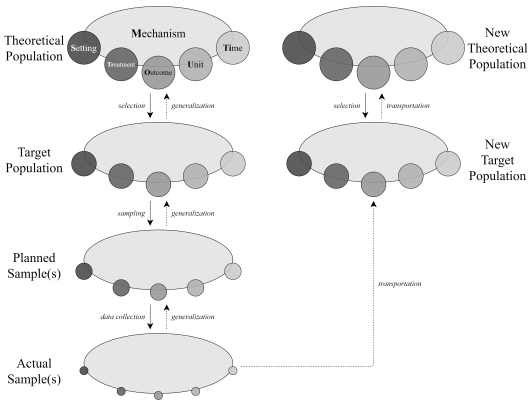
Systematic review of 1,005 articles:

- top journals in econ, poli sci, psych, and soc
- $<$50% make even a minimal EV inference
- only 2.5% include a dedicated EV section

Problem → Solution: From Vague Questions → Theoretical Estimands



Solution: Mechanisms First (M-STOUT)



We reorganize the classical UTOS framework, placing **M first**: mechanisms explain *why* and *how far* findings travel.

- **Observable mechanisms** take the form of modifiers — moderators (pre-treatment) and mediators (post-treatment)
- **Generalizability:** within the data-generating process
- **Transportability:** across different eligibility criteria

Solution: Diagnose 3 Components of EV Bias

$$\underbrace{\Delta}_{\text{EV Bias}} = \underbrace{\Delta_S}_{\text{Setting}} + \underbrace{\Delta_{UT}}_{\text{Unit-Time}} + \underbrace{\Delta_{TO}}_{\text{Treatment-Outcome}}$$

- Δ_S: modifiers differ between sample and target
- Δ_{UT}: sample-vs-target population mismatch — **wrong rows**
- Δ_{TO}: construct mismatch with the theoretical estimand — **wrong columns**

Key takeaway: Reweighting mostly fixes wrong rows. *Less work on wrong columns.*

Solution: Three Evaluative Criteria

Earn external validity inference sequentially:

1. **Model Utility (MU)**
   - Theoretical estimand
   - Mechanism
   - Identifiability
2. **Scope Plausibility (SP)**
   - Target population
   - Planned sample
   - Actual sample
3. **Specification Credibility (SC)**
   - Population-level estimate
   - Uncertainty
   - Abductive prediction

Failures occur when one criterion pretends the previous did its job.

Semi-Supervised Causal Mediation Analysis with Partially Observed Mediators

Tomoki Okuno

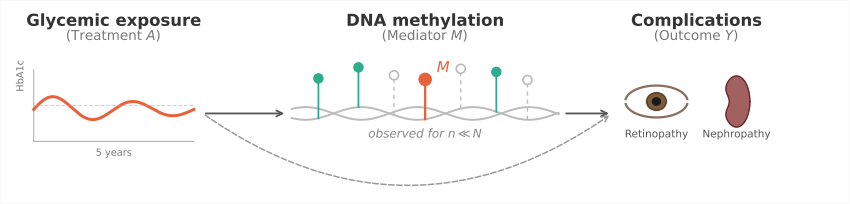
joint work with Gang Li, Sijia Li, Hua Zhou, and Jin Zhou

IMSI · New Horizons on Model Transportability and Data Integration June 25, 2026

Department of Biostatistics, University of California, Los Angeles

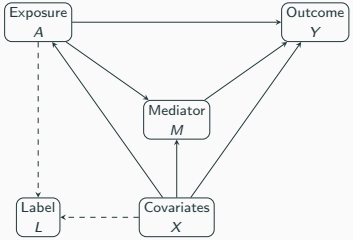
Motivation

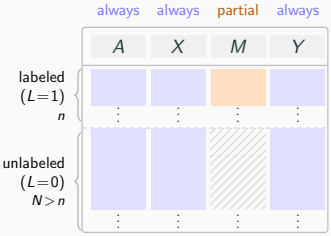
1. We want the natural indirect effect of glycemic exposure on complications through DNA methylation. But methylation is costly, so the mediator is observed only for a subset, while treatment, covariates, and outcomes are available for everyone.



2. Prediction-powered inference (PPI) is a natural tool for unlabeled data, but standard PPI targets a missing outcome. The missing-mediator case is open.

Setting and goal





Under standard identification and MAR labeling, $L \perp\!\!\!\perp (M, Y) \,|\, (A, X)$.

Research question

How can we estimate the natural direct and indirect effects (NDE, NIE) by efficiently integrating labeled and unlabeled data, while preserving valid inference?

$$\mathrm{NDE} = \mathbb{E}[Y(1, M(0))] - \mathbb{E}[Y(0, M(0))], \quad \mathrm{NIE} = \mathbb{E}[Y(1, M(1))] - \mathbb{E}[Y(1, M(0))] \ .$$

Define $\theta_0 = \mathbb{E}[Y(1, M(0))]$.

From full-data EIF to missing-data AIPW

Full-data EIF

If M were observed for all, $\Gamma(A, X, M, Y) - \theta_0$ is the known full-data efficient influence function (EIF) for θ_0 [Tchetgen & Shpitser, 2012].

➤

But Γ needs M

In the observed data M is available only when $L = 1$, so $\Gamma(A, X, M, Y)$ is missing for unlabeled units.

➤

Missing-data problem

Under MAR labeling, $\pi(A, X) = \mathbb{P}(L = 1 \mid A, X)$, estimating $\theta_0 = \mathbb{E}(\Gamma)$ is a missing-data problem for Γ.

Missing-data AIPW, unbiased for any working $h(W)$

$$\theta_0 = \mathbb{E}\Big[h(W) + \frac{L}{\pi(A, X)}\{\Gamma - h(W)\}\Big].$$

The remaining question is how to choose $h(W)$.

Two estimators

Projection (EIF-based)

$h(W) = \mathbb{E}[\Gamma \mid W]$

- Targets the efficient projection directly, via extra nuisances.
- ✓ Attains the semiparametric bound, robust to the labeling model.
- ✗ *Practically*, more involved to compute.

Imputation (PPI-based)

$h(W) = \Gamma(A, X, \widehat{M}, Y), \quad \widehat{M} = f(W)$

- Plugs a predicted mediator into the score.
- ✓ Avoids the projection nuisance. Can use external mediator prediction.
- ✗ Below the bound and loses that robustness.

  Under MCAR, π is not estimated, so this robustness issue disappears.

Asymptotically, $V^{\mathrm{proj}} = V^{\mathrm{eff}} \leq V^{\mathrm{imp++}} \leq V^{\mathrm{cc}}$.

Oracle Mediator Imputation Still Misses the Bound

	Mediation PPI (ours)	**Outcome PPI** (standard)
Missing variable	Mediator M	Outcome Y
Score in the imputed variable	**Nonlinear**	Affine
Oracle plug-in attains bound?	**No**, Jensen gap	Yes
Robust to labeling model (MAR)	**No, even at oracle**	Oracle only
Attains efficiency bound	**No, even at oracle**	Oracle only

Westat®

Improving Lives Through Research®

Sampling and Weighting Strategies for Nonprobability Survey and Serology Data

Laura Gamble[1], Elizabeth Eisenhauer[1], Andreea Erciulescu[1], Rebecca Fink[1], Sarah Smith-Jeffcoat[2], Jefferson Jones[2], Bryan Spencer[3], Eduard Grebe[4], Mars Stone[4], Michael Busch[4]

[1]Westat, [2]Centers for Disease Control and Prevention, [3]American Red Cross, [4]Vitalant Research Institute

June 25, 2026

New Horizons on Model Transportability and Data Integration

Routine surveillance misses many infections

- **Respiratory virus surveillance** is important for public health preparedness.
- **Infections may be missed** when symptoms are mild or absent, or untested.
- **Repeat blood donors** can provide blood samples and survey information over time.

Blood donors are useful, but not a probability sample of the U.S. population

Useful because:

- Repeat blood samples can capture **infection-induced boosting** of antibodies.
- Donors also provide **survey information** from across the United States over time.

Challenging because:

- Repeat donors are a **nonprobability cohort**.
- Demographic and geographic structure may **differ from the target population.**

Sampling and weighting support more generalizable estimates

Sampling: Two-phase probability sampling balanced demographic and regional **representation** goals with **feasibility** for repeated serology testing.

Weighting: Adjustments to account for **sampling, cohort participation, survey response, serology testing, and remaining differences** from population control totals.

Cohort: 25,255 repeat blood donors invited to quarterly surveys.

Serology subset: 11,202 repeat blood donors selected from the cohort for multipathogen antibody testing.

Visit the poster to see how sampling and weighting were integrated

- Our sampling and weighting design addressed **selection bias** and **harmonized geographic and demographic structures** in a nonprobability donor cohort.
- **Weighting adjustments**, while expected to reduce bias, also increased variation.
- **Estimation is ongoing**. Final weighted estimates will describe respiratory virus infection patterns nationally, regionally, and by demographic group.

Thank you

ElizabethEisenhauer@westat.com

Poster #8

westat.com

A General Framework for Inference After Record Linkage

Priyanjali Bukke* Martin Slawski*

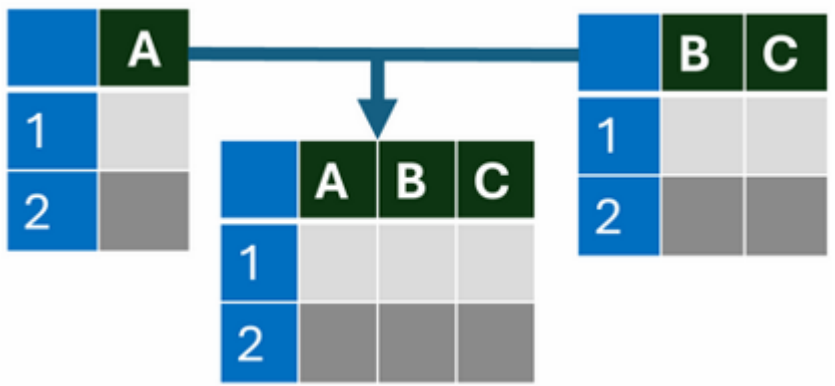
Department of Statistics,

University of Virginia

*Supported NSF grants SES-2120318 and OAC-2411270

Record linkage (RL) is micro-level data integration.

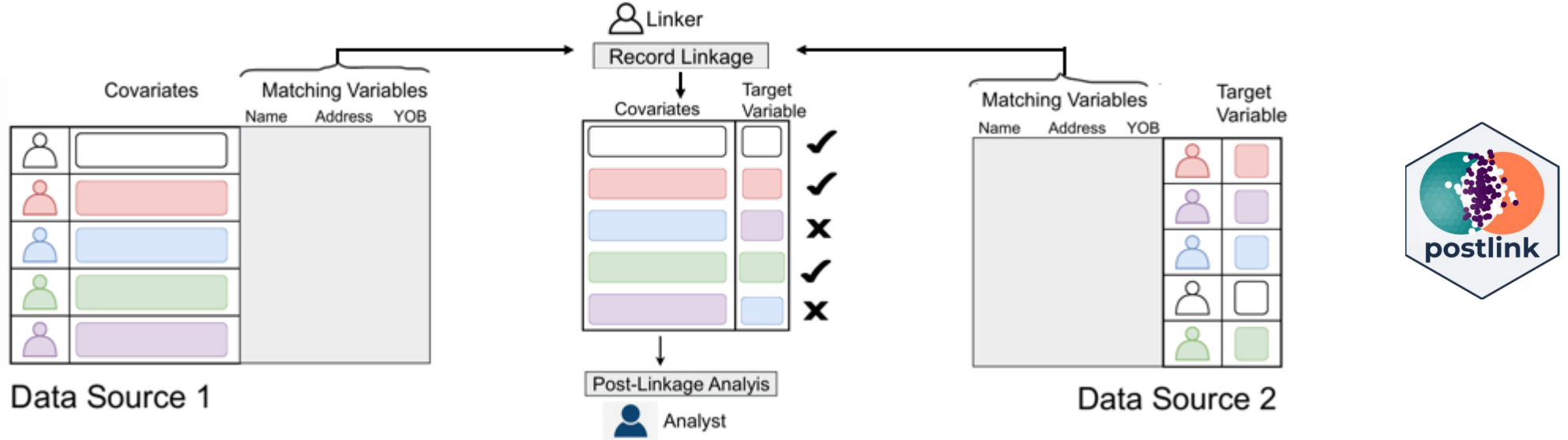
- Data integration is the process of combining multiple existing data sources that contain complementary pieces of information.
- **RL** links records of the same entity across multiple sources.



- RL can be error-prone (*mismatches* or *missed matches*).
- RL can pose privacy concerns.
- Various software packages exist in R and Python for RL.

Post-Linkage Data Analysis (PLDA) Should Account for Linkage Errors for Valid Statistical Inference.

- PLDA can broadly be divided into *primary* and *secondary* settings.

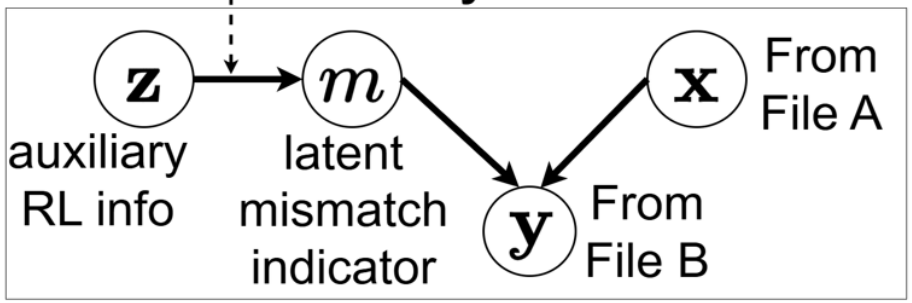


- Linkage error can invalidate statistical inferences.
- Handling linkage vs. mismatch rates is often not straightforward.
- Adjustment methods account for linkage errors and software is being developed in Python and R (**`postlink`**).

Inference with an imperfectly linked dataset can be conducted using a general, extensible framework.

- The framework is based on a two-component mixture model.

Examples:
- Mismatch error rate
- Predictors of match status
- Safe matches
- Intercept-Only



Assumptions:

(A1). $\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{x} | \, m = 1$,

(A2). The mismatch error does not depend on $(\boldsymbol{x}, \boldsymbol{y})$. *Note: This assumption can be relaxed.*[2]

(A3). $P(m_i = 0 | \boldsymbol{z}_i) = h(\boldsymbol{z}_i; \gamma_i)$, $1 \leq i \leq n$, unknown h & γ.

$$\mathbf{y} | \{m = 1\}, \mathbf{x} \sim f(\mathbf{y}) \qquad \mathbf{y} | \{m = 0\}, \mathbf{x} \sim \phi(\mathbf{y} | \mathbf{x}; \theta)$$

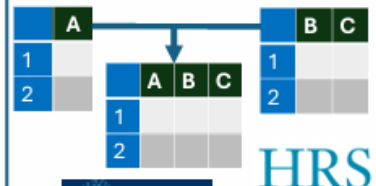
- Inference is via asymptotic theory for composite MLEs.
- Linear regression demonstrated using the **`postlink`** R package.

A General Framework for Inference After Record Linkage

Priyanjali Bukke Martin Slawski Department of Statistics, University of Virginia

Record Linkage (RL)

Record Linkage – link same entity's data across multiple sources



- ❑ Integrate unit-level information across multiple data sets
- ❑ Applicable to large existing and growing volumes of data

LIFE-M — Longitudinal, Intergenerational Family Electronic Micro-Database

HRS — U.S. Health and Retirement Study

CJARS — Criminal Justice Administrative Records System

National COVID Cohort Collaborative

PSID — Panel Study of Income Dynamics

RL based on quasi-identifiers can be error-prone

Which records belong to the same individual?

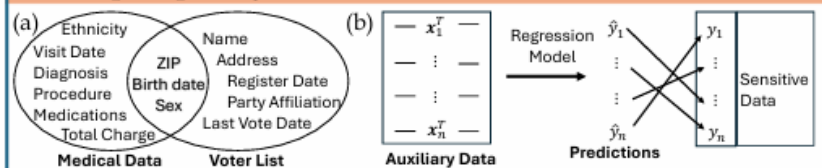
First Name	Middle Name	Last Name	Month of Birth	Lives in
Emanuel	Hyatt	Bendavid	Mar	New York, NY
Emmanuel	Ben	David	Dec	Washington, DC
Emanuel	NA	Ben-Dawid	November	Stanford, CA
Emanuel	NA	Ben-David	3	Oregon
E.	NA	Ben-Davit	Nov	NA

Which records can be linked?

	Age	% White	% Black	% Hispanic	% Native Born	Earnings
Linked (92%)	57.6	68	22	14	87	$43.2k
Unlinked (8%)	56.9	57	24	26	69	$33.3k

Source: Dhiren Patki, ISR RL Seminar Series, University of Michigan, 4/22

RL can pose privacy concerns



Auxiliary data linked to sensitive data using (a) ZIP, DOB, and Sex & (b) "y". Source: P. Samarati, & L. Sweeney (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression

Various software packages exist in R and Python for RL

`fastLink` (in R) and `Splink` (in Python) are popular RL implementations based on the **Fellegi-Sunter** model for fuzzy matching. **Such tools do not address PLDA methods.**

Post-Linkage Data Analysis (PLDA)

PLDA can broadly be divided into two settings

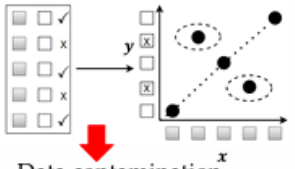
- **Primary Analysis:** Individual files are accessible. RL and data analysis can be performed jointly to directly propagate linkage uncertainty.
- **Secondary Analysis:** Only the linked file is available. Little may be known about the RL process and its error rates. *Increasingly important as data users may not be able or willing to perform the RL.*

Example: Linear regression, outcome variable linked to covariates

Linkage errors can invalidate statistical inferences

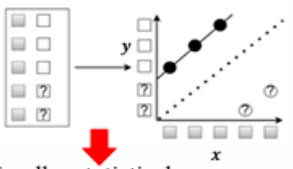
Analyses often rely on ignorable linkage errors.

Mismatches (false links)



- Data contamination
- Attenuation bias

Missed matches (false non-links)



- Smaller statistical power
- Selection bias

Handling linkage vs. mismatch rates is often not straightforward

"Safe" matches only	Maximum Matches
↑ Selection bias danger	↓ Selection bias danger
↓ Statistical power	↑ Data contamination danger

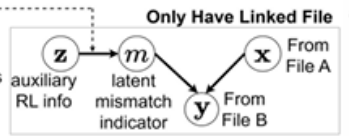
Adjustment methods can help balance between these extremes

"Adjust" the analysis to account for potential linkage errors. **In practice, this may require significant effort and expertise. The newly released `postlink` R software implements PLDA methods [3].**

General Framework for Inference

Model linked data (x,y) relationship based on a mixture model [1]

Examples:
- Mismatch error rate
- Predictors of match status
- Safe matches
- Intercept-Only



Assumptions:
(A1). $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid m = 1$,
(A2). The mismatch error does not depend on $(\boldsymbol{x}, \boldsymbol{y})$. *Note: This assumption can be relaxed [2].*
(A3). $P(m_i = 0 | \boldsymbol{z}_i) = h(\boldsymbol{z}_i; \boldsymbol{\gamma}_i)$, $1 \le i \le n$, known function h & unknown parameter $\boldsymbol{\gamma}$.

$\mathbf{y} | \{m = 1\}, \mathbf{x} \sim f(\mathbf{y})$ $\mathbf{y} | \{m = 0\}, \mathbf{x} \sim \phi(\mathbf{y} | \mathbf{x}; \theta)$

The framework is **extensible** and can be applied to various statistical models: GLMs, semi-parametric regression, Cox regression, contingency table analysis, small area models, causal inference, random forest, penalized regression, . . .

Inference via asymptotic theory for composite MLEs

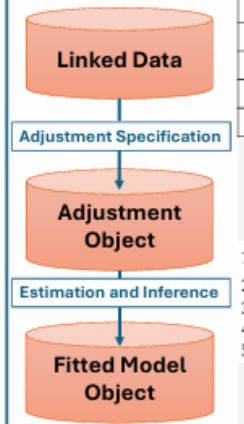
Expectation Maximization (EM) Algorithm maximizes the pseudo-likelihood.

$$L(\theta) = \prod_{i=1}^{n} \{\phi(\mathbf{y}_i | \mathbf{x}_i; \theta) P(m_i = 0 | \mathbf{z}_i) + f_{\mathbf{y}}(\mathbf{y}_i) P(m_i = 1 | \mathbf{z}_i)\}$$

Standard errors via sandwich formula, Louis' method, or multiplier bootstrap.

Linear Regression Demonstration - LIFE-M Project (life-m.org)

LIFE-M has multi-generational data gathered across vital records and censuses.



	yob	age_at_death	hndlnk	commf	comml
1	1905	83	FALSE	0.77	0.45
2	1883	79	TRUE	0.93	0.08
3	1886	58	FALSE	0.89	0.80
…	…	…	…	…	…

- `hndlnk`: hand-linked indicator
- `commf`, `comml`: commonness score of first and last names
- `yob`: predictor (x)
- `age_at_death`: outcome (y)

```
# Constructing the Adjustment Object
adj_object ← adjMixture(linked.data = lifem,
                        m.formula = ~ commf + comml, # (1)
                        m.rate = 0.05, # (2)
                        safe.matches = hndlnk) # (3)
```

1. $m_i | \text{commf}_i, \text{comml}_i \sim \text{Bernoulli}(\frac{\exp(\gamma_0 + \gamma_1 \text{commf}_i + \gamma_2 \text{comml}_i)}{1 + \exp(\gamma_0 + \gamma_1 \text{commf}_i + \gamma_2 \text{comml}_i)})$
2. Anticipated overall mismatch rate is ~5% (from automated probabilistic RL)
3. Hand-linked records are assumed to be correctly matched ("safe") $(m_i = 0)$
4. $y_i | m_i = 0, x_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3, \sigma^2)$
5. $y_i | m_i = 1, x_i \sim N(\mu, \tau^2)$

```
# Model Fitting
fit ← plglm(age_at_death ~ poly(unit_yob, 3, raw = TRUE),
            family = "gaussian", adjustment = adj_object) # (4)
```

	Naïve	Adj[1]	Adj[2]	HL Only
$\hat{\beta}_0$	58.5 (.2)	58.6 (.1)	58.7 (.2)	57.7 (1.3)
$\hat{\beta}_1$	-46.7 (1.8)	-51 (1.5)	-52.5 (1.6)	-44.2 (11.6)
$\hat{\beta}_2$	130.4 (4)	140.3 (3.9)	143.2 (3.9)	118.6 (27.9)
$\hat{\beta}_3$	-72.9 (2.5)	-76.8 (2.6)	-77.7 (2.7)	-59.9 (18.5)
$\hat{\sigma}$	21.2 (.1)	20.7 (.1)	20.4 (.1)	19 (.3)
$\hat{\gamma}_0$	-	-6 (.5)	-4.9 (.4)	-
$\hat{\gamma}_1$	-	-1.5 (.6)	-1.4 (.4)	-
$\hat{\gamma}_2$	-	7.2 (.3)	6.1 (.3)	-

*Adj[1]: `m.rate = 0.05` *Adj[2]: `0.075`

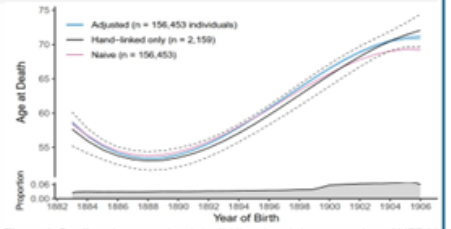


Figure 1. Predicted ages at death (solid lines) and the proportion of LIFE-M individuals (bottom) from 1883-1906. Point-wise 95% CIs (dashes for HL only approach and shaded ribbons for naïve and adjusted approaches).

References & Acknowledgement

[1] Slawski, M., West, B. T., Bukke, P., Diao, G., Wang, Z., & Ben-David, E. (2025). A General Framework for Regression with Mismatched Data Based on Mixture Modeling. *Journal of the Royal Statistical Society (Series A)*.
[2] Bukke, P., Slawski, M. (2025). Relaxing the Assumption of Strongly Non-Informative Linkage Error in Secondary Regression Analysis of Linked Files. https://doi.org/10.48550/arXiv.2510.17553.
[3] Bukke P, Kamat G, Cui J, Gutman R, Slawski M (2026). *postlink: Post-Linkage Data Analysis*. R package. https://CRAN.R-project.org/package=postlink.

We acknowledge support by NSF grants SES-2120318 and OAC-2411270.

Trans-KerKM: Kernel-Weighted Transfer Survival Estimation

Individualized Prediction for Rare Cancers Across the TCGA Pan-Cancer Cohort

Aoran Chen Tun He Yang Feng

Department of Biostatistics, New York University

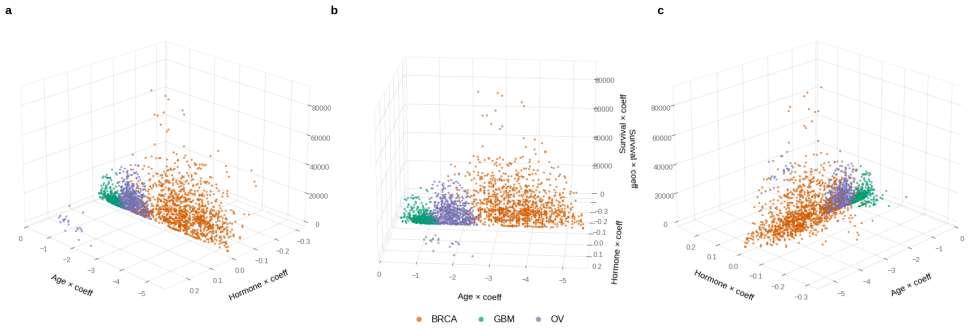
New Horizons on Model Transportability and Data Integration

Institute for Mathematical and Statistical Innovation (IMSI)

June 25, 2026

Small cancer cohorts: too few patients, too much heterogeneity

A rare-cancer target may hold only $\sim$**50** patients — too few for a stable model.
Why not borrow from related cancers?



But TCGA is inherently **heterogeneous** — cancers differ in outcomes, treatments, even the *sign* of effects.

Trans-KerKM

Patients close in covariate space likely share a latent subgroup — whatever their cancer label.

For each target patient x_i, borrow from its neighbors through a kernel-weighted hazard at each event time:

$$\hat{h}(t_\ell \mid x_i) = \frac{\sum_{j \neq i} K_\sigma(x_i, x_j)\, \delta_j \mathbf{1}\{y_j = t_\ell\} \;+\; \frac{1}{\lambda} \sum_{j} K_\sigma(x_i, x'_j)\, \delta'_j \mathbf{1}\{y'_j = t_\ell\}}{\sum_{j \neq i} K_\sigma(x_i, x_j)\, \mathbf{1}\{y_j > t_{\ell-1}\} \;+\; \frac{1}{\lambda} \sum_{j} K_\sigma(x_i, x'_j)\, \mathbf{1}\{y'_j > t_{\ell-1}\}}$$

then accumulate it into a survival curve, $\hat{S}(t_\ell \mid x_i) = \hat{S}(t_{\ell-1} \mid x_i)[1 - \hat{h}]$.

Local

bandwidth σ: weight each patient by covariate similarity to x_i

Global

$\lambda \geq 1$ down-weights the source

Five questions

Is transfer worth it?

Q1 Reliable individualized prediction from a $\sim$50-patient target?

Q2 When does borrowing help — and when does it hurt (*negative transfer*)?

How does it work?

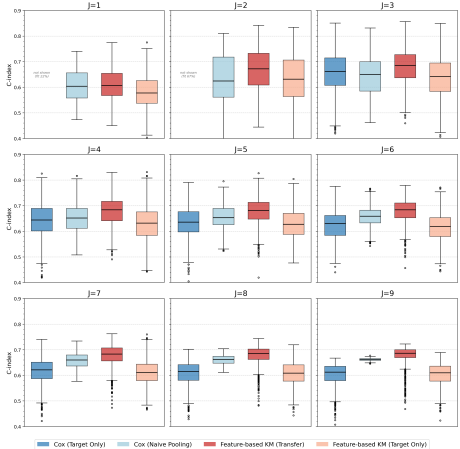
Q3 Which patients and covariates drive the borrowing?

Q4 Does the gain track source–target similarity?

Can we trust it?

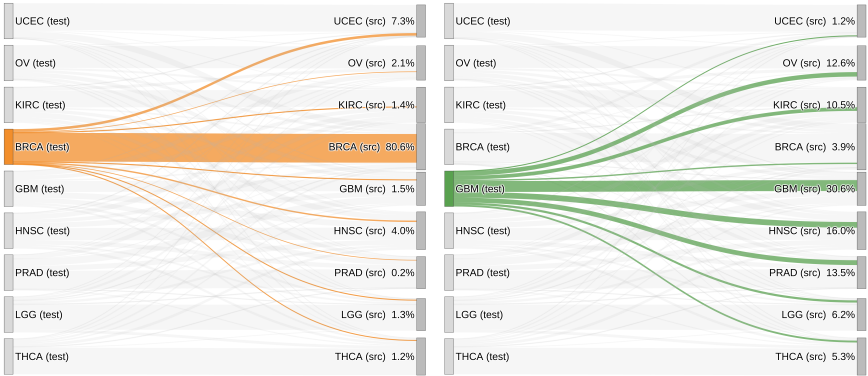
Q5 Robust to bandwidth, transfer strength, target size, covariate set?

Cross-cancer concordance



Test discrimination (C-index) vs. target heterogeneity J (# cancers mixed in).

Case study: BRCA borrows, GBM cannot



BRCA borrows mostly from BRCA-like patients (**81%** of weight);
GBM finds no clear neighbors (**31%**) — the one cancer where borrowing hurts.

We also looked at . . .

- **Which covariates carry the signal** — BRCA's spreads across all subsets; GBM's collapses onto age alone, then gets diluted by the isotropic kernel.
- **Cox coefficient alignment** — the pooled β sits closer to GBM than to BRCA, even flipping the sign of BRCA's radiation effect.
- **Robustness** — stable across bandwidth σ, transfer strength λ, and target size n_T.

Happy to discuss any of this. **Let's talk at the poster session.**

Take-home

On small, heterogeneous data like TCGA, borrow survival information adaptively — **locally** by covariate similarity, **globally** against target–source mismatch — which helps most when a cancer's prognostic signal is spread across covariates rather than collapsed onto one.

Thank You!

Questions & discussion welcome

Cancer Type Prediction Harmonizing AACR GENIE Data with Bayesian Neural Network

IMSI Workshop: New Horizons on Model Transportability and Data Integration

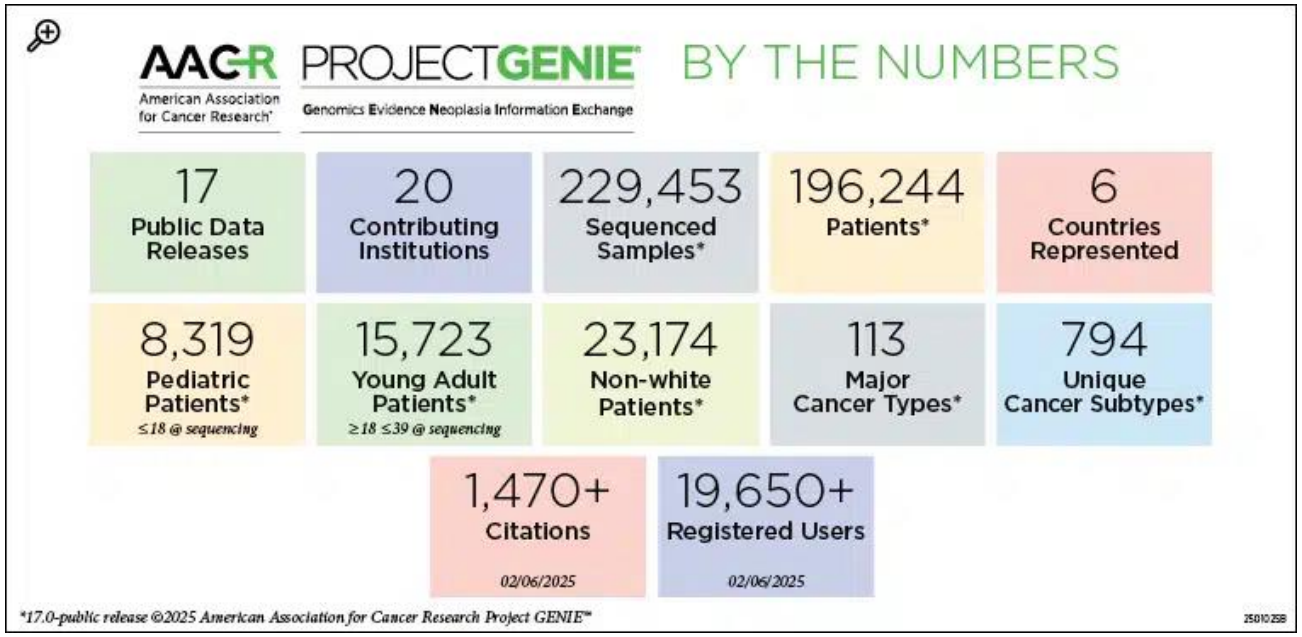
Mookyong Son, Danya Hassan, Katherine Panageas, Ronglai Shen, Yuan Chen

Acknowledgement: Karissa Whiting, and Yufei Deng

June 25th, 2026

Memorial Sloan Kettering
Cancer Center

Motivation of the project - GENIE data

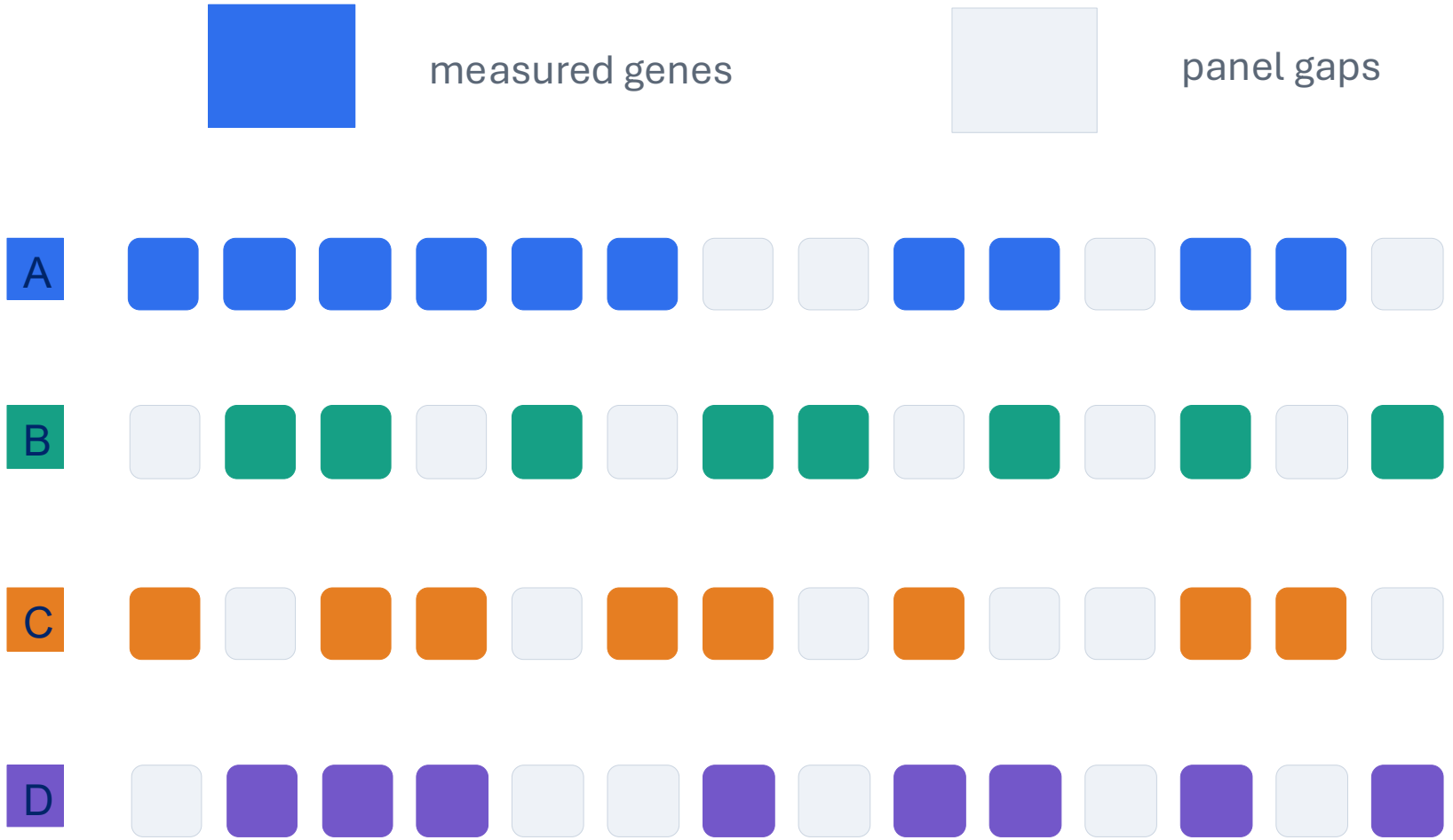


What clinicians can use it for

1. **Predict cancers of unknown primary**
2. **Study rare subgroups**
   Combine small cohorts without forcing identical panels
3. **Strengthen local models**
   Add additional information for center specific analysis.
4. **External validation / generalizability**
   Check whether findings track biology or workflow
5. **Next clinical endpoints**
   Use the extracted mutation features for survival, response, recurrence

1. Large sample size
2. Diverse patient population
3. More opportunities to study rare genes, subtypes, e.t.c.

Motivation of the project
– Challenges 1. panel coverage is structured, not random



Two Traditional Approaches for missing data

GENIE-like data reality: panel coverage is structured, not random

TRADITIONAL APPROACH 1

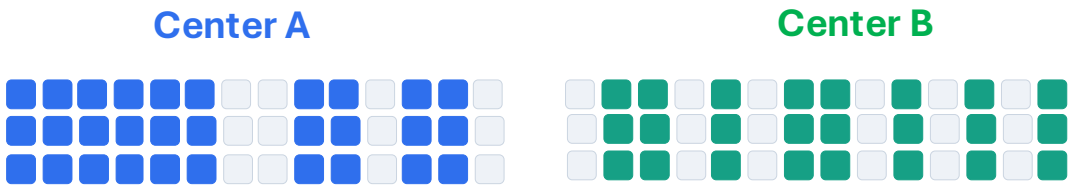
Common gene approach (n↑, p↓)

Pool all institutions → keep only genes assayed in every retained sample

TRADITIONAL APPROACH 2

Center specific approach (n↓, p↑)

Analyze each institution separately → use the center's local complete-case genes

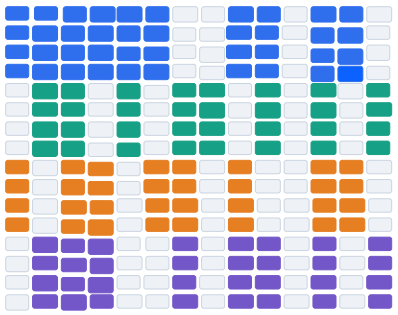


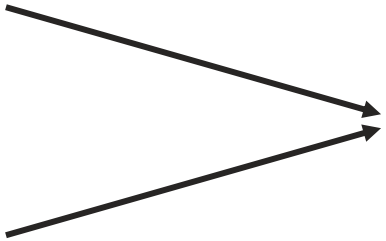
⟶ Both of them subject to substantial information loss

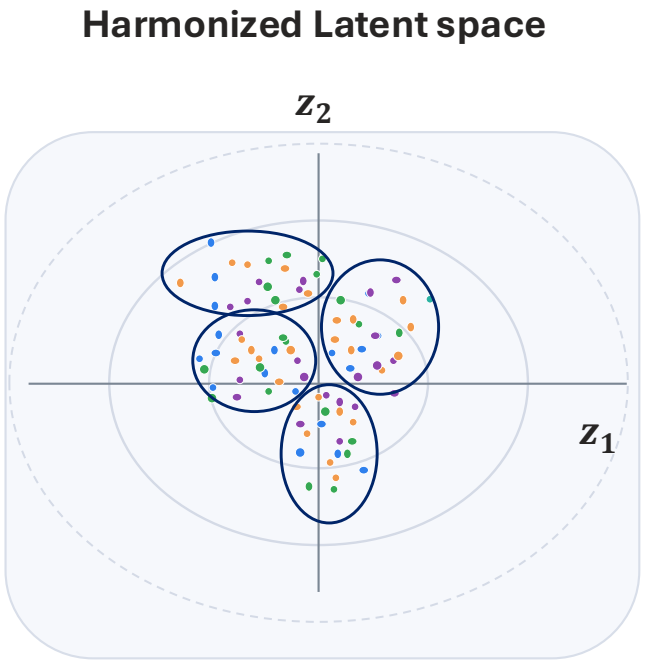
Motivation of the project – Challenges 2. Variations in Sequencing Techniques

SEQ_ASSAY_ID	Sample size	Number of genes	Platform	Specimen tumor Ceullularity	Is_paired_end	Coverage Calling Strategy
VICC-01-DX1	177	324	Illumina	Medium (>20%)	False	tumor_only
VICC-01-SOLIDTUMOR	845	31	Illumina	Small (>10%)	False	tumor_only
VICC-01-T4B	15	279	Illumina	Medium (>20%)	False	tumor_only
VICC-01-T5A	387	327	Illumina	Medium (>20%)	False	tumor_only
VICC-01-T7	1519	432	Illumina	Medium (>20%)	False	tumor_only
VICC-02-XTV2	23	595	Illumina	Medium (>20%)	True	tumor_normal
VICC-02-XTV3	244	649	Illumina	Medium (>20%)	True	tumor_normal
VICC-02-XTV4	388	649	Illumina	Medium (>20%)	True	tumor_normal

Proposed method







1) Comprehensive information
2) Harmonization, denoising

Thank you !

References

- AACR Project GENIE Consortium (2017), ‘AACR Project GENIE: Powering precisionmedicine through an international consortium’, Cancer Discovery 7(8), 818–831.
- Chen, Y., Shen, R., Feng, X. & Panageas, K. (2024), ‘Unlocking the power of multi-institutional data: Integrating and harmonizing genomic data across institutions’, Biometrics 80(4), ujae146.
- Kingma, D. P. & Welling, M. (2014), Auto-encoding variational Bayes, in ‘Proceedings ofthe 2nd International Conference on Learning Representations’. arXiv:1312.6114
- Sohn, K., Lee, H. & Yan, X. (2015), Learning structured output representation using deepconditional generative models, in ‘Advances in Neural Information Processing Systems’,Vol. 28, pp. 3483–3491
- Mark et.al (2021): “VAEs in the presence of Missing Data”

Learning Individualized Treatment Rules for a Target Population with Right-Censored Survival Outcomes

Peiyu He

Department of Biostatistics and Bioinformatics, Duke University

Joint work with Prof. Guanhua Chen

June 25, 2026

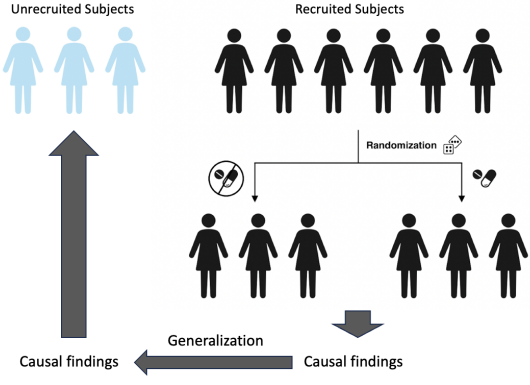
IMSI Workshop Lightning Talk

Motivation and setting

- **The ITRs:** recommend different treatments for different group of patients; e.g, $d(X_i) = \mathbb{1}\{\beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{SOFA}_i > 0\}$
- **Goal:** the optimal within-class ITR in the target population $\mathcal{T}$,

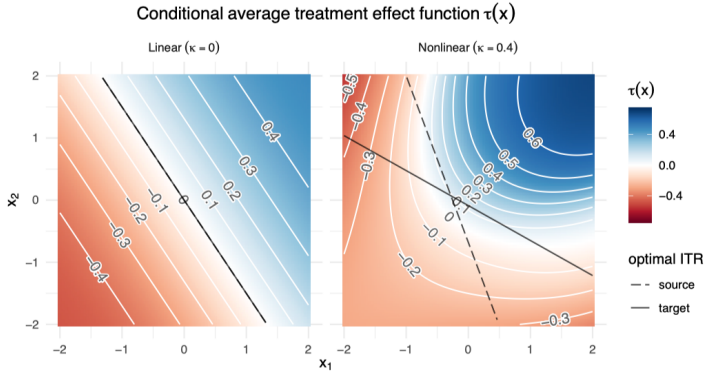
$$d^\star_{\mathcal{T}} \in \arg\max_{d \in \mathcal{D}} \mathcal{V}_{\mathcal{T}}(d),$$
$$\mathcal{V}_{\mathcal{T}}(d) = \mathbb{E}[y(T\{d(X)\}) \mid S = 0].$$

- **Generalization goal:** a source study $(A, X, U, \Delta, S = 1)$ to target population $(X, S = 0)$.



Challenges

- **Covariate shift matters:** with misspecified $\mathcal{D}$, source-optimal rules are not directly transportable to target.



- **Confounding adjustment**: source can be an observational study
- **Right censoring:** survival outcomes are incompletely observed due to discharge/transfer, loss to follow-up, or study end.

Core identification formula

Internal validity in source (positivity, conditional random censoring, ignorability)
Generalizability (survival mean exchangeability, positivity of participation)

$$\begin{aligned}\mathcal{V}_{\mathcal{T}}(d) &= \mathbb{E}[y\{T(d(X))\} \mid S = 0] \\ &= \mathbb{E}[w^*(A, X, U)\, \Delta\, y(U)\, \mathbb{1}\{A = d(X)\} \mid S = 1]\,.\end{aligned}$$

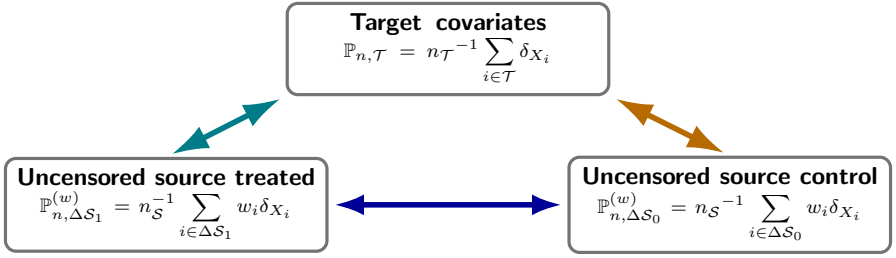
$$w^*(a, x, u) = \underbrace{\frac{\mathbb{E}[S]\{1 - \rho(x)\}}{(1 - \mathbb{E}[S])\rho(x)}}_{\text{population shift}} \underbrace{\left\{\frac{a}{\pi(x)} + \frac{1 - a}{1 - \pi(x)}\right\}}_{\text{confounding}} \underbrace{\frac{1}{S_C(u \mid a, x)}}_{\text{right-censoring}}.$$

- Encodes the three simultaneous challenges
- Multiple instability under conventional approaches

Highlight the need for robust weights

Three-way distributional balancing

Idea (Chen et al., 2023): directly optimize weights to target three-way distributional balance.



$$\hat{w}^s = \arg\min_w \alpha \, \mathrm{MMD}^2(\mathbb{P}^{(w)}_{n,\Delta\mathcal{S}_1}, \mathbb{P}_{n,\mathcal{T}}) + \alpha \, \mathrm{MMD}^2(\mathbb{P}^{(w)}_{n,\Delta\mathcal{S}_0}, \mathbb{P}_{n,\mathcal{T}}) + (1-\alpha) \, \mathrm{MMD}^2(\mathbb{P}^{(w)}_{n,\Delta\mathcal{S}_1}, \mathbb{P}^{(w)}_{n,\Delta\mathcal{S}_0}) + \lambda \|w\|_2^2 / n_S^2.$$

- One step optimization and avoid multiple instability
- No need to specify models for each component or moments to be balanced

Three learning approaches

3DB-IPCW
View censoring as selection of samples

$$\hat{\mathcal{V}}_{n,\mathrm{IPCW}}(d) = \frac{1}{n_S} \sum_{i \in \Delta\mathcal{S}} \hat{w}_i^s \mathbb{I}\{A_i = d(X_i)\} y(U_i)$$

Weighting with uncensored source samples;
$\hat{w}^s$: covariate shift, treatment imbalance, and censoring selection.

3DB-Imp
View censoring as outcome transformation

$$\hat{Y}_i = \Delta_i y(U_i) + (1 - \Delta_i)\hat{\mu}(A_i, X_i, U_i),$$
$$\hat{\mathcal{V}}_{n,\mathrm{Imp}}(d) = \frac{1}{n_S} \sum_{i \in \mathcal{S}} \hat{w}_i^+ \mathbb{I}\{A_i = d(X_i)\} \hat{Y}_i$$

First imputing censored outcomes, then weighting;
$\hat{w}^+$: covariate shift and treatment imbalance.

3DB-Res2W
Combining the two by residual-weighted-learning

$$\hat{\mathcal{V}}_{n,\mathrm{Res2W}}(d) = M(d) + R(d),$$
$$\hat{M}(d) = \frac{1}{n_S} \sum_{i \in \mathcal{S}} \hat{w}_i^+ \mathbb{I}\{A_i = d(X_i)\} \hat{\mu}(A_i, X_i, U_i)$$
$$\hat{R}(d) = \frac{1}{n_S} \sum_{i \in \Delta\mathcal{S}} \hat{w}_i^s \mathbb{I}\{A_i = d(X_i)\} \hat{r}_i, \quad \hat{r}_i = y(U_i) - \hat{\mu}(A_i, X_i, U_i)$$

Weighting working outcome imputations by $\hat{w}^+$, then weighting residual by $\hat{w}^s$ to correct bias;
Take advantage of two robust balancing weights and free of outcome model misspecification

MIMIC-IV → eICU Albumin ITR

- **Single-center to multi-center:** learn from single-center MIMIC-IV ($n = 19{,}108$), and generalize to eICU across 91 U.S. hospitals ($n = 8{,}091$).
- **Substantial covariate shift:** 21/45 covariates have significant difference
- **Highly-imbalanced treatment assignment:** albumin rate are only 19.9% in MIMIC-IV and 4.0% in eICU.
- **High censoring rate:** MIMIC-IV 81.8%; eICU 84.0%.

Results: 3DB-Imp and 3DB-Res2W have the highest estimated values, even better than within-target rule

Rule	$V(d)$	**%Alb**	**Gain**
No albumin	22.93	0.0	–
All albumin	22.00	100.0	-0.93
Source-only	22.89	39.5	-0.04
ACW	22.23	76.0	-0.70
IPW (est.)	22.28	27.7	-0.65
3DB-IPCW	22.30	49.9	-0.63
3DB-Imp	**23.00**	23.9	$+0.07$
3DB-Res2W	**23.06**	30.8	$+0.13$
Within-target	22.98	23.7	$+0.05$