Adaptive and Efficient Learning with Block-wise Missing and Semi-Supervised Data

Yiming Li, Ruoyu Wang, Ying Wei and Molei Liu

Department of Biostatistics, Colombia University

Overview

Background and Introduction

Method: Data-adaptive projecting Estimation approach for data FUsion in the SEmi-supervised setting (DEFUSE)

Theoretical guarantee

Simulation

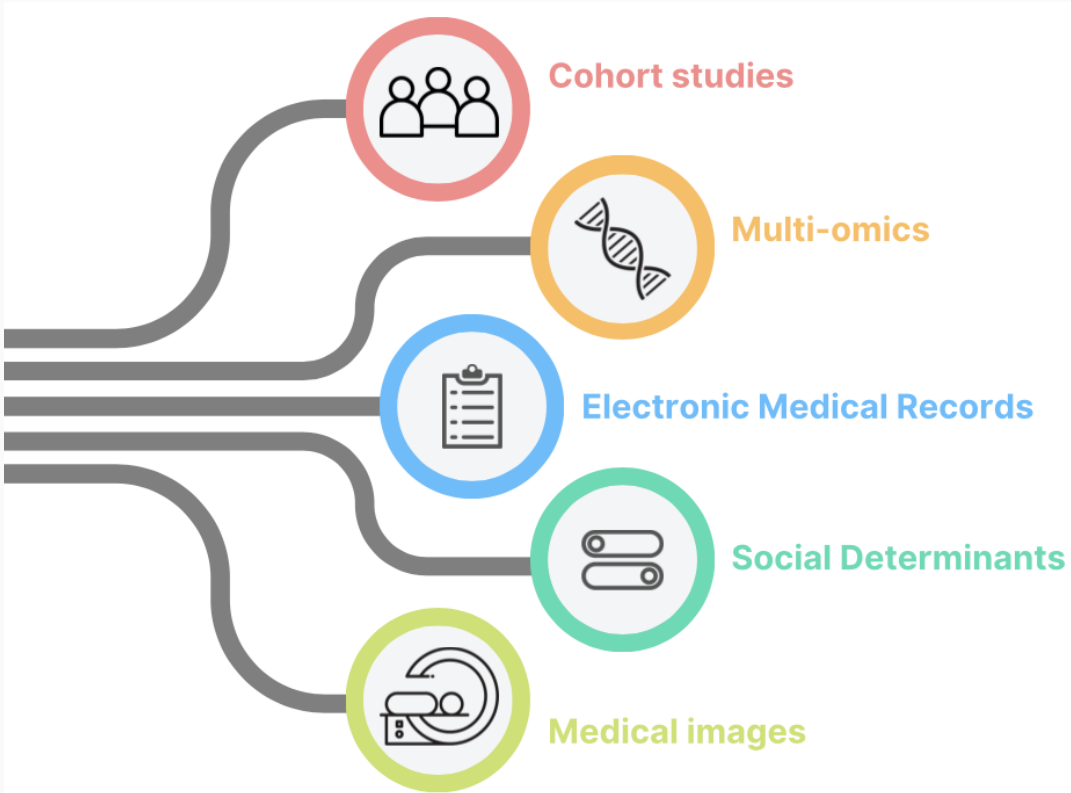
Real Data Analysis

Background and Introduction

Background

- **Data Fusion:** bring different data sets from multiple resources together to **enable a more powerful and comprehensive analysis**

- Examples: linking electronic medical records (EMRs) with biobank data to tackle complex health issues
    - At national-level: UK Biobank / All of US
    - At institutional-level: Harvard, Columbia, Vanderbilt...
    - ...



EHR-based data fusion: complementary strengths

EHR / EMR	Research cohorts / clinical trials
Advantages	**Advantages**
Large sample sizes; heterogeneous real-world patients Longitudinal care trajectories (labs, meds, diagnoses) Reflects actual clinical practice and workflows	Deep phenotypes; standardized measurements High-quality, gold-standard labels/outcomes Carefully curated covariates and follow-up
Limitations	**Limitations**
Sparse/noisy per patient; missing-by-design measurements Limited deep phenotyping; less standardized protocols Gold labels are costly (manual chart review)	Smaller samples; limited statistical power Narrow inclusion criteria; limited external validity Often fewer "real-world" care signals

What can we unlock by linking EHR + research cohorts/trials?

- **Efficiency:** use deep phenotypes / gold labels from research data to reduce variance in EHR-scale analyses.
- **Generalizability:** transport models/knowledge to heterogeneous real-world populations and clinical practices.
- **Better and more precision knowledge discovery & statistical learning:** combine "large+real-world" with "small+deep".

$\Longrightarrow$ connect deep, high-fidelity science to real-world impact at population scale.

Introducing Columbia Brain Health DataBank

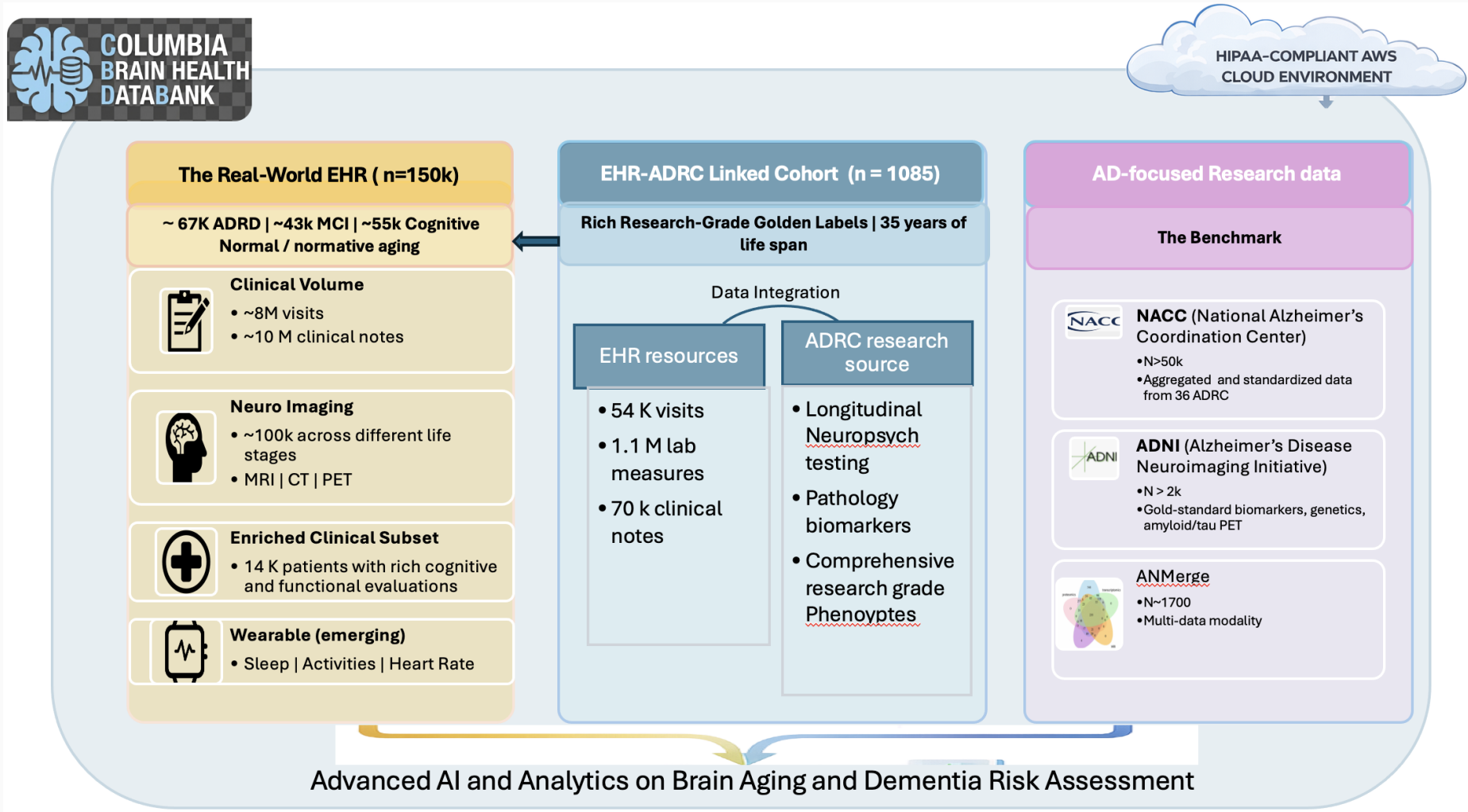


Figure 1: https://www.trail4health.org/Columbia-Brain-Health-DataBank

From Midlife to Dementia: A Linked EHR + ADRC Patient Journ

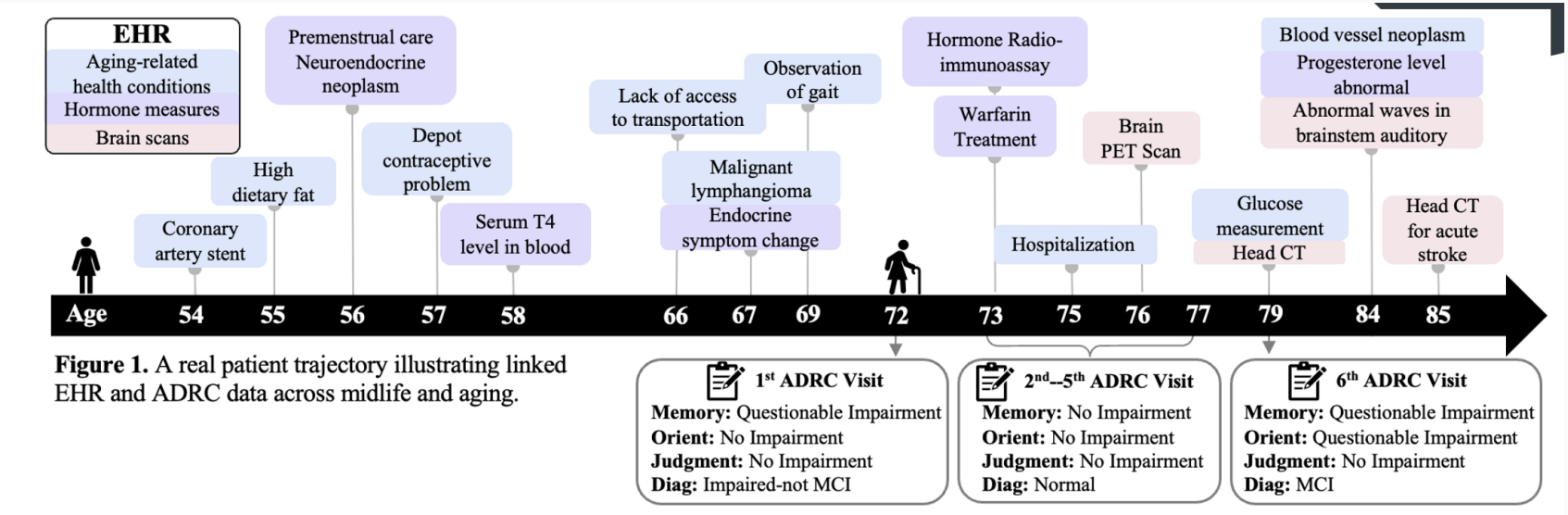


Figure 1. A real patient trajectory illustrating linked EHR and ADRC data across midlife and aging.

Figure 2: A snapshot of a linked patient

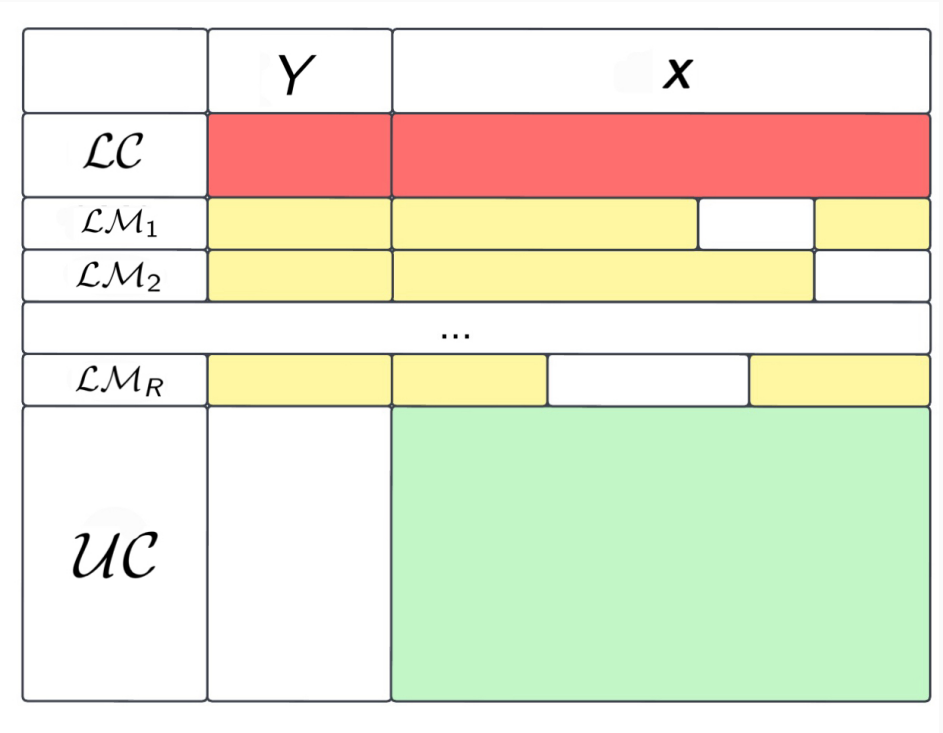
What statistical challenges to fully optimize the power of data fusion?

- **Blockwise Missingness (BM)** when combining data from different cohorts or studies.
- **Semi-supervised learning (SS):** gold-standard labels exist only for a small subset of EHR.
- **Distributional shift:** cohorts/trials and EHR populations differ (selection, measurement, practice patterns).
- **Misalignment / transportability violations:** some sources may not be compatible with the target population.
- **High-dimensionalities:** EHR codes/clnical notes features are high-dimentional, often require ML;

This talk: DEFUSE addresses BM + SS + shift, with screening for misaligned sources.

Problem Setup and Target Estimation

- $\underline{\mathcal{LC}}$: $\mathcal{L}$abeled data with $\mathcal{C}$ompletely observed covariates $\rightarrow$ EHR-linked ADRC
- $\underline{\mathcal{LM}}$: $\mathcal{L}$abeld data with $\mathcal{M}$issing covariates $\rightarrow$ AD related research cohorts
- $\underline{\mathcal{UC}}$: $\mathcal{U}$nlabeled data with $\mathcal{C}$omplete covariates $\rightarrow$ EHR/CBDB



- Our primary interest: enhance the power to learn a generalized linear model (GLM) from a target population (in this talk: EHR/$\mathcal{U}$):

$$\mathrm{E}(Y|\boldsymbol{X}) = g(\gamma^\top \boldsymbol{X}),$$

- by leveraging $\mathcal{LC}$, $\mathcal{LM}$, and $\mathcal{UC}$

Distributional shift + transportability via CAM

Conditional Alignment under Missingness (CAM) = "align what matters, given what you can see".

- **Component 1 ($\mathcal{LC} \to \mathcal{UC}$): transport the outcome model.** Allow $P_{\mathcal{LC}}(\boldsymbol{X}) \neq P_{\mathcal{UC}}(\boldsymbol{X})$, but require the same conditional relationship:

$$P_{\mathcal{LC}}(Y \mid \boldsymbol{X}) = P_{\mathcal{UC}}(Y \mid \boldsymbol{X}).$$

  So we learn how Y depends on the *full* covariates $\boldsymbol{X}$ in labeled-complete data and reuse it in the target.

- **Component 2 ($\mathcal{LM}_r \to \mathcal{UC}$): transport after conditioning on an "alignment" summary.** For each labeled-missing source, pick observed $Z_r \subseteq (\boldsymbol{X}_{\Gamma_r}, Y)$ so that

$$P_{\mathcal{LM}_r}(Y, \boldsymbol{X} \mid Z_r) = P_{\mathcal{UC}}(Y, \boldsymbol{X} \mid Z_r).$$

  Intuition: within strata of Z_r, the source looks like the target, so its partial information can be safely fused.

Two different questions: missingness vs. transportability

- **Q1: Why are covariates missing in $\mathcal{LM}_r$?** (a within-source issue)
  - We do *not* need to model every driver of the missing blocks.
- **Q2: Can $\mathcal{LM}_r$ be validly fused into the target $\mathcal{UC}$?** (a cross-source issue)
  - $\mathcal{UC}$ defines the target covariate distribution $P_{\mathcal{UC}}(\boldsymbol{X})$ (fully observed $\boldsymbol{X}$).
  - $\mathcal{LC}$ + bridge $P_{\mathcal{LC}}(Y \mid \boldsymbol{X}) = P_{\mathcal{UC}}(Y \mid \boldsymbol{X})$ define how Y relates to $\boldsymbol{X}$ in the target.
  - CAM tells when $\mathcal{LM}_r$ is *comparable enough* to $\mathcal{UC}$: after conditioning on observed Z_r.
- **What goes into Z_r and Why Z_r may include Y?**
  - Include a variable only if conditioning on it is needed for alignment.
  - If missingness depends on Y, then the distribution of the observed data in $\mathcal{LM}$ is distorted relative to the target $\mathcal{UC}$

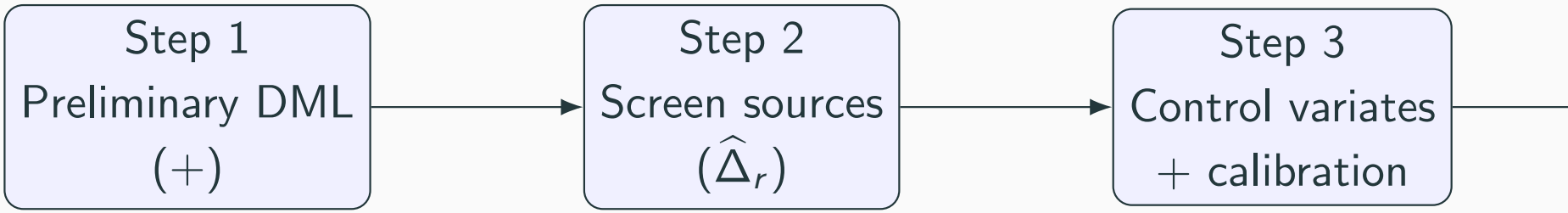
Exist Literature

- Learning with block-wise missing covariates:
  - **Imputation-based method** for integrating multi-source block-wise missing data in linear models (Xue & Qu, 2021).
  - **Optimal Sparse Linear Prediction for Block-missing Multi-modality Data Without Imputation** (Yu et al, 2020)
- Semi-supervised learning with large auxiliary unlabeled data:
  - **Efficient and adaptive semi-supervised regression** (e.g., Kawakita & Takeuchi, 2014; Chakrabortty & Cai, 2018; Azriel et al., 2022; Gronsbell et al., 2022).
- Data fusion under distribution shift: **transportability** (Pearl and Bareinboim,2014; Yang and Ding, 2019; Graham et al., 2025).
- **DUFUSE:** combines all three and adds and *screening* to avoid bias from misaligned sources.

Method: Data-adaptive projecting Estimation approach for data FUsion in the SEmi-supervised setting (DEFUSE)

DEFUSE at a glance (pipeline) and High-level idea

1. Build a **preliminary** estimator using $(\mathcal{LC}, \mathcal{UC})$ that handles covariate shift.
2. Use each $\mathcal{LM}_r$ as a **control variate** to reduce variance *without introducing bias* under CAM.
3. **Screen** and exclude misaligned $\mathcal{LM}_r$ sources to avoid fusion bias.



Key inputs: choice of Z_r (alignment covariates), anchors $\widehat{\phi}_r$, and reweighting features W_r.

Step 1: Preliminary Estimator – Target Population & Bias Correction

Goal: Estimate γ defined w.r.t. the **target population** $\mathcal{UC}$:

$$\mathrm{E}_{\mathcal{UC}}\left[\boldsymbol{X}\{Y - g(\gamma^\top \boldsymbol{X})\}\right] = \mathbf{0}$$

Problem: Y is not observed in $\mathcal{UC}$.

Initial Solution: Leverage

- the completely observed $\mathcal{LC}$
- CAM (1): $P_{\mathcal{LC}}(Y|\boldsymbol{X}) = P_{\mathcal{UC}}(Y|\boldsymbol{X})$
- Double/Debiased Machine Learning (DML) operator(Chernozhukov et al. 2018).

Step 1: Preliminary Estimator – Combining $\mathcal{LC}$ and $\mathcal{UC}$

Double/Debiased Machine Learning (DML) operator(Chernozhukov et al. 2018) converts any $\mathcal{LC}$ average into a $\mathcal{UC}$-consistent one.

$$\mathfrak{M}_{\mathcal{LC}}(D) = \underbrace{h(\boldsymbol{X})\big(D - \mathrm{E}_{\mathcal{LC}}[D \mid \boldsymbol{X}]\big)}_{\text{IPW-corrected residual}} + \underbrace{\mathrm{E}_{\mathcal{UC}}\big[\mathrm{E}_{\mathcal{LC}}[D \mid \boldsymbol{X}]\big]}_{\text{average over } \mathcal{UC} \text{ covariates}}$$

- D can be **any quantity** to be averaged (here: D is the score function

$$D = \boldsymbol{X}\{Y - g(\gamma^\top \boldsymbol{X})\}$$

- $h(\boldsymbol{X}) = p_{\mathcal{UC}}(\boldsymbol{X})/p_{\mathcal{LC}}(\boldsymbol{X})$: density ratio (ML-estimated)
- $m(\boldsymbol{X}) = \mathrm{E}_{\mathcal{LC}}[D \mid \boldsymbol{X}]$: ML-estimated
- Second term in practice: fit the model on $\mathcal{LC}$, then **predict at $\mathcal{UC}$ covariate values** and average $\Rightarrow$ transports the LC model to the UC population
- Enjoy the double-robustness

Step 1: Preliminary Estimator $\tilde{\gamma}$ – Formal Definition

Solve the DML-corrected estimating equation using the $\mathcal{LC}$ and $\mathcal{UC}$ samples:

$$\widetilde{\mathbf{U}}_{\mathcal{LC}}(\gamma) \;=\; \widehat{\mathrm{E}}_{\mathcal{LC}}\Big[\widehat{\mathfrak{M}}_{\mathcal{LC}}\big(\boldsymbol{X}\{Y - g(\gamma^{\top}\boldsymbol{X})\}\big)\Big] = \mathbf{0}$$

Things to note:

- $\tilde{\gamma}$ uses $\mathcal{LC}$ *and* $\mathcal{UC}$: $\mathcal{UC}$ enters the DML operator to **correct bias** (transport the LC model to the UC population), but **does not improve efficiency** – the variance is still governed by LC-level variation, which is governed entirely by $\mathrm{Var}[\mathfrak{M}_{\mathcal{LC}}(\bar{\boldsymbol{S}})]$, where

  $\bar{\boldsymbol{S}} \equiv \bar{\boldsymbol{J}}^{-1}\boldsymbol{X}\{Y - g(\bar{\gamma}^{\top}\boldsymbol{X})\}$ is the score function

Step 2: DEFUSE$_{\mathcal{LM}}$ – Borrowing Strength from Partial Data

The gap: $\tilde{\gamma}$ uses only $\mathcal{LC}$ labels – but $\mathcal{LM}_r$ sources have many more labeled subjects! They observe $(\boldsymbol{X}_{\Gamma_r}, Y)$ but are *missing* $\boldsymbol{X}_{\Gamma_r^c}$.

Control variate idea Glasserman (2004): add a **mean-zero** correction term that is **correlated** with the estimation error $\Rightarrow$ variance decreases.

Choice of Control variate: Partial score function

$$\phi_r = \mathrm{E}[\boldsymbol{c}^\top \bar{\boldsymbol{S}} \mid \boldsymbol{X}_{\Gamma_r}, Y]$$

It is a proxy for the full score $\boldsymbol{c}^\top \bar{\boldsymbol{S}}(\boldsymbol{X}, Y)$, by *integrate out the missing* $\boldsymbol{X}_{\Gamma_r^c}$ and using only what $\mathcal{LM}_r$ observes.

Step 2: DEFUSE$_{\mathcal{LM}}$ – Estimator & Intrinsic Efficiency

Estimator: Augment $\boldsymbol{c}^\top \tilde{\gamma}$ with DML-corrected control variates from each $\mathcal{LM}_r$:

$$\widehat{\beta}(\phi) = \boldsymbol{c}^\top \tilde{\boldsymbol{\gamma}} + \sum_{r=1}^{R} \underbrace{\left\{ \widehat{\mathrm{E}}_{\mathcal{LM}_r}\left[\widehat{\mathfrak{M}}_{\mathcal{LM}_r}(\phi_r)\right] - \widehat{\mathrm{E}}_{\mathcal{LC}}\left[\widehat{\mathfrak{M}}_{\mathcal{LC}}(\phi_r)\right] \right\}}_{\text{mean-zero control variate from source } r \text{ (DML-corrected for its own shift)}}$$

- Their **difference** ≈ 0 under CAM $\Rightarrow$ augmentation is **unbiased**
- When $\phi_r \approx \boldsymbol{c}^\top \bar{\boldsymbol{S}}$: correction **cancels estimation noise** in $\boldsymbol{c}^\top \tilde{\gamma} \Rightarrow$ variance $\downarrow$

Asymptotic variance of $\sqrt{n}(\widehat{\beta}(\phi) - \boldsymbol{c}^\top \bar{\boldsymbol{\gamma}})$:

$$Q(\phi) = \underbrace{\mathrm{Var}_{\mathcal{LC}}\left[\mathfrak{M}_{\mathcal{LC}}(\boldsymbol{c}^\top \bar{\boldsymbol{S}}) - \textstyle\sum_r \mathfrak{M}_{\mathcal{LC}}(\phi_r)\right]}_{\text{residual after ``explaining away'' with } \phi_r\text{'s}} + \sum_r \frac{1}{\rho_r} \underbrace{\mathrm{Var}_{\mathcal{LM}_r}\left[\mathfrak{M}_{\mathcal{LM}_r}(\phi_r)\right]}_{\text{cost of using source } r}$$

Step 2: DEFUSE$_{\mathcal{LM}}$ – Estimator & Intrinsic Efficiency

Data-adaptive Reweight – what is $\tilde{\phi}$?

$\phi_r = \mathrm{E}[\boldsymbol{c}^\top \bar{\boldsymbol{S}} \mid \boldsymbol{X}_{\Gamma_r}, Y]$ is unknown. In practice, estimate it from data:

1. **Get a candidate:** fit $\hat{\phi}_r$ on $\mathcal{LC}$ by regressing $\boldsymbol{c}^\top \widetilde{\boldsymbol{S}}$ onto $(\boldsymbol{X}_{\Gamma_r}, Y)$; or use any ML/AI predictor of the partial score
2. **Reweight:** re-weight $\hat{\phi}_r$ as $\tilde{\phi}_r = w_r(_r; \tilde{\boldsymbol{\delta}}_r) \cdot \hat{\phi}_r$, choosing the calibration parameter $\tilde{\boldsymbol{\delta}}_r = \arg\min_{\boldsymbol{\delta}_r} \widehat{Q}((\boldsymbol{\delta}))$ (closed-form quadratic program)

$\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_1, \ldots, \tilde{\phi}_R)^\top$ is the **calibrated control function vector** used in DEFUSE$_{\mathcal{LM}}$.

- **Intrinsic efficiency** : $\mathrm{AVar}(\widehat{\beta}(\tilde{\boldsymbol{\phi}})) \leq \mathrm{AVar}(\boldsymbol{c}^\top \tilde{\boldsymbol{\gamma}})$ *always* – never hurts
- Only $o_{\mathrm{P}}(1)$ consistency of $\tilde{\phi}_r$ needed – robust to imperfect or misspecified ML models

Reweighting examples for $w_r(W_r; \delta_r)$

Three options considered:

1. **Constant:** $w_r \equiv \delta_r$ (simple source allocation).
2. **Linear:** $w_r = (1, W_r^\top)\delta_r$ (lets weights adapt to covariates).
3. **KRR:** w_r in an RKHS (nonparametric, regularized).

In all cases, calibration minimizes $\widehat{Q}$ (variance proxy) over δ.

Screening Misaligned Data Sources

- Before fusing $\mathcal{LM}_r$, we must check whether it is **aligned** with the target $\mathcal{UC}$ (i.e., CAM holds). Misaligned sources can introduce bias.
- We introduce the **Mean Squared Discrepancy (MSD)** measure:

$$\bar{\Delta}_r^2 = \mathrm{E}_{\mathcal{UC}}\left[m_r(\mathbf{Z}_r) - m_r^*(\mathbf{Z}_r)\right]^2$$

  where $m_r(\mathbf{Z}_r) = \mathrm{E}_{\mathcal{LM}_r}[\phi_r \mid \mathbf{Z}_r]$ and $m_r^*(\mathbf{Z}_r) = \mathrm{E}_{\mathcal{UC}}[\phi_r \mid \mathbf{Z}_r]$.
- $\bar{\Delta}_r = 0$ iff CAM holds; $\bar{\Delta}_r > 0$ signals violation.
- We construct a **debiased estimator** $\widehat{\Delta}_r$ and define the well-aligned set:

$$\widehat{\mathcal{A}} = \{r : \widehat{\Delta}_r < \tau\}$$

  Only sources in $\widehat{\mathcal{A}}$ are used for data fusion.
- MSD is **strictly more sensitive** than the commonly used mean difference measure $|\bar{V}_r|$, since $|\bar{V}_r| \leq \bar{\Delta}_r$.

Theoretical guarantee

Guarantee 1 (validity): Asymptotic normality (informal statement)

Under CAM and DML-style rate conditions on nuisance learners:

$$\sqrt{n}\big(\widehat{\beta}_{\mathrm{DEFUSE}} - c^{\top}\bar{\gamma}\big) \;\Rightarrow\; \mathcal{N}\big(0,\; Q(\bar{\phi})\big).$$

Note:

- Estimation of Density ratios / conditional means: about $o_p(n^{-1/4})$ in L_2.
- Estimation of Control functions ϕ_r: only need $o_p(1)$.

Guarantee 2 (efficiency): why variance can drop a lot

- DEFUSE uses *control functions*

$$\phi_r(\boldsymbol{X})$$

  as variance-reduction devices:

  - partial estimation scores to soak up residual variation in the score
  - data adaptive reweighting function to bring flexibility,
  - leveraging information in (i) unlabeled $\mathcal{UC}$ and (ii) partially observed $\mathcal{LM}_r$.

- **Intrinsic efficiency**: if you restrict yourself to a class $\mathcal{F}$ (e.g., linear, RKHS/KRR), DEFUSE chooses the best-in-class

$$\phi \in \mathcal{F},$$

  so no other method in that class can beat its asymptotic variance.

Guarantee 2 (efficiency): why variance can drop a lot

- **Semiparametric efficiency**: if $\mathcal{F}$ is rich enough to approximate the *optimal* augmentation, DEFUSE reaches the semiparametric efficiency bound (i.e., the information-theoretic minimum variance).

Guarantee 3 (screening consistency): when fusion is safe vs risky

- Misaligned sources violate CAM/transportability, so naively fusing them induces *bias*.
- DEFUSE estimates a source-specific discrepancy (MSD) $\bar{\Delta}_r^2$ that measures how much $\mathcal{LM}_r$ implies a different relationship than the target.
- **Screening consistency** means:
  - if a source is truly aligned ($\bar{\Delta}_r = 0$), it is kept with probability $\to 1$,
  - if it is separated from zero by a detectable margin, it is excluded with probability $\to 1$.
- Interpretation: in large samples, screening acts like an automatic "do-no-harm" guardrail before borrowing strength.

Simulation

Simulation settings (Table 1): what each setting stresses

Setting	(I)	(II)	(III)	(IV)	(V)	(VI)
Missing mechanism	MCAR	MCAR	MCAR	CAM	Misaligned	CAM
R	1	1	2	1,2	4	1
Y type	Cont.	Binary	Cont.	Cont.	Cont.	Cont.
$\boldsymbol{X}_{\Gamma_r} \mid \boldsymbol{X}_{\Gamma}$	Lin.	Nonlin.	Lin.	Lin.	Lin.	Lin.
$Y \mid \boldsymbol{X}$	Lin.	Nonlin.	Lin.	Nonlin./Lin.	Lin.	Lin.
High-dim $\boldsymbol{X}$	–	–	–	–	–	Yes

- (I)–(III): MCAR sanity checks (single vs multiple BM; continuous vs binary).
- (IV),(VI): CAM + shift (nonlinearity; high-dimensional auxiliaries).
- (V): some BM sources violate CAM $\Rightarrow$ need screening.

Methods compared & evaluation metrics

DEFUSE variants: anchor $\widehat{\phi}_r$ = AI-predictor of $c^\top \widetilde{S}$ using observed variables in $_r$; reweighting examples: const / linear / KRR.

Baselines (when applicable): Preliminary ($\widetilde{\gamma}$), MICE, MBI, HTLGMM, SC (and SC-PW under CAM), SSB, SSL.

Metrics reported:

- **RE** to Preliminary (variance/MSE reduction; higher is better).
- **Bias vs SE** (bias-to-SE sanity check for inference).
- **Screening** (Sensitivity / F1 for detecting misaligned sources).

Benchmark

- **[MICE:]** Imputing the missing covariates in $\mathcal{LM}$ using Multiple Imputation by Chained Equations (MICE), and combining it with $\mathcal{LC}$ data.
- **[BMI:]** Blockwise Multiple Imputation proposed in (Xue & Qu, 2021) combining $\mathcal{LC}$ and $\mathcal{LM}$ data.
- **[SSL:]** Semi-Supervised Learning proposed in (Gronsbell, 2022) using the $\mathcal{LC}$ and $\mathcal{UC}$ data sets.
- **[HTLGMM:]** Semi-supervised Learning proposed in (Zhao & Nilanjan, 2023) using the $\mathcal{LC}$ and $\mathcal{LM}$ data.
- **[SSB:]** Proposed in (Song, 2024) that allows the integration of $\mathcal{LC}$, $\mathcal{LM}$ and $\mathcal{UC}$ data.
- **[SC/SC-PW:]** Sample-reweighting for Calibration of the target model in Wang et al., 2023. SC-PW is its adaption under CAM.

Results (Tables 2): MCAR, single vs multiple BM (Settings II–III)

Component-wise RE to Preliminary (higher is better):

	Setting (II)					Setting (III)				
Method	γ_1	γ_2	γ_3	γ_4	γ_5	γ_1	γ_2	γ_3	γ_4	γ_5
DEFUSE (const)	7.05	6.92	3.76	2.71	2.90	2.13	2.13	2.12	1.43	1.57
DEFUSE (linear)	7.24	6.80	4.31	2.69	2.83	2.18	2.18	2.17	1.48	1.63
DEFUSE (KRR)	7.24	6.89	4.71	2.77	2.92	2.41	2.40	2.39	1.57	1.81
MICE	4.86	10.01	4.92	0.18	0.58	0.52	0.63	0.69	0.20	0.22
MBI	–	–	–	–	–	1.06	1.09	1.06	0.95	1.01
SSL	1.01	1.04	1.04	0.97	1.29	1.01	1.00	1.00	0.99	1.02
SSB	–	–	–	–	–	1.07	1.12	1.51	0.60	0.81
HTLGMM	4.30	4.70	2.05	1.14	1.06	–	–	–	–	–
SC	4.34	4.75	2.03	1.12	1.04	–	–	–	–	–

- DEFUSE dominates BM-fusion baselines; SSL (using only $+$) gives little gain when outcome model is correct.
- Within DEFUSE, KRR typically gives the best (or close-to-best) RE.

Results (Table 3): CAM shift (Settings IV & VI)

Setting (IV): increasing nonlinearity (nl). **Setting (VI):** high-dimensional auxiliaries; vary # signals (s).

Setting (IV): Bias_{SE} and RE

	nl=0.4	nl=0.6	nl=0.8
Prelim (RE)	1.00	1.00	1.00
DEF const (RE)	1.74	1.89	1.97
DEF linear (RE)	1.75	1.90	1.98
DEF KRR (RE)	1.80	1.87	1.89
MBI (RE)	1.55	1.18	0.98
SC-PW (RE)	1.45	1.10	0.93

Setting (VI): Bias_{SE} and RE

	s=5	s=10	s=15
Prelim (RE)	1.00	1.00	1.00
DEF const (RE)	1.28	1.38	1.36
DEF linear (RE)	1.27	1.34	1.36
DEF KRR (RE)	1.25	1.30	1.39

- Under CAM + nonlinearity, DEFUSE keeps bias small relative to SE and improves RE; SC-PW/MBI degrade as nonlinearity strengthens.
- With high-dimensional auxiliaries, DEFUSE still provides stable gains over Preliminary.

Results (Table 4): screening under misalignment (Setting V)

Screening misaligned BM sources as discrepancy increases (α_{dis}):

α_{dis}	MD sens.	MD F1	MSD sens.	MSD F1
0.4	0.71	0.78	0.99	0.97
0.6	0.88	0.90	1.00	0.97
0.8	0.97	0.96	1.00	0.97

- MSD is consistently more sensitive (especially at moderate discrepancy), so it is safer as a "don't-fuse-bad-sources" guardrail.

Simulation takeaways: what we learn

- **MCAR (I–III):** in the reported settings, DEFUSE achieves higher RE than the BM-fusion baselines shown; SSL using only + tends to have RE close to 1 when the outcome model is correctly specified.
- **CAM shift (IV, VI):** DEFUSE maintains small bias relative to SE in the reported CAM scenarios; its RE remains above 1 across the nonlinearity / high-dimensional variants considered.
- **Misalignment (V):** screening improves robustness to including non-transportable sources; MSD shows higher sensitivity than MD at moderate discrepancies in Table 4.

Real Data Analysis

Application 1: Predicting Cognitive Status with the NACC data

- **National Alzheimer's Coordinating Center (NACC) Data:**
  - Established by the National Institute on Aging (NIA) in 1999 to foster collaborative research on Alzheimer's disease and related disorders.
  - Maintains a large, standardized database of clinical and neuropathological information collected from more than 40 Alzheimer's Disease Research Centers (ADRCs) across the United States.
  - Since 2005, has utilized the Uniform Data Set (UDS) protocol across all ADRCs, ensuring consistent, longitudinal collection of cognitive, functional, behavioral, and demographic data.

Application 1: Predicting Cognitive Status with the NACC data

- **Y** : A binary indicator for cognitive impairment based on the Mini-Mental State Examination (MMSE) – $Y = 1$ if MMSE < 25, otherwise $Y = 0$.
- **X** include demographic features (gender, race, primary language, age, weight, height) and modifiable behavioral factors (living alone, alcohol consumption, smoking).
- **Data Structure:**
  - $N_{\mathcal{LC}}$: Fully observed centers (with language score)
  - $N_{\mathcal{LM}}$: Centers with missing language score (BM)
  - $N_{\mathcal{UC}}$: Patients without MMSE label (Unlabeled)
- **BM pattern:** some centers have missing language score block.
- **Screening:** start with 8 centers with missing language; MSD screening excludes 3 centers as misaligned.
- Fusion is performed on the remaining (screened-in) centers.

Application 1: Predicting Cognitive Status with the NACC data

Method	Sex	Race	Age	Weight	Height	Life	Alcohol	Smoke	Language
DEFUSE (const)	1.72	1.57	2.94	2.28	1.46	2.00	1.59	1.62	1.01
DEFUSE (linear)	2.04	1.82	3.55	3.72	1.65	2.42	2.05	1.91	1.15
DEFUSE (KRR)	2.82	2.12	4.51	4.30	2.34	3.33	2.76	1.87	1.56

NACC application: interpretation & takeaways

- **Efficiency gains are substantial:** e.g., KRR reweighting yields RE > 2 for many covariates and ≈ 4.5 for age.
- **Screening matters:** excluding misaligned centers avoids bias from non-transportable sources.
- **Practical message:** DEFUSE is useful when a key clinical variable is missing by design in some centers but you still want a target-population model.
- **Method takeaway:** more flexible reweighting (KRR) tends to deliver stronger gains than constant/linear allocations.

Application 2: Risk of Advanced Heart Diseases from MIMIC-III

- MIMIC-III (Medical Information Mart for Intensive Care III) is a large, freely available database comprising de-identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.
- Y: Validated golden label on whether the patient has advanced heart disease ($n = 1045$)
- $X_{\mathcal{P}_{mc}}$: age; High-density lipoprotein (HDL) cholesterol; Type II Diabetes

Application 2: Risk of Advanced Heart Diseases from MIMIC-III

- $X_{\mathcal{P}_m}$: HD-related laboratory (lab) variables

itemid	label	fluid	category	loinc code
50801	Alveolar-arterial Gradient	Blood	Blood Gas	19991-9
50802	Base Excess	Blood	Blood Gas	11555-0
50816	Oxygen	Blood	Blood Gas	19994-3
50820	pH	Blood	Blood Gas	11558-4
50821	pO2	Blood	Blood Gas	11556-8
50993	Thyroid Stimulating Hormone	Blood	Chemistry	3016-3
51137	Anisocytosis	Blood	Hematology	702-1
51233	Hypochromia	Blood	Hematology	728-6
51246	Macrocytes	Blood	Hematology	738-5
51252	Microcytes	Blood	Hematology	741-9
51260	Ovalocytes	Blood	Hematology	774-0
51267	Poikilocytosis	Blood	Hematology	779-9
51274	PT	Blood	Hematology	5902-2
51279	Red Blood Cells	Blood	Hematology	789-8

  - Single BM: A combined risk score from 14 lab values
  - $N_{\mathcal{LC}} = 467$ / $N_{\mathcal{LM}} = 578$ / $N_{\mathcal{UC}} = 7037$
  - Multiple BM: Thyroid-stimulating hormone (TSH)+ Hypochromia (HPO)
  - $N_{\mathcal{LC}} = 486$ / $N_{\mathcal{LM}} = 231 : 175 : 153$ / $N_{\mathcal{UC}} = 7014$

Incoporating Auxiliary HD-related features W

- A pre-screening procedure to select all the codified features that are associated with the ICD encount of HD $\rightarrow$ 140 features selected
- Those auxiliary features are used in the imputation and calibration models to better inform and Y.
- Incorporating $\mathbf{W}$ into the workflow:
  - Extended imputation models: $\mu(\boldsymbol{X}_{\Gamma^c}|\boldsymbol{X}_{\Gamma}, \mathbf{W}, Y; \boldsymbol{\eta})$ and $\nu(Y|\boldsymbol{X}, \mathbf{W}; \boldsymbol{\theta})$
  - Extended calibration models: $w(\boldsymbol{X}, \mathbf{W}, Y; \boldsymbol{\delta})$ and $h(\boldsymbol{X}, \mathbf{W}, Y; \boldsymbol{\zeta})$
- Bias-Immune

Results from Application 2: Risk of Advanced Heart Diseases from MIMIC-III

Method	Age	HDL	Diabetes	LRS	TSH	Hypochromia
Single BM scenario						
DEFUSE (const)	2.11	2.37	2.80	1.11	–	–
DEFUSE (linear)	3.02	2.55	2.89	1.14	–	–
DEFUSE (KRR)	3.61	2.80	3.36	1.17	–	–
HTLGMM / SC	1.76	1.55	1.56	1.02	–	–
Multiple BM scenario						
DEFUSE (const)	1.96	2.21	2.85	–	1.86	1.25
DEFUSE (linear)	3.28	2.43	3.36	–	3.01	1.54
DEFUSE (KRR)	3.65	3.61	5.72	–	5.38	1.70
SSL (LC+UC only)	1.00	0.99	1.06	–	0.94	1.13

MIMIC-III application: interpretation & takeaways

- **Large gains for key predictors:** e.g., in multiple BM, DEFUSE (KRR) achieves RE $\approx$ 5–6 for diabetes and TSH.
- **Unlabeled data alone isn't enough:** SSL (using only $+$) provides near-1 RE for many coefficients.
- **BM sources add real signal:** incorporating labeled BM cohorts via control variates is what drives efficiency.
- **Practical message:** DEFUSE is well-suited to EHR phenotyping settings with scarce gold-standard labels and heterogeneous lab availability.

Summary and Discussions

- A new algorithm using **data-adaptive partial-score-based control variates** to handle block-wise missing, covariate-shifts and semi-supervised learning during data fusions.
- Ensured efficiency gain: intrinsic efficiency in a chosen class; semiparametric efficiency in the ideal limit.
- Robust against imperfect control models. Flexible to accommodate ML-based distribution estimation
- Practical safeguard: debiased MSD screening for misaligned sources.

Acknowledge and Advertisement

Thanks!

This research is supported by NSF (DMS-1953527) and NIH (R01AG087496, RF1-AG072272)

Columbia Biostatistics is recruiting faculty; please stay tuned for upcoming ads in UF and ASA!