Efficient Inference for Distributed Data with Structural Missingness

Huiyuan Wang, Ph.D.
Email: huiyuanw@upenn.edu
University of Pennsylvania

ICSA, Shenzhen

June, 2026

Joint work

Jingyue Huang

Postdoc, University of Pennsylvania

Yong Chen

Professor, University of Pennsylvania

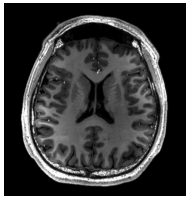
Goal: use fragmented external data without asking sites to share individual records.

Motivation

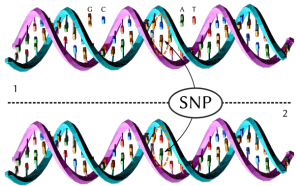
Modern biomedical studies are rich, but the data are rarely rich in the same place.

	Age	Gender	BMI
patient 1	65	M	22
patient 2	54	F	25
...	...	...	...
patient n	50	M	30

Demographics



Imaging

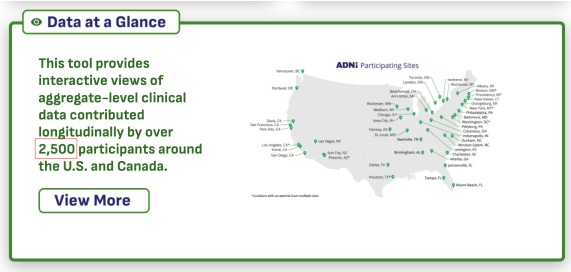


Omics

Different sites may collect different subsets of these modalities. The target site has complete data, but usually not enough of it.

Distributed data

- Sites hold complementary pieces of the same scientific question.
- Legal, privacy, and governance constraints limit record-level sharing.
- Summary statistics are often feasible.



https://adni.loni.usc.edu/

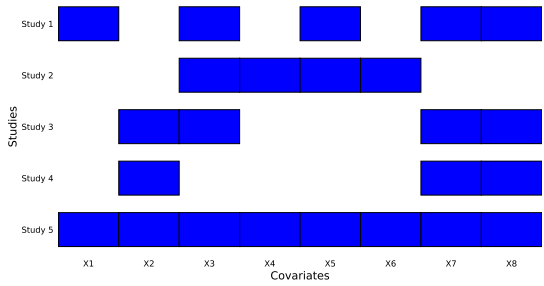
Can external sites improve inference at Site 0 using summaries only?

Structural missingness

Site 0: complete variables and outcome.

External sites: large samples, but only selected blocks of variables.

Target: a parameter $\theta(P)$ at the complete-data distribution.



Representative: Kundu et al. (2019)

The missingness pattern is structural, not accidental.

Kundu, P., Tang, R., and Chatterjee, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika*, 106(3), 567–585.

A running template

Site	Distribution	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
Site 0	P	$Z_{\mathcal{G}_1}$	$Z_{\mathcal{G}_2}$	$Z_{\mathcal{G}_3}$	$Z_{\mathcal{G}_4}$	$Z_{\mathcal{G}_5}$	$Z_{\mathcal{G}_6}$
Site 1	Q_1	$Z_{\mathcal{G}_1}$	–	$Z_{\mathcal{G}_3}$	–	–	–
Site 2	Q_2	–	–	$Z_{\mathcal{G}_3}$	–	$Z_{\mathcal{G}_5}$	–
Site 3	Q_3	$Z_{\mathcal{G}_1}$	–	–	$Z_{\mathcal{G}_4}$	–	$Z_{\mathcal{G}_6}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Site J	Q_J	–	$Z_{\mathcal{G}_2}$	–	$Z_{\mathcal{G}_4}$	$Z_{\mathcal{G}_5}$	$Z_{\mathcal{G}_6}$

Site 0 estimates $\theta(P)$ using individual complete records.

External sites return only low-dimensional summaries of observed blocks.

What is hard?

External data are useful only through what they share with the target influence function.

Statistical tension

Site 0 has the right variables, but limited sample size.

Privacy tension

External sites may be large, but can only report summaries.

- We want valid confidence intervals for $\theta(P)$.
- We want smaller variance than the Site 0-only estimator.
- We do not want to pay a price when an external site is weak or irrelevant.

Formal setup

The target is a complete-data functional, but external sites observe only marginal blocks.

- Site 0: $Z = (Z_1, \ldots, Z_p)^\top \sim P$, sample size n.
- Site j: $Z_{B_j} = (Z_b)_{b \in B_j}^\top \sim Q_j = P_{B_j}$, sample size n_j.
- Target: $\theta^* = \theta(P) \in \mathbb{R}^d$.
- Influence function $\varphi(Z; \theta, \eta)$ satisfies, for a regular submodel with score s,

$$\frac{d}{dt}\theta(P_t)\Big|_{t=0} = \mathbb{E}_P\{\varphi(Z; \theta^*, \eta^*)s(Z)\}, \qquad \mathbb{E}_P\{\varphi(Z; \theta^*, \eta^*)\} = 0.$$

Starting point

The one-step estimator is the internal-data benchmark.

$$\hat{\theta}^{\mathrm{os}} = \hat{\theta}^{\mathrm{init}} + \frac{1}{n} \sum_{i \in S_0} \hat{\varphi}(Z_i)$$

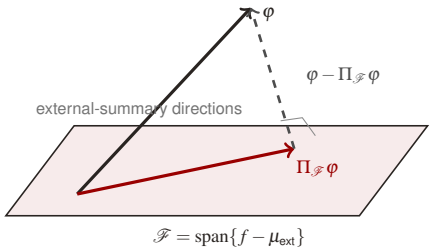
- $\hat{\varphi}$ is the estimated influence function.
- Under standard nuisance-rate conditions,

$$n^{1/2}(\hat{\theta}^{\mathrm{os}} - \theta^*) \to_d N\{0, \mathrm{Var}_P(\varphi)\}.$$

- If φ is the efficient influence function, this is the local minimax benchmark using Site 0 only.
- The question is how to add external summaries without changing the target.

Main idea

Use external summaries as control variates for the influence function.



Choose transfer functions $f_j(Z_{B_j})$ that can be evaluated at Site 0 and summarized at Site j.

AOS replaces φ by the residual after projecting onto the external-summary space:

$$\varphi_{\mathrm{aos}} = \varphi - \Pi_{\mathscr{F}}\varphi.$$

Estimator

The augmented estimator is still a one-step estimator, but with a variance-reducing correction.

$$\hat{\theta}^{\text{aos}} = \hat{\theta}^{\text{init}} + \frac{1}{n} \sum_{i \in S_0} \left\{ \hat{\varphi}(Z_i) - \hat{\beta}^\top (\hat{f}_i - \tilde{\mu}_{\text{ext}}) \right\}.$$

- $\hat{f}_i$ is the vector of transfer functions evaluated on Site 0 records.
- $\tilde{\mu}_{\text{ext}}$ is estimated at external sites.
- $\hat{\beta}$ learns how much each external summary should correct the influence function.

Theorem 1: validity and do-no-harm

Theorem. Suppose

$$\|\hat{\beta} - \beta^{\star}\|_1 + \|\hat{\mathrm{var}}(f) - \mathrm{var}_P(f)\|_{\max} + \|\tilde{\Sigma}_{\mathrm{ext}} - \Sigma_{\mathrm{ext}}\|_{\max} = o_P(1).$$

Under regularity conditions,

$$n^{1/2}\{\hat{\theta}^{\mathrm{aos}} - \theta^*\} \to_d N(0_d, \Sigma_f^*),$$

where

$$\Sigma_f^* = \mathrm{var}(\varphi) - \{\mathrm{cov}(f, \varphi)\}^\top \{\Sigma_{\mathrm{ext}} + \mathrm{var}(f)\}^{-1} \mathrm{cov}(f, \varphi).$$

External summaries reduce the variance by a positive semidefinite term.

Goldberg et al. (2024). "To do no harm–and the most good–with AI in health care." *Nature Medicine*, 30(3), 623–627.

Why regularization enters

The target site chooses which external summaries are worth using.

At Site 0, estimate the projection coefficient

$$\hat{\beta} = \underset{b}{\operatorname{argmin}} \sum_{i \in S_0} \{\hat{\varphi}_i - b^\top (\hat{f}_i - \tilde{\mu}_{\text{ext}})\}^2 + \lambda \|b\|_1.$$

- Many sites or basis functions make p large.
- Some summaries add little incremental signal.
- The ℓ_1 penalty screens weak or redundant corrections.

The theory only needs $\hat{\beta}$ to be close to $\beta^\star$; lasso is one practical route when the useful correction is sparse.

Scalability

The regularized correction remains valid when many external summaries are screened.

- If $f_j(Z_{B_j})$ are sub-Gaussian, concentration gives

$$\|\hat{\mathrm{var}}(f) - \mathrm{var}_P(f)\|_{\max} + \|\tilde{\Sigma}_{\mathrm{ext}} - \Sigma_{\mathrm{ext}}\|_{\max} \lesssim \sqrt{\frac{\log p}{n}} + \min_j \sqrt{\frac{\log p_j}{n_j}}.$$

- Under standard sparse-regression conditions,

$$\|\hat{\beta} - \beta^*\|_1 \lesssim \|\beta^*\|_0 \sqrt{\frac{\log p}{n}}.$$

- Therefore the theorem condition can hold even when p grows exponentially with n, provided the useful correction is sparse.

Optimal transfer functions

Proposition. The variance-minimizing transfer vector

$$f^* = \{f_1^*(Z_{B_1})^\top, \ldots, f_J^*(Z_{B_J})^\top\}^\top$$

solves, for $k = 1, \ldots, J$,

$$\mathbb{E}\left\{\varphi(Z) - \sum_{j=1}^{J} \gamma_{j,k} f_j^*(Z_{B_j}) \mid Z_{B_k}\right\} = 0_d, \qquad \mathbb{E}\{f_k^*(Z_{B_k})\} = 0_d,$$

where $\gamma_{j,k} = 1 + I(j = k) \lim_{n\to\infty} n/n_k$.

For one external site, this reduces to $f_1^*(Z_{B_1}) \propto \mathbb{E}\{\varphi(Z; \theta^*, \eta^*) \mid Z_{B_1}\}$.

Choosing transfer functions

The proposition says what to learn: the predictable part of the influence function.

- Simple choice: identity functions or prespecified basis functions.
- Data-driven choice: learn f_j from Site 0, then send it to Site j for summary evaluation.
- Flexible choices include basis expansions, kernels, or neural networks, depending on the scientific problem.

Theorem 2: data-driven optimality

Theorem. Let $\tilde{f}_j^*(Z_{B_j})$ be the probability limit of the learned transfer function $\hat{f}_j(Z_{B_j})$. For $d = 1$, under the conditions of Theorem 1 and additional regularity conditions,

$$n^{1/2}\{\hat{\theta}_{\text{dd}}^{\text{aos}} - \theta^*\} \to_d N(0, \Sigma_{\tilde{f}^*}^*),$$

where

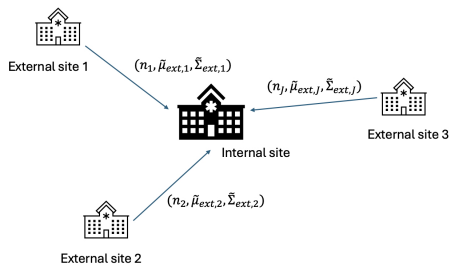
$$\Sigma_{\tilde{f}^*}^* = \text{var}\{\varphi(Z; \theta^*, \eta^*)\} - \beta^{*\top}\{\text{var}(\tilde{f}^*) + \Sigma_{\text{ext}}^*\}\beta^*.$$

If the working model for f_j^* is correctly specified, the data-driven AOS estimator achieves the semiparametric efficiency bound.

Workflow and communication

All modeling happens at Site 0; external sites only evaluate and summarize.

1. Fit the initial estimator and influence function at Site 0.
2. Choose or learn f_j for each external site's observed variables.
3. Send the fitted transfer function to Site j.
4. Site j returns summary statistics.
5. Site 0 computes $\hat{\theta}^{\text{aos}}$ and its standard error.



Site j returns only n_j, $\tilde{\mu}_{\text{ext},j}$, $\tilde{\Sigma}_{\text{ext},j}$.

No outcomes, no individual records, no cross-site joins.

Simulation 1

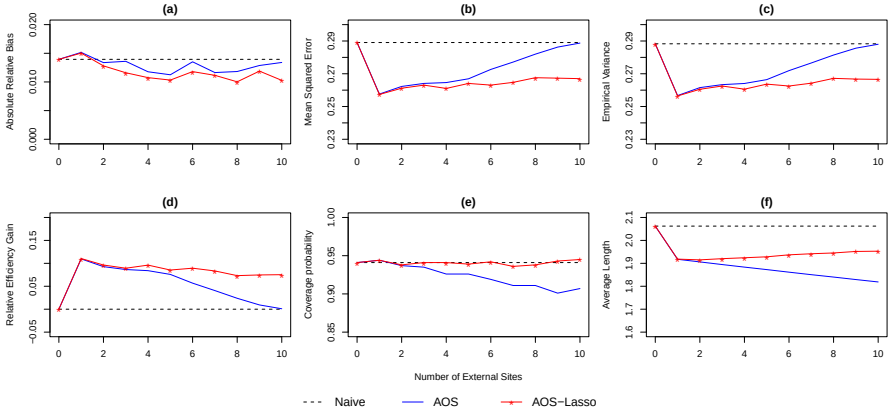
When external covariates are more predictive, the efficiency gain becomes larger.

Correlation ρ	Method	MSE $(\times 10^2)$	Gain
0.3	Naive	149.4	0.0%
	AOS-Lasso	141.9	5.0%
0.5	Naive	53.6	0.0%
	AOS-Lasso	42.6	20.5%
0.7	Naive	27.2	0.0%
	AOS-Lasso	15.3	43.8%

Coverage remains close to nominal; the main change is interval length.

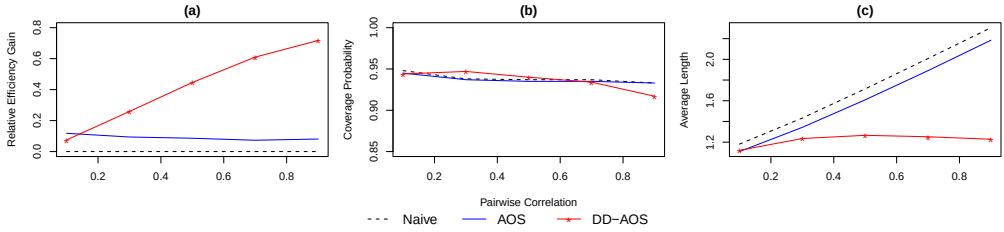
Simulation 2

Regularization protects against uninformative external sites.



Simulation 3

Learning transfer functions can recover nonlinear information that identity summaries miss.



In this example, the outcome depends on pairwise interactions. Learned summaries are designed to see beyond the raw marginal means.

How to read the theory

The formal results support three practical promises.

Valid

Standard errors remain usable.

Helpful

External summaries reduce variance when informative.

Robust

Weak summaries can be downweighted.

The assumptions mainly ensure that the correction term is estimated accurately enough and the external summaries target the same marginal distribution.

When this is useful

- The complete-data site is scientifically central but relatively small.
- External sites observe meaningful subsets of the variables.
- Record-level sharing is not feasible, but function evaluation and summary reporting are feasible.
- The target parameter has an influence-function representation.

The framework turns distributed structural missingness into a variance-reduction opportunity.

Takeaways

- Start from a valid Site 0 one-step estimator.
- Use external summaries as control variates for its influence function.
- Keep communication summary-level.
- Use lasso or data-driven transfer functions when many sites or nonlinear signals are expected.

Better precision, same target, no individual-level data exchange.

Thanks!

Any Questions?