

MOSAIC-H: One-Shot Federated Inference for Heterogeneous EHR Networks via Tensor-Train Loss Compression

Qiaoshi Yang¹, Jie Hu¹, Yudong Wang², Jianqing Fan³, Yong Chen¹

¹ University of Pennsylvania

² University of Macau

³ Princeton University

Background

Multi-site electronic health record (EHR) networks offer unprecedented opportunities for population-scale biomedical research, enabling broader population coverage, improved generalizability, and increased statistical power for studying drug safety, disease progression, and comparative treatment effectiveness.¹ However, two fundamental constraints complicate federated statistical inference in distributed EHR networks. First, patient-level records cannot leave local sites due to privacy regulations and institutional data governance requirements, preventing direct data pooling. Second, existing communication-efficient methods—including ODAL² and similar surrogate likelihood approaches³—transmit low-order derivative summaries (gradients, Hessians) and either assume all parameters are homogeneous across sites or require prior knowledge of which effects are site-specific versus globally shared. When the parameter structure is unknown or misspecified, these methods yield biased estimates that may be of limited clinical value to individual institutions. A federated approach that jointly identifies the shared-parameter structure in a data-driven manner and achieves lossless recovery of the pooled estimator—without transmitting patient-level records—would directly address both constraints and is particularly relevant to the OHDSI community’s mission of enabling trustworthy evidence generation from distributed EHR networks.

Methods

We propose MOSAiC-H (Multi-site One-Shot Approximation via Compressed loss functions for Heterogeneous models), a one-shot federated inference framework for heterogeneous parametric models. Each local site approximates its empirical loss function over a Chebyshev parameter grid using tensor-train (TT) cross approximation,⁴ producing compact low-rank TT cores that are transmitted to a central server in a single communication round. The server

reconstructs continuous surrogate loss functions via barycentric interpolation and optimizes the surrogate pooled objective. When the partition of shared and site-specific parameters is unknown, MOSAiC-H introduces a SCAD-fused contrast reparameterization: each site's parameter is written as a reference value plus a site-specific deviation, and a SCAD penalty shrinks deviations to zero, identifying the shared feature set in a fully data-driven manner. We established three theoretical guarantees: (i) the surrogate loss, gradient, and Hessian uniformly approximate their pooled counterparts to arbitrarily small error controlled by TT rank and grid resolution; (ii) MOSAiC-H consistently identifies the homogeneous coordinate set; and (iii) the shared-parameter estimator achieves the $\sqrt{N}$ convergence rate and the same asymptotic distribution as the infeasible pooled oracle. We validated MOSAiC-H through simulations ($K=10$ sites, $n=5,000$ per site, $B=100$ Monte Carlo replications) and applied it to a Penn Medicine multi-site EHR cohort ($N=163,308$ outpatient encounters across $K=5$ hospital sites: Cedar Creek Hospital, Hospital of the University of Pennsylvania, Medical College of Pennsylvania, Pennsylvania Hospital, and Presbyterian Medical Center) to study the association between telemedicine utilization and 30-day follow-up visit rates, adjusting for age, sex, race/ethnicity, travel distance, and mental health comorbidity.

Results

In simulation, MOSAiC-H point estimates and 95% Wald confidence intervals were essentially indistinguishable from the pooled oracle across all seven parameters. The SCAD-fused penalty correctly identified the heterogeneous intercept and all shared regression coefficients in every one of the $B=100$ replications. ODAL-H with a misspecified partition produced displaced, misleadingly narrow confidence intervals for the affected parameters, with empirical coverage collapsing to 0% for coefficients forced into the wrong partition. MOSAiC-H confidence intervals closely matched those of the pooled oracle throughout, and SE calibration ratios were near 1.0 for all coordinates, confirming accurate uncertainty quantification. In the Penn Medicine EHR application, MOSAiC-H identified demographic covariates (age, sex, race/ethnicity) as globally shared across all five hospitals—consistent with the expectation that fundamental patient characteristics influence follow-up behavior independently of institutional context—and identified telemedicine encounter type, travel distance, and mental health comorbidity as site-specific, reflecting differences in local care infrastructure and protocols. The telemedicine coefficient varied substantially: from -0.065 at HUP (telemedicine utilization

rate 25.6%, the highest across sites) to -0.617 at MCP and -0.515 at PPMC. At HUP, where telemedicine is deeply integrated into care pathways, raw 30-day revisit rates were nearly identical between telemedicine (74.1%) and in-person visits (73.1%), suggesting that mature telemedicine adoption preserves care continuity. At sites with lower telemedicine utilization, the negative association was substantially stronger, indicating that telemedicine at those sites primarily serves acute or episodic needs that substitute for in-person follow-up. MOSAiC-H point estimates and standard errors closely matched the pooled oracle throughout; ODAL-H variants with misspecified partitions produced substantially biased or inflated estimates for the affected parameters.

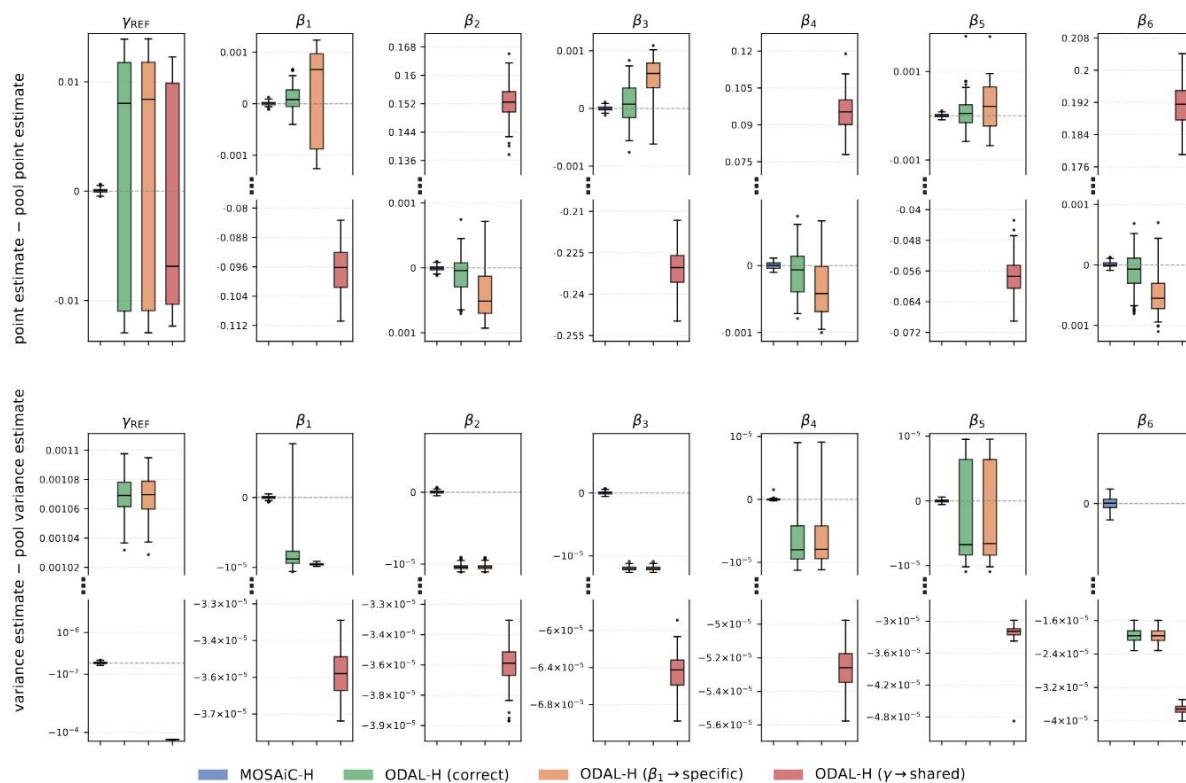


Figure 1. Distribution of point-estimate (top) and variance-estimate (bottom) deviations from the pooled oracle across 100 Monte Carlo replications, stratified by parameter and method.

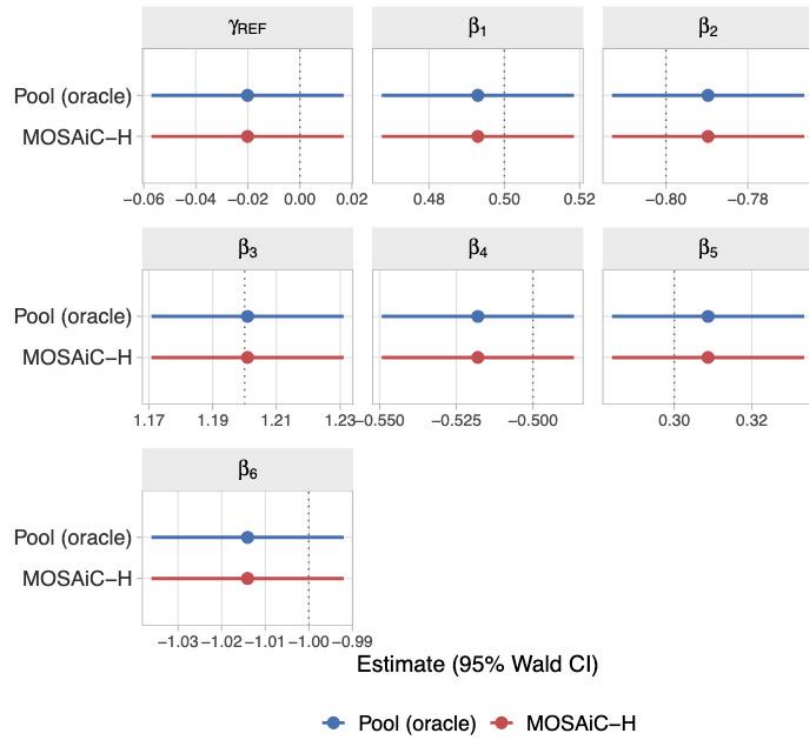


Figure 2. 95% Wald confidence intervals from a representative simulation replication.

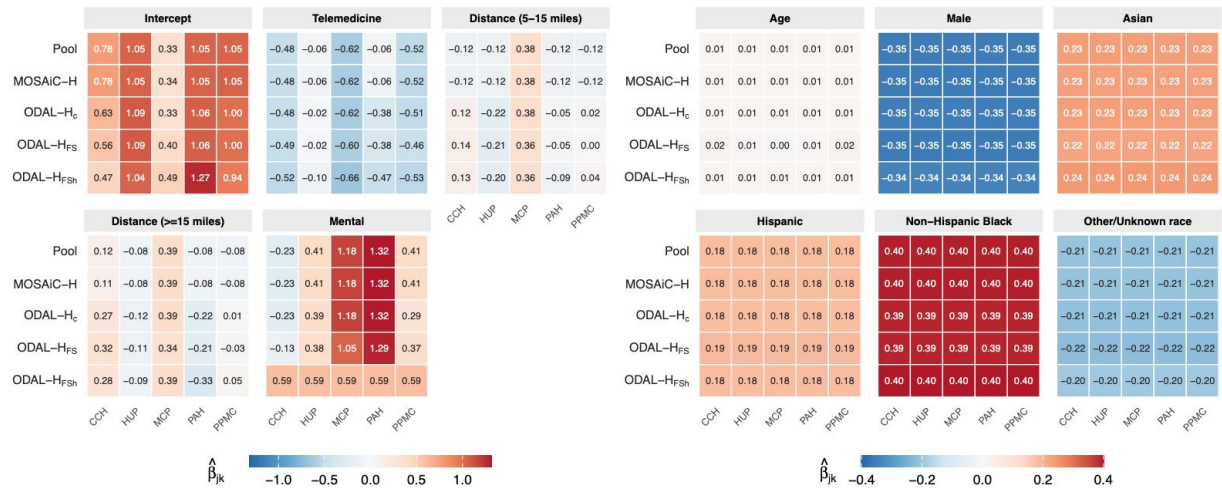


Figure 3. Support set recovery of specific (left) and shared (right) parameters in Penn EHR data, stratified by methods.

Variable	Pool	MOSAiC-H	ODAL-H	ODAL-H _{FS}	ODAL-H _{FS_h}
Site-specific parameters (HUP reference)					
Intercept	1.046 (0.005)	1.048 (0.005)	1.091 (0.014)	1.093 (0.014)	1.044 (0.014)
<i>telemedicine</i> [HUP]	−0.065 (0.007)	−0.065 (0.007)	−0.019 (0.019)	−0.024 (0.020)	−0.096 (0.018)
Distance 5–15mi [HUP]	−0.123 (0.005)	−0.125 (0.005)	−0.216 (0.018)	−0.212 (0.018)	−0.197 (0.018)
Distance ≥15mi [HUP]	−0.083 (0.007)	−0.084 (0.007)	−0.117 (0.021)	−0.113 (0.021)	−0.093 (0.021)
<i>mental</i> [HUP]	0.410 (0.007)	0.409 (0.007)	0.393 (0.020)	0.383 (0.020)	0.592 (0.015)
Shared parameters					
AGE	0.011 (<0.001)	0.011 (<0.001)	0.011 (<0.001)	0.010 (0.001)	0.010 (0.001)
Gender = Male	−0.353 (0.005)	−0.354 (0.005)	−0.352 (0.012)	−0.354 (0.012)	−0.340 (0.012)
Race = Asian	0.233 (0.010)	0.231 (0.010)	0.230 (0.027)	0.224 (0.027)	0.244 (0.026)
Race = Hispanic	0.183 (0.010)	0.183 (0.010)	0.180 (0.029)	0.187 (0.029)	0.182 (0.029)
Race = NHBlack	0.397 (0.006)	0.396 (0.006)	0.392 (0.016)	0.391 (0.016)	0.402 (0.016)
Race = Other	−0.208 (0.008)	−0.210 (0.008)	−0.212 (0.019)	−0.219 (0.019)	−0.196 (0.019)

Note: ODAL-H_{FS}: forced site-specific specification; ODAL-H_{FS_h}: forced shared specification.

Wald *p*-values are smaller than 0.001 for all reported coefficients, except for *telemedicine* [HUP] under ODAL-H and ODAL-H_{FS}.

Table 1. Estimated coefficients (SE) for the real-data analysis.

Conclusion

MOSAiC-H is a lossless, one-shot federated inference framework that requires no prior knowledge of the shared-parameter structure and transmits no patient-level data. By communicating compressed functional representations of local loss functions rather than gradients or Hessians, it avoids the finite-sample discrepancies inherent in derivative-based surrogate methods and achieves statistical equivalence to the infeasible pooled estimator—a property that does not depend on local sample sizes being large. Applied to the Penn Medicine EHR network, MOSAiC-H reveals that the telemedicine–visit association is clinically meaningful but structurally heterogeneous: the negative association between telemedicine

encounters and 30-day follow-up rates attenuates substantially at sites with higher telemedicine adoption and integration, while remaining strong at sites where telemedicine is less established. A pooled estimate would have obscured these site-level differences, which carry direct implications for hospital-specific telemedicine policy and care continuity planning. The OHDSI community can leverage MOSAiC-H to enable trustworthy, privacy-preserving federated analyses across distributed EHR networks without requiring advance knowledge of which clinical effects vary across sites—addressing a practical gap that is common to multi-institutional observational studies.

References

1. Tong J, et al. DISC2O: distributed and scalable causal inference for observational data. *J Am Med Inform Assoc.* 2025.
2. Duan R, et al. Learning from electronic health records across multiple sites: a communication-efficient and privacy-preserving distributed algorithm. *J Am Med Inform Assoc.* 2020;27(3):376–385.
3. Jordan MI, Lee JD, Yang J. Communication-efficient distributed statistical inference. *J Am Stat Assoc.* 2019;114(526):668–681.
4. Oseledets IV. Tensor-train decomposition. *SIAM J Sci Comput.* 2011;33(5):2295–2317.