# Distributed inference under arbitrary cross-site heterogeneity

Fast-CAR: one-shot recovery of pooled GLM inference


Huiyuan Wang[1]

with Jie Hu[1], Yiwen Lu[1], Jingmei Qiu[2], Jianqing Fan[3], and Yong Chen[1]

[1]University of Pennsylvania · [2]University of Delaware · [3]Princeton University


Joint Statistical Meetings 2026 · Boston, MA

# Population scale is already possible—without pooling records

FDA SENTINEL DISTRIBUTED DATABASE · 2000–2025

**138.7M**
members currently accruing new data

**1.5B**
person-years of data

**13.5M**
mother–infant linked deliveries

**WHY SCALE MATTERS**

**Rare adverse events and drugs used in small populations become studyable.**

**HOW THE NETWORK OPERATES**

**Partners keep records behind their firewalls and return de-identified results.**

Sources: FDA Sentinel, *Key Database Statistics* and *How Sentinel Gets Its Data*, updated October 2025.

# Heterogeneity makes local estimates a fragile unit of exchange

$$P_1, \ldots, P_K \text{ may differ arbitrarily} \quad \Longrightarrow \quad R_1(\theta), \ldots, R_K(\theta) \text{ have different shapes.}$$

**Different populations**
Local optima reflect different case mix, treatment patterns, and outcome processes.

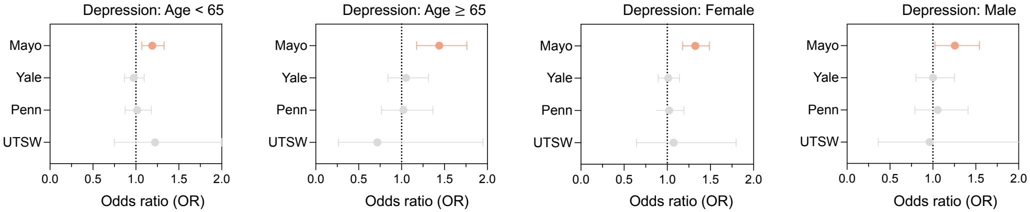
**Sparse or zero events**
Separation can make a local coefficient or standard error unstable or undefined.

**Many downstream targets**
Overall, subgroup, and interaction analyses can trigger repeated site-side fitting.

**Research question**
Can one communication round recover the pooled empirical-risk minimizer—without sharing individual records or modeling how sites differ?

The next example shows that these are not hypothetical edge cases.

# A harmonized protocol does not make sites homogeneous




**One network target**

Four health systems follow a common protocol, while patient-level records remain local.

**Case mix shifts sharply**

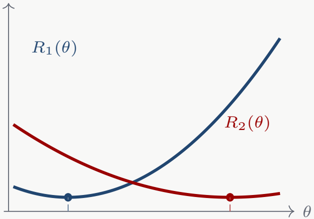
Mean age spans about 57–67 years; Black representation spans about 4%–41% across site-arm cells.

**Sparse strata matter**

Sparse events and separation can make local subgroup fits unstable or undefined.

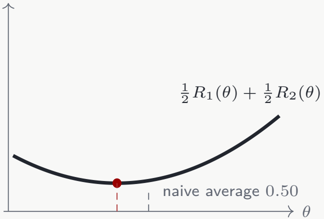
# Averaging local estimators does not recover the pooled fit

### What estimator-level aggregation keeps




Local minima: $\widehat{\theta}_1 = -1.2$, $\widehat{\theta}_2 = 2.2$
Naive average: $\frac{1}{2}(\widehat{\theta}_1 + \widehat{\theta}_2) = 0.50$

### What pooled analysis actually uses




Pooled minimizer: $\widehat{\theta}_{\mathrm{P}} \approx -0.17$
$\arg\min_\theta \sum_k w_k R_k(\theta) \neq \sum_k w_k \arg\min_\theta R_k(\theta)$

# The pooled target is an empirical-risk minimizer

## Gold-standard pooled analysis

$$\widehat{\theta}_{\mathrm{P}} = \arg\min_{\theta \in \Theta} R(\theta), \qquad R(\theta) = \sum_{k=1}^{K} w_k \left\{ Q_k(\theta) - C_k^{\top}\theta \right\}, \quad w_k = \frac{n_k}{N}.$$

## Nonlinear local risk

For logistic regression,

$$Q_k(\theta) = \frac{1}{n_k} \sum_{i=1}^{n_k} \log\left\{1 + \exp(z_{ki}^{\top}\theta)\right\}.$$

This function changes shape with the local joint distribution.

## Low-dimensional linear summary

$$C_k = \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ki} z_{ki}.$$

The vector $C_k$ is directly shareable; the challenge is representing $Q_k(\cdot)$ compactly.

Pooling aggregates local risk shapes—not local estimates.

# CAR shares risk functions, not local estimates

Behind each site firewall

Coordinating center

Reusable outputs

**one transmission**

**Site 1** local records
build $\widehat{Q}_1$; send $(n_1, C_1,$ TT cores$)$

**Site 2** local records
build $\widehat{Q}_2$; send $(n_2, C_2,$ TT cores$)$

**Site $K$** local records
build $\widehat{Q}_K$; send $(n_K, C_K,$ TT cores$)$

**Reconstruct and aggregate the risk**

$$\widehat{R}_{\mathrm{CAR}}(\theta) = \sum_{k=1}^{K} w_k \{\widehat{Q}_k(\theta) - C_k^\top \theta\}$$

$$\widehat{\theta}_{\text{Fast-CAR}} = \arg\min_{\theta \in \Theta} \widehat{R}_{\mathrm{CAR}}(\theta)$$

- overall association
- age / sex subgroups
- joint subgroup contrasts
- profile likelihood CIs

No local GLM fit • no site-selection model • no iterative gradient exchange

**The change of primitive**

$\mathrm{CAR}$ turns distributed inference under heterogeneity into a controlled function-approximation problem.

# One upload supports many coefficient-restricted submodels

**Build and upload once.** Each site approximates the risk for one sufficiently rich design:

$$z_{\text{full}}(A,X) = \left(1, A, X^\top, M^\top, (AM)^\top, h(M)^\top, g(A,M)^\top\right)^\top \implies \widehat{R}_F(\theta_{\text{full}}).$$

## A submodel is a constrained re-fit

$$\mathcal{M}(C) = \{\theta_{\text{full}} \in \Theta : C\theta_{\text{full}} = 0\},$$

$$\widehat{\theta}_F(C) = \arg\min_{\theta \in \mathcal{M}(C)} \widehat{R}_F(\theta).$$

Re-minimize on $\mathcal{M}(C)$; do not take restricted coefficients from the unrestricted fit.

## Changing $C$ changes the question

- **Overall:** zero subgroup and treatment–subgroup blocks.
- **One subgroup:** retain $M_j$ and $AM_j$.
- **Intersectional:** retain selected main effects, interactions, and higher-order terms.

**Reuse.** New $C$ at the center; same upload, no site re-fit. Since $\mathcal{M}(C) \subseteq \Theta$,

$$\sup_{\theta \in \mathcal{M}(C)} |\widehat{R}_F(\theta) - R(\theta)| \le \sup_{\theta \in \Theta} |\widehat{R}_F(\theta) - R(\theta)|.$$

# Uniform risk approximation yields near-lossless estimation

### Aggregated approximation error

$$\Delta_Q = \sum_{k=1}^{K} w_k \, \|\widehat{Q}_k - Q_k\|_\infty, \qquad \|f\|_\infty = \sup_{\theta\in\Theta} |f(\theta)|.$$

### Deterministic perturbation guarantee

If the pooled risk is $\kappa$-strongly convex on $\Theta$, then

$$\boxed{\|\widehat{\theta}_{\mathrm{CAR}} - \widehat{\theta}_{\mathrm{P}}\|_2^2 \le \frac{4\Delta_Q}{\kappa}}$$

**Finite sample**
Deterministic for the realized distributed dataset.

**No heterogeneity penalty**
No dependence on $K$ or a model for how the site distributions differ.

**Tunable error**
Accuracy is governed by risk approximation and pooled curvature.

# Fast-CAR compresses GLM risk without a dense grid

**Local GLM risk**

$$Q_k(\theta) = \frac{1}{n_k} \sum_i b(z_{ki}^\top \theta)$$

high-dimensional in $\theta$, but every term depends on one scalar $z_{ki}^\top \theta$

→

**1. Polynomialize**

$$b(t) \approx p_m(t)$$

$$p_m(t) = \sum_{\ell=0}^{m} c_\ell T_\ell(t)$$

analytic GLM links: error $\lesssim e^{-\rho m}$

→

**2. Explicit TT cores**

$$p_m(z^\top \theta) = G_1\, G \cdots G_p$$

$\boxed{G_1}$ - $\boxed{G}$ - $\cdots$ - $\boxed{G_p}$

exact for the polynomial; intermediate rank $\leq m + 1$

→

**3. Grid + TT-rounding**

$$\widetilde{\mathcal{F}}_k \xrightarrow{\text{round}_\varepsilon} \widehat{\mathcal{F}}_k$$

$$\widehat{Q}_k = \mathcal{J}_B[\widehat{\mathcal{F}}_k]$$

if $B \geq m$, there is no extra interpolation bias

Dense grid: $(B+1)^p$ entries ×

TT cores: $\mathcal{O}\{p(B+1)r^2\}$ entries ✓

The scalar approximation is one-dimensional; the multivariate risk stays in a rank-controlled tensor format.

# One compressed risk preserves inference and supports many targets

**Fast-CAR error for analytic GLM links**

$$\|\widehat{\theta}_{\text{Fast-CAR}} - \widehat{\theta}_{\text{P}}\|_2^2 \;\lesssim\; e^{-\rho m} + (\log B)^p \varepsilon.$$

If $\Delta_Q = o_p(N^{-1})$, then for smooth $\psi$,

$$\sqrt{N}\,\|\psi(\widehat{\theta}_{\text{Fast-CAR}}) - \psi(\widehat{\theta}_{\text{P}})\|_2 = o_p(1),$$

and the profile likelihood-ratio statistics satisfy

$$|\Lambda_{\text{Fast}}(\tau_0) - \Lambda_{\text{P}}(\tau_0)| \le 4N\Delta_Q = o_p(1).$$

one rich pooled risk $\widehat{R}_{\text{Fast-CAR}}$

↓

overall adjusted association

↓

age or sex subgroup contrast

↓

intersectional contrast

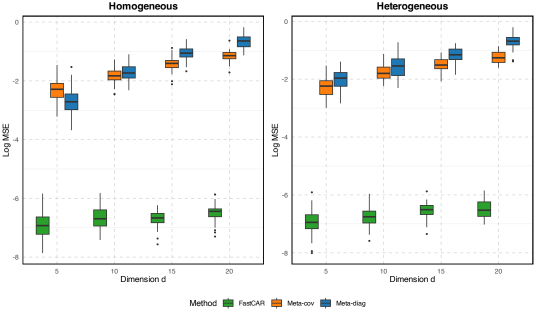
↓

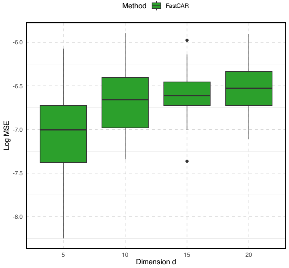
profile likelihood confidence set

**Reuse.** Central restrictions and contrasts require no site rerun.

# Fast-CAR stays close to pooling as heterogeneity intensifies

Logistic model $\cdot$ $N = 500 \cdot K = 5 \cdot d = 5, 10, 15, 20 \cdot m = 9,\ B = 20,\ \varepsilon = 10^{-6}$



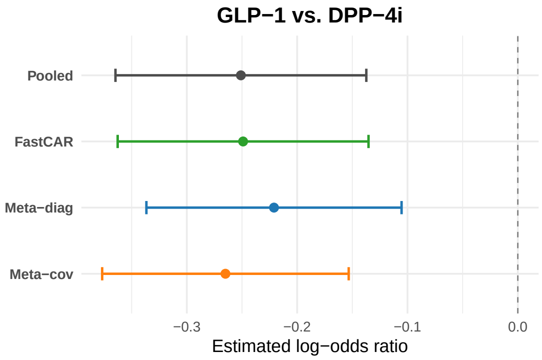

Extreme single-arm separation

Random partition and outcome-dependent heterogeneous partition

**Main reading**

Fast-CAR tracks pooling; estimator-level meta-analysis deteriorates under selection as dimension grows.

# In five hospitals, Fast-CAR reproduces the pooled analysis




Adjusted log-odds ratio for GLP-1RA versus DPP-4i

**Internal validation design**

- Five-hospital network
- Three-year MACE outcome
- Centralized pooled fit

**Treatment coefficient (SE)**

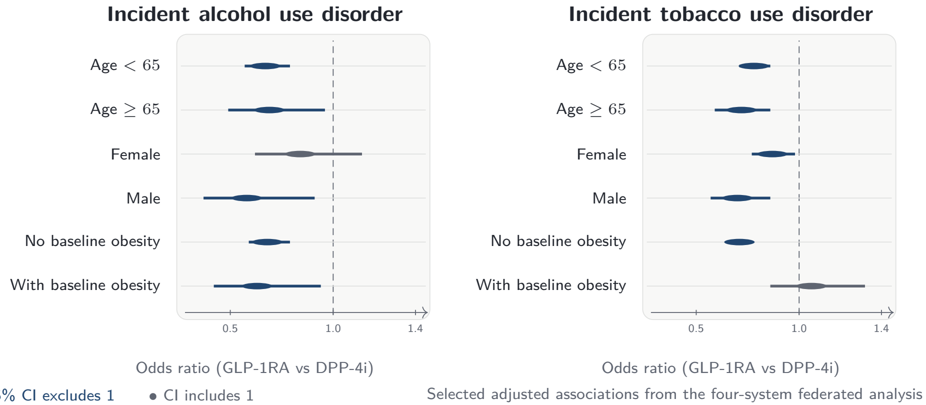
| | |
|---|---|
| Pooled | $-0.251$ (0.058) |
| Fast-CAR | $-0.249$ (0.058) |

Both $p < 0.001$.

Agreement extends to all coefficients and standard errors.

# One federated fit supports many clinically relevant subgroups

Four health systems · records remain local · selected prespecified contrasts




Selected adjusted associations from the four-system federated analysis

These are adjusted associations; subgroup findings are suggestive rather than confirmatory.

# Treat heterogeneity as structure to absorb, not a model to guess

1. **Target the pooled risk.** Averages of local estimates generally do not recover the pooled empirical-risk minimizer.
2. **Approximate and communicate once.** Fast-CAR combines Chebyshev approximation with tensor-train compression and achieves a controllable error relative to full pooling.
3. **Reuse the same object.** One compressed risk supports point estimation, likelihood inference, subgroup contrasts, and restricted models.

**Scope and boundary**

The current construction targets single-linear-predictor ridge losses such as GLMs. It inherits the interpretation of the pooled working model, does not create causal identification, and is not by itself a formal differential-privacy guarantee.

**Under heterogeneity, share risk—not local estimates.**

# Thank you!

Huiyuan Wang

**hyw.stat@outlook.com**

# Add Gaussian noise once; Fast-CAR proceeds unchanged

**1. LOCAL RISK EVALUATIONS**

$$q_k = \{L_k(\theta_j)\}_{j=1}^{J}$$

$\Longrightarrow$

**2. PRIVATE RELEASE**

$$Y_k = q_k + \xi_k, \quad \xi_k \sim N(0, \sigma_k^2 I_J)$$

**CALIBRATE ONCE USING THE QUERY SENSITIVITY $\Delta_{2,k}$**

$$\sigma_k^2 \;\geq\; \frac{2\Delta_{2,k}^{2}\log(1.25/\delta)}{\epsilon^2} \quad\Longrightarrow\quad Y_k \text{ is } (\epsilon,\delta)\text{-DP.}$$

**3. CONTINUE WITH THE USUAL PIPELINE**

**noisy evaluations $Y_k$ $\longrightarrow$ Fast-CAR compression $\longrightarrow$ aggregate and reuse**

**One privacy step.** Post-processing preserves privacy, so all downstream Fast-CAR outputs inherit the same guarantee.

# One rich design encodes many scientific targets

**Overall treatment association**

$$z_{\mathrm{main}}(A, X) = (1, A, X^\top)^\top, \qquad \mathrm{OR} = \exp(\beta_A^\star).$$

**One subgrouping variable $M_j$**

$$z_j(A, X) = (1, A, X^\top, M_j, AM_j)^\top, \qquad \tau_j^\star(m) = \beta_A^\star + m\beta_{AM,j}^\star.$$

**Two subgrouping variables $M_r, M_s$**

$$\tau_{rs}^\star(m_r, m_s) = \beta_A^\star + m_r\beta_{Ar}^\star + m_s\beta_{As}^\star + m_r m_s\beta_{Ars}^\star.$$

Reduced models are obtained by linear coefficient restrictions on the same rich pooled risk.

# Why not reweight across sites?

**Pairwise density ratios would be needed**

$$r_{jk}(X, A, Y) = \frac{dP(X, A, Y \mid S = j)}{dP(X, A, Y \mid S = k)}.$$

- The ratios are unknown and must be estimated from high-dimensional joint distributions.
- Misspecification can produce bias that does not vanish with increasing sample size.
- Support mismatch generates extreme weights and unstable variance.
- A site cannot diagnose a pairwise ratio model using only its own local data.

**Positioning**

CAR targets a pooled empirical risk without modeling site selection; transportability methods remain valuable when their assumptions are credible.

# Chebyshev approximation controls the scalar error

### Degree-$m$ approximation

$$p_m(t) = \sum_{\ell=0}^{m} c_\ell T_\ell(t), \qquad T_\ell(x) = \cos\{\ell \arccos(x)\}.$$

The coefficients depend only on the scalar link and are shared across sites.

### Analytic link

$$\|b - p_m\|_\infty \le \frac{2V}{\rho(M) - 1}\,\rho(M)^{-m}.$$

### Effect of the parameter box

If $|z^\top \theta| \le M$, then for a strip of analyticity of width $\sigma$,

$$\rho(M) = \frac{\sigma}{M} + \sqrt{1 + \frac{\sigma^2}{M^2}}.$$

A wider box slows the geometric rate.

### Logistic link

The nearest complex singularities are at $t = \pm i\pi$, so $\sigma = \pi$.

# Explicit tensor-train cores

Write $x_j = z_j\theta_j$ and $p_m(t) = \sum_{\ell=0}^{m} a_\ell t^\ell$. Then

$$p_m(z^\top\theta) = G_1(x_1)G(x_2)\cdots G(x_{p-1})G_p(x_p),$$

with intermediate TT ranks at most $m+1$.

**Boundary cores**

$$G_1(x) = (\psi_0(x), \dots, \psi_m(x)),$$
$$\psi_s(x) = \sum_{\ell \geq s} a_\ell \frac{\ell!}{s!(\ell-s)!} x^{\ell-s},$$
$$G_p(x) = (1, x, \dots, x^m)^\top.$$

**Interior binomial update**

$$\{G(x)\}_{ir} = \begin{cases} \dfrac{i!}{r!(i-r)!} x^{i-r}, & r \leq i, \\ 0, & r > i. \end{cases}$$

This propagates the powers of the running partial sum across coordinates.

No observation-level numerical tensor factorization is required.

# Two tuning errors determine the final accuracy

**Local surrogate and parameter error**

For $B \geq m$,

$$\|\widehat{Q}_k - Q_k\|_\infty \leq C\{e^{-\rho m} + (\log B)^p \varepsilon\},$$

and therefore

$$\|\widehat{\theta}_{\text{Fast-CAR}} - \widehat{\theta}_{\text{P}}\|_2^2 \leq C\{e^{-\rho m} + (\log B)^p \varepsilon\}.$$

**$m$: scalar approximation**

Increase until the one-dimensional link error is negligible.

**$B$: grid level**

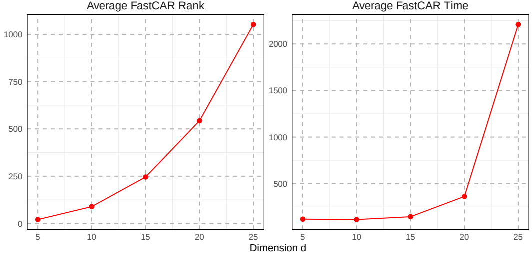
Choose $B \geq m$; the polynomialized risk then has no interpolation bias.

**$\varepsilon$: TT rounding**

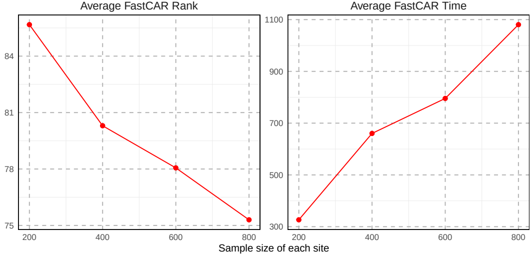
Smaller values preserve more information but increase rank and storage.

**Reported settings:** $m = 9$, $B = 20$, $\varepsilon = 10^{-6}$.

# TT rank and runtime scaling




Increasing model dimension

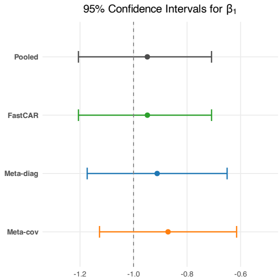



Increasing sample size per site

**Observed pattern**

Rank and runtime increase with dimension; larger site samples are more compressible and yield lower TT rank.

# Likelihood inference tracks the pooled benchmark




| Method | $K$ | Size | Power .2 | Power .4 |
|---|---|---|---|---|
| Pooled LRT | – | .051 | .633 | .990 |
| Fast-CAR LRT | 10 | .051 | .633 | .990 |
| Fast-CAR LRT | 20 | .051 | .633 | .990 |
| Meta-diag Wald | 10 | .030 | .422 | .920 |
| Meta-diag Wald | 20 | .039 | .271 | .806 |
| Meta-cov Wald | 10 | .020 | .382 | .945 |
| Meta-cov Wald | 20 | .013 | .161 | .819 |

**Theory**

$$|\Lambda_{\mathrm{Fast}}(\tau_0) - \Lambda_{\mathrm{P}}(\tau_0)| \leq 4N\Delta_Q.$$

**Heterogeneous setting**

$N = 500$, $d = 5$; power columns correspond to $\beta_1 = 0.2$ and $0.4$. Fast-CAR reproduces pooled LRT size and power for both $K = 10$ and $K = 20$.

## Agreement across the coefficient vector

| Covariate | Pooled | Fast-CAR | Meta-diag | Meta-cov |
|---|---|---|---|---|
| Intercept | $-2.400$ (.042) | $-2.388$ (.041) | $-2.237$ (17.91) | $-2.320$ (17.91) |
| Treatment | $-0.251$ (.058) | $-0.249$ (.058) | $-0.221$ (.059) | $-0.265$ (.057) |
| Chronic kidney disease | .111 (.023) | .108 (.023) | .102 (.023) | .102 (.023) |
| Cerebrovascular disease | .347 (.016) | .347 (.016) | .128 (.018) | .123 (.018) |
| BMI | .014 (.019) | .015 (.019) | $-.029$ (.089) | $-.034$ (.089) |
| T2D combination | $-.003$ (.019) | $-.003$ (.019) | $-.010$ ($8.1\times10^2$) | $-.007$ ($8.1\times10^2$) |

**Fast-CAR**

Near-identical estimates and standard errors across the full model.

**Estimator-level aggregation**

Larger discrepancies and unstable variance for low-prevalence covariates.

# The cohort remains behind site governance

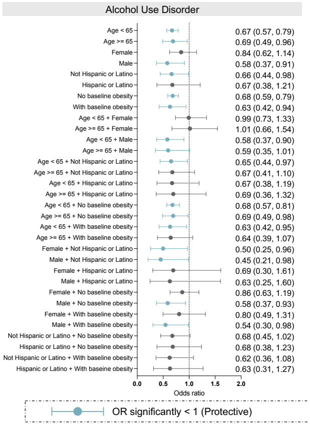
| Site | Protocol eligible | Analysis cohort | GLP-1RA | DPP-4i |
|---|---|---|---|---|
| YNHHS | 27,520 | 16,667 | 9,500 | 7,167 |
| Mayo | 24,347 | 15,284 | 12,270 | 3,014 |
| UPHS | 21,011 | 14,805 | 7,644 | 7,161 |
| UTSW | 7,984 | 4,908 | 3,704 | 1,204 |

**Harmonized construction**
Common eligibility criteria and outcome-specific washout are executed locally.

**Governance**
Individual records remain at each institution; small cells are suppressed in descriptive summaries.

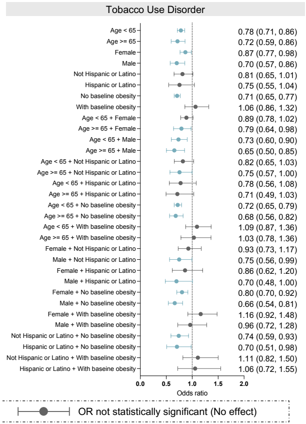
**Privacy boundary**
Averaged risks are summary-level objects, but formal privacy requires secure aggregation or differential privacy.

# Alcohol-use-disorder subgroups




Adjusted odds ratios and 95% confidence intervals; blue intervals exclude 1.

# Tobacco-use-disorder subgroups




Adjusted odds ratios and 95% confidence intervals; blue intervals exclude 1.