Supplementary Material
A Federated Probabilistic Digital Twin for Adverse-Event Risk in Aesthetic Medicine

Damini Rijhwani[0009−0000−2253−6111]

Automation Core Inc., Cambridge, MA, USA
damini@automationcoreinc.com

S1 Cohort and Data Pipeline

Query and completeness. The cohort is the complete openFDA MAUDE history for product code LMH ("Implant, Dermal, For Aesthetic Use"), pulled with `device.device_report_product_code:LMH` on 2026-07-20 (openFDA `meta.last_updated` 2026-07-07): 22,354 reports received 1993-12-13 to 2026-06-30, retrieved over 224 pages at `limit=100`, sorted `date_received:asc`, with 22,354 distinct `mdr_report_key` values and verified ascending page ordering. The first-device product code is LMH for 22,351 of 22,354 reports (39 multi-device reports), so first-device manufacturer attribution is safe.

An earlier 5,000-record pull used the unquoted multi-token value `device.generic_name:implant+dermal+for+aesthetic+use`, which openFDA matches as loose Lucene terms, so any report mentioning "implant" matched. That cohort's site table contained non-filler manufacturers (dental implants, orthopaedics, neuromodulation, dialysis) and is superseded by the product-code query used here. Product code LZN is excluded: its openFDA device name is "Cement Obturator" and all local LZN records are hip-arthroplasty cement restrictors.

Deduplication. Two passes: by `mdr_report_key`/`report_number` (0 duplicates), then by exact normalized event description (393 reports dropped, 22,354 → 21,961). The second pass removes the case where one procedure is reported on two devices with identical narratives.

Classification. A regex classifier is applied to the "Description of Event or Problem" field only, labelling 13,014 of 21,961 deduplicated records. Classifying on all `mdr_text` fields would label 14,460 and change 2,728 labels, because the manufacturer-narrative field carries evaluation boilerplate. Multi-class hit histogram: $\{0{:}8919, 1{:}9806, 2{:}2897, 3{:}290, 4{:}20, 5{:}1\}$. Class distribution: nodule 4,945, infection 4,069, vascular occlusion 2,927, bruising 688, blindness 385. Manual review of a 35-record sample gives $27/35 = 77\%$ precision.

Table S1. Site-partition sensitivity. Held-out Poisson score per test event (higher is better), 10 seeds. Independent run with its own RNG streams (`code/run_camera_ready_supplement.py`); the high-variance σ=48 rows differ in point estimate from Table 1 of the main paper while the ordering is unchanged.

K	Centralized	EB shrinkage	BFed non-priv.	BFed-RDP σ=48	One-shot DP	FedAvg +DP
2	+4.85	+4.87	+4.83	−4.33	+1.94	+4.87
4	+4.16	+4.43	+3.89	−12.68	+1.91	+4.43
6	+3.75	+4.14	+3.40	−28.07	+1.90	+4.14
8	+3.46	+3.92	+3.10	−45.56	+1.85	+3.92
12	+3.05	+3.66	+2.55	−6.39	+1.81	+3.66
16	+2.77	+3.51	+2.23	+1.70	+1.72	+3.50

Exposure. Uniform observation-window exposure of 32.545 reporting-years per site. The window spans eras in which modern hyaluronic-acid fillers were not marketed and in which manufacturers entered and exited; the benchmark compares methods on identical sufficient statistics, so this choice does not affect the relative comparison, but absolute rates must not be read as incidence.

Federation. K=8 sites (top manufacturers by event volume plus OTHER), minimum 5 events per site: ALLERGAN PRINGY 5,263; PRINGY 1,377; ALLERGAN 743; MERZ NORTH AMERICA 679; Q MED 589; TEOXANE 561; GALDERMA Q MED 519; OTHER 3,283.

S2 Site-Partition Sensitivity

Reviewers asked whether a manufacturer partition is a reasonable proxy for clinic heterogeneity and how sensitive the conclusion is to the number and size of sites. Table S1 sweeps K. The one-shot release and FedAvg+DP beat iterative BFed-RDP at every K from 2 to 12; at K=16 BFed-RDP recovers to roughly the one-shot level, since per-site noise $\sigma/\sqrt{K}$ is then smallest relative to the aggregate. The magnitude of the failure depends on the partition; its direction does not, over the range tested.

S3 Posterior-Predictive Coverage Across Privacy Budgets

Under the Gamma-Poisson twin a held-out count y with exposure n has predictive law NegBin(α, $\beta/(\beta + n)$). Table S2 reports empirical coverage of central 90% and 95% predictive intervals over held-out (site, class) cells.

Two findings. First, coverage degrades monotonically as the privacy budget tightens. Second, and more consequentially, the pooled twin under-covers badly with no privacy noise at all: a single global rate per class cannot express between-site variation, so its intervals are too narrow for site-level counts. A non-private

Table S2. Posterior-predictive coverage and held-out score, K=8, 10 seeds. Widths are in events.

Method	ε	cov@90	cov@95	width@90	Poisson score
BFed-RDP, σ=0	non-private	0.357	0.438	15.68	+3.10
BFed-RDP, σ=6	21.55	0.367	0.425	16.31	+3.11
BFed-RDP, σ=12	9.39	0.315	0.378	18.97	+2.89
BFed-RDP, σ=24	4.35	0.268	0.318	104.23	−14.78
BFed-RDP, σ=48	2.09	0.203	0.242	249.79	−45.56
BFed-RDP, σ=96	1.02	0.162	0.172	310.31	−61.14
One-shot DP (T=1)	21.55	0.450	0.517	23.59	+2.35
One-shot DP (T=1)	9.39	0.428	0.497	22.73	+2.22
One-shot DP (T=1)	4.35	0.398	0.455	21.97	+2.04
One-shot DP (T=1)	2.09	0.330	0.380	22.30	+1.85
One-shot DP (T=1)	1.02	0.297	0.335	23.99	+1.64
Centralized pooled Gamma	non-private	0.195	0.225	24.62	+3.46
EB shrinkage (hierarchical)	non-private	**0.965**	**0.988**	24.10	**+3.92**

Table S3. Decomposition of held-out score, K=8, 10 seeds.

Quantity	Poisson score
Centralized, non-private	+3.46
Federated, non-private	+3.10
cost of federation	−0.36
Federated, private, iterative (T=200, ε≈2.09)	−45.56
cost of privacy, iterative	−48.66
Federated, private, one-shot (T=1, ε≈2.09)	+1.85
cost of privacy, one-shot	−1.25

empirical-Bayes hierarchical model, fitted by method of moments per class across sites and then updated conjugately, attains 96.5% coverage at nominal 90% and the best Poisson score in the study. Calibrated site-level uncertainty is a property of the hierarchical structure, not of the privacy mechanism.

S4 Cost of Federation Versus Cost of Privacy

Reviewers asked for the two costs to be separated. Table S3 decomposes them on the K=8 federation. Federating without privacy is nearly free; composing 200 private rounds is not; a single private release costs little more than federating.

S5 Full Benchmark Table and Convergence Verification

Table S4 reports every method on the complete LMH cohort, including the gradient baselines omitted from the main paper for space. Learning rates (and

Table S4. Held-out Poisson score per test event, complete LMH cohort, K=8, 10 seeds, 95% bootstrap CI.

Method	Privacy	Score	Comm./ round/site
Centralized pooled Gamma	None	$+3.46\,[+3.46, +3.47]$	—
FedAvg	None	$+3.92\,[+3.91, +3.92]$	40 B
FedProx	None	$+3.88\,[+3.79, +3.92]$	40 B
SCAFFOLD	None	$+3.92\,[+3.91, +3.93]$	80 B
FedPA	None	$+3.92\,[+3.91, +3.92]$	80 B
BFed non-private	None	$+3.10\,[+3.09, +3.11]$	80 B
FedAvg + DP	$\varepsilon\approx 2.09$	$+3.92\,[+3.91, +3.92]$	40 B
Central-DP pooled Gamma	ε=2.09, T=1	$+2.77\,[+2.63, +2.93]$	80 B
One-shot DP suff. stats	ε=2.09, T=1	$+1.99\,[+1.90, +2.09]$	80 B
BFed-RDP, σ=6	ε=21.55	$+3.09\,[+3.05, +3.14]$	80 B
BFed-RDP, σ=12	ε=9.39	$+3.02\,[+2.93, +3.12]$	80 B
BFed-RDP, σ=24	ε=4.35	$-7.47\,[-19.22, +2.70]$	80 B
BFed-RDP, σ=48	ε=2.09	$-50.88\,[-73.85, -26.51]$	80 B
BFed-RDP, σ=96	ε=1.02	$-30.81\,[-45.55, -16.15]$	80 B

FedProx's μ) were selected per seed on a validation split carved from the training data only; the test set was never used for selection. Every final run tracked the regularized training objective each round.

Convergence was verified per method by the relative change in the training objective over the last 10 rounds ($< 10^{-3}$), or, for the DP variant whose objective is a noisy process, by the coefficient of variation of the final 20 rounds ($< 5\times10^{-2}$). FedAvg, FedAvg+DP, SCAFFOLD and FedPA pass on all 10 seeds. FedProx does not pass on all seeds (worst-seed relative change 3.46×10^{-2}) and its row should be read as a lower bound on tuned FedProx performance. This is reported rather than suppressed because the failure of the reviewed submission's benchmark was precisely an unchecked convergence assumption.

Fed-Heart-Disease stress test. On the cardiac benchmark (K=4, N=920) BFed-RDP's expected calibration error degrades monotonically in σ: 0.180 [0.137, 0.225] at $\varepsilon\approx$21.55; 0.299 [0.219, 0.375] at 9.39; 0.369 [0.284, 0.445] at 4.35; 0.418 [0.350, 0.481] at 2.09; 0.421 [0.352, 0.489] at 1.02. FedAvg+DP attains 0.084 [0.077, 0.096] at $\varepsilon\approx$2.09 and the non-private conjugate ablation attains 0.048, so privacy noise rather than aggregation is the limiting factor on this benchmark.

S6 Privacy Accounting

A mechanism is (α, ρ)-RDP if $D_\alpha(\mathcal{M}(D)\|\mathcal{M}(D')) \le \rho$ for neighbouring D, D'. For a Gaussian mechanism with ℓ_2-sensitivity C and noise σ, the per-release cost is $\rho(\alpha) = \alpha C^2/(2\sigma^2)$, composing additively over T rounds, and converting via

Table S5. Rényi-DP composition. Left: across rounds at fixed σ=48. Right: across σ at fixed T=200, giving the ε values used throughout.

T	α^*	$\rho(\alpha^*)$	ε	σ	α^*	ε
10	52.50	0.023	0.45	6	2.44	21.5506
50	24.03	0.010	1.02	12	3.88	9.3864
100	17.29	0.0075	1.46	24	6.76	4.3460
200	12.52	0.00543	2.09	48	12.52	2.0862
500	8.28	0.0036	3.38	96	24.03	1.0214

$\varepsilon(\delta) = \min_{\alpha>1}[\rho_{\text{tot}}(\alpha) + \log(1/\delta)/(\alpha-1)]$. Table S5 reports the composition. A closed-form evaluation and a grid search over α agree to four decimals at every σ tested. Advanced composition at σ=48, T=200 yields $\varepsilon = 18.99$ against 2.0862 under RDP, a 9.1× tightening. One-shot calibration at $\varepsilon = 2.0862$ gives $\sigma_1 = 3.3941 = 48/\sqrt{200}$.

Effective sample size under subsampling. Each BFed-RDP round releases the privatized statistic of a Bernoulli-subsampled batch (fraction 1/20). Counts and exposures accumulate together, so posterior means stay unbiased, but the effective sample size is inflated by $T/20$ and must be renormalized before posterior credible intervals are interpreted. The accountant charges every release and claims no subsampling amplification.

S7 One-Shot Optimality: Proof and Verification

Theorem 1 (One-shot optimality). *Let a conjugate exponential-family model have sufficient statistic $\Delta\boldsymbol{\eta} \in \mathbb{R}^d$ with $\|\Delta\boldsymbol{\eta}\|_2 \le C$, and let the global posterior natural parameter be $\boldsymbol{\eta}^{\text{global}} = \boldsymbol{\eta}^{\text{prior}} + \sum_{k=1}^{K} \Delta\boldsymbol{\eta}_k$. Compare a one-shot Gaussian release of the full sum with noise σ_1 against an iterative mechanism that releases partial sums over T rounds with per-round noise σ_T and accumulates them. If both are calibrated to the same total Rényi-DP cost $\rho(\alpha)$, then $\sigma_1^2 = \sigma_T^2/T$, the accumulated iterative noise variance is $T\sigma_T^2$, and the one-shot release therefore carries T^2 times less noise variance. No schedule of T releases at equal total cost attains lower variance than the one-shot release.*

Proof. The global posterior depends on the data only through $S = \sum_k \Delta\boldsymbol{\eta}_k$. For the Gaussian mechanism with ℓ_2-sensitivity C, a single release with noise σ costs $\rho_1(\alpha) = \alpha C^2/(2\sigma^2)$, and T releases with per-round noise σ_T cost $\rho_T(\alpha) = T\alpha C^2/(2\sigma_T^2)$ by additive composition. Equating the two budgets gives $\sigma_1^2 = \sigma_T^2/T$. The iterative mechanism accumulates T independent draws, so its released sum carries noise variance $T\sigma_T^2$, while the one-shot release carries $\sigma_1^2 = \sigma_T^2/T$. The ratio is $1/T^2$.

For the general claim, let a schedule assign per-round noise variances σ_t^2 and combine the T noisy releases by weights w_t with $\sum_t w_t = 1$. The combined estimator variance is $\sum_t w_t^2\sigma_t^2$ subject to the budget constraint $\sum_t 1/\sigma_t^2 = 1/\sigma_1^2$.

Minimising the former under the latter by Cauchy–Schwarz gives $\sum_t w_t^2\sigma_t^2 \geq \sigma_1^2$, with equality only for inverse-variance weighting, which reduces to a single effective release. Hence no schedule beats one-shot, and the accumulate schedule ($w_t = 1$ for all t, no normalisation) attains the T^2 penalty above.

Numerical verification. At $\alpha^* C^2/(2\rho)$ the one-shot variance is 11.5200. The accumulate-final-iterate schedule attains theoretical variance 460,800.0 and Monte-Carlo variance 462,310.7, a ratio of exactly $T^2 = 40{,}000$. Inverse-variance-weighted iteration attains 11.5200 in theory and 11.5578 by Monte Carlo, a ratio of 1.000: it ties and never beats. The minimum accumulate-to-one-shot ratio over 200 random matched-cost schedules is 107,976, above the T^2 floor as Cauchy–Schwarz requires. Verification code: `code/verify_theory.py`.

S8 Protocol Pseudocode and Topology

Algorithm 1 BFed-RDP round

1: **Input:** site data $\{\mathcal{D}_k\}$, prior $\boldsymbol{\eta}_z^{(t-1)}$, noise σ, clip C
2: Server broadcasts $\boldsymbol{\eta}_z^{(t-1)}$ to all sites.
3: **for** each site k in parallel **do**
4: Compute $\Delta\boldsymbol{\eta}_{z,k} = (y_{z,k}, -n_{z,k})^T$; clip to C.
5: Draw $\mathbf{b}_k \sim \mathcal{N}(0, \sigma^2 I/K)$; exchange DH seeds for masks $\mathbf{m}_{k,k'}$.
6: Upload the masked update $\tilde{\boldsymbol{\eta}}_{z,k}$.
7: **end for**
8: Server aggregates $\boldsymbol{\eta}_z^{(t)} = \boldsymbol{\eta}_z^{(t-1)} + \sum_k \tilde{\boldsymbol{\eta}}_{z,k}$; accumulate $\rho_t(\alpha)$.

Secure aggregation. Pairwise masks cancel exactly in the aggregate: the maximum norm of the summed masks over all rounds and classes is 4.46×10^{-15}, and masked versus unmasked runs at the same seed differ by at most 7.11×10^{-15} in rate. Secure aggregation hides individual uploads but is not itself differentially private.

S9 Code and Data Availability

All data are public. The MAUDE cohort is retrievable from openFDA with the query and pull date in Sect. S1; Fed-Heart-Disease is distributed with FLamby. The reference implementation, the audited benchmark pipeline, the supplementary experiment driver (`run_camera_ready_supplement.py`), the theory verification (`verify_theory.py`) and all per-seed result files are released with the paper.

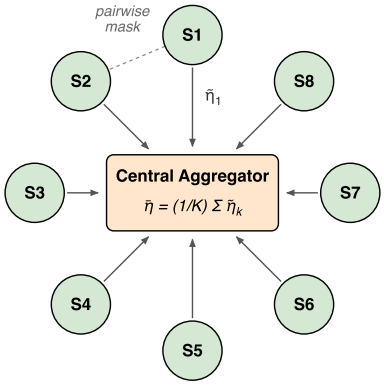

Fig. S1. Federated topology. Each site computes local conjugate updates and transmits only noised natural parameters; raw imagery remains on-device.