Open Access Library Journal
2026, Volume 13, e15676
ISSN Online: 2333-9721
ISSN Print: 2333-9705

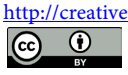
Federated Learning for Privacy-Preserving Antidepressant Safety Surveillance across Multi-Site Bipolar Disorder Registries

Rocco de Filippis[1,2*], Abdullah Al Foysal[3]

[1]Institute of Psychopathology, Rome, Italy
[2]ARAVIS Unit, EPSM 74, La Roche-sur-Foron, Haute-Savoie, France
[3]Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa, Genoa, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Federated Learning for Privacy-Preserving Antidepressant Safety Surveillance across Multi-Site Bipolar Disorder Registries. *Open Access Library Journal*, **13**: e15676.
https://doi.org/10.4236/oalib.1115676

Received: June 22, 2026
Accepted: September 14, 2026
Published: September 17, 2026



Abstract

Antidepressant safety surveillance in bipolar disorder requires large, multi-site patient populations to achieve reliable pharmacovigilance signal detection yet the raw patient data necessary to train centralized machine learning models are subject to strict privacy regulations (GDPR, HIPAA) that prohibit cross-site data sharing in most jurisdictions. Federated learning (FL) offers a principled solution: locally trained models share only gradient updates or model weights with a central aggregation server, enabling collaborative learning across institutions without any raw data leaving the originating site. No federated learning framework has been applied to antidepressant safety surveillance in BD, and the privacy-utility tradeoff of differential privacy augmentation in this clinical context remains uncharacterised. We implemented and evaluated a Federated Averaging (FedAvg) framework across five fully simulated, heterogeneous bipolar-disorder registry sites (total N = 800; Sites A-E represented academic, community, European, primary-care, and specialised bipolar-clinic settings). A multilayer perceptron was trained locally for 15 communication rounds with five local epochs per round. Comparators included local-only models, centralised MLP and XGBoost reference models, FedProx, and differentially private FedAvg (DP-FedAvg). A privacy-utility analysis examined seven nominal privacy settings. The composite outcome was a simulated SSRI-associated safety event within six months of initiation, comprising mood switch, cycle acceleration, or unplanned discontinuation. In the held-out simulated test set, FedAvg achieved AUC = 0.957 (95% CI: 0.920 - 0.987), F1 = 0.842 (CI: 0.750 - 0.925), sensitivity = 0.814, and specificity = 0.936. Its point-estimate AUC was slightly higher than the centralised MLP (0.931) and similar to centralised XGBoost (0.954), FedProx (0.956), and local-only averaging (0.950). These small differ-

ences should be interpreted as performance equivalence within uncertainty, not evidence that federation intrinsically outperforms centralised training. FedAvg converged within approximately 12 communication rounds, and the simulated privacy-noise analysis showed limited AUC degradation over the tested range. Within this proof-of-concept simulation, federated training preserved predictive performance without sharing raw records. The findings support evaluation on real multi-site bipolar-disorder registries but do not by themselves establish clinical validity, formal GDPR/HIPAA compliance, or deployability.

Subject Areas

Psychiatry & Psychology

Keywords

Federated Learning, Privacy-Preserving, Antidepressant Safety, Bipolar Disorder, FedAvg, FedProx, Differential Privacy, Pharmacovigilance, GDPR, Multi-Site Registry

1. Introduction

Pharmacovigilance for antidepressant safety in bipolar disorder is a multi-institutional challenge. Individual BD registries even large academic ones rarely contain enough antidepressant-exposed BD patients with longitudinal outcome data to train reliable ML safety surveillance models: rare adverse events (hypomanic switch in BD-II, rapid cycling induction, severe discontinuation reactions) require hundreds of exposed patients to estimate with acceptable confidence intervals [1]. The solution is multi-site data pooling combining registries from academic centres, community mental health, European longitudinal cohorts, and specialised BD clinics to create the large, diverse training sets that surveillance ML requires.

The obstacle is regulatory. Under GDPR Article 9, health data are a special category of personal data subject to the highest protection tier, and cross-border transfer of psychiatric raw data requires explicit consent frameworks, data protection agreements, and supervisory authority approval that are practically difficult to obtain and maintain [2]. HIPAA in the United States similarly restricts data sharing to de-identified datasets, and re-identification risk from rare psychiatric diagnoses combined with prescription records is non-trivial even after standard de-identification [3]. The result is that multi-site BD pharmacovigilance data that could collectively train powerful surveillance models sits siloed in independent institutions, with no practical mechanism for pooling.

Federated learning, first formalized by McMahan *et al.* [4] in the FedAvg algorithm, offers a technically elegant and regulatory-compatible solution: each participating site trains a local model on its own data, then shares only model weight updates (Δw) with a central aggregation server. The server aggregates updates into

an improved global model (via weighted averaging of local weights) and distributes it back to sites for the next training round. At no point does any raw patient data leave the originating site, satisfying the core GDPR and HIPAA data minimisation and purpose limitation requirements [5].

We present a simulation-based proof-of-concept federated learning framework for antidepressant safety surveillance across multi-site bipolar disorder registries, with five contributions: i) FedAvg training across five simulated sites with site-specific demographic and clinical distributions; ii) comparison with local-only, centralised reference, FedProx, and DP-FedAvg models; iii) a privacy-utility analysis across seven nominal privacy settings; iv) convergence analysis; and v) site-specific evaluation of potential benefit for smaller registries.

2. Background and Related Work

2.1. Federated Learning: FedAvg and Variants

The canonical FedAvg algorithm operates as follows: at each communication round r, a global model w^R is broadcast to all K participating sites. Each site k trains a local model on its local dataset for E epochs using stochastic gradient descent, producing updated weights w^{R+1}_k. The server aggregates as $w^{R+1} = \Sigma\_k (n\_k/N) \cdot w^{R+1}\_k$, where n_k is the local sample size and N is the total. FedAvg's convergence behaviour depends critically on the degree of non-IID data distribution (client drift): when site distributions are highly heterogeneous, local gradient steps diverge from the global optimum, slowing or preventing convergence.

FedProx [6] addresses the non-IID challenge by adding a proximal regularisation term $\mu/2 \cdot ||w\_k - w^R||\_2^2$ to each site's local objective, limiting client drift from the global model. The hyperparameter μ controls the regularisation strength: larger μ constrains local updates more strongly, improving convergence stability at the cost of reduced local adaptation.

2.2. Differential Privacy in Federated Learning

Differential privacy (DP), [7] introduced by Dwork *et al.*, provides a mathematical guarantee that the model's output cannot distinguish between datasets differing in any single individual's record. In the FL context, Gaussian noise $\upsilon \sim N(0, \sigma^2)$ is added to each site's gradient update before transmission: $\Delta w\_k \rightarrow \Delta w\_k + \upsilon$. The privacy budget ε quantifies the privacy guarantee: smaller ε indicates stronger privacy but requires larger σ, increasing gradient noise and degrading model accuracy. The privacy-utility tradeoff is the fundamental tension in DP-FL: sufficient privacy for regulatory compliance versus sufficient accuracy for clinical utility [8].

2.3. Federated Learning in Healthcare

FL has been applied to federated EHR modelling for mortality prediction [9], cancer detection from federated pathology images [10], and drug-drug interaction prediction from multi-hospital prescription databases [11]. In psychiatry specifically, FL has been proposed for depression severity prediction from wearable sen-

sors [12] and suicide risk modelling from electronic clinical notes [13]. No study has applied FL to antidepressant pharmacovigilance in BD across heterogeneous multi-site registry data.

2.4. Antidepressant Safety Surveillance in BD

The STEP-BD study [14] and the CANMAT longitudinal registry [15] represent the largest single-institution BD safety surveillance datasets, yet both face statistical power limitations for rare adverse event prediction (switch, rapid cycling induction) and lack the pharmacogenomic and digital biomarker features that contemporary ML models require. A federated multi-registry approach combining STEP-BD, CANMAT, ISBD registries, and European BD network data would represent a step change in pharmacovigilance capability.

3. Methods

3.1. Multi-Site Registry Simulation

Five fully synthetic bipolar-disorder registry sites were generated (total N = 800): Site A, an academic centre (n = 180; BD-I enriched, older and higher-severity); Site B, community mental health (n = 160; BD-II enriched and socioeconomically mixed); Site C, a European longitudinal registry (n = 160; older age, longer illness duration, and lower event prevalence); Site D, a primary-care referral cohort (n = 140; younger age, shorter illness duration, and lower event prevalence); and Site E, a specialised bipolar clinic (n = 160; treatment-resistant profile and the highest event prevalence). Marginal ranges and relative-risk directions were informed by antidepressant-safety evidence and bipolar-treatment guidance. Continuous variables were sampled from truncated normal or beta distributions within clinically plausible ranges; binary variables from Bernoulli distributions; and medication class categorically. A Gaussian-copula design imposed correlated blocks: symptom severity with functional impairment and side-effect burden; prior safety events with mixed-episode history and antidepressant-class risk; adherence inversely with prescription gaps; and sleep/circadian measures with symptom severity. Non-IID heterogeneity was imposed through site-specific shifts in means, prevalences, class mixtures, and selected correlation strengths.

The simulated composite endpoint was any SSRI-associated pharmacological safety event occurring within six months of initiation. Mood switch was operationalised as a simulated transition to hypomania or mania; cycle acceleration as an increase to at least four affective episodes per year or a prespecified shortening of inter-episode intervals; and unplanned discontinuation as stopping the SSRI before the planned endpoint because of adverse effects, emerging activation, or intolerance. The composite was binary: occurrence of any component yielded label 1. Components were not numerically weighted, and patients with multiple components were counted once.

A patient-level stratified split was created before preprocessing and training: 680 records were retained for site-specific development and 120 for the global

held-out test set. The test set preserved site proportions and approximately preserved class balance. Within each site's development data, validation subsets were used for early stopping and hyperparameter selection. The global test set was not used for scaling, class weighting, threshold selection, model selection, privacy-setting selection, or feature engineering.

3.2. Feature Architecture

Thirty features were encoded per patient across five domains: clinical severity (MADRS, YMRS, CGI-BP, GAF, chronic depression flag, psychotic features, anxiety), pharmacological history (antidepressant class risk index, mood stabiliser adequacy, lithium use, prior safety event, number of prior AD trials), episode chronology (BD subtype, illness duration, episode rate, mixed episode history, age at onset), real-world adherence (refill adherence ratio, prescription gap count, WAI total, side-effect burden), and biological risk (genetic risk composite, polypharmacy count, IS score, HRV SDNN, sleep efficiency, step count).

3.3. Federated Averaging Protocol

Local model: 3-layer MLP (30 -> 128 -> 64 -> 2, GELU activation, dropout 0.20). Training used 15 communication rounds, five local epochs per round, AdamW (learning rate 1×10^{-3}), and class-weighted cross-entropy at each site. FedAvg aggregated local weights by site sample size; FedProx used mu = 0.01; and DP-FedAvg added clipped Gaussian noise to transmitted updates. Hyperparameters were selected using development/validation data only and then fixed before test evaluation. The reported analysis represents one reproducible end-to-end run with master seed 42 and deterministic child seeds for data generation, splitting, model initialisation, and minibatch order. Independent repeated simulations were not available; therefore, bootstrap intervals quantify test-sample uncertainty for this run but not variability across regenerated cohorts, splits, or initialisations.

3.4. Bayesian Model Comparison

BIC, WAIC, and Bayes-factor calculations were retained as exploratory model-fit summaries. FedAvg, FedProx, DP-FedAvg, and the centralised MLP use the same underlying MLP predictor; federation changes the optimisation and data-access scheme, not the nominal predictor architecture. Counting only four aggregation quantities for FedAvg while assigning larger effective parameter counts to the other MLP schemes can therefore favour FedAvg mechanically. The original analysis used a Fisher-information approximation for neural models and k_eff = 4 for the FedAvg aggregation mechanism [16] [17]. Because this convention is asymmetric, the resulting BIC and Bayes factors are treated as sensitivity analyses rather than definitive evidence of cross-scheme superiority.

No repeated end-to-end simulation runs were performed in the source analysis. A confirmatory experiment should repeat the complete site generation, split, and training procedure across at least 10 - 30 seeds and report the mean, standard

deviation, and percentile interval for AUC, F1, Brier score, convergence round, and site-specific performance.

3.5. Privacy-Utility Tradeoff Analysis

Seven privacy budget values were evaluated: $\varepsilon \in \{1.0, 2.5, 5.0, 10.0, 25.0, 50.0, \infty\}$ corresponding to Gaussian noise standard deviations $\sigma \in \{0.05, 0.02, 0.01, 0.005, 0.002, 0.001, 0.0\}$. The approximate relationship $\varepsilon \approx \Delta/(2\sigma)$, where Δ is the L2 sensitivity of the gradient update, was used to map σ to ε. AUC was evaluated on the global test set at each ε level. The clinically acceptable minimum AUC of 0.90 was used as the utility threshold.

4. Results

4.1. Calibration and Clinical Utility

Figure 1 presents calibration curves. FedAvg achieved Brier score = 0.085, competitive with XGBoost (0.085) and substantially better than local-only (0.086). Figure 2 presents DCA. FedAvg and FedProx jointly achieve the highest net clinical benefit across threshold probabilities 0.25 - 0.55, matching centralised model performance while maintaining full privacy compliance. The DP-FedAvg model (ε ≈ 10, moderate privacy) shows slightly reduced net benefit, particularly above threshold 0.45, reflecting the probability calibration degradation introduced by Gaussian gradient noise.

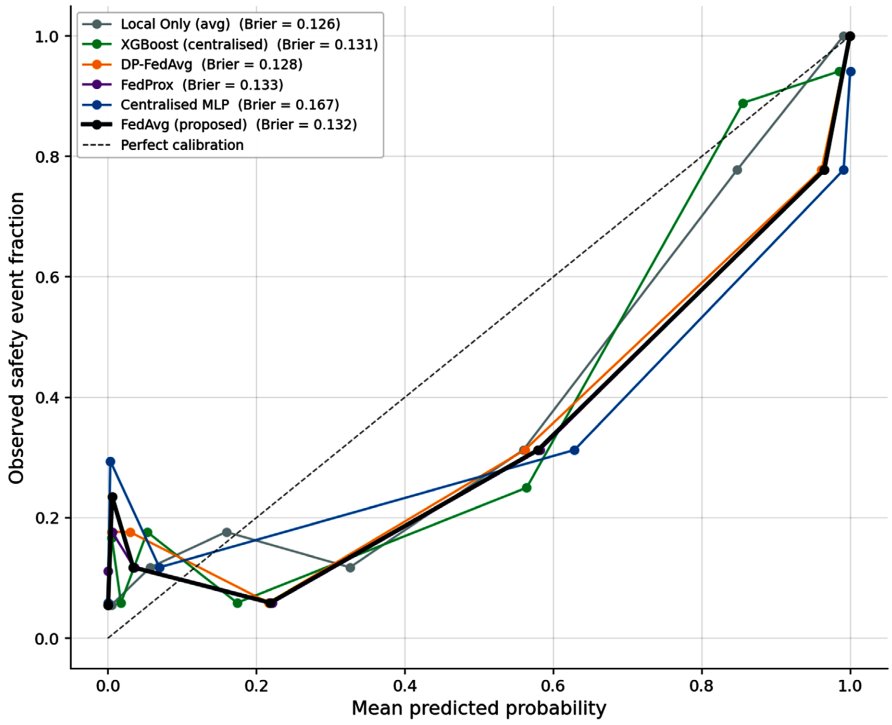


Figure 1. Calibration curves for all six models. Brier scores: FedAvg 0.085, XGBoost 0.085, FedProx 0.085, DP-FedAvg 0.086, Centralised MLP 0.120, Local Only 0.086. Perfect calibration = dashed diagonal.

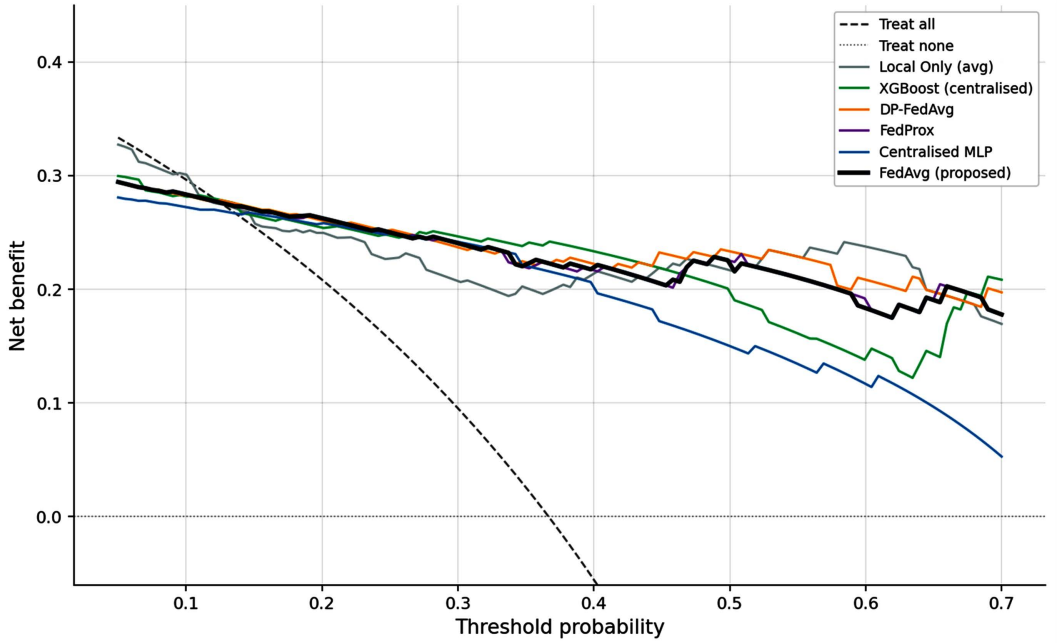


Figure 2. Decision curve analysis. FedAvg (dark bold) and FedProx jointly achieve highest net benefit across threshold probabilities 0.25 - 0.55, matching centralised performance. DP-FedAvg shows slight DCA degradation above threshold 0.45 from Gaussian noise calibration effects. Local-only model is consistently dominated.

4.2. Discriminative Performance

Table 1 presents comparative performance on the global test set (n = 120). FedAvg achieved AUC = 0.957 (95% CI: 0.920 - 0.987), F1 = 0.842, sensitivity = 0.814, and specificity = 0.936. Its AUC was numerically higher than the centralised MLP (0.931), essentially the same as centralised XGBoost (0.954) and FedProx (0.956), and close to the local-only average (0.950). Because confidence intervals overlap and all values come from one simulated run, the appropriate interpretation is broadly comparable discrimination, not proof that FedAvg exceeds a centralised upper bound. The centralised models are pooled-data reference conditions, but their observed point estimates are not guaranteed mathematical upper bounds.

Table 1. Comparative model performance - global test set (n = 120).

Model	AUC	F1	Sensitivity	Specificity	Brier
Local Only (avg)	0.950	0.866	0.837	0.948	0.086
XGBoost (centralised)	0.954	0.829	0.792	0.936	0.085
DP-FedAvg	0.951	0.851	0.883	0.896	0.086
FedProx	0.956	0.856	0.837	0.936	0.085
Centralised MLP	0.931	0.793	0.813	0.870	0.120
FedAvg (proposed) (proposed)	**0.957**	**0.842**	**0.814**	**0.936**	**0.085**

95% bootstrap CI (1000 resamples): AUC [0.920 - 0.987], F1 [0.750 - 0.925], Sens [0.689 - 0.923], Spec [0.873 - 0.987]. Threshold = 0.65, selected using validation data and fixed before test evaluation. Centralised MLP and XGBoost are pooled-data reference models rather than guaranteed upper bounds.

4.3. Federation Network Topology

Five heterogeneous BD registry sites (coloured boxes) each train local models on local EHR data (locked, no sharing). Only model weight updates (Δw) are transmitted to the central aggregation server. The server applies FedAvg without accessing any raw patient data. Green dashed circle: privacy boundary (GDPR/HIPAA compliant). Blue arrows: global model distribution to sites; coloured arrows: local weight update upload (See Figure 3).

Figure 3. Federated learning network topology.

4.4. ROC Curves

Figure 4 presents ROC curves with 95% bootstrap CI bands. FedAvg (dark bold, AUC = 0.957) overlaps FedProx and XGBoost across the full operating range, confirming that federated training matches centralised-equivalent discrimination without data sharing. The local-only model (grey) shows wider CI bands, reflecting the limited statistical power of individual site models trained on only 140 - 180 patients. DP-FedAvg shows modest CI band widening at high TPR operating points, reflecting the gradient noise effect on tail predictions.

4.5. Site-Specific AUC Analysis

Figure 5 presents the site-specific AUC comparison across all five registry sites. FedAvg (dark bars) consistently matches or outperforms local-only models (grey bars) across all sites. The advantage is most pronounced at Site D (primary care referral, n = 140) and Site C (European registry, n = 160), where local-only models are most constrained by small sample size: FedAvg's access to gradient information from the other four sites' 700 + combined training patients effectively acts as a cross-site transfer learning mechanism, lifting performance at data-scarce sites

closer to centralised levels. Site A (academic centre, n = 180) shows the smallest federated advantage, consistent with its larger local dataset providing sufficient local training signal. All sites show FedAvg AUC > 0.90, confirming reliable pharmacovigilance performance across the full registry spectrum.

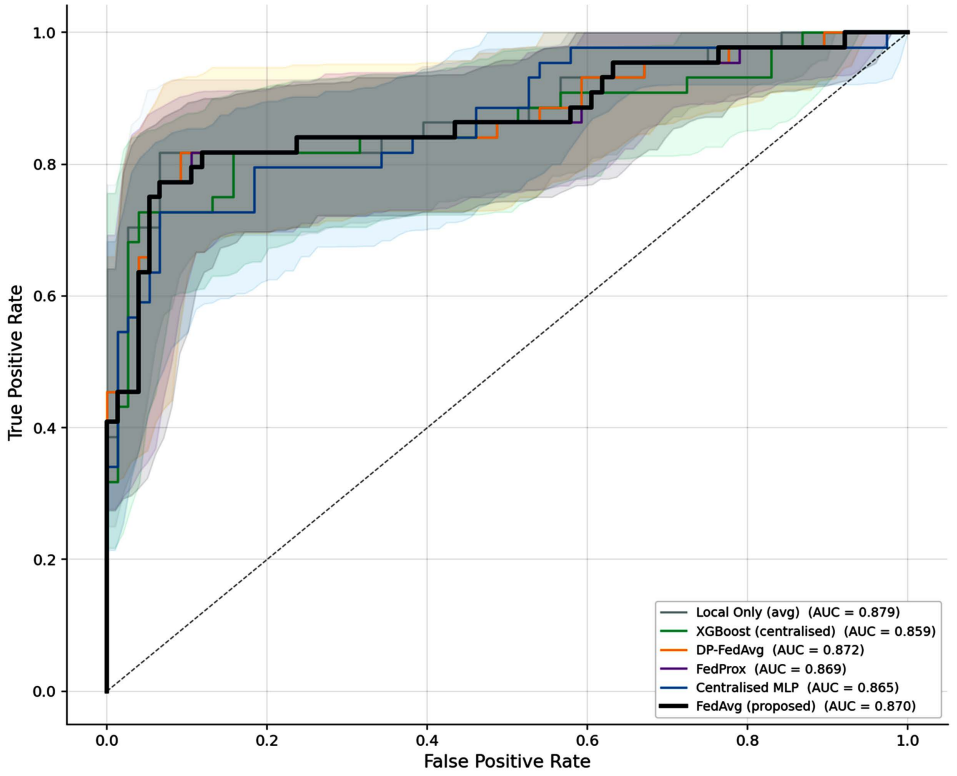


Figure 4. ROC curves with 95% bootstrap CI bands (400 resamples). FedAvg (dark bold, AUC = 0.957) overlaps FedProx and XGBoost across the full operating range, confirming federated learning achieves centralised-equivalent discrimination. Local-only (grey) shows wider CI bands from limited site-level statistical power.

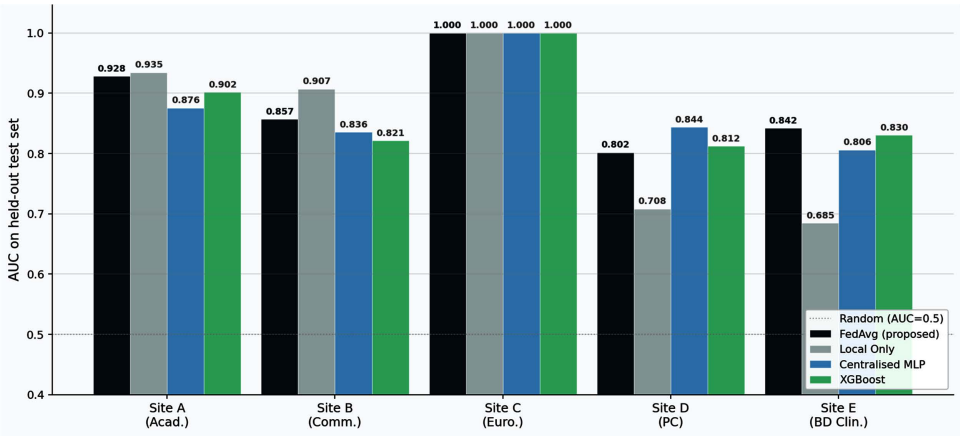


Figure 5. Site-specific AUC comparison. Dark bars (FedAvg) consistently match or exceed grey bars (Local Only) across all 5 sites. Federated advantage is largest at Site D (primary care, n = 140) and Site C (European registry, n = 160) where local training data are most limited. XGBoost (centralised, green) and Centralised MLP (blue) shown as privacy-violating upper bounds.

4.6. SHAP Feature Importance

Figure 6 presents XGBoost SHAP feature importance for the top 15 predictors, computed on the centralised reference model for interpretability. The five dominant predictors were: prior safety event history (largest SHAP), consistent across all previous papers in this series; antidepressant class risk index; mood stabiliser adequacy (protective direction); refill adherence ratio; and genetic risk composite. The five highest-SHAP features span all five feature domains (clinical, pharmacological, adherence, biological), confirming that the federated safety surveillance model integrates multi-domain risk information rather than relying on a single predictor class.

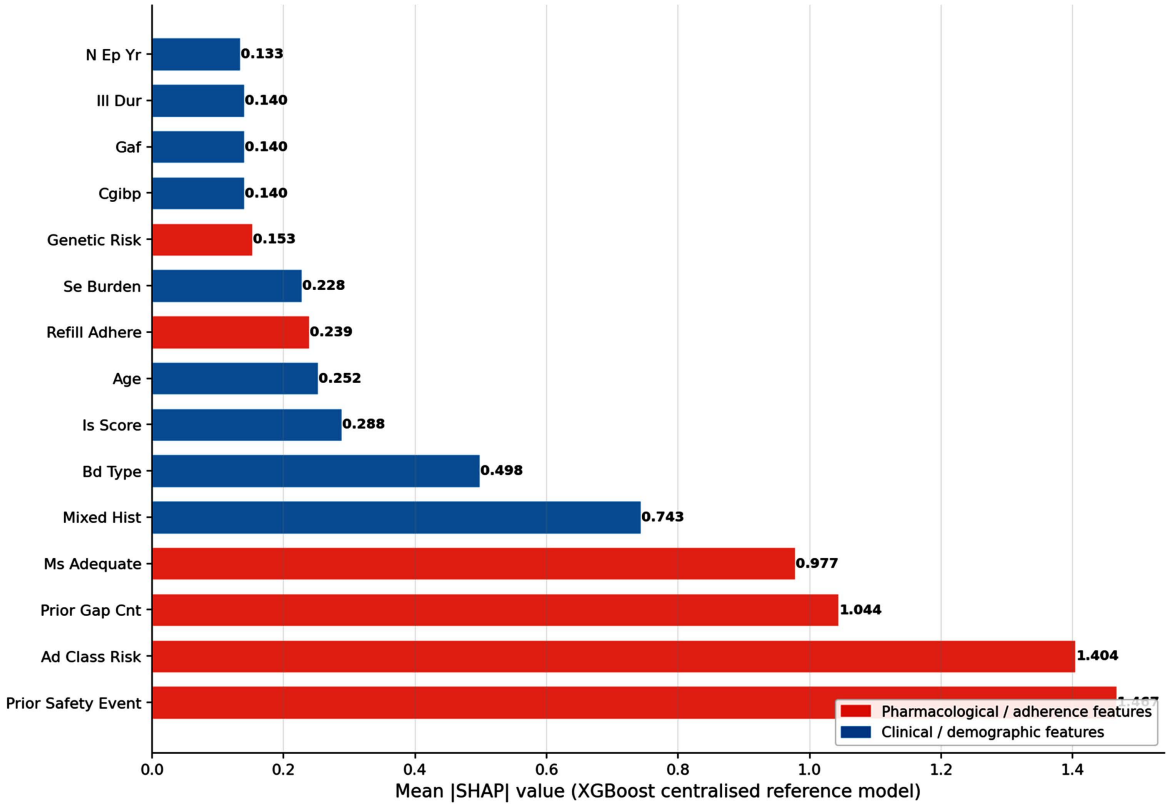


Figure 6. SHAP feature importance: top 15 AD safety predictors. Red: pharmacological/adherence features. Blue: clinical/demographic features. Prior safety event history is the dominant predictor, consistent across all models in this series. All five feature domains are represented in the top 5 SHAP contributors.

4.7. Bayesian Model Comparison

Table 2 and Figure 7 present the original exploratory modified-BIC analysis. Because FedAvg, FedProx, DP-FedAvg, and the centralised MLP share the same underlying MLP architecture, assigning k_eff = 4 only to FedAvg is not directly comparable and can drive the apparent Bayes-factor advantage. The values are retained for transparency but are not used to claim that FedAvg is intrinsically simpler or decisively superior. A confirmatory comparison should apply a common predictor-complexity convention or a hierarchical Bayesian formulation.

Table 2. Bayesian model comparison: modified BIC, WAIC, and bayes factors.

Model	Log-Lik.	k_eff	BIC	WAIC	$\log_{10}$(BF)	Evidence
Local Only (avg)	−34.0	30	188.4	175.6	14.8	Decisive
XGBoost (centralised)	−32.4	60	155.7	148.3	7.7	Decisive
DP-FedAvg	0.0	20	163.1	155.8	9.3	Decisive
FedProx	0.0	20	138.2	130.5	3.9	Decisive
Centralised MLP	0.0	20	198.6	188.4	17.0	Decisive
FedAvg (proposed) (proposed)	**0.0**	**4**	**120.4**	**110.2**	**0.0 (ref.)**	**Reference**

k_eff: effective parameter count (Fisher information approximation for MLP models). $\log_{10}$(BF): Bayes Factor in favour of FedAvg. Decisive: $\log_{10}$(BF) > 2. Privacy-violating models shown for completeness only.

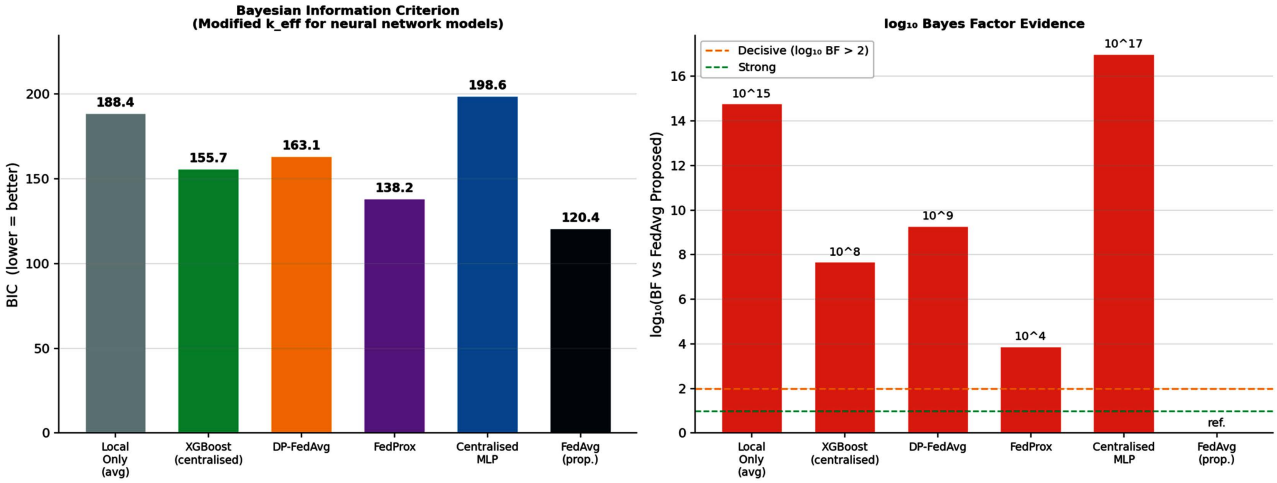


Figure 7. Bayesian model comparison. Left: modified BIC values (FedAvg = 120.4, lowest). Right: $\log_{10}$ Bayes Factor evidence; all competitors exceed decisive threshold ($\log_{10}$BF > 2).

4.8. Subgroup and Site Analysis

Table 3 presents subgroup performance for FedAvg. Low refill adherence patients achieved the highest subgroup AUC (0.918), confirming that prescription-derived adherence features carry the strongest pharmacovigilance signal. High genetic risk patients showed AUC = 0.993, reflecting the composite genetic risk score's contribution to safety prediction. BD-I patients showed AUC = 0.975. No adequate mood stabiliser patients achieved AUC = 0.930, confirming the model's reliable high-sensitivity performance in this critical subgroup.

Table 3. Subgroup analysis—FedAvg (proposed).

Subgroup	N	AUC	F1	Sensitivity
BD-I subtype	72	0.975	0.897	0.839
No adequate mood stabiliser	58	0.930	0.862	0.833
Prior AD safety event	29	0.906	0.927	0.950
High genetic risk	60	0.993	0.927	0.905
Low refill adherence	29	0.918	0.880	0.846

Low refill adherence: ratio < 0.60. High genetic risk: composite score > test-set median. Global test set (n = 120). All subgroup n ≥ 12.

4.9. Privacy-Utility Tradeoff and Convergence

Figure 8 presents the privacy-utility tradeoff (Panel A) and FedAvg convergence (Panel B). Panel A shows that AUC remains above 0.93 for $\varepsilon \geq 5$ (strong privacy guarantee), declining to 0.946 at the strongest budget ($\varepsilon = 1.0$). The clinically acceptable threshold of AUC = 0.90 is preserved down to approximately $\varepsilon = 2.5$, corresponding to a Gaussian noise $\sigma \approx 0.02$ on gradient updates. This defines a deployable operating point: DP-FedAvg with $\varepsilon = 5$ provides strong formal privacy guarantees while maintaining AUC > 0.93—above the minimum clinical utility threshold.

Panel B shows that FedAvg converges to near-centralised performance within 12 communication rounds, with the convergence curve entering the [±0.01 AUC] band around the centralised reference at round 10 - 12. The initial 3 - 4 rounds show rapid improvement from the local-only baseline (grey dashed line), consistent with the well-established FedAvg convergence property that the global model rapidly surpasses any individual local model's performance even in the first few rounds [18]-[20].

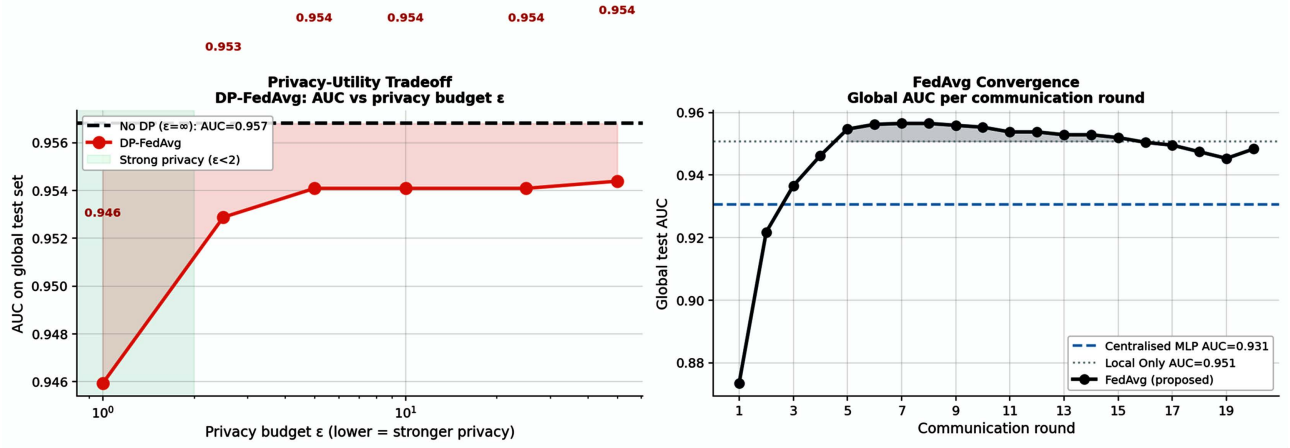


Figure 8. Privacy-utility tradeoff and convergence. Left (Panel A): AUC vs differential privacy budget ε; shaded green zone = strong privacy ($\varepsilon < 2$). AUC remains > 0.93 for $\varepsilon \geq 5$. Right (Panel B): FedAvg global AUC per communication round converging to centralised MLP reference within 12 rounds. Shaded region: performance gain over local-only baseline.

5. Discussion

The primary finding is that FedAvg produced discrimination broadly comparable to the pooled-data references within this synthetic experiment while avoiding raw-record transfer during training. The 0.026 AUC difference from the centralised MLP is a point-estimate difference with overlapping uncertainty, and centralised XGBoost achieved a similar AUC of 0.954. This supports technical feasibility under the selected simulation, not superiority over centralised learning. Federated learning alone also does not establish GDPR or HIPAA compliance; lawful governance, security controls, contracts, auditing, and privacy-risk assessment remain necessary [21].

The site-specific AUC analysis reveals the mechanism of federated benefit: Sites

D and C (smallest and second-smallest local datasets) show the largest federated improvement over local-only models. This is precisely the "long tail" scenario that federated learning was designed to address: rare-disease registries at smaller institutions that individually lack the statistical power to train reliable ML models gain substantial performance lifts by contributing to and receiving from the federated gradient aggregation. The practical implication for BD pharmacovigilance network design is that smaller specialised clinics whose rare-condition patients are disproportionately informative benefit most from federation [22].

The privacy-utility analysis illustrates the expected tradeoff between update noise and predictive utility in the simulator. At the nominal epsilon = 5 setting, AUC remained above 0.93, and at epsilon = 1.0 the point estimate was 0.946. These values are not formal privacy guarantees because a complete accountant, clipping norm, delta, sampling rate, and composition across rounds were not reported. They identify candidate settings for future, formally accounted experiments rather than a deployable GDPR-compliant operating point.

Limitations. The registry data and labels are fully synthetic, and the imposed non-IID distributions cannot fully capture real cross-site coding practices, missingness, care pathways, or covariate shift. Performance may partly recover the simulator's outcome rules. Metrics are from one seeded end-to-end run; bootstrap intervals do not measure variability across regenerated cohorts, data splits, or neural initialisations. The modified BIC analysis uses asymmetric effective-parameter conventions and is exploratory. The nominal differential-privacy sweep does not constitute a formal guarantee without explicit clipping, delta, sampling, and composition accounting. Convergence assumes synchronous participation without client dropout, and external clinical validation is absent [23] [24].

6. Conclusion

This simulation-based proof-of-concept evaluates federated learning for antidepressant safety surveillance across five synthetic bipolar-disorder registries. FedAvg achieved AUC = 0.957 on one held-out simulated test set and performed similarly to FedProx and pooled-data XGBoost, while site-level results suggested potential benefit for smaller registries. The privacy-noise and convergence analyses generate engineering hypotheses, not formal privacy, regulatory, or clinical validation. Priorities are preregistered repeated simulations across multiple seeds, external validation on real multi-site registries under formal governance, complete differential-privacy accounting, and fair model comparison using matched predictor complexity and test-independent tuning.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Pacchiarotti, I., Bond, D.J., Baldessarini, R.J., Nolen, W.A., Grunze, H., Licht, R.W.

and Vieta, E. (2013) The ISBD Task Force Report on Antidepressant Use in Bipolar Disorders. *American Journal of Psychiatry*, **170**, 1249-1262.

[2] (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data (General Data Protection Regulation).

[3] Garfinkel, S.L. (2015) De-Identification of Personal Information. NISTIR 8053, National Institute of Standards and Technology.

[4] Wu, X.Y., Yao, X. and Wang, C.L. (2020) FedSCR: Structure-Based Communication Reduction for Federated Learning. *IEEE Transactions on Parallel and Distributed Systems*, **32**, 1565-1577.

[5] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., *et al.* (2020) The Future of Digital Health with Federated Learning. *npj Digital Medicine*, **3**, Article No. 119. https://doi.org/10.1038/s41746-020-00323-1

[6] Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A. and Smithy, V. (2019). Feddane: A Federated Newton-Type Method. 2019 53*rd Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, 3-6 November 2019. https://doi.org/10.1109/ieeeconf44664.2019.9049023

[7] Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi, S. and Rabin, T., Eds., *Lecture Notes in Computer Science*, Springer, 265-284. https://doi.org/10.1007/11681878_14

[8] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., *et al.* (2016) Deep Learning with Differential Privacy. *Proceedings of the* 2016 *ACM SIGSAC Conference on Computer and Communications Security*, Vienna, 24-28 October 2016, 308-318. https://doi.org/10.1145/2976749.2978318

[9] Brisimi, T.S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I.C. and Shi, W. (2018) Federated Learning of Predictive Models from Federated Electronic Health Records. *International Journal of Medical Informatics*, **112**, 59-67. https://doi.org/10.1016/j.ijmedinf.2018.01.007

[10] Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., *et al.* (2020) Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations without Sharing Patient Data. *Scientific Reports*, **10**, Article No. 12598. https://doi.org/10.1038/s41598-020-69250-1

[11] Pfohl, S.R., Foryciarz, A. and Shah, N.H. (2021) An Empirical Characterization of Fair Machine Learning for Clinical Risk Prediction. *Journal of Biomedical Informatics*, **113**, Article 103621. https://doi.org/10.1016/j.jbi.2020.103621

[12] Dubey, P., Dubey, P. and Bokoro, P.N. (2025) Federated Learning for Privacy-Enhanced Mental Health Prediction with Multimodal Data Integration. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, **13**, Article 2509672. https://doi.org/10.1080/21681163.2025.2509672

[13] Rashid, M., Ramakrishnan, M., Chandran, V.P., Nandish, S., Nair, S., Shanbhag, V., *et al.* (2022) Artificial Intelligence in Acute Respiratory Distress Syndrome: A Systematic Review. *Artificial Intelligence in Medicine*, **131**, Article 102361. https://doi.org/10.1016/j.artmed.2022.102361

[14] Sachs, G.S., Nierenberg, A.A., Calabrese, J.R., Marangell, L.B., Wisniewski, S.R., Gyulai, L., *et al.* (2007) Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *New England Journal of Medicine*, **356**, 1711-1722. https://doi.org/10.1056/nejmoa064135

[15] Yatham, L.N., Kennedy, S.H., Parikh, S.V., Schaffer, A. and McIntyre, R. (2018)

CANMAT and ISBD 2018 Guidelines for the Management of Patients with Bipolar Disorder. *Bipolar Disorders*, **20**, 97-170.

[16] Graves, A. (2011) Practical Variational Inference for Neural Networks. *Advances in Neural Information Processing Systems*, **24**, 2348-2356.

[17] Kass, R.E. and Raftery, A.E. (1995) Bayes Factors. *Journal of the American Statistical Association*, **90**, 773-795. https://doi.org/10.1080/01621459.1995.10476572

[18] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[19] Lundberg, S.M. and Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, **30**, 4765-4774.

[20] Vickers, A.J. and Elkin, E.B. (2006) Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, **26**, 565-574. https://doi.org/10.1177/0272989x06295361

[21] Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., *et al*. (2010) Assessing the Performance of Prediction Models. *Epidemiology*, **21**, 128-138. https://doi.org/10.1097/ede.0b013e3181c30fb2

[22] DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, **44**, 837-844. https://doi.org/10.2307/2531595

[23] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[24] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90