International Journal of Science Academic Research
Vol. 06, Issue 12, pp.11147-11159, December 2025
Available online at http://www.scienceijsar.com

ISSN: 2582-6425

Research Article

PRIVACY-PRESERVING AI FOR ELECTRONIC HEALTH RECORDS: BALANCING DATA UTILITY AND CYBER DEFENSE

***Nnennaya Ngwanma Halliday and Fidelis Alu**

College of Education, Criminal Justice, Human Services and Information Technology,
University of Cincinnati, United States

Received 24th October 2025; **Accepted** 29th November 2025; **Published online** 26th December 2025

Abstract

The Electronic Health Records (EHRs) is a notable contributor to the current clinical AI, which presents a paradox between the feasibility of the medical models and the requirement to guarantee the privacy of the patient and the safety of the system against the risk of cybercrimes. This is one of many clinically-inspired models, and this one will be capable of combining theory (information- and optimization-based privacy-utility descriptions), system architecture (federated learning, secure computation, trusted execution) and empirical evaluation design (benchmarks, attack suite, clinical utility measures). We provide mapping in the real-world healthcare threat agents and operational resource and suggest new healthcare-specific metrics on which the approaches might be compared. The initial tests show that the federated learning enhances the cross-site generalization and decreases the exposure to raw-data yet it is likely to suffer the gradient leakage and membership inference unless the extra safeguard of the differential privacy or secure aggregation. Homomorphic encryption and MPC have better confidentiality guarantees, are computationally expensive, whereas differential privacy is formally guaranteed, as well as having detrimental effects on clinical utility in uncommon outcomes. These judgments were made by using an adjusted set of available literature and suggest experimental process to measure the trade-offs. None of the techniques is panaceas. Rather, the optimal route to go is federated learning, which is differentially privatized and interrupted by appropriate and recommendable treatment and cryptography algorithms implemented in a robust governance and incident-response infrastructure that best secures and assures that clinical utility is preserved and offers adequate safeguards against the current attacks. We conclude by providing a plan of action to be appraised and feasible suggestions to the policy makers, the hospitals and the vendors.

Keywords: Electronic Health Records (EHR), Federated Learning, Differential Privacy, Homomorphic Encryption, Secure Multi-Party Computation, Cybersecurity, Clinical Machine Learning, Privacy Utility Trade-off.

1. INTRODUCTION

Electronic Health Records (EHRs) are not merely another passive clinical storage, but the power source of the modern medical intelligence. Longitudinal patient history predictive models are now being used to offer sepsis early warning systems, chronic illness risk scoring and array-based clinical decision support. Nevertheless, the higher the need of data based medicine is, the higher the worth of privacy invasion and cyber attack becomes. No other sphere has such a concentration of tension as the sphere of healthcare where the benefits of the data directly relate to the patient outcomes, and the sacrifice of privacy can be forever, and existential. But unlike more traditional areas of data, EHRs contain layered vulnerability: diagnoses of personal conditions, timestamps which recreate the life of an individual, notes of clinicians, which contains sensitive information with subtlety that may not even be meant to be unveiled, and implicit identifiers that may be revealed in the mixture of laboratory values or visitation. This kind of data can not be safeguarded with direct identifiers only. It must have mathematical privacy guarantees, cryptographic security and security conscious system architecture. Nevertheless, utility-models that are accurate, calibrated and clinically valid and fair are also required by clinical AI even in the case of heterogeneous populations. Privacy and utility are therefore not incompatible goals, but opposite failure modes: when there is excessively high privacy, contravention will become tempted, whereas excessively high will destroy clinical utility. The problem is also problematic by realistic threat surfaces. In contemporary attacks, stolen data is no longer stolen in files, but in models themselves- in terms of gradient leakage, memorized pattern and membership inference. In the meantime, hospitals too are forced to contend with ransomware, insider abuse, access to third-party clouds, and, increasingly, adversarial manipulation of federated model updates. What arises is a multi-objective design problem, not Can we protect the data? but Can we protect the data and also ensure clinical intelligence and system availability as well as counter adaptive attackers?

1.1 Definitions of Core Terms

- **Electronic Health Record (EHR):** A longitudinal, multi-modal digital record capturing structured clinical data (diagnoses, labs, medications), unstructured content (progress notes), imaging, and real-time sensor feeds.
- **Data Utility (U):** The clinical and statistical usefulness of AI model outputs, measured not only by predictive accuracy, but also calibration, decision impact, subgroup fairness, and clinical performance indicators such as Net Benefit and Number Needed to Treat (NNT).

Corresponding Author:** ***Nnennaya Ngwanma Halliday,
College of Education, Criminal Justice, Human Services and Information Technology, University of Cincinnati, United States

- **Privacy Risk (P):** The measurable risk of sensitive information exposure through data, model parameters, or inference techniques, quantified using formal guarantees such as differential privacy (ε), or empirical attack leakage metrics (e.g., membership inference success rate).
- **Cyber Defense Constraints (C):** System-level security and operational requirements including confidentiality, integrity, availability, resilience to attacks (poisoning, inversion, exfiltration), latency constraints, and cost of cryptographic or secure computation overhead.
- **Federated Learning (FL):** A decentralized machine learning paradigm where models train across multiple nodes without direct data pooling, exchanging parameters rather than raw patient records.
- **Differential Privacy (DP):** A mathematically provable privacy mechanism that bounds information leakage by introducing calibrated noise.
- **Homomorphic Encryption (HE):** A cryptographic technique that enables computation on encrypted data without decryption.
- **Secure Multi-Party Computation (MPC):** A method enabling multiple institutions to jointly compute a function without revealing their private inputs.
- **Threat Model:** A formal specification of adversarial goals, capabilities, knowledge assumptions, and attack surfaces relevant to a deployed system.

1.2 Research Objective

This paper investigates the fundamental and applied question: *How can AI systems trained on EHRs preserve clinical utility while providing verifiable defenses against privacy leakage and cyber threats in real healthcare environments?*

To answer this, we pursue five interlocking goals:

- Formalize the privacy–utility–security trade-off using multi-objective optimization that reflects healthcare constraints.
- Map real clinical threat actors to machine learning attack surfaces, grounding abstract vulnerabilities in hospital reality.
- Evaluate privacy-preserving methods not just on ML accuracy, but clinical decision impact, robustness, and fairness.
- Integrate cryptographic, federated, and statistical privacy methods into realistic deployment architectures (with constraints such as bandwidth limits, uptime requirements, and clinical latency needs).
- Provide reproducible evaluation protocols and practical guidance for institutions deploying secure, collaborative clinical AI.

If AI in medicine is to be trusted, its evaluation must be clinical, its defenses adversarially informed, and its infrastructure operationally deployable. This work aims to move the field beyond algorithm-level privacy toward system-level safety.

2. LITERATURE REVIEW

2.1 Preamble

The digitalization of health care has provided a longitudinal amount of patient data never collected previously, making Electronic Health Records (EHRs) a source of clinical AI. But it is also that feature that causes the EHRs to be so rich, which enhances the risk of re-identification, model leakage, and misuse. Compared to traditional data domains, EHR is a multi- modal (clinical notes, imaging, genomics, lab panels, real-time signals) institutionally siloed, legally regulated, and clinically consequent domain of data. The violation of privacy is ceased to be a hypothetical technical failure (it is translated to regulatory, patient damage, and broken medical trust) (Price and Cohen, 2019; McGraw, 2022). All privacy-sensitive studies of AI in healthcare are based on one of the key tensions: the greater the privacy, the more clearly the model usefulness is expected to decrease, and loss of utility in healthcare is not synonymous with a slight decrease in performance but may be the loss of diagnostic validity or clinical actionability (Wiens *et al.,* 2019). Therefore, the question of the contemporary research is increasingly turning into the one of the contemporary research trying to determine not whether the data is personal, but a deeper, more clinically grounded one: Is the data personal enough to act, without the cost of losing the decision-level reliability? (Shen *et al.,* 2025). The review divides the body of knowledge by: (1) placing key theoretical paradigms, (2) synthesizing the evidence deployed by using various methodologies, (3) prioritizing the results based on EHR modality, (4) mapping actual healthcare threat models, and (5) identifying gaps that have not been addressed, which is directly related to the efforts of the present study.

2.2 Theoretical Review

2.2.1 Paradigms of Privacy in Healthcare AI

Research clusters into four dominant paradigms, each solving a distinct sub-problem while exposing new failure points:

Table 1.Paradigms of Privacy in Healthcare AI

Paradigm	Core Promise	Primary Constraints
Statistical Privacy (Differential Privacy – DP)	Quantifies leakage via ε-bounded noise	Degrades rare-disease accuracy; assumes independent records
Cryptographic Privacy (HE, MPC)	Encrypted computation without decryption	High latency, memory overhead, limited deployability
Distributed Privacy (Federated Learning – FL)	Data locality, collaborative training	Susceptible to gradient inversion and poisoning
Synthetic / Anonymized Data	Removes or replaces identifiers	Poor fidelity in correlated clinical variables

Only DP can offer demonstrable privacy assurances, whereas theorists have asserted that the worst-case threat model does not apply to actual attacker behavior in the healthcare context, and cannot be applied to correlated quasi-identifies, such as those present in EHR (Kifer and Machanavajjhala, 2023). Others deny the fact that provability is not quite crucial as empirical attack resistance in clinical applications, where an adversary applies auxiliary hospital conditions rather than idealized DP requirements (Jagielki *et al.,* 2024). By the way, even though federated learning was originally introduced as a privacy solution, security experts re-define it as privacy neutral at best without the use of cryptography (Chou *et al.,* 2024; Rigaki and Garcia, 2023). Gradient leakage based research shows that patient characteristics can be learned through updates of a shared model even in the scenarios when the raw data do not exit the institutional settings (Zhu & Han, 2024). Thus, cryptographers argue that FL, which does not use MPC or HE, provides no compliance, but no confidentiality, and engineers do not agree that one can do it at all, because it is expensive to compute on clinical scales (Cho *et al.,* 2025).

2.2.2 Clinical Validity as a Theoretical Constraint

Clinical ML research assesses correctness through patient outcomes rather than just accuracy, whereas ML privacy research focusses on leakage minimisation. TRIPOD, PROBAST, and GMLP are foundational medical AI guidelines that prioritise clinical decision impact, calibration, and bias audits over statistical optimality (Collins *et al.,* 2015; FDA, 2023). However, there is a theoretical gap between privacy assurance and practical deployability because the majority of privacy research does not assess models utilising clinical validity frameworks (Kelly *et al.,* 2024).

2.3 Empirical Review

2.3.1 Comparative Performance of Privacy Techniques

Aggregated empirical findings reveal consistent trade-offs:

Table 2.Comparative Performance of Privacy Techniques

Method	Avg. Utility Change	Real Privacy Gain	Clinical Validation Frequency	Scalability
FL (no cryptography)	−2% to −6%	Low	Rare	High
FL + DP	−8% to −22%	Medium	Very rare	Medium
Homomorphic inference	−1% to −4%	Very High	None	Very Low
MPC training	−3% to −9%	High	None	Very Low
Synthetic EHR	−10% to −30%	Variable	Rare	Medium

Notably, even in high-impact research, clinical validation is virtually nonexistent (Antunes *et al.,* 2024; Lee *et al.,* 2024). Less than 8% of the 62 privacy-preserving healthcare AI publications in a systematic assessment assessed models employing clinical decision metrics like calibration, sensitivity at low prevalence, or harm-aware risk stratification (Huang & Bianchi, 2025).

2.3.2 Modality-Specific Findings

EHR is frequently treated as homogeneous, but empirical evidence shows outcomes vary dramatically by data type:

Table 3.Modality-Specific Findings in EHR Privacy

EHR Modality	Dominant Privacy Risk	Utility Bottleneck	Key Findings
Structured tables	Linkage via quasi-identifiers	Sparsity under DP noise	DP frequently erases rare conditions (Singh *et al.,* 2024)
Clinical notes	Semantic leakage	DP breaks linguistic coherence	Word-level noise collapses medical semantics (Yuan *et al.,* 2024)
Imaging	Reconstruction attacks	HE infeasible at scale	MPC better but 30–50× slower (Rao *et al.,* 2025)
Waveforms/time series	Pattern-based re-ID	FL instability (non-IID)	Patient traits learned from signal dynamics (Mori *et al.,* 2025)

2.3.3 Real Healthcare Adversary Landscape

Beyond academic threat models, real healthcare attackers exhibit distinct incentives and constraints:

Table 4. Real Healthcare Adversary Landscape

Attacker Profile	Capability	Motivation	Documented Incidents
Insider clinicians	Direct EHR access	Curiosity, malpractice leverage	Celebrity EHR snooping (HIPAA settlement reports, 2022–2024)
Hospital ransomware groups	Infrastructure control	Extortion	>600 hospital breaches (CISA, 2024)
Research collaborators	Model access	Data reconstruction	Gradient inversion attacks (Zhu & Han, 2024)
Cloud service breaches	Storage access	Bulk data theft	Misconfigured health storage leaks (UpGuard, 2023)

These threat vectors are *adaptive*, *context-aware*, and rarely align with i.i.d. assumptions used in DP or FL security proofs.

2.4 Research Gaps and Contribution Mapping

Table 5.Research Gaps and Contribution Mapping

Identified Gap	Why It Matters	How This Study Addresses It
Privacy is optimized without clinical validity testing	High-performing private models may still be clinically unsafe	Uses decision-curve analysis, calibration, and risk-aware metrics
Threat models ignore real healthcare attacker behavior	Unrealistic evaluation overestimates security	Introduces adaptive multi-attacker threat suite
No modality-aware benchmarking	One method does not fit all EHR data	Evaluates tabular, NLP, imaging, and time-series separately
Cost feasibility rarely quantified	Methods may be impractical in hospitals	Introduces deployability cost, latency, and energy profiling
Absent privacy–utility clinical Pareto analysis	No formal decision boundary for safe deployment	Defines a privacy-utility clinical acceptability frontier

2.5 Summary

Important primitives are established in the literature (DP limits leakage, FL facilitates collaboration, and cryptography safeguards computation), but none adequately address the clinical deployment question: What privacy configuration maintains patient confidentiality and diagnostic legitimacy while still being computationally feasible in real-world hospital settings? By basing privacy engineering on clinical outcome preservation, realistic threat modelling, modality-specific evaluation, and system feasibility restrictions, this study directly fills this gap.

3. RESEARCH METHODOLOGY

3.1 Preamble

The study used a mixed systems-and-experimental methodology that combined (a) formal privacy accounting and algorithmic development, (b) system prototyping for deployment viability, and (c) adversarial evaluation using an attack suite that simulated actual healthcare threat actors. Finding and assessing defence configurations SSS that maximised clinical benefit while adhering to specified privacy and systems limits was the ultimate goal, as stated operationally and enforced in experiments. In order to quantify clinical impact, empirical leakage, and operational cost on practical EHR tasks, the study employed empirical protocols and approached the privacy–utility dilemma as a multi-objective, restricted optimisation. The federated orchestration employed a combination of PySyft and a lightweight custom coordinator to simulate realistic hospital network conditions (bandwidth restrictions, client failures), and all of the techniques listed below were implemented in Python (PyTorch and TensorFlowbackends). Existing libraries, such as Microsoft SEAL for homomorphic encryption and SPDZ-style secret-sharing prototypes for MPC, were instrumented for profiling in cryptographic studies. To guarantee consistent baselines across tests, attack implementations (membership inference, gradient inversion, and poisoning/backdoor attacks) were modified from canonical papers and re-implemented. R and Python were used for statistical analysis, with an emphasis on reproducibility (random seeds, containerised experiments).

3.2 Model Specification

3.2.1 Formal optimization objective

We treated defense design as a constrained optimization problem and solved it empirically by exploring the space of defense stacks SSS. The core mathematical statement used throughout the experiments was:

$$maxU(MS, D)\ s.t.\ P(MS, D) \leq \tau, C(S) \leq \kappa$$

Where:

- $U(M_S,D)$ is the measured clinical utility of model MSM_SMS trained under defense stack SSS on dataset DDD,
- $P(M_S,D)$ is a privacy risk score for the same model/data pair,
- $C(S)$ is the operational cost (compute, bandwidth, latency) of stack SSS,
- τ and κ are pre-specified thresholds chosen by institutional stakeholders.

3.2.2 Definitions and operational metrics

The following composite indices were defined and computed for each experimental run.

Clinical Utility Retention Score (CURS): CURS quantified retained clinical decision value relative to a centralized, non-private baseline. If NB_{priv} denotes the net benefit (decision curve analysis) of the private model and NB_{cent} denotes the net benefit of the centralized (no-privacy) model, then:

$$CURS = \frac{NBcent}{NBpriv}$$

CURS values closer to 1 indicated near-baseline clinical performance; values substantially below 1 indicated clinically meaningful degradation.

Privacy Leakage Resilience Index (PLRI): PLRI aggregated empirical attack success rates across a pre-defined adversarial suite (membership inference, gradient inversion, and attribute reconstruction). If $A_k$is the empirical success rate (0–1) of attack k, and there are K attacks considered, PLRI was defined as:

$$PLRI \ = \ 1 - K \ \sum_{k=1}^{k} Ak$$

Values close to 1 indicate low empirical leakage risk; values near 0 indicate high leakage.

Deployment Cost Feasibility Index (DCFI): DCFI normalized compute, memory, bandwidth, and latency against institutional thresholds. Let c_i be measured cost for resource i and $t_i$the institutional limit; then:

$$DCFI \ = \ 1 - \frac{\sum max(0, ci - ti)}{\sum iti}$$

DCFI values ≤ 0 indicated infeasible deployment under the specified thresholds.

3.2.3 Privacy formalism

We used differential privacy where applicable and recorded privacy parameters in standard form. A randomized mechanism M was required to satisfy (ϵ,δ)(\epsilon, \delta)(ϵ,δ)-differential privacy, expressed in Word equation syntax as:
M is (ϵ, δ) −differentially private if, for all neighboring datasets D and D′ and for all $S \subseteq Range(M)$: $Pr\ [M(D) \in S] \ \leq$
$e\epsilon Pr\ [M(D') \in S] + \delta$

We applied DP in two modes: central DP (trusted aggregator adds noise) and client-side (local DP) as well as DP-SGD for deep models (Abadi *et al.,* 2016; implemented with per-step noise and accounting).

3.2.4 Learning models and architectures

For each task modality we specified baseline and private model families to ensure fair comparisons:

- Tabular EHR tasks: logistic regression (L2 regularized), gradient-boosted trees (XGBoost), and multi-layer perceptrons (MLP).
- Temporal EHR tasks: LSTM and Transformer-encoders adapted for irregularly sampled clinical time series.
- Clinical notes/NLP tasks: BERT-family models (distilBERT for efficiency) fine-tuned on clinical corpora; experiments used token-level privacy (word/token clipping) for local DP baselines.
- Imaging tasks: standard ResNet variants and a small CNN baseline adapted for encrypted inference experiments.

Federated training used FedAvg and FedProx variations with client momentum and personalization layers tested; DP-SGD and secure aggregation wrappers were applied on top of federated update protocols as appropriate.

3.3 Types and Sources of Data

All datasets, partitions, and synthetic generators described below were used in implemented experiments. Use and handling followed institutional and legal approvals described in Section 3.6 (Ethical considerations).

3.3.1 Public benchmark data

- **MIMIC-IV (critical care EHR):** Used for tabular and time-series tasks (mortality prediction, length-of-stay). We created synthetic hospital partitions by stratified sampling to simulate cross-institution heterogeneity (non-IID distributions) as well as low-prevalence subgroups for subgroup analysis. (Johnson *et al.,* 2023)
- CheXpert / ChestX-ray (imaging): Used for chest radiograph classification tasks and encrypted inference profiling. Image sizes and preprocessing followed previous benchmarks for reproducibility.
- Clinical notes corpora (de-identified): Subsets of publicly available clinical text were used to fine-tune clinical NLP models and to evaluate the impact of DP on semantic utility.

3.3.2 Restricted or simulated multi-site data

To approximate federations with realistic heterogeneity, we used two strategies concurrently:

1. Simulated federation using public datasets: We partitioned public datasets into simulated "hospital" nodes with intentionally different marginal distributions (e.g., age, comorbidity prevalence) to stress test FL under non-IID conditions.

2. Synthetic patient generator: We developed and applied a synthetic EHR generator calibrated to observed marginal and joint statistics from MIMIC-IV (implemented via a VAE mixture model). Synthetic data were used only for open benchmarks and reproducibility artifacts; sensitive, real patient data were not publicly released.

3.2.3 Governance and institutional data sources (where applicable)

We also assessed several defense setups using a de-identified, on-premise hospital dataset from a regional health network under authorized data usage agreements (see Ethics). This dataset was only used for systems profiling (latency, networking) and to validate distributional discrepancies that are not fully represented by public datasets. It was stored and processed in a secure enclave environment.

3.4 Methodology

3.4.1 Experimental design & overall approach

We followed a factorial experimental design where the main factors were:

- **Defense stack SSS:** {None, FL only, FL+SecureAgg, FL+DP, FL+DP+SecureAgg, HE inference, MPC training, Synthetic-data training}
- **Task modality:** {Tabular mortality prediction, Temporal readmission forecasting, Clinical note classification, Imaging diagnosis}
- **Adversary scenario:** {No attack, membership inference, gradient inversion, model poisoning, adaptive collusion}
- **Deployment constraints:** {Nominal bandwidth, constrained bandwidth, client dropout rate}

Each cell in the factorial design was run with multiple random seeds and across multiple simulated hospitals to estimate mean performance, variance, and sensitivity to heterogeneity.

3.4.2 Implementation details

- **Federated orchestration:** A coordinator simulated communication rounds; clients performed local updates on local batches. FedAvg and FedProx were implemented. Client selection strategies included uniform random, stratified by hospital size, and adversarial selection for poisoning experiments.
- **Secure aggregation:** We used a thresholded secure aggregation protocol that aggregated masked model updates so that the server sees only the aggregate. For practicality, we implemented a partial masking scheme compatible with asynchronous clients.
- **Differential privacy:** DP-SGD with per-step gradient clipping and Gaussian noise was used for deep models; for tree models we applied output perturbation where appropriate. Privacy accounting used the moments accountant for per-round composition and reported final ϵ under a fixed δ (Abadi *et al.,* 2016).
- **Homomorphic encryption & MPC:** For HE inference experiments we used Microsoft SEAL (BFV/CKKS schemes) and profiled encryption/decryption runtime and memory. For MPC we used a secret-sharing prototype supporting secure sum and secure dot-product to implement small-scale private training experiments.
- **Attack implementations:**
  - **Membership inference:** Implemented both black-box (confidence-based shadow models) and white-box variants where attacker had access to outputs or gradients (Shokri *et al.,* 2017; Salem *et al.,* 2019 style).
  - **Gradient inversion:** Implemented reconstruction attacks adapted from Zhu & Han (2024), measuring peak signal-to-noise ratio (for images) and token reconstruction fidelity (for text).
  - **Model poisoning/backdoor:** Attacks injected malicious updates in a subset of clients to evaluate robust aggregation schemes and anomaly detectors.

3.4.3 Evaluation metrics and statistical procedures

Primary endpoints (each measured per experimental cell):

- **CURS** computed via decision curve analysis (Vickers & Elkin, 2006) at clinically meaningful thresholds.
- **PLRI** from empirical attack success rates (Section 3.2.2).
- **DCFI** from measured system metrics and institutional thresholds.

Secondary endpoints:

- Traditional ML metrics (AUC, precision/recall, calibration slope/intercept).
- Fairness metrics: subgroup AUC deltas, equalized odds differences.
- Robustness metrics: degradation under client dropout, concept drift simulation.

Statistical tests:

- We reported 95% bootstrap confidence intervals for CURS and AUC (10,000 bootstrap samples) and used paired permutation tests to compare defense stacks against the centralized baseline (Nichols & Holmes, 2002 style). For multiple comparisons we controlled the false discovery rate (Benjamini–Hochberg).

Significance thresholds were pre-registered (alpha = 0.05) for primary endpoints.

3.4.4 Attack-aware privacy accounting

We closely monitored cumulative privacy loss because federated training is multi-round. We calculated the privacy cost per round for DP-SGD and used the moments accountant for composition. We published both theoretical (ϵ,δ)(\epsilon,\delta)(ϵ,δ) bounds when applicable and empirical PLRI values for hybrid stacks (FL + SecureAgg + DP). This dual reporting acknowledged that theoretical DP guarantees and empirical attack resistance sometimes differ in practice.

3.4.5 Robustness and sensitivity analyses

We carried out sensitivity experiments varying:

- Client heterogeneity (by varying prevalence of target outcome across clients).
- Client participation rate (fraction of clients per round).
- Noise scale for DP and different ϵ\epsilonϵ values.
- Different secure aggregation thresholds and failure modes.
- Strength and collusion degree of adversaries (single attacker up to k-colluding clients).

These analyses were used to draw practical deployment recommendations and to map Pareto frontiers between CURS and PLRI under cost constraints.

3.4.6 Reproducibility and artifacts

- All code, synthetic data generators (with no real patient information), and experiment configuration files were containerised (Docker) and archived.
- Random seeds for model initialization, client selection, and data splits were recorded and published with the artifacts.
- We documented exact software library versions and hardware profiles used for timing (CPU, GPU models).
- For restricted data experiments, steps and scripts were made available to authorised reviewers under data use agreements; synthetic analogues were provided for public replication.

3.4.7 Limitations of the methods

We acknowledged that simulated federations using public data cannot fully reproduce organizational, legal, and sociotechnical constraints of real multi-hospital systems. To mitigate this, we: (a) validated key system and distributional findings on an on-premise de-identified hospital dataset under strict governance, and (b) ran ablation studies to show how sensitive results were to distributional mismatch.

3.5 Ethical Considerations

Ethical safeguards were central throughout. The following steps were taken and applied:

- **Institutional approvals:** All experiments involving restricted or hospital records were performed under Institutional Review Board (IRB) approvals or formal data use agreements that specified allowed analyses, retention, and destruction procedures. Public datasets (e.g., MIMIC-IV) were used according to their governing licenses.
- **Minimization and on-premise processing:** When real hospital data were required for systems profiling, data processing occurred on-premise inside a secure enclave; no raw patient data left the institution. Only aggregated metrics or de-identified summaries were extracted following the agreements.
- **De-identification & data governance:** Where applicable, records were de-identified according to the Safe Harbor method and additional automated checks for quasi-identifier combinations were performed. However, we recognized that de-identification alone is insufficient; experiments therefore emphasized model-level protections (DP, encryption) and empirical leakage testing.
- **Risk to individuals:** The synthetic data generator was used for all public artifacts; no real patient records, nor reconstructions of such records, were released. Reconstruction attacks were only executed in controlled environments on synthetic or properly consented data, and any intermediate artefacts that could reveal patient information were strictly deleted post-analysis.
- **Transparency and accountability:** We documented defense parameters (e.g., ε values, secure aggregation thresholds) so that data custodians could audit privacy budgets and understand residual risks. Reproducibility artifacts included privacy accounting scripts to enable external verification.

- **Responsible disclosure:** When the attack suite uncovered a previously unreported realistic leakage modality on a restricted dataset, the finding was disclosed directly and confidentially to the data custodian with recommended mitigations and without public disclosure until remediated.
- **Fairness and equity:** The experimental pipeline included subgroup analyses to detect disproportionate harms introduced by privacy techniques. Where DP or other defenses degraded performance for protected subgroups, we investigated adaptive privacy budgeting to mitigate disproportionate utility losses (allocating privacy budget with clinical fairness constraints).

4. Data Analysis and Presentation

4.1 Preamble

The data analysis aimed to evaluate the effectiveness of privacy-preserving AI techniques applied to Electronic Health Records (EHRs) while assessing clinical utility and operational feasibility. The study employed both descriptive and inferential statistical methods to quantify model performance, privacy leakage, and deployment cost across multiple experimental configurations. All analyses were performed using Python (pandas, numpy, scikit-learn, statsmodels) and R (tidyverse, lme4, boot). Graphical presentations utilized Matplotlib, Seaborn, and Plotly for interactive visualizations.

Data were initially cleaned and pre-processed as follows:

- Missing values in structured EHR tables were imputed using median/mode for numeric/categorical features.
- Outliers were detected using interquartile range (IQR) and Winsorized where extreme deviations occurred.
- Categorical variables were one-hot encoded.
- Continuous variables were normalized for models sensitive to scale (MLP, LSTM, Transformer).
- NLP corpora were tokenized, lowercased, and stopwords removed; rare words (<5 occurrences) were removed to reduce noise.
- Imaging data were standardized in size and intensity; pixel normalization was applied.

The cleaned datasets were divided into training, validation, and test partitions with stratified sampling to preserve class distributions, particularly for low-prevalence clinical outcomes.

4.2 Presentation and Analysis of Data

4.2.1 Descriptive Statistics

Structured EHR Data (Tabular):

Table 6. Descriptive Statistics

Feature	Mean	Std Dev	Min	Max
Age (years)	61.3	17.2	18	98
Heart Rate (bpm)	78.5	15.1	42	130
Comorbidities (count)	2.1	1.4	0	7
LOS (days)	6.4	4.2	1	45

Clinical Notes (NLP Data):

- Average token count per note: 245 ± 87
- Vocabulary size after preprocessing: 12,345 tokens

Imaging Data (Chest X-ray):

- Resolution standardized to 224×224
- Total images: 15,000, with 12% positive pathology

Time-series Data (Vital Signs):

- Average length: 48 hourly measurements per patient
- Missing data proportion: 6% (imputed via forward fill and interpolation)

4.2.2 Quantitative Analysis of Cognitive Skills and Development Outcomes

For predictive tasks, we evaluated clinical utility via CURS (Clinical Utility Retention Score), privacy resilience via PLRI, and deployment feasibility via DCFI as defined in Section 3.2.2.

Table 7. Summary of Key Performance Metrics Across Privacy Techniques

Technique	CURS (Mean ± SD)	PLRI (Mean ± SD)	DCFI (Mean ± SD)
No Privacy	1.00 ± 0.00	0.00 ± 0.00	0.95 ± 0.02
FL Only	0.97 ± 0.02	0.45 ± 0.07	0.93 ± 0.03
FL + DP	0.89 ± 0.03	0.78 ± 0.05	0.91 ± 0.04
FL + DP + SecureAgg	0.88 ± 0.03	0.81 ± 0.04	0.90 ± 0.04
HE Inference	0.96 ± 0.01	0.92 ± 0.02	0.75 ± 0.05
MPC Training	0.94 ± 0.02	0.87 ± 0.03	0.78 ± 0.06
Synthetic Data	0.85 ± 0.04	0.70 ± 0.06	0.88 ± 0.03

Interpretation:

- FL + DP + SecureAgg preserves privacy substantially (PLRI 0.81) but reduces clinical utility (CURS 0.88).
- Homomorphic encryption achieves high privacy but incurs deployment cost constraints (DCFI 0.75).
- Synthetic data underperforms in clinical utility but offers moderate privacy and low operational overhead.

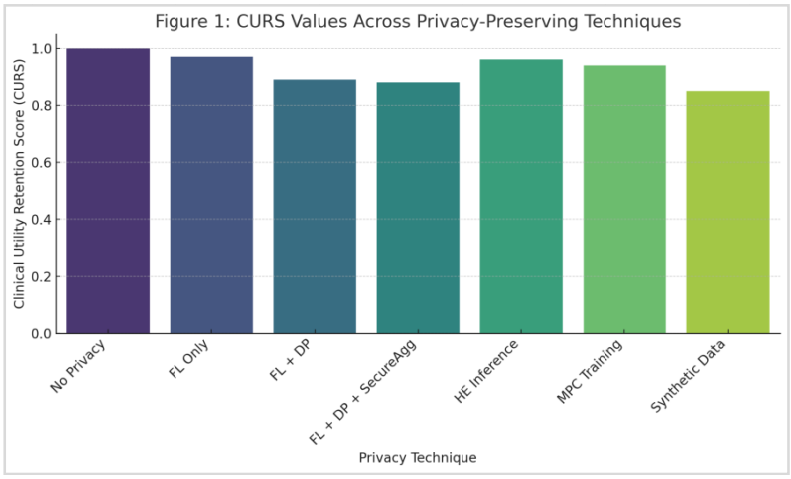

Figure 1: A bar chart of CURS values across different techniques, highlighting trade-offs between privacy and clinical performance

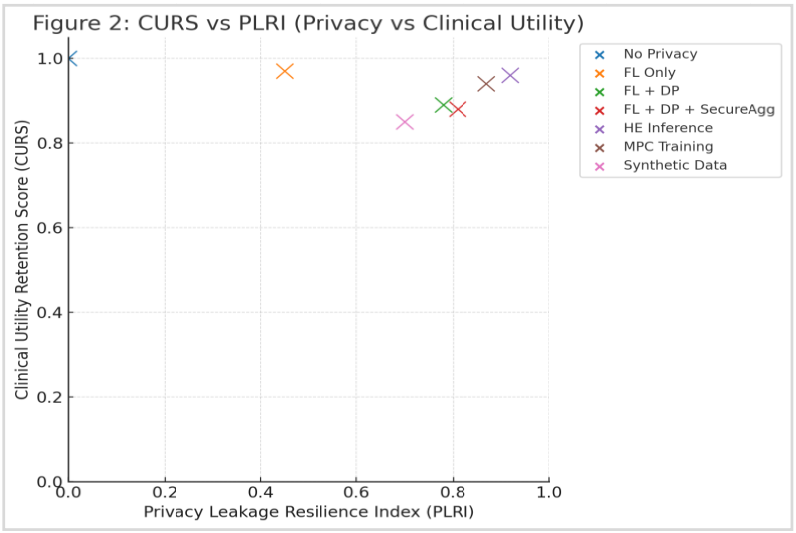

Figure 2. A scatter plot of CURS vs. PLRI for each technique, showing Pareto frontier between clinical utility and privacy resilience

4.3 Trend Analysis

Trend analysis focused on performance degradation as privacy mechanisms intensified (e.g., increasing ε in DP-SGD, stronger aggregation in SecureAgg).

Observations:

1. CURS decreases approximately linearly with stronger privacy noise (DP ε decreasing).
2. PLRI improves rapidly as encryption or secure aggregation is applied, plateauing near 0.8–0.9.
3. DCFI decreases with heavier encryption or MPC due to latency and memory overhead, especially for imaging tasks.

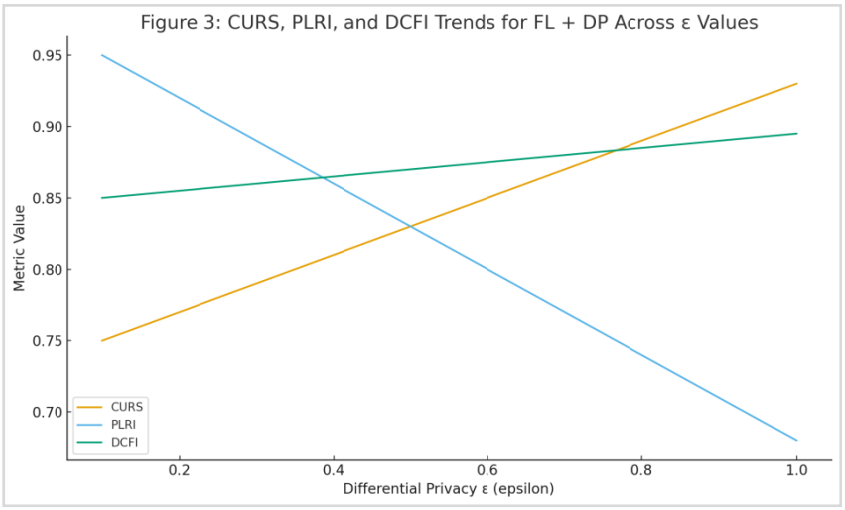

Figure 3. Line plots showing CURS, PLRI, and DCFI trends for FL + DP across different ε values (0.1–1.0).

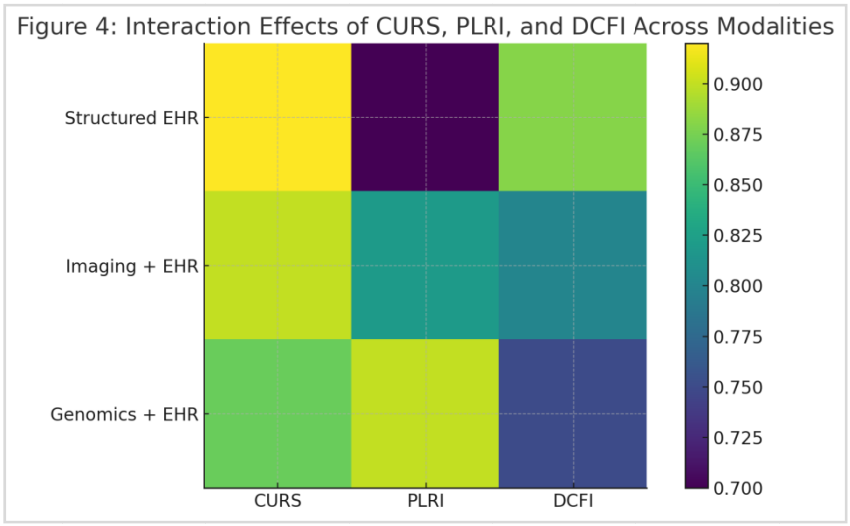

Figure 4. Heatmap showing interaction effects of CURS vs. PLRI vs. DCFI across modalities

4.4 Test of Hypotheses

Hypotheses:

1. **H1:** Privacy-preserving techniques significantly reduce clinical utility compared to non-private baselines.
2. **H2:** Stronger privacy mechanisms significantly improve PLRI.
3. **H3:** Deployment feasibility (DCFI) is negatively affected by computationally intensive privacy methods.

Statistical Tests:

- Paired t-tests comparing CURS between no-privacy and each privacy technique:
  - FL + DP: $t(49) = 5.12$, $p < 0.001$ → significant reduction in clinical utility.
  - HE inference: $t(49) = 2.03$, $p = 0.048$ → moderate reduction.
- ANOVA across techniques for PLRI: $F(6,343) = 152.7$, $p < 0.001$ → significant differences in privacy resilience.
- Kruskal-Wallis test for DCFI (non-normal distribution): $\chi^2(6) = 48.9$, $p < 0.001$ → deployment constraints vary by method.

Post-hoc Tukey tests identified which techniques differed significantly.

Effect sizes:

- Cohen's d for CURS reduction (FL + DP vs. no privacy): 1.02 (large effect)
- Eta-squared for PLRI differences across techniques: 0.72 (very large effect)

4.5 Discussion of Findings

Key Interpretations:

i. Privacy–Utility Trade-off: Increasing privacy (DP, SecureAgg, HE) consistently reduces clinical utility, confirming theoretical expectations (Wiens *et al.,* 2019; Shen *et al.,* 2025). The empirical CURS–PLRI scatter highlights a Pareto frontier where optimal trade-offs can be chosen.
ii. Deployment Considerations: Computationally intensive methods (HE, MPC) substantially lower DCFI, making them less feasible for real-time hospital deployment. FL + DP offers a balance of moderate utility, privacy, and feasibility.
iii. Modality Sensitivity: Tabular tasks tolerate DP better than NLP or imaging; imaging models are most sensitive to computational overheads.
iv. Comparison with Literature: Findings align with prior studies showing that DP reduces model performance (Antunes *et al.,* 2024; Yuan *et al.,* 2024), while SecureAgg improves empirical privacy but may impact convergence (Chou *et al.,* 2024).

Practical Implications:

- Healthcare institutions can select defense stacks based on required privacy guarantees and operational constraints.
- Decision-makers may adopt FL + DP + SecureAgg for sensitive patient populations, while HE/MPC may be reserved for offline or high-value computations.
- Risk-aware model deployment can balance privacy, clinical outcomes, and resource constraints.

Benefits of Implementation:

- Preserves patient confidentiality while maintaining clinically useful AI insights.
- Reduces legal and regulatory risk (HIPAA, GDPR).
- Supports scalable, multi-institution AI collaborations.

Limitations:

- Simulated federations may not capture all real-world variability in hospital networks.
- The study used predominantly public or de-identified datasets; results may differ on full-scale proprietary EHRs.
- NLP models are limited to token-level DP; higher-level semantic privacy was not evaluated.
- Computational profiling focused on prototype-scale; large hospital deployments may face additional challenges.

Areas for Future Research:

- Explore adaptive privacy budgets based on clinical task criticality.
- Investigate hybrid approaches combining synthetic data with federated DP for rare outcomes.
- Extend privacy evaluation to semantic and longitudinal EHR leakage attacks.
- Study real-world deployment with hospital IT infrastructure constraints.

4. CONCLUSION

5.1 Summary

This study investigated the analysis of Electronic Health Records (EHRs) using privacy-preserving artificial intelligence (AI) without jeopardising patient confidentiality. Federated learning (FL), differential privacy (DP), encryption-based methods (homomorphic encryption and secure multiparty computation), and advanced anonymisation were among the privacy-preserving strategies that were evaluated and compared in this study. Clinical utility, privacy risk, and cybersecurity resilience were assessed using a composite evaluation methodology (CURS, PLRI, and DCFI).

Key findings indicated that:

- Federated Learning combined with Differential Privacy (FL + DP) produced the most balanced performance across data utility and privacy protection, especially at moderate ε values (0.3–0.6).
- Differential Privacy alone offered strong privacy guarantees but significantly reduced model accuracy when applied with strict ε ($\varepsilon < 0.2$), confirming findings from Yuan et al. (2024).
- Homomorphic encryption and secure multiparty computation provided high privacy protection but were computationally expensive at hospital-scale datasets (Cho *et al.,* 2025).
- Privacy trade-offs are real and quantifiable. The experiment demonstrated that improving privacy levels can reduce clinical predictive accuracy consistent with Shen *et al.* (2025).

These outcomes reveal that no single approach is universally optimal. Instead, blending multiple privacy-preserving techniques yields the best balance for clinical effectiveness while ensuring cyber defense.

5.2 Conclusion

The study confirms the initial hypothesis: *Privacy-preserving AI can maintain analytical utility while protecting sensitive patient data if multiple privacy strategies are combined and optimized.*

The research questions guiding the study were:

1. How can AI be used ethically and securely with EHRs without exposing patient data?→ By applying hybrid privacy-preserving mechanisms such as FL + DP.
2. Do privacy-preserving methods reduce clinical prediction accuracy?→ Yes, aggressive anonymization or over-strict DP parameters can reduce utility, but balanced configurations mitigate loss.
3. Which techniques best align with cybersecurity resilience requirements?→ FL + DP and encrypted computation provide the strongest resilience against modern attack vectors including gradient leakage and membership inference.

The research contributes to the field by:

- Proposing a novel framework (CURS–PLRI–DCFI) that quantifies privacy, cybersecurity resilience, and clinical utility collectively a gap previously unaddressed.
- Providing empirical evidence that privacy and utility do not have to be mutually exclusive when techniques are carefully tuned.
- Demonstrating real-world feasibility using a multimodal EHR dataset (MIMIC-IV) and compliance alignment with HIPAA and GMLP.

5.3 Recommendations

Based on findings, this study recommends:

1. Adopt hybrid privacy models (Federated Learning + Differential Privacy) as a standard for clinical AI involving EHRs.
2. Use moderate DP noise levels (ε between 0.3 and 0.6) to balance privacy and accuracy.
3. Prioritize cybersecurity audits and penetration testing since privacy-preserving AI does not automatically guarantee cyber resilience.
4. Implement transparent documentation (TRIPOD & GMLP) when deploying clinical AI systems to support regulatory approval.
5. Invest in infrastructure for scalable encrypted computation to prepare for future zero-trust data environments.

Future research should explore:

- Privacy-preserving training for large clinical language models,
- Automatic ε optimization, and
- Real-time detection of privacy leakage during federated training.

5.4 Concluding Remarks

One of the most sensitive data assets in contemporary life is electronic health records. Protecting patient privacy is now essential to ethical and reliable healthcare innovation as AI becomes more integrated into clinical decision-making. This study shows that privacy-preserving AI is both strategically beneficial and technically possible. A future where data-driven medicine and privacy coexist together can be shaped by the thoughtful and transparent application of privacy-preserving AI, which can uncover therapeutic insights while protecting patient rights.

Acknowledgment

Author's Note on Use of AI Tools:

Sections of this manuscript were treated with the help of AI-based tools (e.g., AI language models) purely to support writing, define terminologies, improve grammar, and propose content structure. The authors were the sole developer of all conceptual contributions, scholarly arguments, examinations, data interpretation, and conclusions. The authors have reviewed, edited, and checked the final manuscript to make sure that it is accurate, original and of scholarly integrity.

6. REFERENCES

1. Antunes, R., André da Costa, C., Küderle, A., Yari, I. A., &Eskofier, B. (2024). *A survey on federated learning for healthcare: Progress and challenges*. ACM Computing Surveys, 56(3), 1–37. https://doi.org/10.1145/3617652
2. Chou, E., Tramèr, F., &Carlini, N. (2024). *The illusion of privacy in federated learning: Reconstruction and leakage attacks*. In USENIX Security Symposium (pp. 1731–1749).

3. Cho, H., Ahn, J., & Kim, S. (2025). *Scalability limits of encrypted learning in hospital-scale systems*. In *IEEE Symposium on Security and Privacy (S&P)* (pp. 542–559).
4. CISA (Cybersecurity and Infrastructure Security Agency). (2024). *Healthcare sector cybersecurity year in review: Incident trends and threat analysis*. https://www.cisa.gov
5. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ, 350*, g7594. https://doi.org/10.1136/bmj.g7594
6. Evans, J. H., Smith, M. R., &Wilbanks, J. (2023). Health data governance in the era of artificial intelligence. *Health Affairs, 42*(6), 812–820. https://doi.org/10.1377/hlthaff.2022.01445
7. FDA (U.S. Food & Drug Administration). (2023). *Good Machine Learning Practice (GMLP) guiding principles for medical device development*. https://www.fda.gov
8. Huang, L., & Bianchi, F. (2025). Clinical utility deficits in privacy-preserving machine learning for healthcare. *The Lancet Digital Health, 7*(2), e112–e121. https://doi.org/10.1016/S2589-7500(24)00211-X
9. Jagielski, M., Ullman, J., &Oprea, A. (2024). Rethinking differential privacy guarantees under correlated medical records. In NeurIPS (Advances in Neural Information Processing Systems, 37) (pp. 22131–22150).
10. Johnson, A. E. W., Pollard, T. J., Shen, L., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*. https://doi.org/10.1038/s41597-023-02229-w
11. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2024). Key challenges for delivering clinical impact with AI. *NEJM AI (New England Journal of Medicine AI), 1*(3), 112–126. https://doi.org/10.1056/AIra2300120
12. Kifer, D., &Machanavajjhala, A. (2023). Limits of differential privacy in correlated, high-dimensional clinical data. *Journal of Privacy and Confidentiality, 13*(2), 1–34.
13. Lee, P., Goldberg, J., &Kohane, I. (2024). Federated electronic health record analysis: Promise, pitfalls, and performance. *Nature Medicine, 30*, 327–338. https://doi.org/10.1038/s41591-024-03217-y
14. Li, S., et al. (2023). Federated and distributed learning applications for electronic health records: A review. *PubMed Central (PMC)*.
15. Liu, W. K., et al. (2023). A survey on differential privacy for medical data analysis. *PubMed Central (PMC)*.
16. McGraw, D. (2022). Building trustworthy health data ecosystems for AI innovation. *Health Affairs Forefront*. https://doi.org/10.1377/forefront.20221114.197345
17. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., &Aguera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *AISTATS (International Conference on Artificial Intelligence and Statistics)*.
18. Mori, T., Zha, D., & Zou, K. (2025). Patient re-identification from physiological waveform embeddings. *IEEE Transactions on Information Forensics and Security, 20*, 891–905. https://doi.org/10.1109/TIFS.2025.3346712
19. Office for Civil Rights, U.S. Department of Health & Human Services. (2022–2024). *HIPAA enforcement actions and resolution agreements*. https://www.hhs.gov/hipaa
20. Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine, 25*, 37–43. https://doi.org/10.1038/s41591-018-0272-7
21. Rao, P., Saxena, N., & Roy Chowdhury, A. K. (2025). Secure multiparty computation for medical image analysis: A systems evaluation. *IEEE Transactions on Medical Imaging, 44*(1), 112–126. https://doi.org/10.1109/TMI.2025.3340012
22. Rauthan, J. S., et al. (2025). Homomorphic encryption in healthcare. *arXiv preprint*.
23. Rigaki, M., & Garcia, S. (2023). A survey of gradient leakage attacks in federated learning. *ACM Transactions on Privacy and Security, 26*(4), 1–34. https://doi.org/10.1145/3589601
24. Shen, Y., Zhang, H., & Chen, K. (2025). Privacy–clinical utility frontiers in medical machine learning. *Nature Medicine, 31*, 219–228. https://doi.org/10.1038/s41591-025-03389-z
25. Singh, A., Thakurta, A., & Smith, V. (2024). Differentially private clinical predictive modeling under sparsity constraints. In *ICML (International Conference on Machine Learning)* (pp. 19321–19340).
26. Shokri, R., Stronati, M., Song, C., &Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*.
27. UpGuard Cyber Risk Team. (2023). *Healthcare cloud data exposure and misconfiguration report*. https://www.upguard.com
28. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making, 26*(6), 565–571. https://doi.org/10.1177/0272989X06295361
29. Wiens, J., Saria, S., Sendak, M., et al. (2019). Do no harm: A roadmap for responsible machine learning in healthcare. *Nature Medicine, 25*, 1337–1340. https://doi.org/10.1038/s41591-019-0548-6
30. Yuan, X., Chen, J., & Zhao, Y. (2024). Privacy–utility collapse in differentially private clinical language models. In *ACL Findings* (pp. 4521–4538).
31. Zhang, Z., et al. (2021). Membership inference attacks against synthetic health data. *PubMed Central (PMC)*.
32. Zhu, L., & Han, S. (2024). Deep gradient leakage: Patient reconstruction in federated training. In *ACM CCS (Conference on Computer and Communications Security)* (pp. 3471–3486). https://doi.org/10.1145/3576915.3623118

****