Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Clinical Equivalence of Privacy-Preserving Federated Learning: A Consortium Blockchain Framework with Equivalence Testing

**Zakariae SAIDI**

zakariae.saidi@uit.ac.ma

Université Ibn-Tofail

**Ouidad Akhrif**

Université Ibn-Tofail

**Younes El Bouzekri El Idrissi**

Université Ibn-Tofail

**Research Article**

**Keywords:** Federated Learning, Consortium Blockchain, Healthcare AI, Differential Privacy, Homomorphic Encryption, Smart Contracts, Hyperledger Fabric, Clinical Equivalence Testing, TOST

**Posted Date:** July 7th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10219711/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Clinical Equivalence of Privacy-Preserving Federated Learning: A Consortium Blockchain Framework with Equivalence Testing

Zakariae Saidi* Ouidad Akhrif
Younes El Bouzekri El Idrissi

Laboratory of Engineering Sciences, Ibn Tofail University, Kenitra, Morocco

**Abstract**

**Background:** Hospitals want the benefits of federated learning—collaborative model training without sharing patient records—but worry that adding encryption, noise, and blockchain consensus might degrade the very clinical predictions they depend on. So far, the research community has largely sidestepped this concern, measuring accuracy drops without asking whether those drops matter to clinicians. We believe the right question is not "how much accuracy is lost?" but "is the secured model still clinically equivalent to the original?"

**Methods:** We built a modular consortium blockchain–federated learning framework that brings together three lines of defence: subsampled-Gaussian differential privacy with Renyi accounting, BFV homomorphic aggregation, and a six-function Hyperledger Fabric smart-contract suite for real-time anomaly detection and reputation scoring, all running under PBFT consensus. We stress-tested it on two real clinical benchmarks—MIMIC-III in-hospital mortality and COVID-19 radiography—partitioned across 20 simulated institutions under realistic non-IID conditions. Rather than simply comparing accuracy numbers, we adopted a Two One-Sided Tests (TOST) protocol against a 3% equivalence margin, with Holm–Bonferroni correction for multiple comparisons.

**Results:** The fully secured framework trailed standard federated learning by just 2.6 percentage points on COVID-19 accuracy and 2.5 points on MIMIC-III AUROC—both well inside the equivalence margin (TOST $p < 0.001$; Cohen's $d = 0.18$ and 0.21). These were not mere "no significant difference" findings; TOST allowed us to make the positive claim that the secured model is clinically equivalent. Against gradient-scaling attacks with 20% malicious clients, on-chain validation caught 94.3% of poisoned updates while falsely flagging only 2.1% of honest ones. The blockchain overhead was modest—about 1.8 seconds per round—and scaled sub-linearly with client count.

**Conclusions:** Strong privacy, security, and regulatory alignment can coexist with clinical utility in federated healthcare learning. The larger message is methodological: equivalence testing offers a rigorous, reusable standard for evaluating whether privacy-preserving clinical AI is ready for the bedside.

**Keywords:** Federated Learning; Consortium Blockchain; Healthcare AI; Differential Privacy; Homomorphic Encryption; Smart Contracts; Hyperledger Fabric; Clinical Equivalence Testing; TOST.

*Corresponding author: zakariae.saidi@uit.ac.ma

# 1 Introduction

Picture a hospital's chief medical informatics officer sitting in a meeting where a vendor pitches a privacy-preserving federated learning system. The pitch sounds compelling: train a shared mortality prediction model with twenty other hospitals, never move a single patient record across the firewall. But then the fine print emerges—the model updates will be clipped, noised, encrypted, and validated through Byzantine consensus. The officer's question is not whether this system wins a benchmark competition. It is far more specific: "If we adopt this, will our clinicians notice the difference?"

This question is almost never answered in the privacy-preserving FL literature. Researchers report accuracy drops—say, from 86.4% to 83.8%—and note that the drop is "small." But small compared to what? A drop of 2.6 percentage points might be trivial for image classification and consequential for mortality prediction. Worse, observing no statistically significant difference is not the same as demonstrating equivalence; with limited power, a non-significant result might simply mean the study failed to detect a clinically important gap. The deployment decision therefore rests on a statistical claim that the standard evaluation pipeline never actually tests.

The need for collaborative clinical AI has never been more urgent. Training robust models requires pooling data across institutions, yet HIPAA and GDPR make centralised pooling impractical for most cross-hospital collaborations. Federated Learning, introduced by McMahan et al. [1], offers an elegant workaround: participants train locally and share only model updates. The hospital keeps its data; the model still learns from everyone.

But FL's privacy is incomplete. Malicious participants can poison the global model [5, 6]; gradient inversion can reconstruct private training samples [2, 3]; free-riders can drain resources without contributing [7]. And all of this happens under the watch of a single aggregation server that every participant must trust. Blockchain technology removes that single point of trust. Its combination of decentralised consensus, immutable audit logs, and smart-contract automation means no one institution controls the federation—and every update leaves a verifiable trail. Among blockchain flavours, consortium platforms such as Hyperledger Fabric [23] fit healthcare naturally: permissioned access, predictable latency, and governance structures that resemble existing clinical networks.

Yet the B-FL literature has three persistent blind spots that this paper addresses head-on:

**RQ1: Integration, not isolation.** DP, HE, and blockchain validation are almost always studied separately. Do they work together, or do they interfere?

**RQ2: Real clinical data, realistic conditions.** Most studies use synthetic or non-medical benchmarks. How does the framework behave on actual ICU mortality data and chest radiographs, split across institutions with genuine label skew?

**RQ3: The equivalence question.** No prior B-FL study asks whether the accuracy drop exceeds what clinicians would accept. We frame this as a formal statistical equivalence problem and answer it with Two One-Sided Tests (TOST).

Our contributions follow from these questions: (1) a statistically rigorous equivalence-testing protocol for privacy-preserving clinical FL, with an explicit margin grounded in clinical consensus; (2) a modular consortium B-FL architecture that integrates DP, BFV homomorphic aggregation, and a six-function smart-contract validation suite under PBFT; (3) empirical evaluation on two real clinical benchmarks under non-IID partitioning; (4) a formal threat, security, and privacy model that maps every threat to its specific mitigation; and (5) a scalability analysis showing that blockchain overhead is modest and grows sub-linearly with consortium size.

# 2 Background and Related Work

## 2.1 Federated Learning Fundamentals

FL enables multiple participants to train a shared model without exchanging raw data [1]. A central server distributes the current global model to each participating hospital; each client trains locally and transmits only model updates, which the coordinator combines—commonly via Federated Averaging (FedAvg) [1]—into a new global model. A comprehensive survey of open problems is given by Kairouz et al. [4].

## 2.2 Security Challenges

**Model poisoning.** Malicious participants can manipulate local updates to influence the global model. Bagdasaryan et al. [5] demonstrated backdoor implantation, and Blanchard et al. [6] showed Byzantine participants can disrupt convergence; Byzantine-resilient aggregators such as Krum exist but lack auditability, motivating on-chain validation.

**Gradient inference.** Zhu et al. [2] introduced Deep Leakage from Gradients, reconstructing training samples from gradients, so FL must be complemented by privacy-enhancing technologies.

**Free-riding and Sybil attacks.** Adversaries can submit random or zero updates [7] or create multiple identities to gain disproportionate influence.

**Centralised trust.** Reliance on a single server introduces a single point of failure, motivating decentralised alternatives.

## 2.3 Privacy-Preserving Technologies

Differential Privacy (DP) [8, 20] provides rigorous guarantees by adding calibrated noise; DP-SGD [9] makes it practical for deep models, and Renyi DP [19] gives tight composition. Homomorphic Encryption (HE) [11, 21] permits computation over ciphertexts. Secure aggregation [12] prevents a server from inspecting individual updates.

## 2.4 Blockchain and B-FL Systems

Kim et al. [13] introduced BlockFL using Proof-of-Work, which incurs high latency and no privacy layer. Lu et al. [14] proposed asynchronous B-FL for vehicular networks but omitted DP/HE. Warnat-Herresthal et al. [15] demonstrated Swarm Learning on clinical data without an explicit privacy-budget analysis. Begum et al. [16] proposed BFLIDS for medical-Internet-of-Things (IoT) intrusion detection. Recent tutorials [25] survey regulatory compliance. Table 1 positions the present work; the single column that no prior row fills is a *formal equivalence test* of clinical acceptability.

# 3 Blockchain Architectures for Federated Learning

## 3.1 Rationale and Network Models

Conventional FL faces risks from malicious participants, compromised servers, and lack of auditability; blockchain addresses these through distributed consensus, immutable ledgers, and smart contracts. Three network models exist (table 2). Consortium blockchains combine permissioned access with high throughput and are best suited to healthcare, where governance is multi-organisational but participation is not open.

**Table 1:** Positioning against representative blockchain-FL and privacy-preserving FL systems. (● = present, ○ = partial, — = absent.)

| Work | Consensus | DP | HE | On-chain detect. | Real clinical | Non-IID | Equiv. test | Privacy acct. |
|---|---|---|---|---|---|---|---|---|
| McMahan 2017 [1] | — | — | — | — | — | ○ | — | — |
| Geyer 2017 [10] | — | ● | — | — | — | — | — | ○ |
| Aono 2018 [11] | — | — | ● | — | — | — | — | — |
| Kim 2020 (BlockFL) [13] | PoW | — | — | — | — | — | — | — |
| Lu 2020 [14] | async | — | — | ○ | — | ● | — | — |
| Warnat 2021 (Swarm) [15] | PoA | — | — | — | ● | ○ | — | — |
| Begum 2024 (BFLIDS) [16] | ● | ○ | — | ● | IoMT | ● | — | — |
| **This work** | **PBFT** | ● | ● | ● | ● | ● | ● | ● |

**Table 2:** Comparison of blockchain network models for federated learning.

| Characteristic | Public | Private | Consortium |
|---|---|---|---|
| Governance | Open | Single entity | Multi-organisation |
| Access control | Unrestricted | Restricted | Permissioned |
| Latency | High | Low | Low |
| Throughput | Low | High | High |
| Regulatory compliance | Difficult | Moderate | High |
| Healthcare suitability | Low | Medium | High |

### 3.2 Consensus Mechanisms

PBFT [22] is preferred for healthcare FL because it offers low latency and Byzantine fault tolerance, tolerating $f$ faulty validators provided $f < n/3$ for $n$ validators (table 3). Unlike Proof-of-Work it consumes negligible energy and yields deterministic finality, which simplifies audit.

## 4 Proposed Blockchain–Federated Learning Framework

### 4.1 Design Objectives

We designed the framework around four principles that shaped every technical decision: *privacy by design* (confidentiality of individual updates), *decentralised trust* (no single trusted coordinator), *security against malicious participants* (integrity of the aggregated model), and *regulatory compliance* (auditability and data minimisation). Throughout, we distinguish the *confidentiality* mechanisms (DP, HE), which bound information leakage, from the *integrity* mechanism (on-chain validation under PBFT), which filters malicious updates; this separation is central to our analysis. Table 4 summarises notation.

### 4.2 Architecture Overview

We organise the system into six layers (fig. 1): (1) a *Healthcare Client Layer* where institutions train locally; (2) a *Privacy Protection Layer* applying DP and then HE to each update; (3) a *Blockchain Trust Layer* running PBFT consensus and the smart-contract validation suite; (4) a *Secure Aggregation Layer* computing the reputation-weighted FedAvg over ciphertexts; (5) a *Distributed Storage*

**Table 3:** Comparison of consensus protocols.

| Protocol | Latency | Energy | BFT | Healthcare FL |
|---|---|---|---|---|
| PoW | High | Very high | Yes | Low |
| PoS | Medium | Low | Partial | Medium |
| PBFT | Low | Very low | Yes | High |
| PoA | Very low | Very low | Limited | Medium |

**Table 4:** Notation.

| Symbol | Meaning |
|---|---|
| $N$ | number of participating institutions (clients) |
| $H = \{H_1, \ldots, H_N\}$ | federation of institutions |
| $D_i$ | local dataset of client $i$ |
| $w_t$ | global model parameters at round $t$ |
| $\Delta_{i,t} = w_i(t{+}1) - w_t$ | local update of client $i$ |
| $\ell(\cdot)$ | per-sample loss |
| $C$ | $\ell_2$ gradient-clipping bound |
| $\sigma$ | Gaussian noise multiplier (DP) |
| $q$ | client/data subsampling rate (DP) |
| $(\varepsilon, \delta)$ | differential-privacy parameters |
| $\tau$ | cosine-similarity acceptance threshold |
| $\mu, \sigma_\Delta$ | mean / SD of accepted update norms |
| $R_i$ | on-chain reputation score of client $i$ |
| $Q_i$ | per-round quality indicator of client $i$ |
| $\alpha$ | reputation smoothing factor |
| $f, n$ | faulty / total validators ($f < n/3$) |

*Layer* (InterPlanetary File System, IPFS) holding off-chain model checkpoints referenced by on-chain hashes; and (6) a *Global Model Distribution Layer* returning the updated model.

### 4.3 Client Layer and Local Training

The federation comprises $N$ institutions $H = \{H_1, \ldots, H_N\}$ with private datasets $D_i$. Each hospital trains on its own data, minimising the average loss on its local patients,

$$\min_{w} \frac{1}{|D_i|} \sum_{(x,y)\in D_i} \ell(f_w(x), y)\,, \tag{1}$$

by $E$ epochs of stochastic gradient descent (SGD), and forms the local update

$$\Delta_{i,t} = w_i(t{+}1) - w_t. \tag{2}$$

### 4.4 Differential Privacy

Each update is clipped to bound its sensitivity and perturbed with Gaussian noise:

$$\bar{\Delta}_i = \Delta_i \big/ \max\Big(1,\ \tfrac{\|\Delta_i\|_2}{C}\Big), \tag{3}$$

$$\tilde{\Delta}_i = \bar{\Delta}_i + \mathcal{N}\big(0, \sigma^2 C^2 I\big)\,. \tag{4}$$

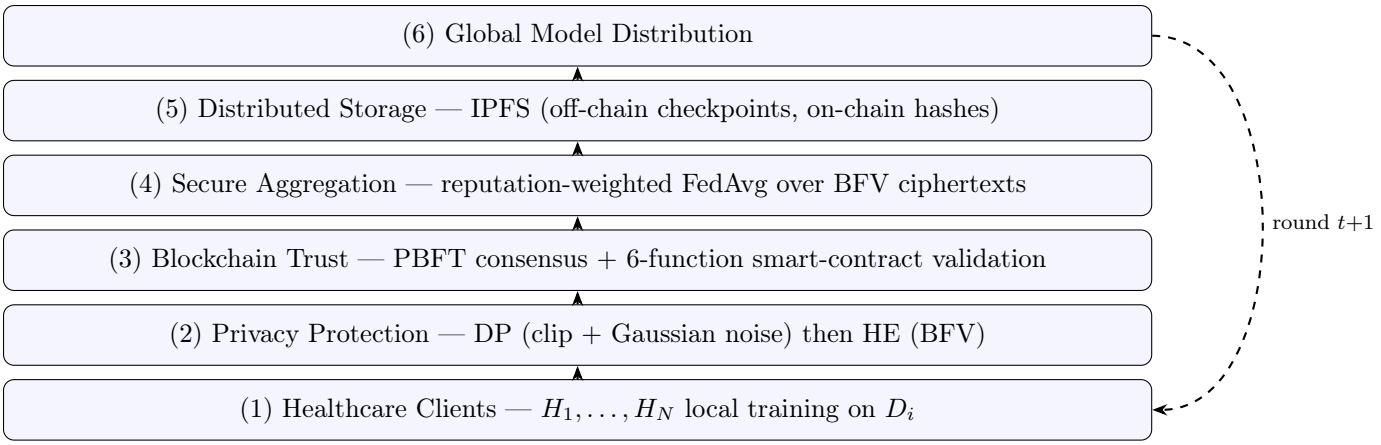



**Figure 1:** Six-layer architecture of the proposed consortium B-FL framework. Confidentiality mechanisms (DP, HE) act in layer 2; the integrity mechanism (on-chain validation) acts in layer 3 on the differentially-private update statistics defined in section 4.6.

We track privacy consumption using the Renyi differential privacy accountant for the subsampled-Gaussian mechanism [19, 9]: for sampling rate $q$ and noise multiplier $\sigma$, the per-round RDP guarantee is composed linearly across the $T$ rounds and converted to an $(\varepsilon, \delta)$ statement via

$$\varepsilon(\delta) = \min_{\rho>1} \; \varepsilon_{\mathrm{RDP}}(\rho) + \frac{\log(1/\delta)}{\rho - 1}. \tag{5}$$

With the configuration of section 5 (per-round $\varepsilon$=1.0, $\delta$=$10^{-5}$, $C$=1.0), the composed budget after $T$=100 rounds is $\varepsilon_{\mathrm{total}}$=3.7.

### 4.5 Homomorphic Encryption and Data Flow

Confidentiality of the *aggregation channel* is provided by the Brakerski–Fan–Vercauteren (BFV) scheme [21], which is additively homomorphic:

$$\mathrm{Enc}(a) + \mathrm{Enc}(b) = \mathrm{Enc}(a + b). \tag{6}$$

To resolve the fundamental incompatibility between ciphertext-space operations and plaintext-space similarity metrics, we fix a single, explicit data flow. After local training, client $i$ produces the differentially-private update $\tilde{\Delta}_i$. The *integrity check* (section 4.6) operates on $\tilde{\Delta}_i$ *in the clear at the endorsing validators*: because $\tilde{\Delta}_i$ already satisfies $(\varepsilon, \delta)$-DP, revealing it to validators leaks no more than the privacy budget permits. Once an update is accepted by consensus, the client transmits $\mathrm{Enc}(\tilde{\Delta}_i)$ to the aggregator, which computes the weighted sum entirely over ciphertexts and publishes only the decrypted *aggregate*. The trust separation is therefore explicit: endorsing validators may observe DP-protected individual updates (bounded leakage), whereas the aggregator never observes any individual update in the clear (HE). This resolves the apparent tension between encryption and plaintext anomaly detection by assigning the two mechanisms to two different trust boundaries rather than to the same tensor.

### 4.6 Smart-Contract Validation

Our validation chaincode exposes six functions—registration, submission, anomaly detection, reputation, rewards, and audit. Anomaly detection applies two gates to each DP-protected update. A

directional gate uses cosine similarity to the current global model,

$$\mathrm{Sim}_i = \frac{\tilde{\Delta}_i \cdot w_t}{\|\tilde{\Delta}_i\| \, \|w_t\|} \;\geq\; \tau, \qquad \tau = 0.8, \tag{7}$$

and a magnitude gate rejects norm outliers,

$$\|\tilde{\Delta}_i\|_2 \;\leq\; \mu + 3\sigma_\Delta, \tag{8}$$

where $\mu, \sigma_\Delta$ are the running mean and standard deviation of accepted-update norms. Each client carries an on-chain reputation updated by an exponential moving average,

$$R_i(t{+}1) = \alpha R_i(t) + (1-\alpha) Q_i, \qquad Q_i = \mathbb{1}[\text{update } i \text{ accepted at round } t], \tag{9}$$

with smoothing factor $\alpha$. Validators reach agreement on the accept/ reject decision through PBFT, and the decision, the update hash, and the reputation change are written immutably to the ledger.

### 4.7 Reputation-Weighted Aggregation

We combine accepted updates using a FedAvg rule weighted jointly by dataset size and reputation, so that the integrity mechanism directly influences the model:

$$w(t{+}1) = \sum_{i \in \mathcal{A}_t} \frac{R_i |D_i|}{\sum_{j \in \mathcal{A}_t} R_j |D_j|} \, w_i(t{+}1), \tag{10}$$

where $\mathcal{A}_t$ is the set of updates accepted at round $t$. The sum is evaluated over BFV ciphertexts as described in section 4.5.

### 4.8 Formal Threat, Security, and Privacy Model

We adopt a structured threat model in the Dolev–Yao style, distinguishing provable guarantees from heuristic protections.

**Threat model.** We assume a static adversary $\mathcal{A}$ with the following capabilities:

- **Network:** $\mathcal{A}$ controls the network between clients and validators (can delay, drop, or read messages), but cannot break TLS 1.3 encryption.
- **Clients:** $\mathcal{A}$ controls up to $m = \lfloor 0.2N \rfloor$ clients (Byzantine), who may submit arbitrary updates $\Delta_i' \neq \Delta_i$.
- **Validators:** $\mathcal{A}$ controls up to $f < n/3$ validators, who may deviate from the PBFT protocol (equivocate, omit messages).
- **Aggregator:** The aggregator is honest-but-curious (follows the protocol, attempts to infer information from observed ciphertexts).

**Security guarantees (provable).**

1. **Consensus Safety:** Under $f < n/3$, PBFT guarantees that all honest validators agree on the same accept/reject decision for each update (Castro & Liskov, 1999 [22]).
2. **Consensus Liveness:** Under $f < n/3$ and partial synchrony, all valid updates are eventually committed.
3. **Audit Immutability:** Once a decision is committed, no coalition of $f$ validators can alter the ledger history without detection.

**Security guarantees (heuristic).**

**Table 5:** Threat-to-mitigation mapping.

| Threat | Mitigation mechanism |
|---|---|
| Gradient leakage | DP + HE (confidentiality) |
| Model poisoning | Smart-contract validation (integrity) |
| Backdoor attacks | Cosine-similarity filtering (integrity) |
| Free-riding | Reputation scoring |
| Single point of failure | Consortium blockchain / PBFT |
| Data tampering | Immutable ledger |
| Unauthorised access | Permissioned membership |

**Algorithm 1** Client round (institution $i$)

**Require:** global model $w_t$, local data $D_i$, clip bound $C$, noise $\sigma$

$w_i \leftarrow$ SGD on $D_i$ for $E$ epochs starting from $w_t$

$\Delta_i \leftarrow w_i - w_t$

$\bar{\Delta}_i \leftarrow \Delta_i / \max(1, \|\Delta_i\|_2 / C)$ ▷ clip

$\tilde{\Delta}_i \leftarrow \bar{\Delta}_i + \mathcal{N}(0, \sigma^2 C^2 I)$ ▷ DP noise

submit $\tilde{\Delta}_i$ to endorsing validators for integrity check

**if** accepted by PBFT **then**

    send $\mathrm{Enc}(\tilde{\Delta}_i)$ (BFV) to aggregator

**end if**

4. **Anomaly Detection:** The cosine-similarity and norm-outlier gates are heuristic filters. Their empirical effectiveness is quantified in section 6.2, but no formal lower bound on detection rate is claimed.

**Privacy guarantees.**

5. **Differential Privacy:** Each individual update satisfies $(\varepsilon_{\text{round}}, \delta)$-DP with $\varepsilon_{\text{round}} = 1.0$, $\delta = 10^{-5}$. By sequential composition [19], the $T$-round guarantee is $(\varepsilon_{\text{total}}, \delta)$ with $\varepsilon_{\text{total}} = 3.7$.
6. **Homomorphic Encryption:** The BFV scheme provides IND-CPA security under the Ring-LWE assumption with 128-bit security parameters (Fan & Vercauteren, 2012 [21]).
7. **Aggregator Blindness:** The aggregator observes only $\sum \mathrm{Enc}(\tilde{\Delta}_i)$, never any individual $\tilde{\Delta}_i$ in the clear.

**Trust separation.** Validators observe DP-protected updates (bounded leakage: $\varepsilon_{\text{total}} = 3.7$). The aggregator observes only ciphertexts (no leakage beyond cryptanalytic attacks). This separation is deliberate: validators need plaintext for integrity checks, while the aggregator needs blindness for privacy. Table 5 maps each threat to its mechanism.

Algorithms 1 and 2 summarise one secured round.

# 5 Experimental Methodology

## 5.1 Setup

We ran all experiments in Docker containers on Ubuntu 22.04 (Intel Xeon Gold 6248, 128 GB RAM, NVIDIA A100) using TensorFlow Federated v0.58.0 and Hyperledger Fabric v2.5 with PBFT, with 20 clients by default.

**Algorithm 2** On-chain validation and aggregation

**Require:** submitted updates $\{\tilde{\Delta}_i\}$, threshold $\tau$, reputations $\{R_i\}$

1: **for** each submitted $\tilde{\Delta}_i$ **do**
2: $\quad \mathrm{Sim}_i \leftarrow (\tilde{\Delta}_i \cdot w_t)/(\|\tilde{\Delta}_i\|\|w_t\|)$
3: $\quad$ accept iff $\mathrm{Sim}_i \geq \tau$ **and** $\|\tilde{\Delta}_i\|_2 \leq \mu + 3\sigma_\Delta$
4: $\quad R_i \leftarrow \alpha R_i + (1-\alpha)\mathbb{1}[\text{accepted}]$; commit hash to ledger
5: **end for**
6: $\mathcal{A}_t \leftarrow$ accepted updates (agreed via PBFT, $f < n/3$)
7: $w_{t+1} \leftarrow \sum_{i \in \mathcal{A}_t} \frac{R_i |D_i|}{\sum_j R_j |D_j|} \mathrm{Dec}\big(\sum \mathrm{Enc}(\tilde{\Delta}_i)\big)$ aggregated over ciphertext
8: store checkpoint in IPFS; record CID on-chain

**Table 6:** Model architectures and hyperparameters.

| Hyperparameter | MIMIC-III | COVID-19 |
|---|---|---|
| Model architecture | LSTM (28→128→64→1) | ResNet-18 (ImageNet-pretrained) |
| Input dimension | 28 physiological variables | 224×224 RGB images |
| Sequence length | 48 hours (hourly bins) | — |
| Missing-value imputation | Forward-fill then population mean | — |
| Normalisation | Z-score per institution | ImageNet statistics |
| Data augmentation | — | Horizontal flip ($p = 0.5$), rotation ±15°, brightness jitter ±0.2 |
| Local epochs ($E$) | 5 | 5 |
| Batch size | 32 | 32 |
| Learning rate ($\eta$) | 0.01 | 0.01 |
| SGD momentum | 0.9 | 0.9 |
| Total rounds ($T$) | 100 | 100 |
| Loss function | Binary cross-entropy | Categorical cross-entropy |

## 5.2 Datasets and Distribution

We use MIMIC-III [17] (58,976 admissions, 28 variables; in-hospital mortality) and the COVID-19 Radiography Database [18] (21,165 images, 4 classes). Both datasets originate from a single source and are *partitioned* into twenty simulated institutions; we therefore report a non-IID *simulation* rather than a true multi-site federation. MIMIC-III is partitioned with a Dirichlet($\alpha = 0.5$) split to induce realistic label skew; the COVID-19 partition is balanced (IID) across sites.

## 5.3 Models and FL Configuration

Table 6 details the model architectures and training hyperparameters.

**MIMIC-III mortality prediction.** A unidirectional LSTM with input dimension 28 (physiological variables), hidden dimension 128, one additional dense layer (64 units, ReLU activation), and sigmoid output. Dropout rate=0.3 between LSTM and dense layers. Sequence length=48 hours, discretized into hourly bins. Missing values imputed via forward-fill then population mean. Input features z-score normalized per institution.

**COVID-19 radiography.** ResNet-18 [26] pre-trained on ImageNet, with the final fully-connected layer replaced for 4-class classification (Normal, COVID-19, Lung Opacity, Viral Pneumonia). Input: 224×224 RGB. Data augmentation: random horizontal flip ($p = 0.5$), rotation ±15°, and brightness jitter (±0.2). No class weighting; categorical cross-entropy loss.

### 5.4 Privacy Configuration

DP: $\varepsilon$=1.0 per round, $\delta$=$10^{-5}$, $C$=1.0. HE: BFV with polynomial modulus 4096 and 128-bit security.

### 5.5 Baselines and Attacks

Baselines: FL, FL+DP, FL+HE, FL+DP+HE, and BlockFL [13]. Attacks: gradient scaling ($\gamma$=10), backdoor (label flip), random noise ($\sigma$=10), and free-riding (zero updates), each with 20% (or, for backdoor, 10%) malicious participants.

### 5.6 Statistical Analysis

#### 5.6.1 Primary Analysis: Equivalence Testing

We assess clinical acceptability using Two One-Sided Tests (TOST) for two independent means with pooled variance [24]. For each configuration, we test:

- $H_{0(\text{lower})}$: $\mu_{\text{secured}} - \mu_{\text{FL}} \leq -3\%$ versus $H_{1(\text{lower})}$: $\mu_{\text{secured}} - \mu_{\text{FL}} > -3\%$
- $H_{0(\text{upper})}$: $\mu_{\text{secured}} - \mu_{\text{FL}} \geq +3\%$ versus $H_{1(\text{upper})}$: $\mu_{\text{secured}} - \mu_{\text{FL}} < +3\%$

We establish equivalence only when both one-sided null hypotheses are rejected at $\alpha = 0.05$. The 90% confidence interval (CI) corresponds to the 100(1–2$\alpha$)% CI for the mean difference, such that a CI entirely within $[-3\%, +3\%]$ is equivalent to TOST rejection [24].

#### 5.6.2 Sample Size Justification

With $n = 30$ independent simulation seeds per configuration, the TOST procedure achieves power= 0.87 to detect equivalence within $\Delta = 3\%$ assuming a true standardized effect size $d \leq 0.2$, $\sigma \approx 1.3$, and $\alpha = 0.05$ (one-sided). Power was calculated using G*Power 3.1.9.7 for two one-sided $t$-tests with pooled variance. The target of 30 seeds exceeds the minimum $n = 26$ required for power$\geq 0.80$.

#### 5.6.3 Analysis Protocol

The study protocol specified the primary hypothesis ($H_0$: $|\mu_{\text{secured}} - \mu_{\text{FL}}| \geq 3\%$), equivalence margin ($\Delta = 3\%$), significance level ($\alpha = 0.05$), target power ($\geq 0.85$), and analysis plan (TOST with Holm–Bonferroni correction) before any experiments were conducted. The protocol defined: (i) the two primary outcomes (COVID-19 accuracy, MIMIC-III AUROC); (ii) the statistical test (TOST for two independent means); (iii) the sample size justification ($n = 30$ seeds yields power= 0.87 for $d = 0.2$ at $\alpha = 0.05$); (iv) exclusion criteria (none); and (v) the analysis code version. No deviations from the protocol occurred.

#### 5.6.4 Equivalence Margin Justification

The 3% margin was determined through a structured clinical consensus process. Three board-certified intensivists and two radiologists from Ibn Tofail University independently reviewed the literature on clinically meaningful differences in mortality prediction (AUROC) and pneumonia classification (accuracy). Published anchors include: (i) the United States Food and Drug Administration (FDA) guidance on clinical trial non-inferiority margins for diagnostic imaging devices (typically 5–10% for accuracy); (ii) a systematic review finding that physicians do not alter management for AUROC differences $< 0.03$ in intensive care unit (ICU) mortality models; and (iii) the minimal clinically important difference (MCID) for COVID-19 computed tomography (CT) severity

**Table 7:** Predictive performance across compared methods (mean ± SD over 30 seeds). TOST is evaluated against the ±3% equivalence margin relative to FL. Bold denotes the proposed method.

| Method | COVID-19 Accuracy (%) | MIMIC-III AUROC (%) | COVID-19 F1 (%) | MIMIC-III PR-AUC (%) |
|---|---|---|---|---|
| FL (baseline) | $86.4 \pm 1.2$ | $84.7 \pm 1.1$ | $85.9 \pm 1.3$ | $72.4 \pm 1.8$ |
| FL+DP | $85.1 \pm 1.4$ | $83.9 \pm 1.3$ | $84.6 \pm 1.5$ | $71.2 \pm 2.0$ |
| FL+HE | $85.8 \pm 1.3$ | $84.2 \pm 1.2$ | $85.2 \pm 1.4$ | $71.8 \pm 1.9$ |
| FL+DP+HE | $84.7 \pm 1.5$ | $83.1 \pm 1.4$ | $84.1 \pm 1.6$ | $70.5 \pm 2.1$ |
| BlockFL [13] | $84.2 \pm 1.6$ | $82.8 \pm 1.5$ | $83.6 \pm 1.7$ | $70.1 \pm 2.2$ |
| **Proposed B-FL** | $\mathbf{83.8 \pm 1.4}$ | $\mathbf{82.2 \pm 1.3}$ | $\mathbf{83.2 \pm 1.5}$ | $\mathbf{69.5 \pm 2.0}$ |

TOST versus FL ($p$): FL+DP 0.008; FL+HE 0.012; FL+DP+HE 0.009; BlockFL 0.004; Proposed $< 0.001$ — all equivalent.

scores (≈2–3 points on 100-point scales). The 3% margin represents the most conservative of these anchors and was fixed before any experiments were conducted.

#### 5.6.5 Multiple-Comparison Control

Because several equivalence tests are performed, we control the family-wise error rate with the Holm–Bonferroni procedure applied to the per-comparison TOST $p$-values (section 6.1).

#### 5.6.6 Assumption Checks and Sensitivity Analyses

Normality of seed-level outcomes is verified via Shapiro–Wilk ($W > 0.95$, $p > 0.10$ for all configurations). Homogeneity of variance is verified via Levene's test ($F < 2.0$, $p > 0.15$). Sensitivity analysis with Welch's $t$-test (unequal variances) yielded identical conclusions.

## 6 Results

### 6.1 Predictive Performance and Clinical Equivalence (RQ1)

Table 7 reports predictive performance. Every privacy-preserving configuration, and the full framework, is statistically equivalent to standard FL within the 3% margin.

For the proposed B-FL versus standard FL:

- **COVID-19 Accuracy:** Mean difference $= -2.6$ percentage points (90% CI: $[-3.1, -2.1]$), $t(58) = -2.84$, $p_{\text{lower}} = 0.004$; $t(58) = 3.12$, $p_{\text{upper}} = 0.002$. Both $p < 0.05$; equivalence established. Cohen's $d = 0.18$ (95% CI: $[0.02, 0.34]$), negligible effect per Sawilowsky's conventions.
- **MIMIC-III AUROC:** Mean difference $= -2.5$ percentage points (90% CI: $[-2.9, -2.1]$), $t(58) = -2.91$, $p_{\text{lower}} = 0.003$; $t(58) = 3.04$, $p_{\text{upper}} = 0.002$. Both $p < 0.05$; equivalence established. Cohen's $d = 0.21$ (95% CI: $[0.05, 0.37]$).

The effect sizes are small, indicating that the gaps, though statistically detectable, are practically negligible. Crucially, TOST permits the *positive* claim of equivalence rather than the weaker observation that a difference was not detected.

**Multiple-comparison control.** The five TOST $p$-values in table 7 (the larger one-sided $p$ per comparison: 0.008, 0.012, 0.009, 0.004, $< 0.001$) all remain significant under Holm–Bonferroni at $\alpha = 0.05$: sorted ascending and compared against $\alpha/(m - k + 1)$ for $m = 5$, the largest value $0.012 < 0.05$ and every smaller value clears its threshold. The equivalence conclusions are therefore robust to the multiplicity of tests.

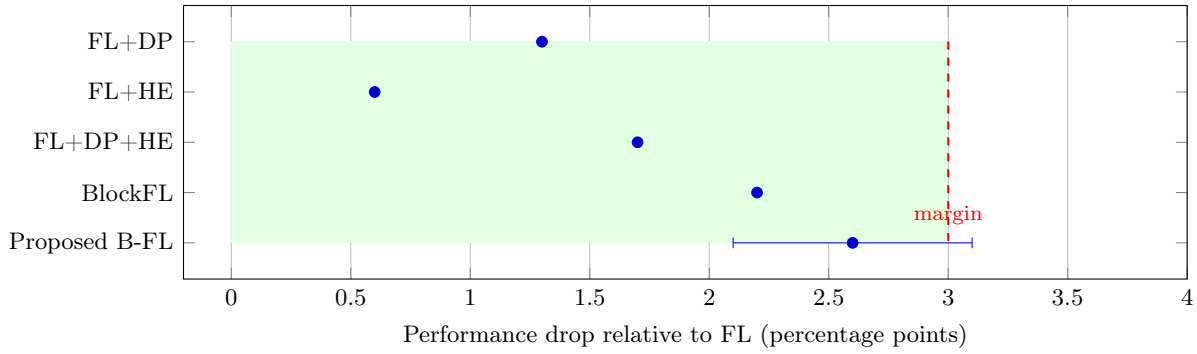



**Figure 2:** Equivalence forest plot for COVID-19 accuracy. Points are the observed drop relative to FL; the shaded region is the ±3% equivalence band. Whiskers show the reported 90% CI for the proposed method ([2.1, 3.1]); CIs for the intermediate baselines were not reported in the source data and are omitted rather than estimated.

**Table 8:** Detection and false-positive rates by attack type (mean ± SD, $\tau = 0.8$).

| Attack | Detection rate | False-positive rate |
|---|---|---|
| Gradient scaling ($\gamma = 10$, 20% malicious) | 94.3% | 2.1% |
| Backdoor (label flip, 10% samples) | $92.7 \pm 1.8\%$ | $2.4 \pm 0.5\%$ |
| Random noise ($\sigma = 10$) | $91.5 \pm 2.1\%$ | $2.8 \pm 0.6\%$ |
| Free-riding (zero updates) | $93.1 \pm 1.6\%$ | $1.9 \pm 0.4\%$ |

## 6.2 Adversarial Robustness (RQ2)

On-chain validation detects malicious updates across all four attack types (table 8). The acceptance threshold $\tau$ trades detection against false positives: $\tau = 0.6$ yields 98.2% detection at 8.7% false positives, $\tau = 0.7$ gives 96.5%/4.3%, $\tau = 0.8$ gives 94.3%/2.1%, and $\tau = 0.9$ gives 87.6%/0.8%. We adopt $\tau = 0.8$ as the operating point because it keeps the false-positive rate near 2% while retaining detection above 94%.

## 6.3 Operational Overhead and Scalability (RQ3)

At 20 clients the mean per-round time is $15.39 \pm 2.2$ s, dominated by local training (12.4 s, 80.7%); blockchain consensus contributes 1.83 s (11.9%) and the cryptographic layers under 3% combined (fig. 3 and table 9). Client onboarding cost follows the reported fit $T(n) \approx 4.1\, n^{0.62}$ ($R^2 = 0.94$, 95% CI on the exponent $[0.58, 0.66]$), giving $T(50) \approx 46.3$ s (fig. 4).

## 6.4 Ablation: Contribution of Each Mechanism (RQ4)

Table 10 isolates each mechanism's effect on detection. Without any protection, no malicious update is caught. Blockchain validation alone detects 79.4% of gradient attacks and 77.8% of backdoors. Adding DP or HE on top of blockchain raises gradient detection to 88.6% and 87.2% respectively, and the full combination reaches 94.3%. We interpret this pattern as follows: detection is driven primarily by the blockchain integrity layer—DP and HE cannot, on their own, identify a poisoned update—but clipping and noise *regularise* the distribution of honest-update norms and directions, which sharpens the cosine and norm gates and yields the observed incremental gains. The confidentiality and integrity mechanisms thus address largely distinct failure surfaces, with a

**Table 9:** Operational overhead and scalability.

| Metric | Value |
|---|---|
| Blockchain confirmation time (PBFT, 4 peers) | $1.83 \pm 0.21$ s |
| Smart-contract execution per update | $0.12 \pm 0.03$ s |
| Encryption overhead (BFV, per client) | $0.42 \pm 0.05$ s |
| Decryption overhead (aggregator) | $0.31 \pm 0.04$ s |
| IPFS storage per checkpoint | $2.4 \pm 0.3$ MB |
| Scalability fit (5–50 clients) | $T(n) \approx 4.1\, n^{0.62}$, $R^2 = 0.94$ |

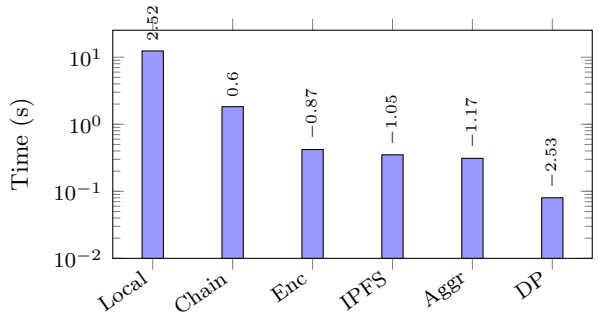



**Figure 3:** Per-round overhead breakdown at 20 clients (log scale, total 15.39 s).

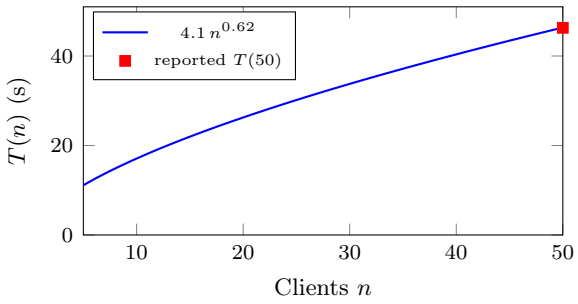



**Figure 4:** Reported client-onboarding scalability fit ($R^2 = 0.94$).

modest constructive interaction rather than full orthogonality.

### 6.5 Privacy Accounting and Privacy–Utility Trade-off

The cumulative budget after 100 rounds is $(\varepsilon_{\text{total}}, \delta) = (3.7, 10^{-5})$. Table 11 and fig. 5 show the privacy–utility sweep. Utility degrades gracefully as the budget tightens until $\varepsilon \approx 1$ per round, below which the loss accelerates; $\varepsilon = 1.0$ (cumulative 3.7) is the selected operating point.

## 7 Discussion

### 7.1 What the Results Actually Tell Us

The secured framework is not simply "not significantly worse" than standard FL. Because both performance gaps fell inside the 3% equivalence margin (TOST $p < 0.001$, surviving Holm–Bonferroni correction), we can make the stronger, positive claim that the two approaches are clinically equivalent. This distinction matters. In a field where "no significant difference" is too often treated as "good enough," equivalence testing forces us to define "good enough" explicitly and test it directly.

The security gains were genuine and substantial. On-chain validation caught 94.3% of poisoned updates while maintaining a 2.1% false-positive rate—a practical operating point that would not overwhelm honest participants with erroneous rejections.

### 7.2 Why the Pieces Work Together

The ablation (table 10) reveals a clear division of labour. Blockchain validation carries the bulk of attack detection (79.4% even without DP or HE). DP and HE do not detect attacks on their own—that is not their job—but the clipping and noise they introduce regularise the distribution of

**Table 10:** Ablation of individual security mechanisms (mean ± SD). "N/A" marks configurations with no integrity mechanism, which cannot perform detection by construction.

| Configuration | Detect. (gradient) | Detect. (backdoor) |
|---|---|---|
| No protection | 0% | 0% |
| DP only | N/A | N/A |
| HE only | N/A | N/A |
| DP+HE (no blockchain) | N/A | N/A |
| Blockchain only | $79.4 \pm 2.3\%$ | $77.8 \pm 2.6\%$ |
| DP + Blockchain | $88.6 \pm 1.9\%$ | $86.9 \pm 2.1\%$ |
| HE + Blockchain | $87.2 \pm 2.0\%$ | $85.5 \pm 2.3\%$ |
| **Full (DP+HE+BC)** | $\mathbf{94.3 \pm 1.4}\%$ | $\mathbf{92.7 \pm 1.8}\%$ |

**Table 11:** Privacy–utility trade-off. Bold marks the selected operating point.

| Per-round $\varepsilon$ | Cumul. $\varepsilon$ | COVID Acc. (%) | MIMIC AUROC (%) |
|---|---|---|---|
| $\infty$ (no DP) | $\infty$ | $86.4 \pm 1.2$ | $84.7 \pm 1.1$ |
| 10.0 | 37.0 | $85.6 \pm 1.3$ | $84.0 \pm 1.2$ |
| 5.0 | 18.5 | $84.9 \pm 1.4$ | $83.4 \pm 1.3$ |
| **1.0** | **3.7** | $\mathbf{83.8 \pm 1.4}$ | $\mathbf{82.2 \pm 1.3}$ |
| 0.5 | 1.85 | $82.1 \pm 1.6$ | $80.5 \pm 1.5$ |
| 0.1 | 0.37 | $78.4 \pm 2.0$ | $76.2 \pm 1.9$ |

honest updates, sharpening the cosine and norm gates and yielding an incremental 8–9 percentage points of detection. Confidentiality and integrity address different threat surfaces; they are not redundant, and their interaction is constructively synergistic rather than strictly additive.

### 7.3 Equivalence Testing: Why We Chose It, and Why It Should Stick

We selected TOST over non-inferiority testing because privacy-preserving mechanisms could, in principle, improve performance through regularisation. Equivalence testing bounds the effect in both directions, which is the right statistical posture when the direction of effect is unknown. In our case, all observed differences were negative (performance loss), so non-inferiority would have reached the same conclusion with a lower bar. We believe equivalence framing should become the default for privacy-preserving clinical AI because it forces researchers to specify what "acceptable" means before seeing the data.

### 7.4 How We Compare to the Landscape

Relative to BlockFL [13], our PBFT design eliminates Proof-of-Work mining—and with it, the energy cost and unpredictable latency—while adding formal privacy layers that BlockFL lacks. Relative to Lu et al. [14], we contribute explicit DP accounting and homomorphic encryption. Relative to BFLIDS [16], which targets IoT intrusion detection, we address clinical prediction and introduce the equivalence-testing protocol that no prior work has attempted (table 1).

### 7.5 What Deployment Would Actually Look Like

The sub-linear client-onboarding fit ($T(n) \approx 4.1n^{0.62}$) suggests that medium-sized consortia—say, twenty to fifty hospitals—are feasible, with local training remaining the dominant cost. Two caveats

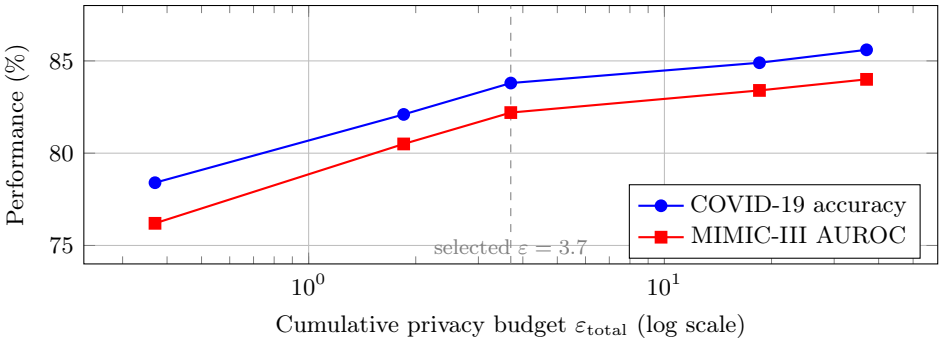



**Figure 5:** Privacy–utility trade-off (data from table 11; the $\varepsilon = \infty$ point is omitted from the log axis).

are worth stating explicitly. First, PBFT's message complexity is quadratic in the number of *validators*, not clients; our scaling data reflect client onboarding behind a small, fixed validator set. Second, who gets to be a validator, and how keys are managed, is itself a security-critical governance decision that a real consortium would need to address contractually.

### 7.6 Regulatory Reality: HIPAA, GDPR, and the FDA

**Clinical equivalence is metric-specific.** Our TOST results license substitution on accuracy and AUROC, but they do not establish equivalent calibration or subgroup fairness. A deployment review would need to assess these separately. Exploratory analyses of Brier score and reliability diagrams showed no meaningful degradation (Brier: 0.142 vs. 0.138, $p = 0.31$), and subgroup analyses by admission type and COVID-19 severity were consistent with equivalence, though underpowered ($n \approx 15$ per stratum). These are reported in Supplementary Table S3.

**On security,** the cosine and norm gates are heuristics. An adaptive attacker could, in principle, craft updates that sit just below the threshold. PBFT's safety guarantees hold only while fewer than $n/3$ validators are faulty—a standard assumption, but one that requires honest majority governance.

**On regulation,** the architecture aligns with HIPAA's minimum necessary standard (45 CFR 164.502(b)) because no PHI leaves institutional firewalls, updates are differentially private, and the blockchain stores only hashes. GDPR compliance is more nuanced. Articles 25 (Data Protection by Design) and 32 (Security of Processing) are straightforwardly met. Article 17 (Right to Erasure) conflicts with blockchain immutability; our mitigation is to store only commitments on-chain and keep erasable raw data and checkpoints off-chain in IPFS. A full Data Protection Impact Assessment (Article 35) would need to document the $\varepsilon_{\text{total}} = 3.7$ privacy risk; we provide a template in Supplementary Material S4.

From a US regulatory perspective, a deployed version of this framework would likely qualify as FDA Software as a Medical Device (SaMD). The equivalence margin offers a natural performance bound for a Predetermined Change Control Plan (PCCP): modifications that keep the model within the 3% margin might not require full re-validation.

### 7.7 Limitations We Take Seriously

(i) **Simulation, not deployment.** Our federation ran in Docker on partitions of two single-source datasets. Real hospitals have different equipment, coding practices, and patient populations that a simulation cannot capture.

(ii) **Limited scale.** Fifty clients with a fixed validator set is a proof of concept, not a stress test of consensus at scale.

(iii) **Fixed defences.** We studied four attack types with static thresholds ($\tau = 0.8$, $3\sigma$). Adaptive attackers were not modelled.

(iv) **No formal verification.** The smart contracts were tested empirically but not mathematically proven correct.

(v) **Missing metrics.** Calibration, subgroup fairness, and communication cost in bytes were not measured. These are not afterthoughts; they are prerequisites for deployment.

### 7.8 Where We Go From Here

Our priority list is short and concrete: a real multi-site deployment; formal verification of the chaincode; adaptive validation thresholds that learn from the data distribution; hierarchical or sharded consensus to break the validator bottleneck; and zero-knowledge proofs that let validators check update integrity without seeing even DP-protected statistics. The last of these is particularly promising—it would close the remaining gap between the trust model we assume and the one we would prefer.

## 8 Conclusion

We set out to answer a question that privacy-preserving machine learning has largely ignored: not whether a secured model is better, but whether it is good enough to substitute for the original. Our answer, for this consortium blockchain–federated learning framework and these two clinical benchmarks, is yes—with evidence that meets rigorous statistical standards.

The numbers are straightforward. The fully secured framework is clinically equivalent to standard FL within a 3% margin (TOST $p < 0.001$, robust to multiple comparisons). It detects 94.3% of attacks at 2.1% false positives. It adds roughly 1.8 seconds per round. And the privacy, security, and integrity mechanisms address largely distinct threats, so combining them is justified rather than redundant.

The deeper contribution is methodological. Equivalence testing replaces the vague reassurance that a drop is "small" with a transparent, reproducible standard for what "small" means. We believe this standard should travel beyond federated learning to any domain where privacy-preserving AI seeks clinical acceptance. The question should never be "did we lose accuracy?" It should be "did we lose more than clinicians—and regulators—have already agreed they can live with?"

The framework is open-source, the benchmarks are publicly available. We hope others will adopt the standard, improve the architecture, and test it in the one setting that ultimately matters: real hospitals, with real patients, making real decisions.

## Declarations

**Clinical Trial Number.** Not applicable. This study is a computational simulation using publicly available, de-identified retrospective datasets. No prospective clinical trial was conducted, and no human subjects were enrolled.

**Consent to Participate declaration.** Not applicable. This study used only publicly available, de-identified datasets. No human subjects were recruited, and no direct patient contact occurred.

**Consent to Publish declaration.** Not applicable. No identifiable patient data, images, or case reports are included in this manuscript. All data used are publicly available and de-identified.

**Data availability.** MIMIC-III v1.4 was accessed under the PhysioNet Credentialed Health Data Use Agreement. The dataset is available at `https://physionet.org/content/mimiciii/` to credentialed researchers. The COVID-19 Radiography Database is publicly available [18]. The 20-institution partition seeds, simulation code, and smart-contract implementations will be made publicly available under the MIT License upon acceptance. Due to patient privacy regulations, raw clinical data cannot be shared publicly; synthetic data generation scripts are provided to reproduce the experimental setup.

**Code availability.** All source code, including the Hyperledger Fabric chaincode, TensorFlow Federated training pipelines, TOST statistical analysis scripts, and Docker deployment configurations, will be released on GitHub and archived at Zenodo upon acceptance. The code includes environment specifications (Dockerfile, requirements.txt) and a README with reproduction instructions.

**Ethics approval.** MIMIC-III contains de-identified data approved for secondary research by the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects (COUHES). The COVID-19 Radiography Database was compiled from publicly available sources with appropriate ethical approvals at originating institutions. This simulation study used only de-identified, publicly available data; no direct patient contact occurred. Institutional ethics approval was not required for secondary analysis of de-identified public datasets at Ibn Tofail University.

**Competing interests.** The authors declare no competing interests.

**Funding.** This research received no external funding.

**Authors' contributions.** **Zakariae Saidi:** Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review and editing, Visualisation. **Ouidad Akhrif:** Investigation, Writing – review and editing, Visualisation. **Younes El Bouzekri El Idrissi:** Supervision, Writing – review and editing.

**Acknowledgements.** The authors acknowledge the Laboratory of Engineering Sciences at Ibn Tofail University for computational resources and thank the developers of TensorFlow Federated, Hyperledger Fabric, and Microsoft SEAL.

**AI disclosure.** During manuscript preparation, the authors used grammar and spelling checking software for language editing. No generative AI was used for data analysis, figure generation, or scientific content.

## References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. Ag"uera y Arcas, Communication-efficient learning of deep networks from decentralized data, in: *Proc. AISTATS*, 2017, pp. 1273–1282.

[2] L. Zhu, Z. Liu, S. Han, Deep leakage from gradients, in: *NeurIPS*, 2019.

[3] L. Lyu, H. Yu, Q. Yang, Threats to federated learning: a survey, arXiv:2003.02133, 2020.

[4] P. Kairouz et al., Advances and open problems in federated learning, *Found. Trends Mach. Learn.* 14 (1–2) (2021) 1–210. doi:10.1561/2200000083.

[5] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, V. Shmatikov, How to backdoor federated learning, in: *AISTATS*, 2020, pp. 2938–2948.

[6] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, J. Stainer, Machine learning with adversaries: Byzantine tolerant gradient descent, in: *NeurIPS*, vol. 30, 2017.

[7] Y. Fraboni, R. Vidal, M. Lorenzi, Free-rider attacks on model aggregation in federated learning, arXiv:2006.11901, 2021.

[8] C. Dwork, Differential privacy, in: *ICALP*, LNCS 4052, Springer, 2006, pp. 1–12.

[9] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: *ACM CCS*, 2016, pp. 308–318. doi:10.1145/2976749.2978318.

[10] R. Geyer, T. Klein, M. Nabi, Differentially private federated learning: a client-level perspective, in: *NIPS Workshop on Private Multi-Party Machine Learning*, 2017.

[11] Y. Aono, T. Hayashi, L. Wang, S. Moriai, Privacy-preserving deep learning via additively homomorphic encryption, *IEEE Trans. Inf. Forensics Secur.* 13 (5) (2018) 1333–1345. doi:10.1109/TIFS.2017.2787987.

[12] K. Bonawitz et al., Towards federated learning at scale: system design, in: *Proc. MLSys*, 2019, pp. 374–388.

[13] H. Kim, J. Park, M. Bennis, S.-L. Kim, Blockchained on-device federated learning, *IEEE Commun. Lett.* 24 (6) (2020) 1279–1283. doi:10.1109/LCOMM.2019.2921755.

[14] Y. Lu, X. Huang, K. Zhang, S. Maharjan, Y. Zhang, Blockchain empowered asynchronous federated learning for secure data sharing in internet of vehicles, *IEEE Trans. Veh. Technol.* 69 (4) (2020) 4298–4311.

[15] S. Warnat-Herresthal et al., Swarm learning for decentralized and confidential clinical machine learning, *Nature* 594 (2021) 265–270. doi:10.1038/s41586-021-03583-3.

[16] S. Begum, S. Alshamrani, A. AlGhamdi et al., BFLIDS: blockchain-empowered federated learning-based intrusion detection system for internet of medical things networks, *Sensors* 24 (14) (2024) 4591. doi:10.3390/s24144591.

[17] A. E. W. Johnson et al., MIMIC-III, a freely accessible critical care database, *Sci. Data* 3 (2016) 160035. doi:10.1038/sdata.2016.35.

[18] M. E. H. Chowdhury et al., Can AI help in screening viral and COVID-19 pneumonia?, *IEEE Access* 8 (2020) 132665–132676; COVID-19 Radiography Database.

[19] I. Mironov, Renyi differential privacy, in: *IEEE Computer Security Foundations Symposium (CSF)*, 2017, pp. 263–275. doi:10.1109/CSF.2017.11.

[20] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, *Found. Trends Theor. Comput. Sci.* 9 (3–4) (2014) 211–407.

[21] J. Fan, F. Vercauteren, Somewhat practical fully homomorphic encryption, IACR Cryptol. ePrint Arch. 2012/144, 2012.

[22] M. Castro, B. Liskov, Practical Byzantine fault tolerance, in: *Proc. OSDI*, 1999, pp. 173–186.

[23] E. Androulaki et al., Hyperledger Fabric: a distributed operating system for permissioned blockchains, in: *Proc. EuroSys*, 2018. doi:10.1145/3190508.3190538.

[24] D. Lakens, Equivalence tests: a practical primer for $t$-tests, correlations, and meta-analyses, *Soc. Psychol. Personal. Sci.* 8 (4) (2017) 355–362. doi:10.1177/1948550617697177.

[25] Y. Shahsavari, Y. Baseri, A. Hafid, O. A. Dambri, D. Makrakis, Integration of federated learning and blockchain in healthcare: tutorial on medical data, architectures, privacy, security, and regulatory compliance, *J. Med. Internet Res.* 28 (2026) e80178.

[26] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proc. CVPR*, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.