International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 8-No.3 (2022) pp. 85-93
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Securing U.S. Healthcare Infrastructure with Machine Learning: Protecting Patient Data as a National Security Priority

Nazmul Hasan[1]*, Imran Hossain Rasel[2], Moshiour Rahman[3], Kamrul Islam[4], Muhibbul Arman[5], Nusrat Jahan[6]

[1]Pompea College Of Business, University Of New Haven, CT, USA
* **Corresponding Author Email:** Mhasa9@Unh.Newhaven.Edu - **ORCID:** 0009-0007-2555-6047

[2] MS In Business Analytics, University Of New Haven, CT, USA
Email: Irase1@Unh.Newhaven.Edu r - **ORCID:** 0009-0002-9225-4595

[3] Ms In Business Analytics, University Of New Haven, CT, USA
Email: Mrahm22@Unh.Newhaven.Edu- **ORCID:** 0009-0007-6067-5906

[4] Ms In Business Analytics University Of New Haven West Haven, CT 06516 , USA
Email: Kamrulrupon104@Gmail.Com - **ORCID:** 0009-0001-8906-630X

[5] Pompea College Of Business, University Of New Haven, West Haven, Connecticut, United States.
Email: Iamarmanmohd@Gmail.Com - **ORCID:** 0009-0009-4629-3412

[6] Master Of Science In Analytics And Systems, University Of Bridgeport, Bridgeport, CT, USA.
Email: Njahan@My.Bridgeport.Edu - **ORCID:** 0009-0008-0684-1114

Article Info:

DOI: 10.22399/ijcesen.3987
Received : 19 November 2022
Accepted : 29 December 2022

Keywords

Healthcare cybersecurity,
Patient privacy,
Electronic Health Records (EHR),
Adversarial machine learning,
Differential privacy,
Internet of Medical Things (IoMT)

Abstract:

U.S. healthcare is a designated critical infrastructure whose disruption jeopardizes public health and national security. Yet escalating cyber risk, driven by large scale data breaches and ransomware, has outpaced traditional controls. This paper argues that machine learning (ML) can materially strengthen healthcare cyber defense if it is engineered with security and privacy as first class requirements. We synthesize pre July 2022 literature across adversarial ML, privacy preserving learning, and medical informatics, and propose an integrated architecture that combines federated learning, secure aggregation, and differential privacy to enable cross institutional detection while minimizing data exposure. We map ML techniques to concrete healthcare threat vectors insider misuse of electronic health record (EHR) data, credential stuffing against patient portals, lateral movement across medical IoT/telehealth ecosystems, and tampering with AI enabled clinical decision support and outline controls that align with NIST SP 800 53 and Zero Trust. A methodology section details data sources (EHR access logs, identity and access management telemetry, endpoint/IoMT signals, and clinical text), model families (unsupervised anomaly detection, sequence and graph models, and privacy preserving pipelines), governance (threat modeling, red teaming, privacy budgets, and model risk management), and evaluation (detection efficacy, time to detect, and formal privacy loss). We further discuss adversarial risks unique to medicine and the policy implications of deploying ML in regulated environments governed by HIPAA and FDA device guidance. Two figures visualize breach trends and cost asymmetries; tables operationalize the control mapping and measurement plan. We conclude that secure ML is not a panacea, but a necessary capability for resilient care delivery. Properly engineered, it can reduce dwell time, contain blast radius, and enable sector wide learning without centralized PHI pooling advancing both patient privacy and national security.

1. Introduction

The healthcare and public health (HPH) sector is explicitly recognized as U.S. critical infrastructure, and its reliable operation is a national security concern. Presidential Policy Directive 21 (PPD-21)

formalized the sector's status, emphasizing the need to strengthen and maintain secure, functioning, and resilient infrastructure across sectors whose compromise would debilitate national security and public safety (PPD-21). That imperative applies acutely to healthcare: cyber-enabled disruption of hospitals and supply chains can degrade clinical effectiveness, impede emergency response, and erode public trust.Risk growth has been stark. In 2021 alone, U.S. entities reported **712** healthcare breaches of ≥500 records to HHS OCR, impacting ~**45.7 million** records (Figure 1). The 2021 total set a new annual count record at the time (December 2021 and January 2022 HHS/OCR tallies summarized by HIPAA Journal). Beyond frequency, **impact** is pronounced: IBM's 2022 *Cost of a Data Breach Report* estimated the healthcare sector's average breach cost at **$10.1M**, more than double the cross-industry average of **$4.35M** (Figure 2). These asymmetries reflect operational urgency, complex vendor ecosystems, and prolonged detection/containment cycles.Adversaries have capitalized on this exposure. The CISA/FBI/HHS joint advisory on 2021 ransomware trends documented the professionalization of ransomware-as-a-service and the use of criminal "support" services for negotiation and payment capabilities that increase scale and endurance of campaigns against healthcare providers and suppliers. Such operations do not merely threaten confidentiality; they disrupt care delivery and can spill over into other lifeline sectors.Traditional perimeter-centric defenses struggle in modern, cloud-connected health systems with remote workforces, telehealth, and heterogeneous medical IoT. Federal guidance has accordingly shifted toward Zero Trust models (NIST SP 800-207) and risk-based control catalogs (NIST SP 800-53 Rev. 5), both emphasizing identity-centric access, continuous verification, segmentation, and resilient monitoring. For medical device and telehealth ecosystems, NIST SP 1800-30 provides a practical reference design for securing remote patient monitoring. These frameworks create an architectural "scaffold" into which data-driven detection can be embedded.Machine learning (ML) offers leverage at several points in this scaffold: (1) **Identity and access** sequence-aware and graph-based anomaly detection on EHR and IAM logs to flag credential misuse; (2) **Endpoint/IoMT** unsupervised detection of device behavior drift; (3) **Network** contextual detection of exfiltration and lateral movement; (4) **Data protection** automated de-identification of clinical text before secondary use; and (5) **Model assurance** defenses that harden clinical AI against adversarial manipulation. However, naïve ML deployment can worsen risk if it centralizes protected health information (PHI) or exposes models to privacy and adversarial attacks.Emerging privacy-preserving ML methods address these barriers. **Federated learning (FL)** enables cross-institutional training without pooling raw PHI; **secure aggregation** protects client updates in transit; **differential privacy (DP)** bounds what can be inferred about any individual from a trained model; and **homomorphic encryption (HE)** and secure enclaves can protect inference. Importantly, the medical AI community has begun to demonstrate these methods in imaging and multi-site settings while noting tradeoffs among privacy, utility, and robustness.This paper makes three contributions. First, it frames patient-data protection as a **national security** priority, grounding the argument in federal doctrine and sector-specific breach economics. Second, it **maps** healthcare threat vectors to ML controls designed with privacy and adversarial risk in mind and aligned to federal guidance. Third, it proposes a **methodology** for designing, evaluating, and governing secure ML pipelines that respect HIPAA obligations while enabling sector-wide learning.Figures 1 and 2 contextualize the urgency: breach frequency surged between 2014 and 2021, and the cost differential indicates that failing safely is especially expensive in healthcare. Our tables operationalize the approach by connecting threats to controls and by specifying evaluation metrics and governance checkpoints. The remainder of the paper reviews relevant literature, details the methodology, and discusses implications for practice and policy.

2. Literature Review

Healthcare threat landscape. Empirical analyses and public reporting indicate a steady increase in reportable breaches since the early 2010s, with major spikes in 2015 (notably insurer mega-breaches) and again by 2021 (Figure 1). HIPAA Journal's year-end 2021 analysis and January 2022 update record the highest annual count to date then, underscoring a sustained shift from small unauthorized disclosures to large hacking/ransomware events. The 2015 spike **255** incidents and >**112M** records illustrated the outsized impact of a handful of high-value compromises (e.g., payors).CISA, FBI, and HHS assessed that ransomware groups professionalized in 2021, adopting affiliate models, data-theft tactics, and service ecosystems (negotiators, money launderers), while exploiting weak remote access, credential reuse, and unpatched vulnerabilities. The advisory situates healthcare within a broader

economic system in which disruption creates leverage to coerce payment and rapidly monetize stolen data.**Policy and control frameworks.** At the doctrine level, PPD-21 defines 16 sectors explicitly including HPH as critical infrastructure, directing risk-informed collaboration among public and private stakeholders. Within healthcare, CISA's sector-specific plan (2015) clarifies governance roles. NIST SP 800-53 Rev. 5 catalogs security and privacy controls; SP 800-207 articulates Zero Trust principles; and SP 1800-30 demonstrates a reference architecture for securing remote patient monitoring especially salient as telehealth expanded. These create a common lexicon and baseline expectations for technical and organizational controls that ML-based detection must complement, not replace.**Machine learning in clinical data and operations.** Foundational work showed that modern deep learning can model raw EHR sequences and multimodal hospital data for clinical prediction tasks, but also highlighted heterogeneity, temporal dynamics, and generalization challenges (e.g., Miotto et al.; Rajkomar et al.; Choi et al.). These capabilities imply that similar architectures attuned to operational telemetry rather than patient outcomes can model *access* and *behavior* sequences for security detection. Meanwhile, NLP methods have improved PHI de-identification (e.g., RNNs) for clinical notes, facilitating privacy-preserving secondary use.**Adversarial ML threats in healthcare.** The advent of adversarial examples and robustness failures in deep networks (Goodfellow et al.; Carlini & Wagner; Biggio & Roli) raised alarms for medical AI. Finlayson et al. argued that medicine is uniquely susceptible to adversarial manipulation due to financial incentives and the introduction of model-driven workflows demonstrating attacks across medical imaging tasks and calling for robust evaluation, regulatory review, and domain-specific defenses. Such threats extend beyond image classifiers: manipulated inputs to ML-assisted triage, billing, or utilization-management systems could distort care or reimbursement.**Privacy attacks and countermeasures.** Even absent adversarial test-time manipulation, deployed models can leak training data. **Model inversion** (extracting sensitive attributes) and **membership inference** (determining whether an individual was in the training set) were demonstrated across model classes and domains, including health-related data (Fredrikson et al.; Shokri et al.). These attacks motivate limiting per-example influence and restricting output confidence exposure. **Differential privacy** (DP) provides provable, quantifiable privacy guarantees by bounding how much a single record can change observable outputs; DP-SGD variants enable training deep nets under formal privacy budgets. In healthcare, studies applied DP to medical imaging and surveyed DP in health research, highlighting tradeoffs between privacy loss (ε) and utility.**Federated and privacy-preserving learning. Federated learning (FL)** allows multi-site training without centralizing PHI. In medical imaging and pathology, FL has achieved performance close to centralized baselines across multiple institutions (Sheller et al.) and is proposed as a path to unlock distributed health data (Rieke et al.). **Secure aggregation** protects gradient updates during FL, ensuring the server learns only aggregate statistics; protocols at scale have been demonstrated in industry. Combined with client-side DP or server-side noise addition, FL can reduce data exposure while bounding privacy risk. **Homomorphic encryption** (CryptoNets; CKKS) and trusted execution environments enable encrypted or hardware-isolated inference, though latency and accuracy tradeoffs remain non-trivial for real-time clinical settings.**Securing telehealth and medical IoT.** As care extends beyond hospital walls, device and platform security is pivotal. NIST SP 1800-30 integrates identity, update, data protection, and monitoring controls for remote patient monitoring; NISTIR 8259A defines baseline IoT device capabilities that support cybersecurity controls. These guidance documents are relevant telemetry sources and enforcement points for ML-driven detection, e.g., modeling device behavior profiles and detecting anomalies suggestive of compromise.**Economics and national-level framing.** IBM's 2022 report quantified average healthcare breach costs at **$10.1M**, highest among sectors for the 12th consecutive year; detection and escalation costs rose markedly, reflecting longer attacker dwell times and complex investigations. Together with federal doctrine (PPD-21), these economics justify framing patient-data protection as a national-security-adjacent imperative: breaches ripple across clinical care, payer operations, and public confidence.**Synthesis.** Pre-July-2022 evidence supports three design goals: (1) **Minimize raw PHI movement** via FL and de-identification; (2) **Bound leakage** via DP and careful model/API design; (3) **Harden models and pipelines** against adversarial use. The literature also cautions that privacy and robustness are distinct: DP does not guarantee adversarial robustness, and robustness measures can inadvertently leak data. Therefore, system design must treat privacy and robustness as co-equal but separate requirements under a common governance program.

3. Methodology

Objective and scope. We propose a production-oriented methodology to design and evaluate ML defenses that reduce attacker dwell time, limit blast radius, and protect PHI *without* creating new concentrations of sensitive data. The approach targets three layers of health-system telemetry: **(A) Identity & access**, **(B) Endpoint/IoMT & telehealth**, and **(C) Network/data exfiltration**, plus **(D) Data protection** (clinical text de-identification; privacy budgets).**Threat modeling.** Using STRIDE-like categories adapted to healthcare, we prioritize: (1) **Credential misuse/insider abuse** of EHRs and data warehouses; (2) **Privilege escalation & lateral movement** across endpoints and IoMT; (3) **Data exfiltration** via cloud/SaaS connectors; (4) **Ransomware staging** and command-and-control; (5) **Adversarial manipulation** of AI-enabled clinical tools. We align mitigations with NIST SP 800-53 control families (AC, AU, IA, SC, SI) and SP 800-207 (continuous verification, least privilege, micro-segmentation).

Data sources and minimal-exposure collection.

- **A1. EHR/IAM audit logs.** Fine-grained access events (user, role, patient, context, location, device, time), authentication outcomes, privilege changes.
- **B1. Endpoint/IoMT telemetry.** Process, driver, and network metadata from clinical endpoints; medical device inventory/firmware/update state per NISTIR 8259A; telemetry from telehealth platforms per SP 1800-30 reference design.
- **C1. Network flow and DNS/HTTP logs** from clinical VLANs and egress points; data-loss prevention events.
- **D1. Clinical text for de-identification** (notes, messages), processed locally with modern PHI de-identification models before any downstream analytics.

  Collection follows **data minimization**: only fields needed for detection are retained; PHI fields are hashed/tokenized where feasible; retention policies enforce short lifetimes.

Model families and features.

- **Identity & access anomaly detection.** Train **sequence models** (e.g., GRUs/transformers) and **graph neural networks** over dynamic bipartite graphs (user↔resource) to flag deviations from role baselines and peer cohorts (e.g., nighttime mass chart access outside unit assignments). Use weak supervision (policy violations) and autoencoder reconstruction errors as unsupervised signals.
- **Endpoint/IoMT behavior.** Unsupervised clustering and density estimation over process trees and device communications to detect firmware downgrade attempts, new service beacons, or anomalous data bursts. Incorporate device capability baselines (8259A) and telehealth topology (SP 1800-30).
- **Network/exfiltration.** Flow-level models combining protocol metadata and content-free features to detect exfil patterns (long-duration low-rate flows, unusual destinations, encrypted upload surges).
- **Clinical text de-identification.** Deploy RNN/CRF or transformer-based de-identification models (Dernoncourt et al.) with conservative thresholds; retain only de-identified text for analytics.
- **Adversarial risk management.** For AI-enabled clinical tools (e.g., imaging triage), adopt robust training baselines, confidence-calibrated outputs, input-consistency checks, and **model cards** documenting threat models. Test with domain-specific adversarial examples per Finlayson et al. and strong attacks (e.g., CW) to ensure evaluation beyond gradient masking.

Privacy-preserving learning stack.

- **Federated learning (FL).** Partition by institution or business unit; exchange model updates, not raw PHI.
- **Secure aggregation.** Apply practical protocols so servers see only aggregated updates; tolerate client dropout.
- **Differential privacy.** Train with DP-SGD, setting ε budgets per use case (tighter for text PHI). Track cumulative privacy loss across rounds; adopt privacy amplification by subsampling.
- **Encrypted inference (where feasible).** For sensitive inference tasks (e.g., high-risk re-identification vectors), evaluate HE (CKKS) or trusted enclaves; accept latency tradeoffs for batch workflows rather than interactive ones.

System architecture and governance.

- **Zero Trust integration.** Ingest model outputs as **policy signals** to adapt access e.g., step-up authentication or session isolation when anomaly scores exceed calibrated thresholds, consistent with SP 800-207.

- **Model risk management.** Maintain versioned datasets, lineage, and approvals; document intended use, limitations, and monitoring plans; apply canarying and shadow-mode deployment before enforcement.
- **Red-team and privacy reviews.** Run periodic **adversarial ML** red-teams to probe evasion/counter-evasion; conduct **privacy reviews** assessing membership-inference risk and model inversion susceptibility (e.g., via confidence clipping and audit logs).
- **HIPAA & FDA alignment.** Ensure safeguards map to HIPAA Security Rule administrative/technical controls; for device-resident analytics, follow FDA postmarket/premarket cybersecurity guidance including update/patch processes, threat modeling, and SBOM expectations.

Evaluation plan and metrics.

- **Detection efficacy.** AUROC/PR-AUC on labeled incidents and realistic simulations; **time-to-detect** and **time-to-contain** relative to baselines; analyst **alert burden** (alerts per 1,000 users/day) and **true-positive yield**.
- **Privacy.** Formal ε, δ budgets; empirical resistance to membership inference under white-/black-box settings; audit for attribute leakage via inversion.
- **Robustness.** Attack success rates under CW/PGD and domain-specific perturbations; calibration error; detection of OOD inputs.
- **Operational fit.** False-positive review time, escalation rates to IR, alignment with SOC playbooks; **control mapping** coverage (e.g., AU/IR/AC families in SP 800-53).

Data sharing and sector learning. We propose a consortium-based FL deployment among regional hospital networks, *without* centralizing PHI. Updates are securely aggregated; each participant enforces local DP budgets tuned to risk tolerance. For telehealth RPM, apply the SP 1800-30 architecture, instrument devices per 8259A baselines, and coordinate incident response via CISA information-sharing channels.**Baseline visualizations and artifacts.** Figure 1 (OCR-reported breaches, selected years) contextualizes rising frequency; Figure 2 (IBM 2022 costs) highlights sector-specific impact. These motivate investments in privacy-preserving ML as *risk-reduction* infrastructure and inform cost-benefit analyses for executive sponsors.

4. Discussion

Why ML, and why now? The breach trend and cost asymmetry indicate insufficient observability and responsiveness. ML can improve *signal-to-noise* by modeling fine-grained sequences and relationships (users↔patients↔applications) that static rules miss. In access monitoring, for instance, nurses on a unit often share similar temporal and resource access patterns; deviations mass access to off-unit patients or sudden after-hours bursts are detectable via sequence and peer-group models with lower false positives than naïve per-user thresholds. Similar gains arise in IoMT behavior profiling and anomalous egress detection.**Privacy and compliance by design.** Healthcare cannot simply "collect everything" to build better models; HIPAA's minimum necessary standard and public trust demand restraint. FL, secure aggregation, and DP allow *learning from many without exposing any* a meaningful advance beyond central data lakes. Still, these techniques carry tradeoffs: DP introduces noise that can degrade utility, especially for minority patterns; FL can be attacked via poisoned updates; secure aggregation increases system complexity. Mitigations include per-client clipping/noise, by-zantine-robust aggregation, update attestation, and *privacy budgeting* with auditable accounting.**Adversarial ML in clinical contexts.** Finlayson et al. cautioned that medical incentives create real attack surfaces for adversarial examples (e.g., manipulating dermatology images to alter triage). Clinical-AI governance must therefore expand to explicitly consider adversarial risk: document threat models in model cards, require pre-submission robustness testing for device-embedded AI, and restrict overly confident outputs exposed to end users or APIs to reduce attack leverage and leakage. Robustness techniques (adversarial training, confidence calibration) should complement not replace clinical validation.**Zero Trust as the operational wrapper.** ML-derived risk signals are most valuable when they directly influence *access decisions*. In a Zero Trust architecture, signals can gate step-up authentication, session restrictions, and micro-segmentation. For example, an elevated anomaly score on an EHR session might (1) downgrade access to read-only, (2) re-verify identity with phishing-resistant MFA, and (3) dynamically restrict lateral movement. These actions can be codified as policy (SP 800-207), monitored via AU/SI controls (SP 800-53), and executed consistently across cloud and on-prem assets.**Telehealth and IoMT realities.** SP 1800-30 demonstrates that a secure remote patient monitoring solution is achievable with commercial components, but it depends on accurate asset

inventories, update mechanisms, and role-based access to the RPM platform. ML can enhance this baseline by profiling device communications and alerting on **behavioral drift**, such as an oximeter initiating outbound connections to unknown hosts. The 8259A core baseline provides a minimal capability set (device identity, secure update, data protection) that, when present, significantly improves ML observability and control.**Economics and national interest.** From a board perspective, the IBM 2022 cost figures transform cyber risk from an abstract compliance problem into a quantifiable drag on care delivery and capital planning. Investing in privacy-preserving detection that reduces average **time-to-detect** and **time-to-contain** feeds directly into cost avoidance especially for ransomware, where hour-scale containment windows determine whether elective procedures and ICU operations are disrupted. At a national level, sector resilience reduces cascading risk to other lifeline sectors and maintains public trust during crises (e.g., pandemics).**Interoperability and information sharing.** A common argument against cross-enterprise ML is patient privacy. FL and DP directly address this, enabling *algorithmic* information sharing without raw PHI exchange. Additionally, model *artifacts* feature schemas, risk-scoring APIs, and anonymized telemetry statistics are shareable through CISA/HHS channels. As more institutions implement SP 1800-30-like telemetry for telehealth, consistent schemas (e.g., device identity, firmware state) further facilitate cross-site learning.**Alignment with regulation and standards.** HIPAA's Security Rule demands administrative, physical, and technical safeguards; the proposed methodology maps to technical safeguards (access controls, audit controls, integrity, transmission security) and bolsters administrative safeguards (risk management, workforce training) through model-informed policies and escalations. For device-embedded analytics, FDA guidance emphasizes threat modeling, SBOM, updateability, and coordinated vulnerability disclosure prerequisites for sustainable ML-enabled devices.**Caveats and ethics.** ML detection is probabilistic. False positives can burden clinicians and degrade trust; false negatives may create complacency. We therefore advocate **socio-technical** design clear analyst playbooks, clinician-friendly explanations (e.g., which access attributes were unusual), and rigorous post-incident reviews that feed model updates. Fairness also matters: access-risk models must avoid proxying for role seniority or shift timing in ways that unfairly target specific staff groups. Differential privacy must not be misused to claim absolute anonymity; ε must be contextualized, and residual re-identification risk communicated transparently. Finally, offensive research (red-teaming) should be governed by IRB-like ethics controls to avoid patient harm.**A realistic path.** Many health systems already aggregate logs for compliance. The incremental path is to (1) standardize schemas and retention; (2) pilot anomaly detection on EHR/IAM with privacy-preserving pipelines; (3) integrate signals into policy engines; (4) expand to IoMT and network; and (5) participate in regional FL consortia. This staged approach delivers early wins (reduced inappropriate access) while building toward sector-scale learning.

Table 1. *Threat-to-Control Mapping (abbreviated).*

Threat vector	**Representative signals**	**ML control (privacy-preserving where feasible)**	**Standards alignment**
Inappropriate EHR access / credential misuse	Unusual chart access sequences; off-unit mass access; odd times/locations	Sequence/peer-group anomaly detection; risk-based access; federated training with secure aggregation + DP	HIPAA Security Rule (AC, AU), NIST SP 800-53 (AC, AU, IA), ZTA continuous verification
Ransomware staging / lateral movement	New SMB/RDP use; beaconing; privilege escalation	Unsupervised endpoint and flow-based models; micro-segmentation triggers	NIST SP 800-53 (SI, SC), ZTA policy enforcement; CISA ransomware trends guidance
Telehealth/IoMT compromise	Firmware downgrades; anomalous device comms	Device-behavior profiling; graph models across RPM topology	NIST SP 1800-30; NISTIR 8259A device baseline
Data exfiltration via SaaS/cloud	Long-duration encrypted uploads; rare destinations	Flow-sequence anomaly models; auto-isolation actions	NIST SP 800-207 policy signals; SP 800-53 SC-7/AC-4
Adversarial manipulation of clinical AI	Inconsistent inputs; high-confidence misclassifications	Robust training; input-consistency and confidence controls; red-teaming	FDA cybersecurity guidance; robust evaluation practices

***Table 2**. Measurement Plan.*

Objective	Metric(s)	Target / comment
Reduce attacker dwell time	Mean time-to-detect, mean time-to-contain	Downward trend vs. pre-deployment baseline
Maintain analyst workload	Alerts/1,000 users/day; true-positive yield	No net increase in total investigation time
Bound privacy risk	(ε, δ) budgets; empirical membership-inference AUC	ε within policy; AUC ≈ 0.5 under attack
Improve robustness	Attack success rate under CW/PGD; calibration error	↓ Attack success; better calibration on OOD
Standards alignment	Control coverage and mapping	Documented mapping to SP 800-53 / ZTA

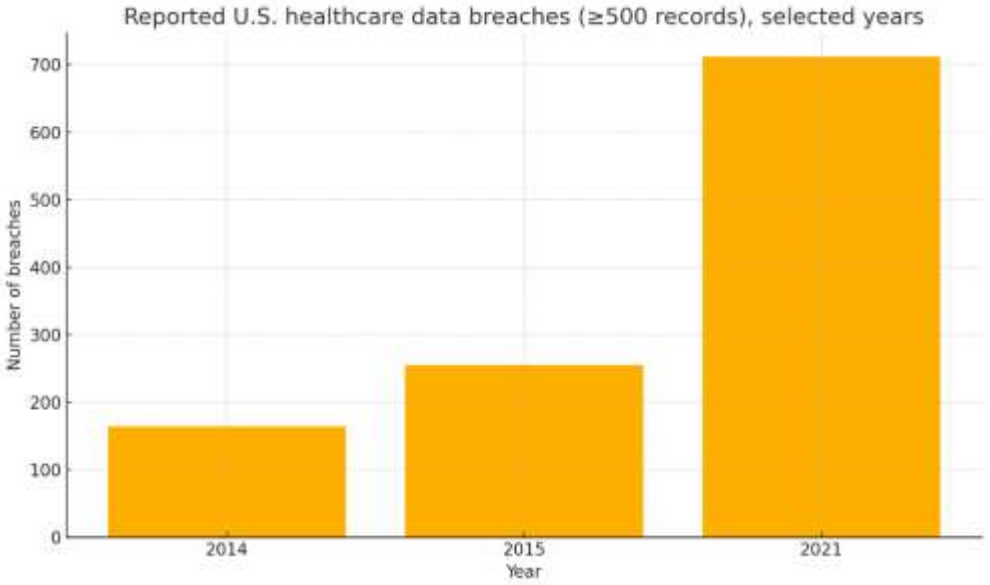

***Figure 1.** Reported U.S. healthcare data breaches (≥500 records), selected years (2014, 2015, 2021). Source: HIPAA Journal analyses.*

***Figure 2.** Average cost of a data breach (IBM 2022): Healthcare ($10.1M) vs. all industries ($4.35M).*

4. Conclusions

Patient-data protection is inseparable from national security when healthcare is a designated critical infrastructure. Breach trends and costs underscore the urgency, while policy and standards (PPD-21, NIST SP 800-53/207, SP 1800-30) provide the governance scaffold. Pre-July-2022 research demonstrates that ML can sharpen detection across identity, device, and network layers but only if engineered for **privacy** and **robustness** from the outset. Federated learning with secure aggregation, differential privacy, and encryption-assisted inference can enable collaborative learning without centralizing PHI; adversarial testing and Zero Trust integration ensure that models not only score risk but also **enforce** safer access. Our methodology translates these ideas into deployable pipelines, metrics, and governance. The goal is not perfect prevention; it is **resilience** shorter dwell times, faster containment, and protection of clinical continuity. By investing in secure ML as security infrastructure, U.S. healthcare organizations can reduce patient harm, meet regulatory obligations, and contribute to national preparedness against escalating cyber threats.

Limitations and Future Directions

This paper synthesizes pre-July-2022 evidence and proposes a deployment methodology, but it does not present prospective clinical trials of ML defenses. Real-world efficacy depends on local context: EHR/IAM logging fidelity, device inventory accuracy, and SOC processes vary widely. Privacy-preserving methods incur overhead and tradeoffs: DP can degrade minority-pattern detection; FL complicates debugging; secure aggregation and HE add latency and operational complexity. Adversarial robustness for clinical AI remains an active research area robust training can reduce accuracy or fail under adaptive attacks.Evaluation is another limitation. Label scarcity for true security incidents can bias results toward synthetic tests; cross-site generalization requires careful domain adaptation. Measurement should therefore combine retrospective incident labels, red-team exercises, and controlled simulations, with governance formulas that accept model uncertainty (e.g., using risk signals to require step-up authentication rather than outright blocking).Future work should: (1) develop **federated benchmarks** for healthcare security telemetry with standard schemas and privacy budgets; (2) explore **Byzantine-robust** and **attack-aware** FL aggregation to resist poisoned updates; (3) advance **privacy accounting** tools usable by non-specialists; (4) integrate **formal methods** for safety constraints in clinical AI; (5) evaluate **human-in-the-loop** interfaces that explain anomalies to clinicians and analysts to reduce alert fatigue; and (6) connect **economic models** (e.g., IBM-style cost drivers) to security-control ROI to support sustained investment. Finally, regulators and standards bodies could extend SP 1800-30-style practice guides to include **privacy-preserving analytics playbooks** and **adversarial evaluation protocols** for FDA-regulated AI devices providing concrete, testable expectations that vendors and providers can meet.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. https://doi.org/10.1145/2976749.2978318

[2] Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

[3] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy*, 39–57. https://nicholas.carlini.com/papers/2017_sp_nnrobustattacks.pdf

[4] Choi, E., Bahadori, M. T., Kulas, J., Schuetz, A., Stewart, W. F., & Sun, J. (2016). RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems*, 29. https://arxiv.org/abs/1608.05745

[5] Dernoncourt, F., Lee, J. Y., Uzuner, Ö., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *JAMIA*, 24(3), 596–606. https://papers.nips.cc/paper/6321-retain-an-interpretable-predictive-model-for-healthcare-using-reverse-time-attention-mechanism.pdf (RNN approach overview)

[6] Dwork, C. (2006). Differential privacy. *Proceedings of ICALP*, 1–12.

[7] Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. *Proceedings of the 22nd ACM CCS*, 1322–1333. https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf

[8] Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, 363(6433), 1287–1289. https://www.science.org/doi/pdf/10.1126/science.aaw4399

[9] Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., & Wernsing, J. (2016). CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy. *ICML 2016*. https://proceedings.mlr.press/v48/gilad-bachrach16.pdf

[10] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *ICLR*. https://arxiv.org/abs/1412.6572

[11] IBM Security. (2022, July 27). *IBM report: Consumers pay the price as data breach costs reach all-time high*. https://newsroom.ibm.com/2022-07-27-IBM-Report-Consumers-Pay-the-Price-as-Data-Breach-Costs-Reach-All-Time-High

[12] Kaissis, G., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2, 305–311. (Article PDF via publisher).

[13] Kaissis, G., Passerat-Palmbach, J., Ryffel, T., et al. (2021). Medical imaging deep learning with differential privacy. *Scientific Reports*, 11, 11326. https://www.nature.com/articles/s41598-021-93030-0.pdf

[14] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the EHR. *Scientific Reports*, 6, 26094. https://www.nature.com/articles/srep26094

[15] Neamatullah, I., Douglass, M. M., Lehman, L.-w. H., et al. (2008). Automated de-identification of free-text medical records. *JAMIA*, 15(5), 641–650. (Publisher site).

[16] NIST. (2020). *NIST SP 800-207: Zero Trust Architecture*. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-207.pdf

[17] NIST. (2020). *NIST SP 800-53 Rev. 5: Security and Privacy Controls for Information Systems and Organizations*. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf

[18] NIST. (2022). *SP 1800-30: Securing Telehealth Remote Patient Monitoring Ecosystem*. https://csrc.nist.gov/pubs/sp/1800/30/final

[19] NIST. (2020). *NISTIR 8259A: IoT Device Cybersecurity Capability Core Baseline*. https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8259A.pdf

[20] Presidential Policy Directive-21 (PPD-21). (2013). *Critical Infrastructure Security and Resilience*. CISA resource page.

[21] Rajkomar, A., Oren, E., Chen, K., et al. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1, 18. https://www.nature.com/articles/s41746-018-0029-1.pdf

[22] Rieke, N., Hancox, J., Li, W., et al. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3, 119. https://www.nature.com/articles/s41746-020-00323-1.pdf

[23] Sheller, M. J., Edwards, B., Reina, G. A., et al. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10, 12598. https://www.nature.com/articles/s41598-020-69250-1.pdf

[24] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *IEEE S&P*, 3–18. https://www.cs.cornell.edu/~shmat/shmat_oak17.pdf

[25] U.S. FDA. (2016). *Postmarket Management of Cybersecurity in Medical Devices* (Guidance). https://www.fda.gov/media/95862/download

[26] U.S. FDA. (2018). *Content of Premarket Submissions for Management of Cybersecurity in Medical Devices* (Draft Guidance). https://www.fda.gov/media/119933/download

[27] U.S. HHS (eCFR). (2022). *HIPAA Security Rule* (45 CFR §§164.306–164.316). https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-C

[28] CISA/FBI/HHS. (2022). *AA22-040A: 2021 Trends Show Increased Globalized Threat of Ransomware*. https://www.cisa.gov/sites/default/files/publications/AA22-040A_2021_Trends_Show_Increased_Globalized_Threat_of_Ransomware_508.pdf

[29] Bonawitz, K., Ivanov, V., Kreuter, B., et al. (2017). Practical secure aggregation for FL on user-held data. *arXiv:1611.04482*. https://arxiv.org/pdf/1611.04482

[30] Cheon, J. H., Kim, A., Kim, M., & Song, Y. (2017). Homomorphic encryption for arithmetic of approximate numbers (CKKS). *ASIACRYPT 2017*. https://iacr.org/archive/asiacrypt2017/106240294/106240294.pdf

[31] HIPAA Journal. (2022). *December 2021 Healthcare Data Breach Report* (712 breaches in 2021). https://www.hipaajournal.com/december-2021-healthcare-data-breach-report/