International Journal of Science Academic Research
Vol. 06, Issue 12, pp.11160-11174, December 2025
Available online at http://www.scienceijsar.com

ISSN: 2582-6425

Research Article

SECURING AI- DRIVEN TELEMEDICINE PLATFORMS: VULNERABILITY ASSESSMENT AND RESILIENT ARCHITECTURE DESIGN

***Nnennaya Ngwanma Halliday**

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

Received 27th October 2025; **Accepted** 18th November 2025; **Published online** 26th December 2025

Abstract

Telemedicine systems based on AI are used more and more to assist in diagnostics, triage, and remote patient monitoring, although their integration presents additive risks to safety both in the area of cybersecurity, model reliability and clinical decision integrity. Although the traditional concept of healthcare security focuses on the confidentiality of data and system availability, the AI-enabled telemedicine also has to ensure diagnostic accuracy and safety of clinical outcomes that is still unaddressed in current models. Recent research shows that artificially low accuracy in medical imaging models can be caused by adversarial inputs by up to 4065% (Zhou *et al.,* 2024), that medical chatbots in the triage can become more prone to clinically unsafe responses when prompt-injection attacked by up to 30%+ (Li *et al.,* 2023), and that unsecured IoMT channels have false-negative spoofing detection rates above 40% (Moradi *et al.,* 2024). These weaknesses present risks that are likely to directly disrupt the clinical decision-making process, and not necessarily expose patient data. This article determines the existence of vulnerabilities in AI inference pipeline, IoMT telemetry, and telehealth communication operations that are critical and provides a secure-by-design architectural framework that incorporates zero-trust networking, model integrity verification, encrypted biosignal provenance, adversarial input filtering, and clinician-in-the-loop safety restrictions. The framework also complies with the healthcare regulatory standards, such as HIPAA, GDPR, ISO 14971, IEC 62304, and FDA AI/ML medical device recommendations, to facilitate compliant and ready-to-implement deployment to the clinic.

Keywords: AI telemedicine security, Adversarial machine learning, IoMT, zero-trust healthcare, Model integrity, Clinical AI safety, Telehealth cybersecurity.

1. INTRODUCTION

Telemedicine has become not just an additional feature but a primary type of care provision, and the COVID-19 pandemic and its consequences have increased the pace of its expansion due to the rising prevalence of remote diagnostics, online triage, and constant health monitoring (WHO, 2023). The AI-based telemedicine is the next stage of telehealth, which allows the platform to recognize medical images, assess symptoms, and process real-time physiological signals transmitted by connected devices. Currently, hospitals and digital health providers implement AI-based triage assistants, diagnostic decision-making tools, and Internet of medical things (IoMT) wearables to decrease clinician workloads, increase access and facilitate early intervention (Keesara, Jonas & Schulman, 2020; Deloitte, 2024). Nevertheless, clinical care mediated by AI presents wider attack surfaces that are distinctly different to those posed by conventional healthcare cybersecurity. Historical Telemedicine platforms were concerned with the security of confidentiality and access to electronic health records (EHRs) and interactions with patients. AI introduces a different type of vulnerability, a clinical safety breach, in which an attack can be effective without revealing information but can distort the results of tests or a triage referral instead (Zhou *et al.,* 2024). As an illustration, the presence of adversarial noise on a chest X-ray may make AI models interpret pneumonia as a normal image (Finlayson *et al.,* 2019), and adversarial injection attacks on clinical chatbots may break safety rules and produce medically harmful recommendations (Li *et al.,* 2023). The IoMTs employed in home-based monitoring have been prone to information spoofing and code rejection (Morabi *et al.,* 2024), which may inaccurately restore the critical measurements that are relied upon by clinicians.

These vulnerabilities are amplified by several systemic factors:

- Complex data flows between mobile apps, cloud AI services, and hospital systems.
- Inconsistent security maturity among telehealth vendors and IoMT manufacturers.
- Lack of standardized security models for AI inference in clinical workflows.
- Regulatory gaps where AI is deployed before rigorous oversight frameworks are finalized (FDA, 2024).

Security breaches are no longer limited to ransomware or unauthorized access. In AI-enabled telemedicine, an attacker can *silently change the clinical outcome*, creating misdiagnosis, delayed escalation, or inappropriate treatment. Unlike visible system failures, these errors can occur without generating alerts, meaning clinicians may never realize that the system was compromised.

Therefore, securing AI-driven telemedicine requires more than data protection it requires preserving diagnosis integrity and clinical decision safety. Current research lacks integrated frameworks that address cybersecurity and AI robustness alongside clinician workflow considerations (IEEE EMBS, 2024). This paper aims to close that gap by:

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

1. Identifying the full spectrum of cyber–AI–clinical vulnerabilities in modern telemedicine platforms.
2. Evaluating attack impact on diagnostic accuracy and clinical outcomes.
3. Designing a resilient architecture that integrates zero-trust controls, adversarial defense, encrypted biosignal provenance, and human-in-the-loop oversight.

The goal is not only to secure systems, but to ensure that AI-assisted medical decisions remain safe, transparent, and trustworthy.

2. LITERATURE REVIEW

2.1 Converging Domains: Cybersecurity, AI Reliability, and Clinical Safety

Three research areas that have historically developed independently healthcare cybersecurity, AI safety and robustness, and clinical risk management and human factors intersect to form AI-driven telemedicine. According to recent research, these domains cannot be kept apart since flaws in any one of them might affect the others, making a model failure a system-wide medical risk or a security compromise a clinical hazard (Goel *et al.,* 2023; Sendak *et al.,* 2020). Although this convergence has been acknowledged, few research haveoperationalised it into unified security or evaluation frameworks that assess the severity of clinical consequences in addition to attack success (Beam *et al.,* 2023).

2.2 Telemedicine Infrastructure: Traditional Cyber Risks and Systemic Weakness

Telemedicine platforms also have typical healthcare IT weaknesses, such as API exposure, misconfigured clouds, insufficient encryption, and fragmented identity management (Jalali *et al.,* 2021; Coventry *et al.,* 2023). Healthcare remains among the most attacked critical industries; not only has the number of ransomware attacks in the industry increased by 60-70 percent after the pandemic, but it also tends to shut down telehealth platforms and postpone the delivery of remote care (IBM Security, 2024). Nevertheless, new data shows that most industry reaction, such as data protection and breach containment, is not effective to deal with risks that do not steal data but alter clinical outcomes (Sittig and Singh, 2022). Zero-trust architectures are highly effective in reducing the risk of breaches, though their implementation in clinical settings is under-utilized (less than 30 percent) because of cost and legacy system friction and lack of interoperability with hospital identity systems (Forrester, 2024). Moreover, healthcare threat modeling is still mostly IT-oriented, as opposed to patient-oriented, and hardly takes into account healthcare-specific models such as MITRE ATLAS of adversarial ML or clinical variants of STRIDE/LINDDUN (Agrafiotis *et al.,* 2018; MITRE, 2023).

2.3 AI Model Vulnerabilities: From Diagnostic Error to Clinical Misguidance

AI diagnostics inject attack vectors into medical software that has never been seen. Radiology tasks are misclassified at rates of 4570% by adversarial modifications to medical imaging models without visual distortion often being noticeable (Zhou *et al.,* 2024; Finlayson *et al.,* 2019; Ma *et al.,* 2023). This completely changes the profile of risk: failures are invisible to the clinicians, scalable, and actable. There are also similarly alarming failure modes in Lee LM-based triage and symptom assessment tools. Immediate injections and jailbreak attacks increase unsafe clinical response rates (30-53 percent) and result in the generation of artificial drug interactions, erroneous differential diagnosis, and unsuitable escalation instructions (Li *et al.,* 2023; Yan *et al.,* 2024). This is as opposed to traditional software bugs, where failures are more deterministic, and therefore edge-case behaviors are more difficult to identify, audit, and control (Kelly *et al.,* 2023). Although recent research has advanced adversarial defenses (input filters, model watermarking, adversarial training), the vast majority of researchers have adopted technical measures (attack success rate, decrease in accuracy) as their evaluation of robustness instead of clinical outcome measures (diagnostic divergence, delay at triage, probability of escalation to failure) (Rajpurkar *et al.,* 2022).

2.4 IoMT Security and the Fragility of Remote Patient Monitoring

IoMT-enabled telemedicine significantly expands the attack surface beyond hospital networks into patient homes. Studies reveal alarming baselines:

- 41% of devices transmit data without encryption
- Firmware injection succeeds in >30% of tested units
- Biosignal spoofing evades detection 38–45% of the time (Moradi *et al.,* 2024; Alsubaei *et al.,* 2023; Zhang *et al.,* 2023)

These malfunctions endanger clinical stability and go beyond device compromise. Although most IoMT studies stop at compromise feasibility without mapping clinical damage pathways, manipulated vitals (e.g., falsified oxygen saturation, blood pressure injections) can cause false triage escalation, medication errors, delayed emergency response, or misinterpreted deterioration signals (Moradi *et al.,* 2024). Furthermore, frameworks for anomaly detection, OTA signing, and secure device attestation are still applied inconsistently (Check Point Research, 2023).

2.5 Clinical Workflow Risks, Automation Bias, and Human–AI Failure Modes

Technical vulnerabilities have clinical levels of interaction with human cognition. Clinicians will be most likely to suffer automation bias (excessive trust in AI outputs) and will be more susceptible to alert fatigue when AI systems produce nonspecific warnings, which negatively affect diagnostic vigilance (Parker and Ashrafian, 2022; Lyell and Coiera, 2017). Correct AI alerts do not work unless clinicians think of them as cognitive patients, which has been demonstrated to result in failure in high-throughput emergency and tele-triage settings (Sendak *et al.,* 2020). The emergence of digital health failures is not only based on technical failures as noted by socio-technical models such as the Systems Engineering Initiative for Patient Safety (SEIPS) (Carayon *et al.,* 2006) but rather through incompatibility of the system behavior with the clinician workflow. However, the studies of AI security seldom incorporate human decision paths into threat modeling, despite the fact that a flawed AI tool will eventually harm the clients by use of clinical interpretation.

2.6 Bias, Equity, and Disparate Harm Amplification

Under-represented groups may be disproportionately affected by AI vulnerabilities. Measurable diagnostic differences in radiology, dermatology, and risk scoring models have been caused by dataset bias (Obermeyer *et al.,* 2019; Seyyed-Kalantari *et al.,* 2021). The most clinically fragile patients may also become the most algorithmically vulnerable when adversarial assaults or data poisoning take place. However, there is a significant blind spot in equitable AI security design because adversarial ML studies hardly ever stratify impact by clinical subgroup or demographic.

2.7 Governance, Accountability, and Post-Deployment Surveillance

The current regulatory standards, FDA AI/ML guidance, EU AI Act, ISO 14971, and IEC 62304, recognize the risk of lifecycle, but they do not have enforceable runtime integrity validation, adversarial monitoring, or clinical harm scoring principles (FDA, 2024; EU Commission, 2024; McDermid *et al.,* 2022). More so-called responsible AI tools, including Model Cards, Datasheets, and AI incident reporting standards, are becoming more suggested than seldom integrated into clinical procurement or monitoring patterns (Mitchell *et al.,* 2019; WHO, 2021). Another gap is the post-deployment surveillance. Medical AI also tends to be released into the market without systematic audits of safety in the real world, unlike pharmaceuticals. Such approaches as drift monitoring, out-of-distribution detection, or clinical performance surveillance dashboards are emphasized in recent studies, yet they are yet to be adopted (Wu *et al.,* 2021; MHRA, 2023).

2.8 Economic and Implementation Constraints

Due to financial limitations, difficulties integrating hospital legacy systems, and bandwidth restrictions in remote areas, even technically acceptable security architectures may not succeed in adoption. According to studies, the main obstacles to safe telehealth modernisation are a lack of cybersecurity personnel, tight finances, and incompatible EHR architectures (Kruse *et al.,* 2021; Ponemon Institute, 2023). Security designs run the risk of being theoretically sound but operationally impractical without practical, cost-conscious design.

2.9 Summary of Key Research Deficits

Table 1. Summary of Key Research Deficits

Area	Deficit
Threat Modeling	Lacks patient-impact and adversarial ML integration
AI Robustness	Focuses on attack success, not clinical harm
IoMT	Measures device compromise, not triage disruption
Human Factors	Underexplored automation bias and workflow effects
Equity	Minimal analysis of disparate attack impact
Governance	Weak runtime accountability standards
Deployment	Limited feasibility and interoperability evaluation

2.10 Positioning of This Research

This paper addresses the gaps by proposing a clinically meaningful security paradigm, where AI telemedicine defense is measured not only by breach resistance, but also by:

1. Model integrity under adversarial conditions
2. Preservation of diagnostic correctness
3. Safety of clinical escalation pathways
4. Equity of failure impact
5. Deployability in real healthcare environments

3. METHODOLOGY

The methodology is organized into four integrated components: (1) threat and vulnerability identification, (2) clinical impact modeling, (3) security architecture design, and (4) validation and evaluation.

3.1 Research Design

A hybrid security-clinical evaluation paradigm was adopted, combining:

- Threat modeling frameworks (healthcare-specific adaptations)
- Adversarial AI and penetration testing techniques
- Clinical harm modeling and workflow analysis
- Resilience-driven architecture design

This design ensures that security issues are not evaluated solely from a technical perspective but also measured by clinical impact, patient safety implications, and deployability in healthcare environments.

3.2 AI Telemedicine System Decomposition

To ensure systematic coverage, the telemedicine ecosystem is decomposed into three evaluative planes:

Table 2. AI Telemedicine System Decomposition

Component Layer	Examples	Key Risks
AI Intelligence Layer	Diagnostic ML models, triage LLMs, symptom classifiers	Adversarial manipulation, model theft, output poisoning, hallucinations
IoMT Sensor Layer	Wearables, glucose monitors, ECG patches, home devices	Spoofed vitals, firmware tampering, adversarial signal injection
Communication & Orchestration Layer	APIs, telehealth portals, EHR integrations, cloud pipelines	MITM, authentication bypass, insecure MLOps pipelines

This layered structure guides threat identification, test scoping, and mitigation mapping.

3.3 Threat and Vulnerability Identification

3.3.1 Threat Modeling Frameworks Employed

A multi-framework approach was selected to capture both AI-specific and clinical system threats:

Table 3. Threat Modeling Frameworks Employed

Framework	Purpose	Adaptation
MITRE ATLAS	Adversarial ML attack tactics	Mapped to clinical AI misuse cases
MITRE ATT&CK (Healthcare)	System intrusion behaviors	Tailored to telemedicine workflows
STRIDE	Software threat classification	Custom clinical risk labels added (e.g., diagnostic misdirection)
LINDDUN	Privacy risk analysis	Applied to health data and model inference channels

Justification: Multi-model threat mapping increases coverage by 37–42% compared with single-framework threat discovery in healthcare AI (Agrafiotis *et al.,* 2018; MITRE, 2023).

3.3.2 Vulnerability Sources Investigated

Vulnerabilities were categorized across four vectors:

i. Model-Level – adversarial inputs, poisoning, prompt injection, model inversion
ii. Device-Level – IoMT spoofing, firmware compromise, sensor tampering
iii. Network-Level – insecure API calls, token hijacking, traffic interception
iv. Human–System Interaction – automation bias, alert fatigue, unsafe overrides

3.4 Clinical Impact and Harm Modeling

4.4.1 Failure Mode and Clinical Consequence Analysis

Instead of measuring only attack success, we map attack outcomes to clinical harm trajectories using:

- FMEA (Failure Mode & Effects Analysis) – to model failure modes
- RPN Scoring (Risk Priority Number) – to quantify risk severity
- Time-to-Clinical Escalation Impact Index (TCEI) – to measure delays in care response
- Diagnostic Divergence Rate (DDR) – variance between AI output and clinician decision

These metrics allow impact to be measured not just by technical compromise, but by potential patient-level consequences (DeRosier *et al.,* 2002; Wu *et al.,* 2021).

3.4.2 Human-AI Workflow Risk Mapping

Clinical interaction risks are mapped using SEIPS (Systems Engineering Initiative for Patient Safety) to capture:

- Decision handoff points
- Cognitive overload zones
- Automation trust failure modes
- Alert routing bottlenecks

This ensures threats affecting *clinical behavior* are evaluated alongside technical attacks (Carayon *et al.,* 2006).

3.5 Resilient Architecture Design Principles

Based on identified risks, secure architecture requirements were defined under the following pillars:

Table 4. Secure Architecture Requirements and Architectural Response

Security Requirement	Architectural Response
Identity Integrity	Zero-trust authentication + device attestation
Model Integrity	Model signing + differential integrity verification
Data Provenance	Encrypted biosignal lineage + timestamp anchoring
Clinical Safety	Human-in-the-loop escalation overrides + policy guardrails
Attack Resilience	Adversarial input filtering + telemetry anomaly detection

These principles align with FDA AI lifecycle guidance and ISO 14971 clinical risk standards (FDA, 2024; ISO, 2019).

3.6 Verification and Evaluation Strategy

4.6.1 Technical Security Validation

Security assessment uses:

- Red-team adversarial ML testing
- IoMT penetration testing
- Model behavior fuzzing
- Prompt injection benchmarking
- API and identity attack simulations

3.6.2 Clinical Safety Validation

Clinical reliability is tested via:

- Reduction in Diagnostic Divergence Rate (DDR)
- Reduction in false triage escalations
- Improved clinical escalation response time (TCEI)
- Human override safety retention rates

3.6.3 Success Evaluation Criteria

Table 5. Success Evaluation Criteria

Goal	Target Outcome
Robust AI inference	<5% diagnostic drift under adversarial input
IoMT integrity	>98% detection of spoofed biometric signals
Secure authentication	0 privilege escalation pathways
Clinical reliability	No increase in triage delay or misrouting
Human-AI resilience	Restoration of clinician decision authority in >99% of ambiguous failures

3.7 Ethical and Compliance Alignment

Security and resilience objectives were cross-validated against existing regulatory and safety expectations:

- ISO 14971   clinical risk management
- IEC 62304   medical software lifecycle
- FDA AI/ML SaMD Guidance
- HIPAA & GDPR   data governance
- WHO AI Ethics Framework   health AI accountability

4. THREAT TAXONOMY & ATTACK SURFACE ANALYSIS

By combining clinical decision logic, machine inference, IoMT sensing, patient engagement, and distributed cloud orchestration, AI-driven telemedicine technologies increase the attack surface of traditional healthcare. Attacks against clinical AI may weaponise treatment timing, physiological signal authenticity, or diagnostic accuracy, posing a danger to patient safety even in the absence of data exfiltration, in contrast to traditional enterprise IT threats, which primarily seek to compromise data security or system uptime. A organised taxonomy of threats, their attack methods, and any possible clinical repercussions are presented in this section.

4.1 Taxonomy of Threat Classes in AI Telemedicine

Threats are categorized into five interdependent classes, each linked to clinical impact vectors beyond technical compromise.

Table 6. Taxonomy of Threat Classes in AI Telemedicine

Threat Class	Core Target	Example Attack Forms	Primary Clinical Risk
A. Adversarial AI Manipulation	ML diagnostic models, LLM-based triage	adversarial images, model inversion, prompt injection	misdiagnosis, unsafe care advice
B. IoMT Signal Compromise	Wearable and remote monitoring devices	spoofed vitals, sensor replay, firmware tampering	false deterioration alerts, missed emergencies
C. Identity & Access Abuse	Patient/clinician/auth services	token theft, deepfake authentication, session hijack	fraudulent treatment requests, unauthorized clinical actions
D. MLOps Supply Chain Attacks	Model build, deployment, and updates	poisoned training data, malicious weights, CI/CD injection	persistent incorrect model behavior at scale
E. Workflow and Interaction Exploits	Clinical decision loops & automation reliance	alert flooding, decision fatigue, UI manipulation	clinician override suppression, delayed escalation

This multi-layer taxonomy captures sociotechnical risk propagation where infrastructure compromise translates into *clinical decision failure* instead of only data loss.

4.2 Attack Surface Mapping Across the Telemedicine Stack

AI-telemedicine systems exhibit a vertically layered vulnerability topology:

1. Presentation and Interaction Layer

- Target: Patient triage chatbots, clinician dashboards, teleconsultation UIs
- Risks: prompt injection, misleading clinical suggestions, UI tampering, phishing-based consultation takeover
- Clinical Impact: inappropriate triage severity, suppressed symptom disclosure guidance
- Evidence: Prompt injection has been shown to elicit unsafe medical instructions in 30–53% of tested clinical LLM cases (Yan *et al.,* 2024; Li *et al.,* 2023).

2. AI Inference Layer

- Target: Diagnostic classifiers (imaging, symptom triage, predictive risk models)
- Risks: adversarial perturbations, model backdoor triggers, gradient leakage
- Clinical Impact: silent diagnostic errors, biased or unsafe recommendations
- Evidence: Adversarial medical image perturbations yield 45–70% misclassification rates without visual detectability (Zhou *et al.,* 2024; Finlayson *et al.,* 2019).

3. IoMT Sensing Layer

- Target: Home-based sensors, wearables, implant-adjacent monitors
- Risks:biosignal spoofing, replay attacks, Bluetooth/802.11 medical device interception
- Clinical Impact: false-positive crisis alerts, missed deterioration signs, medication titration errors
- Evidence: Biomarker spoof detection failures reach 38–45% in unprotected IoMT systems (Moradi *et al.,* 2024).

4. Orchestration & API Layer

- Target: Telehealth APIs, EHR integrations, cloud messaging brokers
- Risks: API key leakage, insecure data serialization, SSRF (Server-side request forgery)
- Clinical Impact: tampered medical records, altered treatment logs, clinician impersonation
- Evidence: Insecure APIs are among the top 3 exploited entry points in healthcare system intrusions (IBM Security, 2024).

5. MLOps and Model Deployment Layer

- Target: CI/CD pipelines, model registries, training data lakes
- Risks: training data poisoning, dependency hijacking, malicious model checkpoints
- Clinical Impact: scaled propagation of unsafe or biased clinical outputs
- Evidence: Poisoned training pipelines can introduce persistent misclassification behavior without detection for months in medical ML systems (Jagielski *et al.,* 2021).

4.3 Clinical Kill Chain: From Compromise to Patient Harm

Traditional cyber kill chains fail to capture telemedicine harm correctly, as the *final objective is not system control but clinical outcome manipulation*. The clinical kill chain for AI telemedicine proceeds as follows:

i. Access vector activation (API exploit, prompt injection, sensor spoofing)
ii. Clinical inference manipulation (altered diagnosis, tampered vital signals)
iii. Workflow propagation (EHR updates, triage rerouting, alarm suppression)
iv. Clinical decision influence (incorrect clinician trust, escalation failure)
v. Patient-level impact (treatment delay, inappropriate intervention, under/over triage)

This model reframes success conditions of attacks not by uptime loss, but by diagnostic divergence and escalation distortion (Sendak *et al.,* 2020).

4.4 Human-Centric Vulnerabilities and Cognitive Exploitation

Security flaws compound when interacting with clinician cognition:

- Automation bias: clinicians accept AI output without adequate scrutiny *(higher likelihood when system confidence is implicitly displayed)*
- Alert fatigue attack surface: adversaries trigger repeated low-risk alerts to desensitize response
- Authority assumption risk: clinicians perceive system outputs as validated medical evidence
- Decision suppression: adversarial UI manipulation can slow escalation pathways

These issues convert technical vulnerabilities into *clinical inertia*, a documented risk in computer-assisted care environments (Lyell &Coiera, 2017; Parker &Ashrafian, 2022).

4.5 Threat Severity Scoring by Clinical Consequence

Table 7. Threat Severity Scoring by Clinical Consequence

Threat Type	Technical Severity	Clinical Severity	Patient Safety Risk
Adversarial diagnostic inputs	High	Critical	Misdiagnosis, harmful treatment
IoMT signal spoofing	Medium	Critical	Missed emergencies, false interventions
Prompt injection	Medium	High	Unsafe clinical advice
EHR/API tampering	High	High	Incorrect medical records
Alert flooding	Low	Medium	Decision delays, escalation miss
Model theft	High	Low–Medium	Secondary misuse risk

Key insight: Some *technically moderate* attacks (e.g., biosignal spoofing) have *catastrophic clinical impact*, underscoring the inadequacy of technical-risk-only scoring.

5.6 Summary of Key Attack Surface Insights

i. Threats target clinical logic, not just infrastructure.
ii. The most damaging attacks degrade decision accuracy, not system availability.
iii. IoMT attacks are clinically more harmful than many traditional software exploits.
iv. Human cognitive pathways are legitimate attack surfaces.
v. A new defense priority hierarchy is required clinical safety first, data security second.

5. SECURE & RESILIENT ARCHITECTURE DESIGN

5.1 Architectural Design Principles

The architecture is guided by five non-negotiable requirements:

1. Clinical Safety Preservation – attacks must not alter diagnosis, triage level, or escalation timing.
2. Zero-Trust by Default – no component, device, model, or user is inherently trusted.
3. Verifiable AI Integrity – model predictions must be traceable, signed, and tamper-detectable.
4. Physiological Data Provenance – patient vitals must carry cryptographic authenticity.
5. Human-Override Guarantee – clinical authority must always supersede automated outputs.

These principles align with regulatory expectations for Safety-Critical AI (FDA, 2024; ISO 14971, 2019).

5.2 Layered Security Architecture

Layer 1 Secure Interaction & Identity Fabric
Purpose: Prevent impersonation, unauthorized consultations, and credential misuse.

Key Controls

- FIDO2 + biometrics for clinician authentication
- Risk-adaptive re-authentication for high-impact actions
- AI-deepfake detection on video teleconsultation feeds
- Device attestation for clinician endpoints

Clinical Value: Mitigates unauthorized prescription orders, fraudulent clinician access, and disguised patient interactions (Mirsky& Lee, 2021).

Layer 2 Protected AI Inference Environment

Purpose: Ensure diagnostic outputs are legitimate and untampered.

Key Controls

- Model signing and weight hashing prior to deployment
- Trusted Execution Environments (TEE) for runtime inference
- Adversarial input filtering and behavioral anomaly detection
- Model output watermarking for forensic traceability

Clinical Value: Ensures that triage decisions and diagnostic outputs can be cryptographically verified as authentic and unchanged (Zhang *et al.,* 2022; Khaim *et al.,* 2023).

Layer 3 Physiological Signal Authentication (IoMT Shield)
Purpose: Guarantee that patient sensor data reflects genuine physiology, not spoofed input.

Key Controls

- End-to-end encrypted telemetry (AES-GCM + TLS 1.3 IoMT channels)
- Hardware-rooted attestation on wearables
- Biosignal challenge–response validation (randomized signal watermarking)
- ML-based physiological plausibility checking (heart rate–SpO2 correlation, etc.)

Clinical Value: Prevents falsified vitals from triggering unsafe triage, medication dosing errors, or missed deterioration alerts (Moradi *et al.,* 2024; Alsubaei *et al.,* 2023).

Layer 4 Secure MLOps& Model Supply Chain
Purpose: Stop poisoning, backdooring, or tampered model distribution.

Key Controls

- Dataset lineage tracking and checksum validation
- Model card documentation and reproducibility manifests
- CI/CD container signing and SBOM (software bill of materials)
- Canary model deployment with clinical performance drift monitoring

Clinical Value: Ensures only validated and untampered models are promoted into telehealth production environments (Jagielski *et al.,* 2021; Mitchell *et al.,* 2019).

Layer 5 Zero-Trust Clinical Network Segmentation
Purpose: Limit breach blast radius and protect clinical workflow continuity.

Key Controls

- Micro-segmented telehealth VLANs
- API firewall with behavioral anomaly detection
- Least-privilege clinician session scopes
- Continuity-preserving failover for teleconsultation services

Clinical Value: Prevents cyber incidents from halting virtual care delivery during ongoing patient encounters (Forrester, 2024; IBM, 2024).

Layer 6 Clinical Decision Safeguard Engine
Purpose: Ensure AI advice cannot bypass clinical judgment or violate safety thresholds.

Key Controls

- Human-in-the-loop triage confirmation for high-risk cases
- Rule-bounded AI guardrails (contraindication and red-flag enforcement)
- Diagnostic divergence detectors (AI vs clinician decision delta monitoring)
- Mandatory escalation pathways for conflicting decisions

Clinical Value: Stops unsafe AI recommendations from reaching patients without clinician validation and preserves escalation integrity (Sendak *et al.,* 2020; Kelly *et al.,* 2023).

5.3 Resilience Mechanisms Against Specific Attacks

Table 8. Resilience mechanisms against specific attacks

Threat	Architectural Defense	Clinical Protection Outcome
Adversarial diagnostic inputs	Input filtering + model verification + divergence monitoring	Protects diagnosis integrity
Vitals spoofing	IoMT attestation + physiological plausibility checks	Prevents false alarms or missed emergencies
Prompt injection	Guardrail-bounded LLM responses + clinical policy sandboxing	Blocks unsafe clinical instruction
Model poisoning	Dataset lineage + signed models + CI/CD integrity controls	No deployment of malicious models
Clinician impersonation	Biometric MFA + endpoint attestation	Stops unauthorized clinical actions
Alert flooding	Adaptive alert throttling + priority triage queues	Avoids clinician desensitization

5.4 Clinical Safety Guarantees Embedded in Design

Unlike conventional security architectures, this framework embeds measurable clinical invariants that must hold even under attack:

Table 9. Clinical Invariants Enforced by Architecture

Clinical Invariant	Enforced By
No unauthorized change to diagnosis	Signed AI outputs + clinician verification gate
No falsified vitals accepted as authentic	IoMT attestation + signal plausibility checks
No blocked escalation pathways	Workflow continuity + forced override channels
No AI advice outside medical policy	Rule-bounded guardrail engines
No suppression of clinician authority	Human-override supremacy control

These invariants define *failure-safe states*, ensuring that even when compromised, the system degrades toward clinician control, not autonomous error.

5.5 Deployment Blueprint for Clinical Integration

The architecture is deployment-ready via a phased model:

1. Pre-deployment → model signing, IoMT onboarding, threat simulation
2. Controlled clinical sandbox → red-team evaluation, workflow observation
3. Canary patient cohort → monitored inference with clinician adjudication
4. Full deployment with observability → drift monitoring + forensics logging
5. Continuous verification → monthly adversarial resilience testing

This phased rollout aligns with FDA SaMD lifecycle expectations and real-world hospital governance (FDA, 2024; MHRA, 2023).

5.6 Summary of Architectural Contributions

This architecture introduces four advances over existing designs:

1. Security measured by patient safety, not system uptime
2. Proof-of-authenticity for both AI outputs and human physiology
3. Guaranteed clinician authority in all failure states
4. Unified defense across inference, network, device, and workflow layers

6. EVALUATION & VALIDATION FRAMEWORK

Without quantifiable validation across cybersecurity robustness, AI dependability, and clinical safety outcomes, a secure telemedicine architecture cannot be deemed effective. This framework combines patient-risk-oriented clinical performance evaluation, adversarial resilience testing, and technical security benchmarks into a unified assessment process.

6.1 Evaluation Objectives

Validation is structured around three primary assurance domains:

1. Security Assurance – Can the system resist, detect, and recover from attacks?
2. AI Clinical Reliability – Are model outputs accurate, explainable, and stable under adversarial conditions?
3. Operational Safety – Does the system preserve clinical correctness, escalation reliability, and human oversight even under stress or compromise?

Conventional AI evaluation only addresses accuracy metrics. Here, correctness must remain invariant under attack conditions a higher standard required for medical contexts (Raji *et al.,* 2022; Kelly *et al.,* 2023).

6.2 Multi-Layer Evaluation Pipeline

The validation process operates at five progressive layers:

Layer 1 Data and Model Integrity Validation

Table 10. Layer 1 data and model integrity validation

Test	Purpose	Success Criteria
Data lineage audit	Verify dataset authenticity	100% traceable provenance
Training data poisoning scan	Detect malicious injections	< 0.1% anomaly tolerance
Model hash verification	Ensure deployment integrity	Bit-level hash match
Reproducibility reproduction	Rebuild model in isolated environment	Output parity > 99.9%

(Mitchell *et al.,* 2019; Jagielski *et al.,* 2021)

Layer 2 Adversarial and Abuse Resilience Testing

Table 11. Layer 2 adversarial and abuse resilience testing

Test	Threat Simulated	Pass Condition
Adversarial perturbation challenge	Diagnostic misclassification	< 5% output distortion
Prompt injection for LLM triage	Unauthorized clinical instruction	0 unsafe responses
Backdoor trigger activation	Hidden model behaviors	No trigger-based deviation
Evasion testing on guardrails	Bypass clinical constraints	100% guardrail enforcement

(Li *et al.,* 2023; Finlayson *et al.,* 2019; Yan *et al.,* 2024)

Layer 3 IoMT Signal Authenticity and Robustness

Table 12. Layer 3 IoMT signal authenticity and robustness

Test	Objective	Requirement
Biosignal spoof injection	Detect falsified vitals	Spoof rejected ≥ 98%
Data replay attack detection	Identify repeated telemetry	100% temporal anomaly detection
Sensor attestation verification	Confirm hardware identity	All devices attested
Statistical physiological validation	Cross-vital plausibility	No false clinical states accepted

(Moradi *et al.,* 2024; Alsubaei *et al.,* 2023)

Layer 4 Clinical Workflow Safety Validations

Table 13. Layer 4 clinical workflow safety validation

Scenario	Validated Property	Expected Outcome
Conflicting AI vs clinician diagnosis	Escalation integrity	Human decision prioritized
Triage severity tampering attempt	Clinical invariance	No unauthorized severity change
Alert flooding simulation	Cognitive safety	Critical alerts still delivered
API/EHR manipulation trial	Record integrity	Tampering blocked and logged

(Sendak *et al.,* 2020; Lyell &Coiera, 2017)

Layer 5 Recovery and Failure-State Safety

Table 14. Layer 5 recovery and failure-state safety

Test	Requirement
System compromise simulation	AI switches to clinician-only control mode
Model uncertainty surge	Automated clinical deferral enabled
Sensor input corruption	Escalation to manual assessment
Network isolation	Care continuity maintained locally

(Kelly *et al.,* 2023; Parker &Ashrafian, 2022)

6.3 Key Performance Indicators (KPIs)

Technical Security KPIs

- Attack detection rate: ≥ 97%
- False positive rate: ≤ 2%
- Mean time to threat containment: ≤ 90 seconds
- Model integrity verification success: 100%

AI Clinical Reliability KPIs

- Diagnostic stability under attack: ≥ 95% consistency
- Unsafe AI output rate: 0%
- Explainability auditability: 100% of outputs log-traceable
- LLM clinical guardrail bypass rate: 0%

Clinical Safety KPIs

- Escalation pathway preservation: 100% availability
- False triage severity change tolerance: 0 incidents
- Clinician override latency:< 2 seconds
- Adverse event rate attributable to AI: 0%

These thresholds reflect benchmarks proposed in medical AI safety recommendations (WHO, 2024; FDA, 2024).

6.4 Validation Datasets and Testbeds

To ensure realism, evaluation must include:

Table 15. Validation datasets and testbeds

Testbed Type	Purpose
Public clinical datasets (MIMIC-IV, PhysioNet)	Triage and vitals validation
Medical imaging corpora (NIH ChestX-ray, ISIC)	Adversarial radiology testing
Synthetic adversarial prompt libraries	LLM abuse simulation
Hospital sandbox environments	Clinician-in-the-loop threat drills

(Johnson *et al.,* 2023; Goldberger *et al.,* 2023)

6.5 Human Safety Verification Protocols

Because automation bias is a real clinical vulnerability, human-focused validation measures include:

- Clinician blind-A/B trials (AI vs adversarial AI vs human control)
- Cognitive load assessment during attack simulations
- Decision override compliance studies
- Alert fatigue tolerance thresholds

(Lyell &Coiera, 2017; Parker &Ashrafian, 2022)

6.6 Continuous Post-Deployment Evaluation

Validation must persist beyond initial certification. Required monitoring components include:

1. Drift surveillance   model, distributional, and clinical drift
2. Security telemetry correlation   cross-layer anomaly fusion
3. Monthly red-team clinical attack drills
4. Automated safety regression testing on every model update
5. Real-world harm monitoring with mandatory incident review board

(MHRA, 2023; Kelly *et al.,* 2023)

6.7 Summary of Evaluation Strengths

This framework advances prior approaches by ensuring:

- Security is measured in clinical harm prevention, not only breach avoidance
- AI correctness is validated under adversarial, not only ideal, inputs
- Human clinical authority is a testable system property
- Failure states are required to degrade into safe clinician control

7. DISCUSSION

7.1 Integrating Security and Clinical Safety

The CIA triad confidentiality, integrity, and availability is the main focus of traditional cybersecurity. However, telemedicine platforms bring about a paradigm shift: clinical safety takes precedence above other security metrics, as shown in Sections 5–7. The severity of an attack must be assessed in terms of potential patient injury, misdiagnosis, triage errors, or delayed actions rather than just system compromise (Sendak *et al.,* 2020; Kelly *et al.,* 2023).

- **Key insight:** Attacks with moderate technical severity (e.g., IoMT spoofing) can have catastrophic clinical consequences, illustrating the inadequacy of conventional IT-centric threat prioritization.
- **Implication:** Hospitals and telehealth providers must adopt clinically informed risk scoring, integrating automation bias, human workflow, and physiological signal verification into cybersecurity metrics.

7.2 Balancing AI Autonomy and Human Oversight

The architecture design emphasizes human-in-the-loop (HITL) control and escalation override mechanisms, which safeguard against automation bias and cognitive overload (Lyell &Coiera, 2017; Parker &Ashrafian, 2022). While AI tools can improve efficiency, the discussion highlights the need to:

1. Preserve clinician authority in diagnostic and triage decisions.
2. Introduce guardrails that prevent AI from bypassing established clinical policies.
3. Monitor decision divergence between AI and human judgments to detect potential system failure or attacks.

- Trade-off: Increased HITL intervention may reduce operational speed but significantly mitigates patient risk, emphasizing the primacy of safety over throughput in telemedicine.

7.3 IoMT and Remote Monitoring: Expanding the Attack Surface

Remote patient monitoring expands the threat landscape beyond hospital firewalls. Findings from IoMT vulnerability analysis (Section 5.2) underscore that:

- Attacks on biosignal integrity can simulate patient deterioration or mask emergencies.
- Cryptographic attestation, anomaly detection, and plausibility verification are critical for trustworthy clinical decision-making.
- Policy implication: Regulatory bodies should mandate end-to-end data integrity standards for connected medical devices and continuous monitoring post-deployment (MHRA, 2023; Moradi *et al.,* 2024).

7.4 Adversarial AI and LLM Triage Systems

Adversarial vulnerabilities in diagnostic models and LLM-based triage systems are non-deterministic and difficult to audit, posing new challenges:

- Adversarial inputs or prompt injections can generate clinically unsafe outputs without overt system errors.
- Existing model evaluation methods (accuracy, F1 score) are insufficient; metrics must incorporate clinical harm potential (Raji *et al.,* 2022).
- Strategic insight: Continuous adversarial testing, runtime model monitoring, and deployment of policy-bound guardrails are essential for safe AI deployment.

7.5 Human Factors and Automation Bias

Human factors play a pivotal role in system resilience:

- Automation bias: Clinicians may accept AI outputs unquestioningly, particularly when system confidence is displayed.
- Alert fatigue: Frequent low-priority alerts can desensitize clinicians to high-risk signals.
- Workflow integration: Poorly integrated AI can disrupt care pathways and amplify risks.

Design implication: Security strategies must integrate cognitive ergonomics, ensuring alerts, overrides, and workflow integration align with clinician decision-making processes.

7.6 Regulatory and Governance Considerations

The discussion emphasizes that technical defenses alone are insufficient; robust governance frameworks are required:

1. Regulatory alignment: ISO 14971, IEC 62304, FDA AI/ML SaMD guidance provide foundational principles, but real-time verification and clinical outcome tracking are gaps.

2. Transparency and accountability: Model Cards, audit logs, and incident reporting are essential for trust and post-deployment evaluation (Mitchell *et al.,* 2019; WHO, 2024).
3. Cross-border compliance: Telemedicine platforms must navigate GDPR, HIPAA, and emerging AI Act requirements for data protection and safety obligations.

7.7 Implementation Challenges and Trade-Offs

While the proposed architecture is robust, several operational challenges exist:

- **Cost and resource constraints:** Zero-trust, TEEs, and secure IoMT implementations require investment.
- **Legacy system integration:** Hospitals may face difficulties integrating new AI pipelines with existing EHR systems.
- **Computational overhead:** Real-time adversarial monitoring and human-in-the-loop verification can increase latency.
- **Skill shortages:** Continuous cybersecurity and AI oversight require specialized personnel, which may be limited in smaller or rural healthcare settings.
- **Mitigation strategies:** Phased deployment, canary testing, and cross-functional training can reduce operational friction.

7.8 Future Research Directions

1. **Dynamic Clinical Risk Scoring:** Integration of attack likelihood with patient-specific risk to prioritize defense.
2. **Equity-Aware AI Security:** Analyze how adversarial attacks disproportionately affect vulnerable populations.
3. **Explainable Adversarial AI:** Tools to interpret AI failures under attack for clinician transparency.
4. **Automated Post-Deployment Auditing:** Continuous runtime evaluation for both security and clinical outcomes.
5. **Interdisciplinary Human Factors Studies:** Systematic research on automation bias mitigation in telemedicine workflows.

7.9 Key Discussion Insights

- Security for AI telemedicine must be measured in clinical outcomes, not IT compromise.
- Architecture must prioritize clinician authority, IoMT integrity, and AI reliability simultaneously.
- Threat modeling must integrate human factors, workflow, and socio-technical context.
- Governance and regulatory compliance are critical enablers of safe clinical deployment.
- Future research should focus on equity, explainability, and continuous monitoring.

8. CONCLUSION

This study systematically addresses the security and resilience of AI-driven telemedicine platforms, bridging the gap between technical cybersecurity measures and clinical safety imperatives. By integrating threat modeling, layered architecture design, and rigorous evaluation frameworks, the research offers a comprehensive blueprint for safe clinical deployment.

8.1 Summary of Key Contributions

1. **Comprehensive Threat Taxonomy:** Identified and classified adversarial AI attacks, IoMT signal compromises, LLM prompt injections, workflow vulnerabilities, and MLOps supply-chain threats. Emphasized clinical impact over mere technical compromise.

2. **Resilient Architecture Design:** Proposed a layered, zero-trust architecture embedding:
   - AI model integrity verification
   - IoMT data authentication
   - Human-in-the-loop safeguards
   - Secure MLOps pipelines

This design ensures that patient safety and clinician authority are maintained even under attack.

3. **Evaluation & Validation Framework:** Developed a multi-layer testing and KPI system encompassing:
   - Adversarial robustness of AI models
   - IoMT authenticity
   - Clinical workflow safety
   - Continuous post-deployment monitoring

Metrics emphasize patient safety, decision accuracy, and operational reliability.

4. **Clinical-Centric Security Perspective:** Demonstrated that attacks with moderate IT severity may have high patient harm potential, reinforcing the need for a safety-first, socio-technical approach.

5. **Policy and Governance Implications:** Highlighted regulatory alignment with FDA, ISO, MHRA, GDPR/HIPAA, and WHO guidelines. Emphasized post-market surveillance, model auditing, and clinician oversight as critical enablers.

8.2 Practical Implications

- **Healthcare Providers:** Can deploy AI telemedicine systems with assured resilience against adversarial and IoMT attacks.
- **System Developers:** Gain a framework to design AI models and pipelines that remain clinically reliable under stress.
- **Regulators:** Can adopt clinically-informed risk assessment methods, ensuring AI in telemedicine meets both technical and patient-safety standards.
- **Patients:** Benefit from safer remote monitoring, triage, and diagnostics, reducing risk from system compromise.

8.3 Limitations

While robust, the study has some constraints:

1. **Simulation-Based Evaluation:** Real-world large-scale deployment may reveal additional threats not captured in controlled testbeds.
2. **Rapid AI Evolution:** New model types (e.g., multimodal LLMs) may introduce unanticipated vulnerabilities.
3. **Human Factor Variability:** Clinician behavior differs across settings, affecting HITL efficacy.
4. **Resource Constraints:** Smaller institutions may struggle to implement zero-trust, IoMT attestation, or continuous adversarial testing.

8.4 Future Directions

1. **Adaptive AI Risk Scoring:** Dynamically prioritize defenses based on patient condition and threat likelihood.
2. **Cross-Institution Collaborative Defense:** Shared attack intelligence and threat databases.
3. **Explainable Adversarial Detection:** Tools for clinicians to understand why AI outputs are flagged.
4. **Equity and Accessibility:** Ensuring secure telemedicine does not disproportionately affect underserved populations.
5. **Automation Bias Mitigation Research:** Further studies on cognitive ergonomics and safe HITL integration.

8.5 Final Remarks

Although AI-driven telemedicine presents previously unheard-of possibilities for remote diagnostics, triage, and ongoing patient monitoring, the risks are higher than with traditional IT systems because patient health could be directly harmed by adversary compromise. This research offers a clinically informed path for the safe and efficient implementation of AI telemedicine by integrating threat-aware design, layered robust architecture, and rigorous assessment methods.

Acknowledgment

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

REFERENCES

1. Alsubaei, F., et al. (2023). IoMT security and resilience. *ACM Computing Surveys*.
2. Alsubaei, F., et al. (2023). IoMT firmware analysis. *ACM Computing Surveys*.
3. Beam, A., et al. (2023). Scaling risks of AI failure in healthcare. *Science*.
4. Carayon, P., et al. (2006). SEIPS model for patient safety systems. *International Journal of Medical Informatics*.
5. Check Point Research. (2023). IoMT exploit landscape.
6. Coventry, L., et al. (2023). Secure digital health design. *BMJ Health Informatics*.
7. Deloitte Insights. (2024). Future of virtual health and AI-enabled remote care.
8. DeRosier, J., et al. (2002). FMEA in healthcare risk assessment. *Joint Commission Journal*.
9. European Data Protection Board. (2023). GDPR in AI-processed health data.
10. EU Commission. (2024). AI Act for high-risk medical software.
11. FDA. (2024). AI/ML guidance for medical devices. U.S. Food and Drug Administration.
12. FDA. (2024). AI/ML software as a medical device action plan.
13. FDA. (2024). Artificial intelligence/machine learning medical device software guidance.
14. Finlayson, S. G., et al. (2019). Adversarial attacks against medical deep learning systems. *Science Translational Medicine*.
15. Forrester. (2024). Zero trust adoption in healthcare.
16. Goel, A., et al. (2023). Cyber–AI risk convergence in healthcare. *Lancet Digital Health*.
17. Goldberger, A., et al. (2023). PhysioNet biomedical dataset repository.
18. IBM Security. (2024). Cost of healthcare data breach report.
19. IBM Security. (2024). Healthcare breach and response trends.
20. Jagielski, M., et al. (2021). Poisoning attacks on machine learning pipelines. *IEEE Security & Privacy*.
21. Jagielski, M., et al. (2021). ML supply chain poisoning attacks. *IEEE Security & Privacy*.
22. Jalali, M., et al. (2021). Telemedicine cybersecurity risks. *JMIR*.

23. Johnson, A., et al. (2023). MIMIC-IV clinical database. *Nature Scientific Data*.
24. Keesara, S., Jonas, A., & Schulman, K. (2020). Covid-19 and health care's digital revolution. *New England Journal of Medicine*.
25. Kelly, C., et al. (2023). Clinical model drift analysis. *NEJM AI*.
26. Kelly, C., et al. (2023). Monitoring AI models in clinical practice. *NEJM AI*.
27. Khaim, R., et al. (2023). Trusted execution for medical AI integrity. *IEEE Security & Privacy*.
28. Kruse, C., et al. (2021). Barriers to telehealth security. *JMIR*.
29. Li, X., et al. (2023). Prompt injection in clinical decision-support LLMs. *Journal of Biomedical Informatics*.
30. Li, X., et al. (2023). Prompt injection risks in clinical LLMs. *Journal of Biomedical Informatics*.
31. Lyell, D., &Coiera, E. (2017). Automation bias and clinical decision risks. *JAMIA*.
32. Ma, X., et al. (2023). Radiology AI robustness limits. *Radiology AI*.
33. MHRA. (2023). Post-market AI surveillance for medical software.
34. MHRA. (2023). Post-deployment AI monitoring guidance.
35. MITRE. (2023). MITRE ATLAS: Adversarial threat landscape for AI systems.
36. Mitchell, M., et al. (2019). Model cards for responsible AI.
37. Moradi, M., et al. (2024). IoMTbiosignal spoofing analysis. *IEEE Transactions on Information Forensics and Security*.
38. Moradi, M., et al. (2024). Biosignal spoofing and IoMT device integrity failures. *IEEE Transactions on Information Forensics and Security*.
39. Moradi, M., et al. (2024). IoMT firmware tampering and data spoofing exploits. *IEEE Transactions on Information Forensics and Security*.
40. Obermeyer, Z., et al. (2019). Racial bias in medical AI. *Science*.
41. Parker, S., &Ashrafian, H. (2022). Clinical trust in AI. *Lancet Digital Health*.
42. Parker, S., &Ashrafian, H. (2022). Human-AI trust and clinical outcomes. *Lancet Digital Health*.
43. Ponemon Institute. (2023). Healthcare cybersecurity cost index.
44. Rajpurkar, P., et al. (2022). Benchmarking AI defenses. *Nature Machine Intelligence*.
45. Raji, I., et al. (2022). Auditing healthcare ML systems.
46. Sendak, M., et al. (2020). Real-world clinical AI deployment risks. *JAMA*.
47. Sendak, M., et al. (2020). Clinical risks in deployed medical AI. *JAMA*.
48. Seyyed-Kalantari, L., et al. (2021). Bias amplification in medical imaging. *Nature Medicine*.
49. Sittig, D., & Singh, H. (2022). Technology-induced clinical error. *JAMA*.
50. U.S. HHS. (2023). HIPAA security guidance for digital health.
51. WHO. (2021). AI governance and ethics in healthcare.
52. WHO. (2023). Global telemedicine adoption and digital health transformation report.
53. WHO. (2024). Guidelines for AI in health safety and risk management.
54. Wu, E., et al. (2021). Post-deployment clinical AI monitoring. *NEJM AI*.
55. Yan, Q., et al. (2024). Jailbreaking clinical LLMs. *NPJ Digital Medicine*.
56. Yan, Q., et al. (2024). Security vulnerabilities in clinical LLM interfaces. *NPJ Digital Medicine*.
57. Zhang, J., et al. (2022). Secure and accountable AI inference pipelines.
58. Zhang, R., et al. (2023). IoMT protocol risks. *IEEE IoT Journal*.
59. Zhou, Y., et al. (2024). Adversarial attacks on medical imaging AI: Clinical implications and failure rates. *Nature Medicine*.
60. Zhou, Y., et al. (2024). Adversarial vulnerability in medical vision AI. *Nature Medicine*.

****