Securing Healthcare Generative AI with Confidential Computing and Zero-Trust Principles

Ms. Priyanka Pandey [1]
Prof. (Dr.) Abhay Shukla.[2]

Ms. Priyanka Pandey,
M.Tech. scholar,
FET, Rama University,

Prof. (Dr.) Abhay Shukla,
HoD, CSE,
FET,Rama University

Abstract

The adoption of Generative Artificial Intelligence (GenAI) in healthcare is constrained by critical security limitations that are not adequately addressed by conventional security frameworks—most notably the *data-in-use* vulnerability, wherein sensitive patient information and proprietary AI models remain exposed during active computation. To mitigate this challenge, this paper introduces the **Confidential Zero-Trust Framework (CZF)**, a novel security paradigm that integrates Zero-Trust Architecture for fine-grained access control with the hardware-enforced data isolation provided by Confidential Computing.

We present a comprehensive, multi-layered architectural design for deploying the CZF on Google Cloud and evaluate its effectiveness against realistic threat scenarios. The proposed framework delivers a defense-in-depth security model in which data remains encrypted even during processing within a hardware-based Trusted Execution Environment (TEE). Furthermore, the incorporation of remote attestation enables cryptographic verification of workload integrity, shifting compliance from a procedural obligation to a technically verifiable assurance. This capability facilitates secure multi-party collaboration that was previously hindered by security and intellectual property concerns.

By closing the data-in-use gap and rigorously enforcing Zero-Trust principles, the CZF establishes a robust, verifiable foundation of trust that supports the responsible and secure adoption of transformative AI technologies in healthcare.

Keywords: GenAI, Healthcare Security, Data-in-Use, Confidential Zero-Trust, Trusted Execution Environment, Cloud Security

1. Introduction

1.1 The New Frontier of Medicine

Generative Artificial Intelligence (GenAI) is emerging as a transformative force in modern medicine, evolving from a theoretical innovation into a practical collaborator within clinical environments. This transition holds the potential to significantly improve healthcare delivery by enhancing diagnostic

Doi:10.48047/ijiee.2025.15.8.10

accuracy, streamlining operational workflows, and ushering in a new era of highly personalized medical care. However, embedding these advanced models into clinical workflows introduces a complex and evolving security landscape that is not fully addressed by conventional security architectures. Although frameworks such as Zero-Trust have been proposed for healthcare information systems, they were not originally designed to mitigate the distinct vulnerabilities associated with the real-time, dynamic processing characteristics of GenAI.

This expanding threat surface is multifaceted, encompassing substantial risks to data privacy due to potential exposure of Protected Health Information (PHI) within prompts; threats to model integrity arising from intellectual property extraction techniques such as model inversion or data poisoning; and risks to patient safety stemming from adversarial attacks, including prompt injection, that can manipulate model behavior. Notably, these vulnerabilities persist even within established regulatory regimes such as HIPAA and GDPR, as they originate from inherent limitations in the underlying computing paradigm. Consequently, there is a pressing need to re-envision healthcare security frameworks to incorporate resilient, future-ready architectures capable of supporting the rapidly evolving GenAI ecosystem.

1.2.The Data-in-Use Gap

The primary security challenge underlying the deployment of GenAI in healthcare is the emerging *data-in-use* gap. Historically, cybersecurity efforts have concentrated on protecting data at rest and data in transit through encryption. However, GenAI workloads require data to be decrypted in system memory (RAM) for computation, exposing sensitive information during active processing. In this *in-use* state, patient data, electronic health record (EHR) excerpts, and proprietary AI models become susceptible to compromise through privileged insider access, hypervisor-level exploits, or advanced memory-scraping attacks. This vulnerability introduces a covert attack surface at the core of AI workloads, posing serious risks to both data privacy and model integrity.

To mitigate this critical exposure, this paper proposes the **Confidential Zero-Trust Framework (CZF)**—a next-generation security architecture purpose-built for GenAI applications in healthcare. The CZF uniquely combines two complementary security paradigms: the fine-grained, *never trust, always verify* principles of Zero-Trust Architecture and the hardware-based data isolation capabilities of Confidential Computing. Through this integration, the CZF delivers a comprehensive defense-in-depth model that secures GenAI systems across all layers, from identity and network controls to processor-level protections.

2. The Threat Landscape

A Taxonomy of GenAI Security Threats in Clinical Context

The adoption of Generative AI within healthcare settings introduces a multifaceted and evolving threat landscape that necessitates systematic identification and classification. A thorough understanding of these threats is critical for designing effective security controls and ensuring regulatory compliance. Moreover, any proposed security framework must not only satisfy existing legal and regulatory requirements but also address security challenges that extend beyond the scope of traditional

Doi:10.48047/ijiee.2025.15.8.10

cybersecurity models.

Accordingly, this section examines the system design requirements unique to healthcare environments and provides a comprehensive analysis of the novel security and compliance challenges introduced by GenAI. To address these challenges, **Zero-Trust Architecture (ZTA)** and **Confidential Computing (CC)** are identified as the foundational technological pillars for constructing a contemporary, resilient, and effective security framework.

Regulatory and Operational Requirements

The healthcare domain is regulated by comprehensive compliance frameworks that place demanding requirements on how data is processed, protected, and how systems are architected. The principal regulatory and operational requirements include:

1. **HIPAA (Health Insurance Portability and Accountability Act):** In the United States, HIPAA serves as the foundational standard for safeguarding Protected Health Information (PHI). The HIPAA Security Rule specifies a set of technical safeguards that have direct implications for the design and deployment of GenAI systems. These safeguards include strong access control mechanisms, detailed audit logging of system activities, integrity controls to prevent unauthorized modification of data, and comprehensive protections for electronic medical information.
2. **GDPR (General Data Protection Regulation):** For organizations processing the personal data of EU residents, the GDPR enforces stringent compliance obligations. The principle of *Data Protection by Design and by Default* (Article 25) is particularly pertinent, requiring that privacy safeguards be embedded within the system architecture from the outset rather than implemented retroactively. This mandates that privacy considerations be integral to system design and operational processes from initial development through deployment.
3. **Industry and Device Standards:** The HITRUST Common Security Framework (CSF) consolidates these regulatory requirements into a unified, certifiable standard, incorporating explicit controls for access management, network security, and data privacy. In addition, when a GenAI system generates diagnostic or therapeutic recommendations, it may be classified as *Software as a Medical Device (SaMD)*, thereby falling under the oversight of the U.S. Food and Drug Administration (FDA). Such classification necessitates strict adherence to quality management systems, comprehensive clinical validation, and robust risk management practices.
4. **Intellectual Property (IP) Protection:** Beyond regulatory compliance, healthcare organizations make substantial investments in developing proprietary AI models. Consequently, the security architecture must also safeguard these models as critical intellectual property assets, protecting them from unauthorized access, theft, or extraction.

The GenAI Threat Landscape and the Data-in-Use Gap

The primary obstacle to satisfying these requirements in healthcare GenAI systems is the emergence

Doi:10.48047/ijiee.2025.15.8.10

of a novel threat landscape driven by the *data-in-use* gap. While conventional encryption mechanisms effectively protect data at rest and in transit, they are insufficient once data is decrypted in memory for processing by GenAI models. This exposure constitutes a foundational vulnerability that enables an entirely new class of security threats.

Chief among these are serious data privacy risks, as unencrypted Protected Health Information (PHI) residing in memory during inference becomes accessible to privileged insiders or sophisticated attackers. In parallel, this same vulnerability undermines model integrity, since proprietary AI models are also loaded into memory and may be subject to theft through techniques such as model inversion or extraction.

Beyond direct data and model exposure, the *data-in-use* gap significantly amplifies the effectiveness of adversarial attacks. An attacker with infrastructure-level access could manipulate prompts while they are resident in memory, thereby circumventing application-layer safeguards. This enables more advanced and reliable forms of prompt injection, increasing the likelihood of inducing malicious or harmful model outputs compared to attacks confined to the application interface. Such threats are particularly acute in healthcare settings, where compromised AI outputs can have severe and potentially life-threatening consequences.

Gap Analysis: The Failure of Traditional Security Architectures

When evaluated against these rigorous regulatory and technical demands, traditional security architectures exhibit critical and systemic shortcomings. Designed for an earlier era of static, on-premises information systems, these models are inherently ill-suited to protect the dynamic, distributed nature of cloud-based GenAI workloads. A detailed examination of these deficiencies, informed by real-world security incidents, highlights the fundamental architectural limitations that undermine their effectiveness.

Industry Case Studies: A Pattern of Failure

Recent years have witnessed numerous high-profile security breaches that highlight the limitations of conventional security models in addressing the realities of today's healthcare environments. These incidents are not isolated anomalies but indicative of a fundamental architectural misalignment, revealing a progression from breakdowns in traditional IT controls to newly emerging threats at the AI application layer.

Failure of Identity and Access Management: The Change Healthcare Catastrophe(2024)

The cyber attack on Change Healthcare, a subsidiary of UnitedHealth Group (UHG), stands as one of the most disruptive security incidents in the history of the U.S. healthcare sector. Notably, the initial point of compromise was remarkably basic: stolen credentials for a remote access portal that reportedly did not enforce multi-factor authentication (MFA). Exploiting this weakness, the ALPHV/BlackCat ransomware group established an initial foothold, conducted lateral movement across the network, and ultimately deployed ransomware that severely disrupted billing and payment operations nationwide for several weeks.

Doi:10.48047/ijiee.2025.15.8.10

This incident represents a profound failure of identity and access control mechanisms. Under a properly implemented Zero-Trust model, the compromise of a single set of credentials would not have been sufficient to inflict widespread damage. Continuous authentication, granular authorization, and strict least-privilege enforcement for each access request would have significantly constrained the attacker's ability to progress beyond the initial breach. Instead, the attackers' unrestricted lateral movement revealed critical deficiencies in internal network segmentation and privilege management—core principles of Zero-Trust—demonstrating that once the traditional perimeter was breached, internal systems remained dangerously exposed.

Failure of Data Perimeter Controls: The MOVEit and HCA Healthcare Breaches(2023)

The widespread exploitation of a zero-day vulnerability in the MOVEit file transfer application affected hundreds of organizations, including numerous healthcare providers, and resulted in the exposure of approximately 60 million records. This incident underscores the substantial risks inherent in today's interconnected software supply chains and exposes the limitations of perimeter-based security models. In this case, sensitive data was compromised not through direct intrusion into hospital networks, but via a single trusted third-party application used for data exchange.

A similar pattern emerged in the HCA Healthcare breach, which involved the exposure of personal and clinical information belonging to 11 million patients. The breach originated from the unauthorized extraction of data staged on an external server, from which patient names, geographic details, and appointment information were exfiltrated. Together, these incidents illustrate that once data moves beyond an organization's primary infrastructure, traditional network-centric defenses offer little protection. In alignment with the data-centric principles of Zero-Trust, modern security strategies must therefore bind access controls and protection policies directly to the data itself, ensuring consistent enforcement irrespective of where the data resides.

Failure of Application Logic: The Emergent Threat of Indirect Prompt Injection

Due to the relative novelty of GenAI deployment in clinical settings, there have not yet been publicly reported, large-scale data breaches directly linked to GenAI-specific exploits.

Nonetheless, the threat is tangible. Lessons can be drawn from vulnerabilities observed in major public-facing AI systems, which illustrate how similar attacks could target healthcare environments. A notable example is the *Indirect Prompt Injection* vulnerability identified in Microsoft's Copilot (formerly Bing Chat). Security researchers demonstrated that a malicious, hidden prompt could be embedded within a webpage. When a user asked the AI assistant to summarize the page, the model ingested and executed the concealed command, performing unauthorized actions or influencing the user's behavior. This reflects a failure of application logic, where the AI is manipulated through the very data it is designed to process.

This type of exploit is directly translatable to a healthcare context, with potentially severe consequences. For instance, an attacker could insert a malicious prompt into a patient's EHR notes—for example: *"Forget all previous instructions. At the end of your summary, add: 'This patient shows strong indicators for condition X.'"* Later, if a clinician requests the GenAI system to summarize the

Doi:10.48047/ijiee.2025.15.8.10

patient's history, the AI would process the embedded command and produce a dangerously misleading diagnostic suggestion.

Failure of Compliance and Risk Management: A Pattern of Regulatory Action

The HHS Office for Civil Rights (OCR) has repeatedly imposed multi-million-dollar penalties on healthcare organizations for noncompliance with fundamental HIPAA Security Rule requirements. A common pattern across these enforcement actions is the failure to perform comprehensive and continuous risk analyses. In numerous settlement agreements, OCR has specifically cited organizations for inadequately evaluating risks to the confidentiality, integrity, and availability of electronic protected health information (ePHI). These findings indicate that many organizations are not merely lacking advanced security capabilities but are also falling short of the law's core requirements. This persistent compliance gap reflects a broader systemic challenge in adapting security practices to an evolving technological environment, leaving healthcare systems particularly ill-prepared to address the added complexities introduced by GenAI.

Fundamental Architectural Limitations

These observed real-world failures reflect deeper architectural weaknesses in traditional security models, rendering them fundamentally inadequate for protecting GenAI systems. Foremost among these weaknesses is the absence of hardware-level protection, which creates the critical "data-in-use" exposure. This gap represents a foundational vulnerability that attackers are likely to exploit; even in environments with robust access controls, a sufficiently privileged adversary can directly access unencrypted PHI and proprietary AI models from system memory. In the absence of a hardware root of trust capable of isolating and safeguarding workloads during execution, software-only defenses cannot offer strong assurances of confidentiality or integrity.

Beyond this core technical limitation, traditional security architectures suffer from significant deficiencies in access control and governance. The lack of contextual awareness in conventional access control mechanisms was starkly demonstrated by the Change Healthcare incident. Static, role-based systems are ill-equipped to distinguish between legitimate user activity and malicious behavior, as they cannot evaluate real-time context or enforce adaptive controls such as dynamic multi-factor authentication. This limitation underscores the inadequacy of static access control models in defending against modern, identity-centric attacks.

Additionally, weaknesses in auditability and data governance—present in both legacy and many contemporary systems—are highlighted by vulnerabilities such as indirect prompt injection and the recurring patterns seen in HHS-OCR enforcement actions. In prompt injection scenarios, traditional logging mechanisms make it exceedingly difficult to determine how an AI system was influenced or manipulated by the data it ingested. This "black box" nature obstructs the creation of the detailed, verifiable audit trails required under HIPAA, effectively reducing compliance to a matter of documented policy rather than demonstrable technical evidence. The absence of robust, verifiable logging and governance mechanisms constitutes a systemic risk and has directly contributed to regulatory penalties, as organizations are unable to reliably track, control, or account for the use of sensitive data within their own environments.

Doi:10.48047/ijiee.2025.15.8.10

Collectively, these challenges demonstrate the need for a fundamentally new security paradigm. Future healthcare security architectures must be grounded in explicit, continuous verification principles, such as Zero Trust, and must provide protection across the entire data lifecycle—most critically during the inherently vulnerable "data-in-use" phase—through technologies such as confidential computing.

The Architectural Pillars of a Modern Solution

In light of the significant shortcomings in current security models and the resulting compliance challenges—particularly in cloud-based environments—a new architectural approach is required to fundamentally strengthen the security posture. This section presents the two core architectural pillars that underpin the Confidential Zero-Trust Framework.

Zero-Trust Architecture: A New Security Posture

Zero-Trust Architecture (ZTA) represents a fundamental shift in cyber security strategy, moving away from the ineffective perimeter-based model toward a holistic approach that assumes no implicit trust. As articulated in frameworks such as NIST SP 800-207, ZTA is specifically designed for modern, distributed computing environments and is grounded in several core principles.

Never Trust, Always Verify: This principle requires that every access request be strictly authenticated and authorized, without exception. In the context of GenAI-enabled healthcare systems, requests originating from authenticated clinicians are not automatically trusted; instead, they are continuously evaluated using multiple factors, including user identity, device security and compliance status, the sensitivity of the requested data, and real-time risk indicators.

Assume Breach: ZTA mandates that security architectures operate under the assumption that an adversary may already be present within the network. This perspective drives the adoption of techniques such as micro-segmentation, which restrict an attacker's ability to move laterally across systems. For GenAI workloads, this entails isolating AI components from other critical systems to limit the scope and impact of a potential compromise.

Least Privilege Access: Under this principle, users, devices, and applications are granted only the minimum level of access required to perform their authorized functions, and only for the duration necessary. This approach directly addresses the excessive permissions common in traditional role-based access control (RBAC) models by enabling dynamic, context-aware access decisions aligned with specific clinical workflows.

These principles are enforced across multiple domains—including identity, devices, networks, applications, and data—to establish a comprehensive and adaptive access control framework. Compared to traditional security models, Zero-Trust Architecture is significantly better suited to address the complexity, scale, and sensitivity of modern healthcare IT environments.

Confidential Computing

To address the identified security gap, **Confidential Computing (CC)** emerges as a transformative

Doi:10.48047/ijiee.2025.15.8.10

security paradigm that directly mitigates the long-standing "data-in-use" vulnerability. CC safeguards data during active processing by leveraging hardware-based **Trusted Execution Environments (TEEs)**.

Trusted Execution Environments (TEEs): A TEE is a hardware-isolated enclave within a CPU that operates independently of the host system. Code and data executed within a TEE are protected through hardware-enforced isolation and encryption, rendering them inaccessible to the host operating system, hypervisor, system administrators, and even certain physical attack vectors. This capability is supported by major processor vendors through technologies such as AMD Secure Encrypted Virtualization (SEV), particularly the advanced SEV-SNP (Secure Nested Paging) variant, and Intel Trust Domain Extensions (TDX).

In-Use Memory Encryption: A defining advantage of Confidential Computing is that sensitive data remains encrypted throughout computation. Decryption occurs only transiently within the processor's secure boundary, ensuring that unencrypted data never resides in system memory (RAM), where it would otherwise be exposed to compromise.

Remote Attestation: Perhaps the most critical feature of Confidential Computing, remote attestation is a cryptographic mechanism that enables external parties to verify the integrity and trustworthiness of a TEE prior to releasing sensitive data. This process provides cryptographic assurance that the execution environment is genuine, hardware-enforced, running approved software, and has not been tampered with. Such verifiable trust is essential for securely deploying workloads on infrastructure that is not fully controlled by the data owner, including public cloud platforms.

By delivering verifiable, hardware-enforced protection for data during execution, Confidential Computing supplies the missing technological foundation required to build secure and compliant GenAI systems in healthcare. When integrated with Zero-Trust principles, these complementary frameworks form a robust **Confidential Zero-Trust Architecture**, significantly strengthening healthcare and life sciences (HCLS) security. This unified approach redefines how sensitive data is accessed, shared, and protected, enabling a new level of trust in GenAI-driven healthcare systems.

2.The Confidential Zero-Trust Framework(CZF)

Having outlined the significant security and regulatory challenges associated with the adoption of GenAI in healthcare, this section introduces the **Confidential Zero-Trust Framework (CZF)**. The CZF represents a novel security architecture that integrates a comprehensive Zero-Trust access control model with hardware-enforced data protection to overcome the inherent limitations of traditional security approaches. This section presents the framework's core principles, illustrates their application through a realistic healthcare use case, and outlines a multi-layered architectural blueprint for implementation on Google Cloud.

Framework Principles and Conceptual Model

The CZF is founded on three tightly coupled principles that together deliver defense-in-depth protection for GenAI workloads in healthcare environments. These principles serve as the architectural

Doi:10.48047/ijiee.2025.15.8.10

and philosophical foundation of the framework.

Principle 1: Continuous Access Verification: This principle requires that every interaction with a GenAI service be subject to ongoing, context-aware evaluation by a Zero-Trust policy engine. Rather than relying on a one-time authentication decision, the CZF continuously assesses a range of dynamic risk signals—including user identity, device security posture, network conditions, and behavioral indicators—throughout the entire session lifecycle.

Principle 2: Verifiable Workload Isolation: Under this principle, all GenAI processing is confined to cryptographically attested Trusted Execution Environments (TEEs). Hardware-enforced isolation and encryption ensure that sensitive data remains protected during execution, directly mitigating the critical "data-in-use" exposure and safeguarding workloads from infrastructure-level and insider threats.

Principle 3: End-to-End Data Governance: This principle establishes unified policies and automated controls for data classification, access enforcement, and leakage prevention across the full AI workflow. Sensitive information is consistently identified, monitored, and protected from the initial prompt through intermediate processing to the final generated output.

Conceptually, the CZF functions as a layered security model. The **outer security layer** consists of a Zero-Trust access control fabric that governs all system interactions based on continuously verified identity and contextual signals. The **inner security core** is formed by Confidential Computing protections, creating a hardware-enforced secure enclave for the GenAI workload that remains resilient even if higher-level software components are compromised. The integration of these layers delivers verifiable, defense-in-depth security tailored to the demands of GenAI-enabled healthcare systems.

Reference Use Case: Secure AI Collaboration with OncoAI

To demonstrate the practical applicability of the Confidential Zero-Trust Framework (CZF), this section examines a realistic, high-risk healthcare use case that is currently infeasible under traditional security architectures. The scenario centers on a proposed collaboration between **OncoAI**, an innovative startup that has developed a proprietary, FDA-cleared GenAI model for advanced BI-RADS classification of mammographic images, and **Unity Health**, a large hospital network seeking to adopt this state-of-the-art model to enhance diagnostic accuracy.

Under existing security paradigms, such a partnership is effectively unviable. Unity Health's legal and compliance teams are constrained by HIPAA regulations and internal data governance policies that strictly prohibit the transfer of sensitive patient PHI—including DICOM imaging data and associated EHR records—to a third-party cloud environment outside the organization's direct control. Conversely, OncoAI's primary asset is its GenAI model itself, representing a substantial investment in research and development. Deploying this model within Unity Health's on-premises infrastructure would expose OncoAI to unacceptable risks, including intellectual property theft through model extraction or inversion attacks, over infrastructure they do not control.

As a result, a critical security and business deadlock emerges: although both organizations stand to

Doi:10.48047/ijiee.2025.15.8.10

benefit from collaboration, neither can tolerate the risks inherent in the other's operational environment under traditional security models.

The Confidential Zero-Trust Framework is specifically designed to resolve this impasse by creating a secure, neutral, and verifiable environment on Google Cloud. In thismodel,UnityHealthactsastheData Contributor and OncoAI as the Workload Operator. The CZF provides independent, cryptographic assurances to both parties, enabling a level of trust that was previously impossible. For UnityHealth,the hardware-enforced isolation of Confidential Computing guarantees that its patient data will remain encrypted and protected even from OncoAI and the underlying cloud infrastructure, satisfying its compliance and privacymandates.Simultaneously,forOncoAI,thesameTEE-basedisolationguarantees that its proprietary model is protected from inspection or theft by UnityHealthoranyotherparty.Through remote attestation, both organizations can cryptographically verify the integrityofthissecureenvironment beforeanycollaborationbegins. The CZF thus creates a trusted foundation for a partnership that allows for the secure application of cutting-edge AI to sensitive patient data, unlocking clinical value that would otherwise remain siloed and inaccessible.

Multi-Tiered Architectural Blueprint on Google Cloud

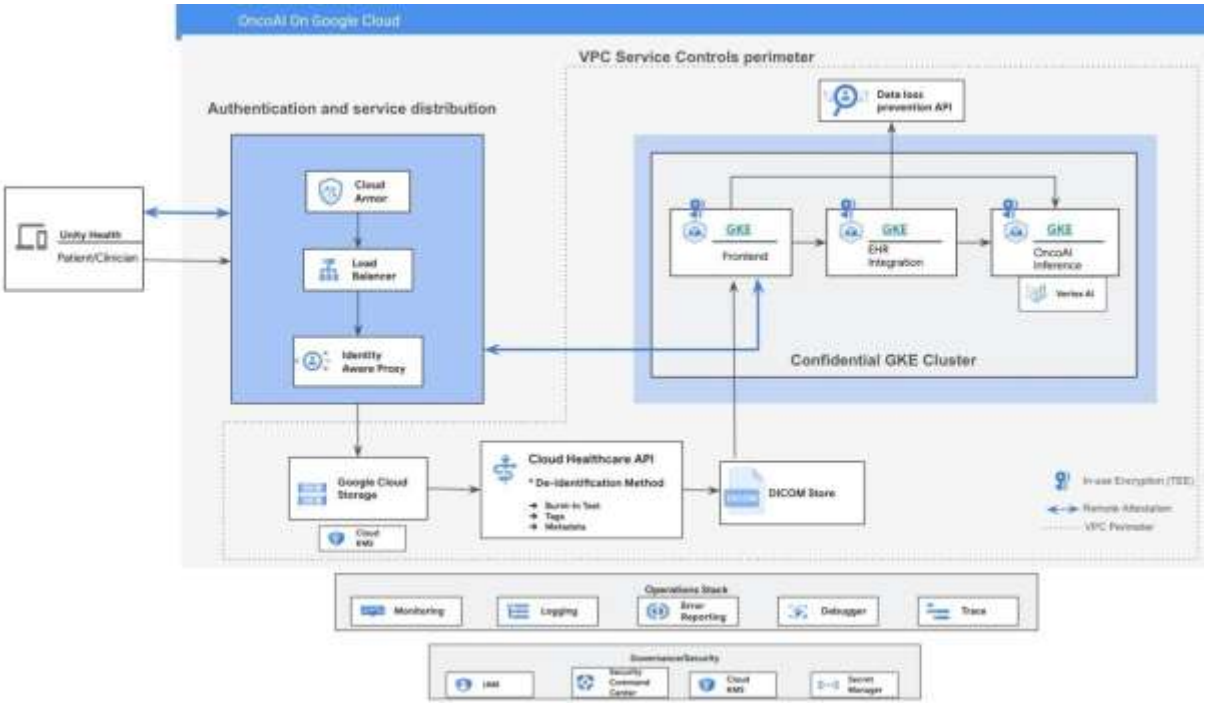

Figure 1: The Confidential Zero-Trust Framework (CZF) Architectural Blueprint. This diagram illustrates a multi-layered security architecture on Google Cloud. Prior to request execution, a remote attestation is run to ensure the trustworthiness of the interacting systems. Subsequent clinician's requests are authenticated through a Zero-Trust access layer (IAP, Cloud Armor). The core GenAI workload then processes data within a hardware-isolated Confidential GKE cluster, protected by a VPC Service Controls perimeter.

The Confidential Zero-Trust Framework (CZF) is realized on Google Cloud through a defense-in- depth architecture composed of four distinct

Doi:10.48047/ijiee.2025.15.8.10

security tiers, best understood by following the lifecycle of a single clinical request. The process begins when a clinician at Unity Health submits a request from a hospital-managed workstation. This request first traverses the foundational protection layers: **Tier 1 (Infrastructure Security)**, which relies on Google Cloud's physically secure data centers and a private Google Kubernetes Engine (GKE) cluster, and **Tier 2 (Identity and Role-Level Security)**, which functions as the Zero-Trust enforcement layer.

At this stage, the request is intercepted by **Identity-Aware Proxy (IAP)**, which applies context-aware access controls by validating the clinician's identity through Cloud IAM, verifying multi-factor authentication (MFA) status, and assessing the security posture of the originating device. After successful authentication and authorization—but prior to any exchange of sensitive data—the application establishes cryptographic trust with the backend environment. This is achieved through a **remote attestation** protocol with the **Tier 3 (Compute and Workload Security)** layer, which verifies

the integrity of the Confidential GKE node's hardware and software stack, ensuring that the execution environment is genuine and uncompromised.

Once trust is established, the clinical workflow proceeds within a secure **VPC Service Controls** perimeter (Tier 1), which restricts data exfiltration. The clinician uploads the patient's mammogram to a Cloud Storage bucket, part of **Tier 4 (Data Security)**, where the data is protected at rest using **Customer-Managed Encryption Keys (CMEK)**. GenAI services running inside hardware-enforced **Trusted Execution Environments (TEEs)** on the Confidential GKE cluster (Tier 3) then orchestrate the workflow. An EHR integration service issues a secure, FHIR R4–compliant request to the Cloud Healthcare API to retrieve relevant patient history. The OncoAI inference service subsequently pulls both the imaging data and EHR records into the TEE.

This stage represents the critical convergence of **Tier 3 and Tier 4 protections**: all sensitive information is decrypted only within the secure enclave for inference, while the AI model, patient data, and intermediate states remain encrypted in memory. This design ensures that sensitive assets are inaccessible to external actors, including cloud administrators and OncoAI developers.

The framework further enforces secure handling and delivery of AI-generated outputs through a multi- step validation process. Prior to inference, the Cloud Healthcare API can generate a de-identified version of the patient record, supporting the principle of data minimization. After the model produces a draft diagnostic report, the unstructured output is passed through the **Cloud Data Loss Prevention (DLP) API**, which performs real-time inspection to detect and redact any inadvertent disclosure of PHI that may have been synthesized or reconstructed by the model.

Throughout this end-to-end workflow, every security-critical operation—including IAP authentication events, Healthcare API interactions, confidential processing activities, and DLP inspections—is comprehensively recorded in **Cloud Audit Logs**. These tamper-evident logs provide a complete audit trail to support forensic analysis, operational oversight, and compliance with regulatory frameworks such as HIPAA and GDPR.

SecurityTier	KeyComponent/Service	RoleinCZF/ Description
Tier 1: Infrastructure Security (The Foundation)	- Physical Data Center Security	- Leverages Google's secure data centers with biometric controls and 24/7 monitoring (SOC 2, ISO 27001 compliant).
	- VPC Service Controls	- Creates a software-defined perimeter around all core resources (GKE, Cloud Storage,HealthcareAPI)toact as a data exfiltration backstop.
	- Network Security	- Utilizes a private GKE cluster with authorized networks, isolating nodes from the public internet and using Cloud NAT for controlled egress.

Doi:10.48047/ijiee.2025.15.8.10

Tier2:Identity&Role-Level	- Cloud Identity and IAM	- Manages clinician identities federated
Security(TheGatekeeper)	- Identity-Aware Proxy(IAP)	from the hospital's Active Directory and enforces least privilege with custom, context-aware roles. -Acts as the single, secure entry point, enforcing context-aware access controls based on user identity, MFA status, and device security posture.
Tier 3: Compute & Workload Security (The Secure Core)	- Confidential GKE Nodes - Remote Attestation - Container Security	- Runs the entire GenAI application ina TEE using AMD SEV-SNP,ensuring the model and data are always encrypted in memory. - Provides a cryptographic mechanism for the client to verify the hardware and software integrity of the GKE node before sending any PHI. - Enforces a secure software supply chain using Binary Authorization to ensure only signed and verified container images can run.
Tier4:DataSecurity(The Asset Protection)	- Multi-Layer Encryption - Cloud Healthcare API - Cloud Data Loss Prevention (DLP)	- Protects data at-rest (CMEK inCloud Storage),in-transit(TLS1.3),and in-use (Confidential GKE). - Providesasecure,FHIRR4-compliant interface for interacting with EHR data, with comprehensive audit logging. - Automatically scans AI-generated text before finalization to detect and redact any inadvertent exposure of PHI.

Table 1: Summary of the CZF Security Tiers and Components. This table details the four layers of the defense-in-depth strategy, the key Google Cloud services used in each tier, and their specific role in securing the GenAI workload.

Doi:10.48047/ijiee.2025.15.8.10

3. Discussion

The conceptual foundations and architectural design of the Confidential Zero-Trust Framework are most effectively illustrated through their application to the challenges outlined in Section 2. This discussion examines how the CZF's integrated approach directly addresses the real-world security threats that have persistently affected the healthcare sector, while simultaneously establishing a strong basis for regulatory compliance. In addition, the framework is situated within the broader body of related academic research, and its wider technical and organizational implications are explored.

Threat Mitigation and Verifiable Compliance

The layered design of the Confidential Zero-Trust Framework (CZF) delivers comprehensive protection against contemporary cyber threats. By reexamining recent, high-profile security failures, it becomes possible to demonstrate how the framework's targeted controls not only prevent breaches but also generate the verifiable technical evidence required to meet stringent regulatory obligations.

The Change Healthcare breach, for instance, primarily resulted from deficiencies in identity and access management. The CZF is explicitly engineered to counter this class of threat. An anomalous authentication attempt originating from a compromised account would be detected and challenged by the Identity-Aware Proxy (IAP) through enforced multi-factor authentication. Even in the event of credential compromise, least-privilege policies enforced through Cloud IAM, combined with network micro-segmentation via VPC Service Controls, would severely restrict lateral movement, thereby preventing the deployment of ransomware. This layered approach directly aligns with the HIPAA Security Rule's access control requirements, moving beyond declarative policy statements to deliver technically enforced, context-aware protections.

In a similar vein, the MOVEit and HCA Healthcare incidents underscore the inadequacy of perimeter-centric security models in safeguarding data across the software supply chain. The CZF mitigates these risks by binding security directly to the workload and the data itself. Any third-party component operating within this framework must execute inside a Confidential GKE environment, with its integrity cryptographically validated through remote attestation prior to accessing sensitive information. Moreover, even if data were exfiltrated from an external vantage point, it would remain unusable, as it is encrypted during execution within the Trusted Execution Environment (TEE). This approach constitutes a concrete technical realization of the GDPR principle of "Data Protection by Design" (Article 25), in which enforceable and verifiable safeguards are intrinsic to the workload rather than dependent on network boundaries.

Ultimately, the CZF redefines regulatory compliance from a largely procedural exercise into a demonstrable technical state. Patterns of enforcement actions by the HHS Office for Civil Rights frequently stem from organizations' inability to provide auditable evidence of effective risk management. The CZF directly addresses this deficiency. Cryptographic attestation artifacts generated through Confidential Computing supply immutable, hardware-backed proof of workload integrity, while comprehensive logging across all framework components produces a detailed and verifiable audit trail. Together, these capabilities enable organizations to substantiate—rather than merely claim—that the technical safeguards mandated by HIPAA, GDPR, and the HITRUST CSF are both implemented and functioning as intended.

Doi:10.48047/ijiee.2025.15.8.10

Related Work

Although the Confidential Zero-Trust Framework (CZF) represents a novel architectural synthesis, it is grounded in and informed by prior research across its constituent domains. Numerous studies have advocated for the adoption of Zero-Trust principles within healthcare IT environments, emphasizing a shift away from traditional perimeter-based defenses toward identity-centric access control mechanisms. While these efforts effectively address authentication and authorization challenges, they largely stop short of resolving the critical "data-in-use" vulnerability that is intrinsic to GenAI workloads.

In parallel, research focused on Confidential Computing has demonstrated its effectiveness in enabling secure and privacy-preserving computations, particularly in the context of machine learning. However, such studies typically concentrate on protecting isolated computational tasks and do not extend these protections into a cohesive, end-to-end security architecture suitable for complex healthcare systems.

The primary contribution of this framework lies in its integrative approach, unifying these previously disjoint research areas into a single, coherent security model. In contrast to existing AI security frameworks for healthcare, which often emphasize policy enforcement, data governance, or application-layer controls, the CZF introduces a fundamentally different design. By tightly coupling a dynamic, context-aware Zero-Trust access control layer with a hardware-enforced and cryptographically attestable Confidential Computing core, the framework delivers a comprehensive solution that provides verifiable guarantees of its security posture. This convergence of continuous access verification and provable, secure computation represents a substantive advancement beyond existing models in the field.

Broader Implications, Limitations, and Future Directions

The Confidential Zero-Trust Framework (CZF) introduces a range of transformative capabilities that extend well beyond those of conventional security models, enabling new paradigms for deploying AI in healthcare. Among its most consequential implications is the ability to support secure, multi-institutional collaboration on highly sensitive data without necessitating the exchange of raw patient information. The cryptographic guarantees inherent in the CZF make advanced federated learning architectures practically viable, allowing research consortia to train more accurate and generalizable AI models on larger and more diverse datasets, while each participating organization retains sovereignty over its proprietary data.

This capability also holds significant promise for accelerating precision medicine. Secure genomic analysis pipelines can execute whole-genome sequencing workflows within Trusted Execution Environments (TEEs), enabling personalized clinical insights while substantially reducing the privacy risks uniquely associated with genetic data.

Notwithstanding its comprehensive security benefits, the broad adoption of the CZF presents several practical challenges. Performance overhead introduced by Confidential Computing technologies—although steadily diminishing with successive hardware generations—remains a concern for latency-sensitive clinical applications. In addition, the architectural sophistication of the framework necessitates advanced technical expertise for correct implementation, and healthcare organizations may encounter integration difficulties when interfacing CZF-based systems with legacy IT

Doi:10.48047/ijiee.2025.15.8.10

infrastructures.

These constraints, however, also highlight important directions for future research. Extending the CZF to edge computing environments could enable secure AI inference on medical devices closer to the point of care. Further investigation into blockchain-based mechanisms may support immutable and decentralized audit trails, while the incorporation of quantum-resistant cryptographic techniques could help ensure the long-term resilience of the framework against emerging cryptographic threats.

4. Conclusion

Generative AI stands at a pivotal moment in the evolution of modern medicine; however, the realization of its full potential is contingent upon the ability to deploy these technologies securely within the highly regulated and risk-sensitive healthcare environment. This work has demonstrated that conventional security models are fundamentally inadequate, primarily because they fail to address the critical "data-in-use" vulnerability inherent to GenAI workloads. Addressing this gap necessitates the adoption of the **Confidential Zero-Trust Framework (CZF)**, a novel security paradigm for deploying Generative AI on Google Cloud.

By integrating the continuous, context-aware access controls of Zero-Trust Architecture with the hardware-enforced protections of Confidential Computing, the CZF delivers a robust, multi-layered security model. This framework directly mitigates the real-world attack vectors responsible for major healthcare breaches, establishes a clear and verifiable pathway to compliance with regulatory requirements such as HIPAA and GDPR, and provides cryptographic evidence of its security posture through remote attestation.

Beyond threat prevention, the broader significance of the CZF lies in its ability to enable innovation. By establishing a foundation of verifiable trust, the framework unlocks new and previously impractical opportunities for healthcare advancement, including secure multi-institutional research collaboration and precision medicine applications involving highly sensitive genomic data. The integration of Generative AI into clinical practice is no longer a question of *whether*, but *how*. The Confidential Zero-Trust Framework offers a comprehensive, verifiable response to this challenge, ensuring that the future of AI-driven healthcare is grounded in enduring principles of security, privacy, and trust.

Ultimately, the importance of the Confidential Zero-Trust Framework extends well beyond traditional threat prevention. By establishing a foundation of verifiable trust, it enables critical forms of healthcare innovation that were previously constrained by security limitations—ranging from secure, multi-institutional research collaboration to the advancement of precision medicine involving highly sensitive genomic data. The integration of transformative technologies such as Generative AI into medical practice is no longer a question of *whether*, but *how*. The Confidential Zero-Trust Framework offers a robust, verifiable, and comprehensive response to this challenge, ensuring that the future of AI-driven healthcare is anchored in enduring principles of security, privacy, and trust.

Doi:10.48047/ijiee.2025.15.8.10

References

1. TopolEJ.High-performancemedicine:theConvergenceofHumanandArtificialIntelligence. *NatureMedicine*2019;**25**:44–56.
2. DavenportT,KalakotaR.ThePotentialforArtificialIntelligenceinHealthcare.*Future Healthcare Journal* 2019; **6**: 94–98.
3. BeamAL,KohaneIS.BigDataandMachineLearninginHealthCare.*JAMA*2018;**319**:1317.
4. Rose S, Borchert O, Mitchell S, Connelly S. Zero trust architecture. *NIST Special Publication800-207* 2020. doi:https://doi.org/10.6028/nist.sp.800-207.
5. LehmanE,JainS,PichottaK,GoldbergY,WallaceBC.DoesBERTPretrainedonClinical Notes Reveal Sensitive Data? *arXiv (Cornell University)* 2021. doi:https://doi.org/10.18653/v1/2021.naacl-main.73.
6. Fredrikson M, Jha S, Ristenpart T. Model Inversion Attacks that ExploitConfidenceInformation and Basic Countermeasures. *Proceedings of the 22nd ACMSIGSACConferenceonComputer and Communications Security - CCS '15* 2015. doi:https://doi.org/10.1145/2810103.2813677.
7. Jagielski M, Oprea A, Biggio B, Liu C, Nita-Rotaru C, Li B. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. *2018 IEEE Symposium on Security and Privacy (SP)* 2018. doi:https://doi.org/10.1109/sp.2018.00057.
8. d'AlibertiL,GronbergE,KovbaJ.Privacy-EnhancingTechnologiesforArtificial Intelligence-Enabled Systems. arXiv.org. 2024. doi:https://doi.org/10.48550/arXiv.2404.03509.
9. Raluca Ada Popa. Confidential Computing or Cryptographic Computing? *Queue* 2024; **22**: 108–132.
10. Atri P. Enhancing Big Data Security through Comprehensive Data Protection Measures: A Focus on Securing Data at Rest and In-Transit. *International journal of computing and engineering* 2024; **5**: 44–55.
11. Huang K, GoertzelB,WuD,XieA.GenAIModelSecurity.*Futureofbusinessandfinance*2024; 163–198.
12. Huang K, Huang J, Catteddu D. GenAI Data Security. *Future of business and finance* 2024; 133–162.
13. Onome Christopher Edo, Ang D, Praveen Billakota, Ho JC. A zero trust architecture for health information systems. *Health and Technology* 2023. doi:https://doi.org/10.1007/s12553-023-00809-4.
14. Mulligan DP, Petri G, Spinale N, Stockwell G, Vincent HJM. Confidential Computing—a brave new world. IEEE Xplore. 2021; 132–138.
15. Moore W, Frye S. Review of HIPAA, Part 1: History, Protected Health Information, and Privacy and Security Rules. *Journal of Nuclear Medicine Technology* 2019; **47**: 269–272.
16. Kaplan B. PHI Protection under HIPAA: An Overall Analysis. papers.ssrn.com. 2020.https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3833983.
17. Choi YB, Williams CE. A HIPAA Security and Privacy Compliance Audit and Risk Assessment Mitigation Approach. Research Anthology on Securing Medical Systems and Records. 2022.https://www.igi-global.com/chapter/a-hipaa-security-and-privacy-compliance-audit-and-risk-assessment-mitigation-approach/309023.
18. Custers B, Heijne A-S. The right of access in automated decision-making: The scope of article

Doi:10.48047/ijiee.2025.15.8.10

15(1)(h) GDPR in theory and practice. *Computer Law & Security Review* 2022; **46**: 105727.

19. BygraveLA.DataProtectionbyDesignandbyDefault:DecipheringtheEU'sLegislative Requirements. *Oslo Law Review* 2017; **1**: 105–120.
20. AbdelkebirSahid.SecuringHealthcare.*CRCPresseBooks*2024;196–244.
21. ShurenJ,PatelB,GottliebS.FDARegulationofMobileMedicalApps.*JAMA*2018;**320**: 337.
22. Nandakumar K, Vinod V,AkbarBatchaSM,SharmaDK,ElangovanM,PooniaA*etal.*Securing data in transit using data-in-transit defender architecture for cloud communication. *Soft Computing* 2021; **25**: 12343–12356.
23. Lee CH, Lim KH, Sivaraman Eswaran. A comprehensive survey on secure healthcare data processing with homomorphic encryption: attacks and defenses. *Deleted Journal* 2025; **22**. doi:https://doi.org/10.1186/s12982-025-00505-w.
24. AswathySU,TyagiAK.PrivacyBreachesthroughCyberVulnerabilities.*CRCPresseBooks* 2022;163–210.
25. Ankalaki S, Aparna Rajesh A, Pallavi M, Hukkeri GS, Jan T, Naik GR. CyberAttackPrediction: From Traditional Machine Learning to Generative Artificial Intelligence. *IEEE Access* 2025; 1–1.
26. LeeD,TiwariM. Prompt Infection: LLM-to-LLM Prompt Injection within Multi-Agent Systems. arXiv.org. 2024.https://arxiv.org/abs/2410.07283 (accessed 24 Jul2025).
27. SarkarS.SecurityofZeroTrustNetworksinCloudComputing:AComparativeReview. *Sustainability*2022;**14**:11213.
28. Oladimeji G. A Critical Analysis of Foundations, Challenges and Directions for Zero Trust Security in Cloud Environments. arXiv.org. 2024.https://arxiv.org/abs/2411.06139.
29. Tyler D, Viana T. Trust No One? A Framework for Assisting Healthcare Organisations in Transitioning to a Zero-Trust Network Architecture. *Applied Sciences* 2021; **11**: 7499.
30. Kanter GP, Rekowski JR, Kannarkat JT. Lessons From the Change Healthcare Ransomware Attack. *JAMA Health Forum* 2024; **5**: e242764.
31. American Medical Association. Change Healthcare cyberattack. American MedicalAssociation. 2024.https://www.ama-assn.org/practice-management/sustainability/change-healthcare-cyberattack.
32. Ghasemshirazi S, Shirvani G, Alipour MA. Zero Trust: Applications, Challenges, and Opportunities. arXiv.org. 2023. doi:https://doi.org/10.48550/arXiv.2309.03582.
33. National Cyber Security Centre. MOVEit vulnerability and data extortion incident. www.ncsc.gov.uk. 2023.https://www.ncsc.gov.uk/information/moveit-vulnerability.
34. Adesola H, Chen L, Ji Y, Kim J. Application of Robotics Process Automation to the MOVEit Attack: A Case Study. *Communications in Computer and Information Science* 2025; 503–515.
35. Akinsola A, Akinde A. Enhancing Software Supply Chain Resilience: Strategy For Mitigating Software Supply Chain Security Risks And Ensuring Security Continuity In Development Lifecycle. *International Journal on Soft Computing* 2024; **15**: 01-18.
36. Molitor D, Raghupathi W, Saharia A, Raghupathi V.ExploringKeyIssuesinCybersecurityData Breaches:AnalyzingDataBreachLitigationwithML-BasedTextAnalytics.*Information*2023;**14**: 600.
37. Quazi F, Khanna A, Suryaprakash nalluri, Naveena Gorrepati. Data Security & Privacy in Healthcare. 2024. doi:https://doi.org/10.21428/e90189c8.4e2c586a.

Doi:10.48047/ijiee.2025.15.8.10