FRACTO

HIPAA-Compliant AI Architecture Blueprint

A Practical Guide for Healthcare Leaders

Cloud Hybrid Edge RACI Policy Templates Checklist

What's Inside

- Reference architecture diagrams for cloud, hybrid and edge deployments
- Data-flow examples for EHR, imaging and monitoring use cases
- RACI matrix for security, compliance, data and clinical owners
- Policy templates for de-identification, access control and audit logging
- 25-point HIPAA-safe AI deployment checklist

2026 Edition | Healthcare AI Series

fracto.ie • rahul@fracto.ie

Table of Contents

SECTION 00

Introduction

Hospitals and health systems are no longer asking *if* they should use AI — they are asking **how to deploy it safely in production** without violating HIPAA or overwhelming clinical workflows. By 2024, an estimated 71% of US hospitals were already using predictive AI integrated with their EHRs, and the FDA had cleared or approved over 1,000 AI/ML-enabled medical devices.

Yet architecture failures — not algorithmic ones — remain the leading cause of compliance breaches. This blueprint distils the patterns, frameworks and templates that Fracto uses with healthcare clients to move from AI concept to HIPAA-compliant production.

Who this is for	CIOs, CTOs, CMIOs, Chief Data Officers, Digital Health leaders, compliance and security teams responsible for deploying or governing AI in hospitals and health systems.
How to use it	Work through each section in order for a full implementation journey, or jump directly to the RACI, policy templates or checklist for immediate operational use.

SECTION 01

Reference Architecture — Seven-Layer Model

A HIPAA-compliant healthcare AI system is not a single product — it is an end-to-end architecture of seven integrated layers. Each layer must be individually secured and governed; weakness in any one layer creates compliance risk for the entire system.

1 Data Sources & Ingestion

- EHR (HL7/FHIR), DICOM imaging, lab systems, IoT/edge sensors
- Ingestion via integration engines (Mirth, Rhapsody) or FHIR APIs
- PHI tagged and classified at point of ingestion

2 Secure Data Lake / Warehouse

- AES-256 encryption at rest; keys in HSM-backed KMS
- Segmentation of raw PHI, de-identified datasets, and analytical marts
- Row- and column-level security with fine-grained access controls

3 De-Identification & Tokenisation Layer

- HIPAA Safe Harbor or Expert Determination depending on use case
- Pseudonymisation or tokenisation for training datasets
- Controlled re-linkage process for approved clinical contexts

4 AI/ML Workbench & Training Environment

- Private subnets inside dedicated VPCs; no public IPs on training nodes
- Access only via bastion hosts, VPN or zero-trust proxies
- Training on de-identified or minimally necessary PHI only

5 Model Serving & Application Layer

- Containerised model servers in private subnets (Kubernetes / serverless)
- API gateway enforcing mTLS, JWT auth and rate-limiting
- EHR-integrated via SMART on FHIR; edge appliances for ICU/OR/bedside

6 Security, Monitoring & Audit

- Centralised SIEM logging for all PHI access, queries and inference calls
- AI-driven anomaly detection on access patterns in real time
- Audit trails retained minimum six years per HIPAA requirements

7 Governance & Business Associate Agreements

- Signed BAAs with all cloud and AI vendors handling PHI
- Vendor risk assessments: encryption, incident history, subcontractors
- Documented shared responsibility models for every control domain

Note: Layers 4 and 6 must never share network segments. A hard boundary between training environments and production inference clusters is a non-negotiable HIPAA control.

SECTION 02

Deployment Models — Cloud, Hybrid & Edge

Healthcare AI deployments are not one-size-fits-all. The right model depends on data sensitivity, latency requirements, connectivity and clinical context. Most enterprise health systems operate a **hybrid** model combining all three.

	Cloud-First	Hybrid	Edge / On-Premises
Best for	Population health, admin AI	Most enterprise health systems	ICU, OR, bedside monitoring
PHI handling	Encrypted in HIPAA-eligible cloud; BAA req.	Sensitive PHI stays on-prem or in VPC	PHI stays on local devices
Latency	100–500 ms	50–200 ms	< 20 ms
Resilience	Depends on WAN/internet	High (local fallback)	Highest (fully offline capable)
Cost profile	OpEx; scales with workload	Mixed: OpEx + CapEx	Higher CapEx; lower egress
Compliance risk	Shared resp.; vendor review	Moderate; clear boundary controls	Lower data movement risk
Recommended BAA	Required	Required	Required

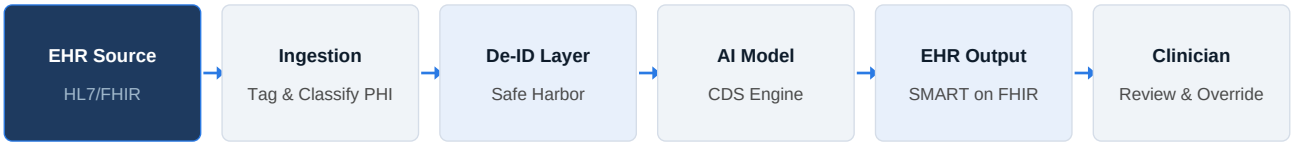
Cloud	Always ensure the cloud provider signs a BAA covering the specific services handling PHI. 'HIPAA-eligible' infrastructure ≠ HIPAA-compliant system. Review shared responsibility models carefully.
Hybrid	Define clear data-boundary controls: which PHI stays on-premises, which moves to cloud, and under what conditions. Network segmentation between environments is mandatory.
Edge	Edge devices holding PHI require hardened OS configurations, encrypted storage, remote device management, and regular security patching — often more operationally demanding than cloud.

SECTION 03

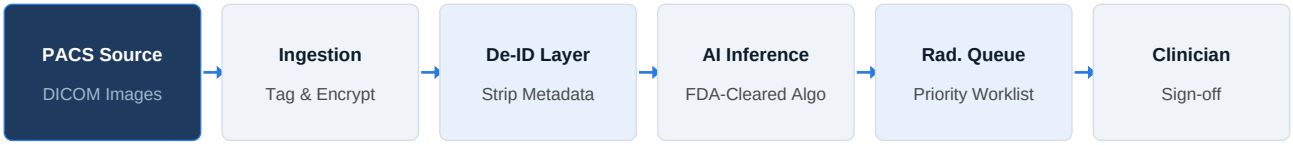
Data Flow Examples

The following diagrams illustrate how PHI moves through a HIPAA-compliant AI pipeline for three core healthcare AI use cases. In each case, PHI is classified, encrypted and de-identified as early as possible in the flow.

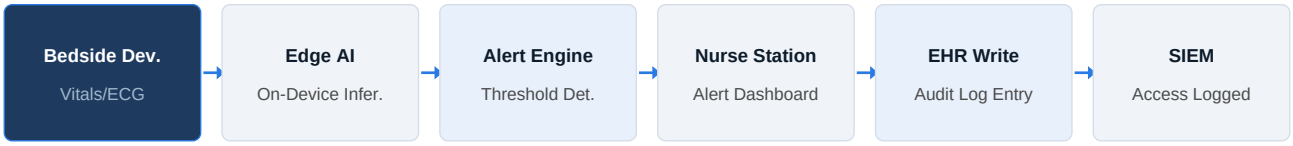
Flow 1 — EHR & Clinical Decision Support



Flow 2 — Medical Imaging (DICOM / PACS)



Flow 3 — ICU / Bedside Monitoring (Edge)



In all flows: PHI must be encrypted in transit (TLS 1.2+), at rest (AES-256), and access must be logged to a SIEM with tamper-evident audit trails retained for ≥ 6 years.

SECTION 04

RACI Matrix

Use this RACI to assign clear accountability across your security, compliance, data engineering and clinical teams for every key control domain.

R = Responsible	A = Accountable	C = Consulted	I = Informed

Activity / Control Area	Security	Compliance	Data	Clinical
PHI classification at ingestion	R	A	C	I
De-identification method selection	C	A	R	C
Network architecture & VPC design	R	A	I	I
Encryption key management (KMS/HSM)	R	A	C	I
BAA execution with vendors	C	A	I	I
AI model training environment setup	R	C	A	I
Model bias & subgroup validation	C	C	R	A
EHR / FHIR workflow integration	C	I	R	A
SIEM logging configuration	R	A	C	I
Audit trail review (6-yr retention)	R	A	C	I
Incident response plan & testing	R	A	C	C
FDA device classification review	C	A	C	R
Clinician training on AI tools	I	C	I	R
Post-deployment monitoring	R	C	A	C
Annual HIPAA risk analysis	C	A	R	I
Vendor security posture review	R	A	C	I
Model change control (PCCPs)	C	A	R	C
Patient safety escalation design	C	C	I	A

SECTION 05

Policy Templates

The following templates provide a starting framework for the three core operational policies required before any AI system handling PHI goes live. Adapt each template to your organisation's specific systems, vendors and risk appetite.

Policy Template 1 — PHI De-Identification & Tokenisation

1. Scope: All AI model training, testing and fine-tuning activities that involve patient data from EHR, PACS, LIS or monitoring systems.
2. Default method: Apply HIPAA Safe Harbor de-identification (45 CFR §164.514(b)) as the baseline. Use Expert Determination only when Safe Harbor is insufficient and a qualified statistician confirms negligible re-identification risk.
3. Tokenisation: Assign a unique, non-reversible token to each patient record used in AI training. Store the mapping in a separate, HSM-protected vault restricted to two named custodians.
4. Re-linkage: Re-linking tokens to PHI is permitted only for a prospective clinical validation under an approved IRB protocol or a covered treatment/operations purpose. Every re-linkage event must be logged.
5. Prohibited: Exporting raw, identifiable PHI to any third-party AI API, cloud storage or development environment not covered by a signed BAA.
6. Review cycle: Policy reviewed annually or upon any material change to data sources, vendors or AI use cases.

Policy Template 2 — Access Control for AI Systems Handling PHI

1. Principle: Least privilege by default. No user, service account or AI system may access more PHI than is strictly necessary for its defined function.
2. Authentication: All human access to AI training environments, data lakes and inference APIs requires MFA and SSO integration with the organisation's identity provider.
3. Role-based controls: Access roles must be defined per architectural layer and reviewed quarterly. Privileged roles require dual-approval for changes.
4. Service accounts: AI pipelines must use short-lived credentials (IAM roles, workload identity). Permanent, embedded secrets in code or containers are prohibited.
5. Emergency access: Break-glass access to PHI environments must be pre-approved, individually logged, and reviewed within 24 hours by the CISO or delegate.
6. Termination: Access must be revoked within four hours of departure. Quarterly access recertification is mandatory for all PHI-touching roles.

Policy Template 3 — Audit Logging & Monitoring for AI Systems

1. Coverage: Logging must capture all PHI access (read, write, query, export), all inference API calls, all administrative actions, and all model deployments.
2. Log format: Logs must include timestamp (UTC), user/service identity, action type, data resource accessed, source IP and session ID. Structured JSON is recommended.
3. Retention: All logs must be retained for a minimum of six years from the date of creation, in an immutable, tamper-evident store (WORM storage or chained logs).
4. Monitoring: Real-time alerting must flag: abnormal query volumes, access from unexpected geographies, after-hours PHI access, and failed authentication above threshold.
5. Incident response: SIEM alerts must be triaged within four hours. Any confirmed PHI access anomaly triggers the breach response protocol within 24 hours.
6. Review: Logging configuration and alert ruleset must be tested at least annually and after any significant system change.

SECTION 06

25-Point HIPAA-Safe AI Deployment Checklist

Complete this checklist before any AI system touching PHI goes live. If you cannot tick the majority of these items, the system is not yet ready for safe, compliant deployment.

GOVERNANCE & OWNERSHIP

- ■ Clinical champion and business owner identified and named
- ■ AI use case classified for FDA relevance and risk tier
- ■ AI governance council or steering committee formally in place
- ■ Organisational policy for acceptable AI use published and communicated

DATA & PRIVACY

- ■ Data inventory completed with all PHI elements mapped and classified
- ■ De-identification or pseudonymisation applied for model training data
- ■ Minimum necessary PHI principle enforced for all inference calls
- ■ Data retention and deletion policies defined, documented and tested

ARCHITECTURE & SECURITY

- ■ AI compute runs in private subnets with no public-facing IP addresses
- ■ Encryption at rest (AES-256) and in transit (TLS 1.2/1.3) enforced on all PHI
- ■ Separate dev / test / prod environments with restricted cross-environment access
- ■ Zero-trust access controls and MFA implemented for all human and service access

VENDORS & BAAs

- ■ BAAs executed with every cloud, AI and integration vendor touching PHI
- ■ Vendor security and compliance posture independently reviewed
- ■ All subprocessors handling PHI disclosed and contractually bound
- ■ Distinction between 'HIPAA-eligible' and 'HIPAA-configured' services documented

MODEL & WORKFLOW

- ■ Model performance validated on local population data before go-live
- ■ Bias and subgroup performance assessed and documented where relevant
- ■ Human-in-the-loop or override paths clearly defined and tested
- ■ AI outputs integrated into clinicians' existing tools (EHR, PACS, dashboards)

MONITORING & INCIDENT RESPONSE

- ■ Comprehensive logging enabled for data access, inference calls and admin actions
- ■ SIEM integration and alerting configured for anomalous access patterns
- ■ Model performance, drift and safety continuously monitored post-deployment
- ■ Incident response plan tested, including simulated PHI breach scenario
- ■ Audit trail retention policy confirmed at minimum six years

SECTION 07

Next Steps with Fracto

This blueprint gives you the architecture reference, deployment models, RACI, policy templates and checklist to begin a structured, compliant healthcare AI programme. The next step is a structured assessment of where your organisation stands today.

Healthcare AI Readiness & Compliance Assessment

A structured 3–5 day engagement with your technology, compliance and clinical leadership teams. Deliverables include:

- Current-state assessment of your data, architecture and governance maturity
- Identification of high-ROI, low-risk AI deployment opportunities
- Mapping of HIPAA, FDA and NIST AI RMF obligations to your roadmap
- A 90-day implementation plan with prioritised, costed actions

Download the Companion Resources

The following resources complement this blueprint and are available on request:

- Editable RACI spreadsheet (Excel)
- Policy template pack (Word)
- Architecture diagram source files (draw.io / Visio)
- Vendor BAA evaluation scorecard

Contact Fracto	**Email:** rahul@fracto.ie **Web:** fracto.ie	Fracto is an AI & digital transformation advisory firm specialising in enterprise AI deployment, governance and compliance for healthcare and regulated industries.