Policy to practice

Building Safely on Healthcare Systems

IMPLEMENTATION SUPPLEMENT · 5 OCTOBER 2026

How to use this edition

This two-page implementation supplement accompanies the original teaching guide. It records a coherence review against selected WHO and ICRC references. It does not certify legal compliance, clinical safety or institutional approval. The original guide follows unchanged; the qualifications below govern how its broad rules and regulatory summaries should be interpreted.

Scope comes before compliance

WHO guidance, ICRC organizational policies and national law have different scopes. Identify intended use, affected people, jurisdiction and organizational responsibilities first. Treat the original legal tables, dated regulatory timelines and categorical agreement statements as questions for qualified reviewers, not automatic conclusions. A vendor, contract, certification claim or passing test alone does not establish compliance.

Synthetic data is our learning boundary

This workspace accepts invented examples and non-personal public information. External guidance may discuss real-world health or humanitarian operations; that does not authorize real patient or confidential institutional data here. A proposed move beyond a demonstration needs the institution’s appropriate data, privacy, security and clinical owners, and research oversight where applicable.

Security without over-collection

Keep secrets out of code and prompts; enforce authorization for each record on the server. Publicly reachable service endpoints are not inherently forbidden, but unprotected data and privileged administrative access must not be exposed. Use minimal, access-controlled event logs. Interpret any instruction to log everything or retain a full trace as an auditable record of necessary events, never raw sensitive payloads.

Rights, disclosure and human control

Record permission to use datasets, models, documents and code separately. Disclosure rules for WHO or ICRC do not authorize disclosure of a learner’s pitch. Check publication-specific licences and commercial-use terms. Explain AI involvement, name a responsible person, and give that person an effective way to inspect, stop, reject and reverse actions where feasible. A decorative approval button is insufficient.

A check for each principle

These are How I Do It implementation suggestions. They are not verbatim requirements from WHO or ICRC. See the dated originals for context; neither organization endorses this product.

Purpose and minimum data

WHO AI ethics, executive summary; ICRC AI policy §6.2

Record one purpose and the minimum fields. Test that unrelated fields and another account's records cannot be returned.

Guidance to implementation

WHO SMART DAK overview, pp.1–2

Trace the source version to workflow, data definitions, requirement and test. Clinical logic needs qualified validation.

Loss of control and cybersecurity

WHO LMM guidance §3.1–3.2; ICRC AI policy §6.2

Treat retrieved content as untrusted. Restrict tool permissions outside the model. Test hostile uploads and human cancellation.

Evidence and deviations

WHO regulatory considerations §5.1; ICRC AI policy §6.4

For each control, record expected behavior, evidence, gap, owner and next check. Label untested claims as unknown.

Permissions and publication

WHO copyright/website terms; ICRC copyright terms

Keep an asset register: source, licence, allowed use, attribution and permission status. Link to originals when rights are unresolved.

Decision record

Source and edition → requirement → design choice → evidence or gap → accountable owner → next review. Status: proposed / implemented (reported) / tested with evidence / deviates / unknown. Keep unresolved risks visible.

Sources and reuse

WHO SMART

WHO AI ethics

WHO LMM guidance

WHO regulatory considerations

ICRC AI policy

WHO rights

ICRC rights

Building Safely on Healthcare Systems: A Security and Architecture Guide for AI-Assisted Development

Oct 5, 2026 · @Salim Afshar

1. Purpose and how to use this guide

This guide sets out what a builder must know, decide, and put in place before software written with AI assistance touches a health system or the people it serves. It covers the United States and Europe, and it follows a staged path from first prototype to production.

AI-assisted development, often called vibe coding, lets a person describe what they want in plain language and receive working software. The speed is real. So is the distance between software that runs and software that deserves the trust of a patient. This guide exists to close that distance.

Three convictions shape everything that follows:

- **Trust is the foundation.** A person who shares health information is extending trust. Every design choice either honors that trust or spends it.
- **Capability carries responsibility.** The ability to build quickly does not reduce the duty to build carefully. The author of a prompt is the author of the product.
- **Safety is built in, not added on.** Protection that is designed from the first day costs little. Protection added after an incident costs a great deal, and the cost falls first on those least able to bear it.

Who this is for. Students and first-time builders, clinicians building inside a health system, founders selling to health systems, and hackathon teams working across countries. Section 13 speaks to each group directly.

How to read it. Sections 2 to 6 explain the rules. Sections 7 to 12 explain how to build. Section 14 gathers the checklists. A companion Markdown file carries the same guidance in a form an AI coding agent can load as standing instructions.

What this guide is not. It is orientation, not legal advice. Laws change, and their application depends on facts. Before real patient data is involved, consult the privacy, security, and compliance officers of the institution concerned, and qualified counsel.

2. Core definitions

Most compliance failures begin as vocabulary failures: a builder does not know that the data in front of them has a legal name and legal consequences. These are the terms to know.

United States

Term	Meaning	Why it matters to a builder
Protected health information (PHI)	Individually identifiable health information held or transmitted by a covered entity or its business associate. Electronic PHI is called ePHI.	If your tool touches PHI, HIPAA applies to you through a contract, whether or not you planned for it.
Covered entity	A health plan, a healthcare clearinghouse, or a provider that bills electronically.	Hospitals and clinics are covered entities. They cannot hand you PHI without an agreement.
Business associate	Any person or company that creates, receives, maintains, or transmits PHI on behalf of a covered entity. Subcontractors count.	A startup serving a hospital is a business associate. So is its cloud host and its AI model vendor.
Business associate agreement (BAA)	The contract that permits a business associate to handle PHI and binds it to HIPAA safeguards.	No BAA, no PHI. This applies to every vendor in the chain, including the AI coding tool and the model API.
De-identified data	Data stripped of identity by one of two methods: Safe Harbor (remove 18 listed identifiers) or Expert Determination (a statistician certifies very small risk).	De-identified data falls outside HIPAA. Removing names alone does not achieve it. Dates, zip codes, and free text carry identity.
Limited data set	PHI with direct identifiers removed but dates and some geography retained, shared under a data use agreement.	Common for research. It is still PHI.
Minimum necessary	The duty to use and disclose only the PHI needed for the task.	Design queries, screens, and prompts to pull only the fields required.
Breach	An impermissible use or disclosure of unsecured PHI, presumed reportable unless a risk assessment shows low probability of compromise.	Notification to individuals within 60 days, to HHS, and sometimes to media.

Term	Meaning	Why it matters to a builder
Consumer health data	Health information held by companies outside HIPAA, such as wellness apps. Defined by the FTC and by state laws such as Washington's My Health My Data Act.	An app with no hospital customer is still regulated.
Software as a medical device (SaMD)	Software intended for a medical purpose that is not part of a hardware device.	Intended use, not technology, decides whether the FDA regulates your product.
Clinical decision support (CDS)	Software that offers recommendations to clinicians. Excluded from device regulation only if it meets four statutory criteria.	Many AI tools cross this line without their authors noticing. See section 4.
Electronic health information (EHI)	The electronic PHI in a designated record set. The object of the information blocking rules.	Health systems and certified EHR vendors must not unreasonably block access to it.

Europe

Term	Meaning	Why it matters to a builder
Personal data	Any information relating to an identified or identifiable living person.	Far broader than PHI. It covers every sector and every holder.
Data concerning health	Personal data about physical or mental health, including the provision of care. A special category under GDPR Article 9.	Processing is prohibited unless a specific Article 9 condition applies, on top of an Article 6 lawful basis.
Controller	The party that decides why and how personal data is processed.	The hospital is usually the controller. A founder with a consumer app is the controller.
Processor	A party that processes data on the controller's behalf under a written contract (GDPR Article 28).	The European counterpart of the business associate. The contract is a data processing agreement (DPA).

Term	Meaning	Why it matters to a builder
Pseudonymised data	Data that cannot be attributed to a person without separately held additional information.	Still personal data. GDPR applies in full.
Anonymised data	Data from which a person cannot be identified by any means reasonably likely to be used.	Outside GDPR. The bar is higher than HIPAA Safe Harbor. Data that is de-identified in the US may remain personal data in Europe.
Data protection impact assessment (DPIA)	A documented assessment required before high-risk processing (GDPR Article 35).	Large-scale health data and new technology both trigger it. Assume one is needed.
International transfer	Movement of personal data outside the European Economic Area, permitted only with a safeguard such as an adequacy decision or standard contractual clauses.	Calling a US-hosted model API with European patient data is a transfer.
Medical device software (MDSW)	Software with a medical purpose under the Medical Device Regulation (MDR). Under Rule 11, most diagnostic or therapeutic decision software is Class IIa or higher.	Class IIa and above require a notified body. Self-certification is rare for clinical software.
AI system, provider, deployer	The AI Act's terms for the system, the party that places it on the market, and the party that uses it professionally.	A builder is usually a provider. A hospital using the tool is a deployer. Each carries duties.
High-risk AI system	An AI system that is a regulated product or its safety component (Annex I, includes medical devices) or falls in a listed use (Annex III).	Triggers risk management, data governance, logging, human oversight, and conformity assessment.
EHR system	Under the European Health Data Space (EHDS), software that stores, imports, exports, or displays priority categories of electronic health data.	EHR systems will need to meet common interoperability and logging requirements and carry CE marking.

Term	Meaning	Why it matters to a builder
Essential or important entity	Under the NIS2 Directive, an organization in a critical sector, including healthcare providers, bound to cybersecurity and incident reporting duties.	Your hospital customer will pass these duties to you through supply chain requirements.

Terms used in both places

- **Synthetic data.** Artificially generated records with no link to a real person. The safe default for every prototype.
- **Authentication and authorization.** Authentication proves who a user is. Authorization decides what that user may do. AI-generated code often delivers the first and omits the second.
- **Encryption in transit and at rest.** Protection of data while it moves across a network and while it is stored.
- **Audit log.** A tamper-resistant record of who accessed what, when, and from where.
- **Software bill of materials (SBOM).** An inventory of every component and library inside a piece of software.
- **Threat model.** A structured account of what could go wrong, who might cause it, and what stands in the way.

3. Triage: which rules apply to what you are building

Six questions, answered honestly before the first line of code, determine nearly every obligation in this guide. More than one answer can be yes, and the obligations add together.

Question	If yes, in the United States	If yes, in Europe
1. Will the tool touch real data about identifiable people?	Go to question 2. If no, you may prototype freely on synthetic data.	GDPR applies. Health data needs an Article 9 condition and almost certainly a DPIA.
2. Does the data come from, or go to, a hospital, clinic, insurer, or their vendors?	HIPAA applies. You are a covered entity's workforce member or a business associate. A BAA is required with every vendor in the path.	You are likely a processor. A data processing agreement with the controller is required.
3. Do you collect health data directly from consumers, outside any provider relationship?	The FTC Act, the Health Breach Notification Rule, and state consumer health data laws apply.	You are the controller. Explicit consent is the usual Article 9 condition.
4. Does the tool inform or drive a clinical decision about an individual patient?	It may be a medical device. Test it against the four CDS criteria in section 4.	It is probably medical device software under the MDR, often Class IIa or above.

Question	If yes, in the United States	If yes, in Europe
5. Does the tool use AI to produce outputs that influence care, access, or eligibility?	Expect health system AI governance review, state AI laws, and FDA expectations if it is a device.	The AI Act applies. Device AI is high-risk. Transparency duties apply now.
6. Will the tool connect to an EHR or exchange records with other systems?	Use the certified FHIR APIs. Information blocking rules shape what the health system must allow.	Plan for the EHDS interoperability format and its phased obligations from 2027.

Three rules of thumb for the undecided.

- If you are unsure whether data is identifiable, treat it as identifiable.
- If you are unsure whether your tool is a medical device, write down its intended use in one sentence and show it to a regulatory specialist. The sentence decides the answer.
- If you serve people in both jurisdictions, build to the stricter requirement on each point. Section 6 shows which is stricter.

4. United States requirements

The United States has no single health data law. Obligations depend on who holds the data, what the software is intended to do, and which state the person lives in.

HIPAA

HIPAA governs covered entities and their business associates. It has three working parts.

- **Privacy Rule.** Sets when PHI may be used or disclosed. Treatment, payment, and operations are permitted. Most other uses, including product development and model training, need patient authorization, de-identification, or a waiver from an institutional review board.
- **Security Rule.** Requires administrative, physical, and technical safeguards for ePHI. The anchor requirement is a documented, organization-wide risk analysis. Access control, audit controls, integrity, authentication, and transmission security follow from it.
- **Breach Notification Rule.** Requires notice to affected individuals without unreasonable delay and within 60 days, with notice to HHS and, above 500 people, to the media.

Status of the Security Rule update. HHS proposed a major strengthening in January 2025. It would make encryption, multi-factor authentication, asset inventories, network segmentation, vulnerability scanning every six months, and annual penetration testing mandatory. As of this guide's date it is still a proposal, with final action now targeted for July 2027. The existing rule remains in force and is actively enforced. Build to the proposed controls anyway. They describe what a reasonable security program looks like, and health system procurement teams already ask for them.

Substance use disorder records. Records from federally assisted treatment programs carry stricter rules under 42 CFR Part 2. Compliance with the aligned rule has been required since 16 February 2026. Segment this data and track consent separately.

Outside HIPAA: the FTC and the states

A health app with no covered entity behind it is not free of regulation.

- **FTC Act, Section 5.** Deceptive or unfair data practices are unlawful. A privacy promise in an app store listing is enforceable.
- **Health Breach Notification Rule.** Vendors of personal health records and related apps must notify users, the FTC, and sometimes the media after unauthorized disclosure. Sharing data with advertisers without authorization has been treated as a breach.
- **State consumer health data laws.** Washington's My Health My Data Act requires separate consent for collection and for sharing, and gives individuals a private right of action. Nevada and other states have followed. California adds the Confidentiality of Medical Information Act and the CCPA.
- **State AI laws.** Several states now regulate AI in clinical communication, utilization review, and consequential decisions. Check the states where your users live.

FDA: when software becomes a medical device

Software intended to diagnose, treat, mitigate, or prevent disease is a device unless an exclusion applies. For decision support, the exclusion requires all four of the following:

1. The software does not acquire, process, or analyze a medical image, a signal from a diagnostic device, or a pattern from a signal acquisition system.
2. It displays, analyzes, or prints medical information about a patient or other medical information.
3. It supports or provides recommendations to a healthcare professional. It does not address patients or caregivers directly.
4. It enables the professional to independently review the basis for the recommendation, so that the professional does not rely primarily on it.

The FDA revised its CDS and general wellness guidance on 6 January 2026. The revision widens enforcement discretion in some cases, including software that offers a single recommendation where only one option is clinically appropriate. The four criteria still stand. A generative model whose reasoning a clinician cannot inspect will struggle with the fourth.

If your product is a device, expect design controls under the Quality Management System Regulation, premarket cybersecurity documentation including an SBOM, and, for adaptive AI, a predetermined change control plan.

Interoperability and information blocking

- **Certified APIs.** Certified EHRs must expose a standardized FHIR API using US Core profiles and SMART on FHIR authorization. This is your sanctioned door into the record.
- **Information blocking.** The 21st Century Cures Act prohibits practices likely to interfere with access, exchange, or use of electronic health information, subject to defined exceptions for privacy, security, and infeasibility.
- **Direction of travel.** The HTI-5 proposed rule of December 2025 would narrow certification to a FHIR-centered core, remove AI model card requirements for decision support, and extend the

definitions of access and use to automated and agentic systems. Confirm its final status before relying on any detail.
- **TEFCA.** The national exchange framework now offers a route to cross-network data, including a FHIR-based path. Participation carries its own security terms.

What a US health system will ask you for

Regardless of what the law strictly requires, expect these in procurement:

- A signed BAA and a list of your subcontractors
- A completed security questionnaire and evidence of SOC 2 Type II or HITRUST certification
- A recent penetration test report
- Proof of encryption, multi-factor authentication, and audit logging
- Review by the institution's AI governance committee if the tool uses a model
- Review by the institutional review board if the work is research

5. European requirements

Europe regulates by layers. One horizontal privacy law covers all personal data, and product, AI, health data, and cybersecurity laws stack on top of it. A single clinical AI tool can answer to four or five of them at once.

GDPR

The General Data Protection Regulation applies to any organization processing personal data of people in the EU, wherever the organization sits. The UK applies a near-identical regime.

- **Two legal gates.** Health data needs a lawful basis under Article 6 and a separate condition under Article 9. Common Article 9 conditions are explicit consent, provision of health care under professional secrecy, public health, and scientific research with safeguards. Member states add their own rules, so the answer can differ between Germany, France, and Ireland.
- **Principles that shape design.** Purpose limitation, data minimisation, accuracy, storage limitation, integrity and confidentiality, and accountability. Data collected for care cannot be quietly reused to train a model.
- **Data protection by design and by default (Article 25).** Privacy protections must be built into the system and switched on without user action.
- **DPIA (Article 35).** Required before processing begins. Treat it as a design document, not a form.
- **Processor contracts (Article 28).** Every vendor handling the data needs a data processing agreement, and the controller must approve sub-processors.
- **Individual rights.** Access, rectification, erasure, restriction, portability, and objection. Article 22 restricts decisions based solely on automated processing that significantly affect a person.
- **Breach notification.** To the supervisory authority within 72 hours of awareness. To individuals without undue delay when the risk is high.
- **International transfers.** Data leaving the European Economic Area needs an adequacy decision, standard contractual clauses with a transfer impact assessment, or another Chapter V safeguard. The EU-US Data Privacy Framework covers certified US companies and has faced continuing legal challenge. Verify its status and prefer EU-region hosting and processing for health data.

- **Penalties.** Up to 20 million euros or 4 percent of worldwide annual turnover, whichever is higher.

The AI Act

Regulation (EU) 2024/1689 sorts AI systems by risk. Its timeline changed in 2026, and many published summaries are out of date.

Obligation	Applies from	Note
Prohibited practices and AI literacy duty	2 February 2025	In force. Staff who operate AI must be trained to a sufficient level.
General-purpose AI model obligations	2 August 2025	In force. Falls on model providers, but shapes what you can obtain from them.
Article 50 transparency duties	2 August 2026	In force. People must be told when they interact with an AI system. Synthetic content must be marked.
High-risk systems listed in Annex III	2 December 2027	Deferred from August 2026. Includes emergency triage and eligibility for health services.
High-risk AI in regulated products, including medical devices (Annex I)	2 August 2028	Deferred from August 2027.

The deferral was made by the Digital Omnibus on AI, adopted in June 2026 and in force since July 2026. Medical device obligations under the MDR are unaffected and apply today.

High-risk obligations, when they apply, include a risk management system, data governance, technical documentation, automatic logging, transparency to deployers, human oversight, and accuracy, robustness, and cybersecurity. Build the logging and oversight now. They are far cheaper to design in than to retrofit.

Medical Device Regulation and IVDR

Software with a medical purpose is a device under Regulation (EU) 2017/745. Rule 11 places software that informs diagnostic or therapeutic decisions in Class IIa at minimum, rising to IIb or III where a wrong output could cause serious harm or death. Class IIa and above need a notified body, a quality management system (ISO 13485), a software lifecycle process (IEC 62304), risk management (ISO 14971), clinical evaluation, and post-market surveillance. Cybersecurity expectations are set out in guidance MDCG 2019-16. A reform of the MDR was proposed in late 2025. Follow its progress.

European Health Data Space

Regulation (EU) 2025/327 creates common rules for the primary use of health data in care and its secondary use in research and innovation. It entered into force in March 2025 and applies in phases.

- **26 March 2027.** General application begins. Key implementing acts and technical specifications are due.
- **26 March 2029.** Exchange of the first priority categories, including patient summaries and electronic prescriptions, and the secondary use framework with health data access bodies.
- **26 March 2031.** Remaining categories, including medical images, laboratory results, and discharge reports.

Builders should note three consequences. EHR systems will need to conform to the European electronic health record exchange format, which is built on HL7 FHIR. Wellness apps that claim interoperability with EHRs will carry labelling duties. And secondary use will run through secure processing environments under a data permit, which offers a lawful and well-governed path to data for research and model development.

NIS2 and the Cyber Resilience Act

- **NIS2 Directive.** Healthcare providers, laboratories, and medical device manufacturers of sufficient size are essential or important entities. Duties include risk management measures, supply chain security, and incident reporting with an early warning within 24 hours and a notification within 72 hours. Management is personally accountable. Your customers will pass these duties to you by contract.
- **Cyber Resilience Act.** Regulation (EU) 2024/2847 sets security requirements for products with digital elements. Reporting of actively exploited vulnerabilities has applied since 11 September 2026, and full requirements, including an SBOM and CE marking, apply from 11 December 2027. Medical devices under the MDR are excluded. Health software that is not a medical device is generally in scope.

What a European health system will ask you for

- A data processing agreement and a list of sub-processors
- A completed DPIA, or the material to complete the controller's own
- Data residency within the EEA, or a documented transfer mechanism
- ISO 27001 certification, and where relevant national schemes such as France's HDS hosting certification, Germany's C5, or the UK's DSPT and DCB0129 clinical safety case
- CE marking and the declaration of conformity if the product is a device
- AI Act role analysis: who is the provider, who is the deployer, and what each will document

6. United States and Europe side by side

The two regimes protect the same thing by different means: the United States regulates by who holds the data, and Europe regulates the data itself. A team building for both should adopt the stricter column on each row.

Topic	United States	Europe	Build to
What is protected	PHI held by covered entities and business associates. Consumer health data under FTC and state law.	All personal data, with health data as a special category.	Europe
Who is bound	Covered entities, business associates, and some consumer app vendors.	Every controller and processor, in every sector.	Europe
Legal basis for use	Treatment, payment, operations, or authorization.	Article 6 basis plus Article 9 condition.	Europe
Vendor contract	Business associate agreement.	Data processing agreement under Article 28.	Both, each where it applies
De-identification	Safe Harbor (18 identifiers) or Expert Determination.	Anonymisation judged by all means reasonably likely to be used. Pseudonymised data stays regulated.	Europe
Risk assessment	Security risk analysis, required.	DPIA for privacy, plus security risk management under NIS2.	Do both
Breach notice to regulator	Within 60 days. Annual for breaches under 500 people.	Within 72 hours.	Europe
Individual rights	Access and amendment.	Access, rectification, erasure, portability, objection, limits on automated decisions.	Europe
Cross-border data	No general restriction under HIPAA. Contracts often require US hosting.	Restricted. Needs a transfer mechanism.	Keep data in its region

Topic	United States	Europe	Build to
Security controls	Flexible and risk-based today. Prescriptive controls proposed.	Appropriate technical and organisational measures. NIS2 names specific measures.	The proposed HIPAA controls plus NIS2 measures
Clinical software	FDA device rules. CDS exclusion with four criteria.	MDR. Rule 11 puts most decision software in Class IIa or higher.	Europe is stricter. More software is a device.
AI regulation	No federal AI statute. FDA guidance, state laws, institutional governance.	AI Act with dated obligations.	Europe
Interoperability	FHIR US Core, SMART on FHIR, information blocking rules, TEFCA.	EHDS and the European EHR exchange format, built on FHIR.	FHIR R4 with regional profiles
Product cybersecurity	FDA premarket cybersecurity for devices.	CRA for non-device software. MDR guidance for devices.	SBOM and vulnerability handling everywhere
Maximum penalty	Tiered civil penalties per violation, with annual caps, plus state and FTC action.	4 percent of worldwide turnover under GDPR. Up to 7 percent under the AI Act for prohibited practices.	n/a

One practical conclusion. A team that builds to GDPR principles, the proposed HIPAA security controls, FHIR standards, and the AI Act's logging and oversight requirements will find little left to add for either market.

7. The staged path from prototype to production

A tool earns access to more sensitive data by demonstrating more discipline. Each stage below has a gate, and no one passes a gate by enthusiasm alone.

Stage	Data allowed	Environment	Gate to enter	Who approves
0. Explore	None. Ideas, mock screens, public information.	Any AI tool. Personal laptop.	None.	The builder.
1. Prototype	Synthetic data only.	Consumer or team AI coding tools. Free-tier	Triage in section 3 completed. Intended use	The builder and a mentor or supervisor.

Stage	Data allowed	Environment	Gate to enter	Who approves
		hosting is acceptable.	written in one sentence.	
2. Validate	De-identified or anonymised data, or a vendor sandbox.	Institution-approved environment. Private repository. Secrets in a vault.	Data use agreement. Proof of de-identification. Threat model drafted. Security baseline from section 9 in place.	Data owner. Privacy office. IRB or ethics committee if research.
3. Pilot	Real patient data, limited to a defined group and period.	Compliant cloud under BAA or DPA. Enterprise AI tools with zero data retention. Full logging.	BAA or DPA signed with every vendor. Risk analysis and DPIA complete. Code reviewed by a qualified engineer. Penetration test passed. Device and AI Act classification confirmed. Clinical safety review.	Privacy, security, and compliance officers. AI governance committee. Clinical sponsor.
4. Production	Real patient data at scale.	Hardened, monitored, segmented production with tested backup and recovery.	Certifications held or in progress. Incident response plan rehearsed. Regulatory clearance if a device. Monitoring for model drift and misuse. Named owner for ongoing maintenance.	Executive sponsor. Contracting. Regulators where required.

Rules that hold at every stage

1. **Real data never enters a lower stage.** No patient record, screenshot, export, or pasted note goes into a tool or environment approved only for stages 0 to 2.

2. **Synthetic first.** Use generators such as Synthea, or the public sandboxes offered by EHR vendors and SMART Health IT, before asking for anything real.
3. **A prototype is evidence, not a product.** Code written to prove an idea is usually rewritten before a pilot. Plan for that from the start.
4. **Moving up is a decision by others.** The builder proposes. The institution's officers decide.
5. **Moving down is always open.** When something is uncertain, step back a stage, resolve it, and return.

What changes when the code is AI-generated

The gates do not change. The evidence required to pass them does. Because the builder may not have read every line, gates 2 and 3 need independent proof: automated scanning results, a human review by someone who can read the code, and tests that demonstrate each security control works. Section 10 sets out how.

8. Reference architecture

A sound architecture places every route to patient data behind the same three checks: who is asking, whether they may, and a record that they did.



reference architecture · 4 layers and a foundation

Read the drawing from top to bottom. The application server is highlighted because it is the single point where authorization is decided, and it is the component AI-generated code most often gets wrong.

Architectural principles

1. **Zero trust.** No request is trusted because of where it comes from. Each is authenticated and authorized on its own merits.
2. **Least privilege.** Every user, service, and agent holds the narrowest access that lets it do its work, for the shortest time.
3. **Defence in depth.** No single control is relied upon. A failure in one layer is caught by the next.
4. **Data stays home.** Leave data in the system of record where possible. Fetch what is needed, when it is needed, and keep little.
5. **Separation of environments.** Development, test, and production are distinct accounts with distinct credentials. Real data lives only in production.

6. **Regional boundaries.** European data is stored and processed in Europe, US data in the US, unless a documented mechanism permits otherwise.
7. **Everything is logged.** Access to data, administrative actions, and model exchanges all leave a trace that cannot be quietly altered.
8. **Infrastructure as code.** Define servers, networks, and permissions in reviewed files, so the environment can be inspected, scanned, and rebuilt.
9. **Graceful failure.** When a component fails, the system denies access and care continues by other means.

Choices that reduce risk for small teams

Need	Prefer	Avoid
Hosting	A major cloud's healthcare-eligible services under a BAA or DPA, in the correct region	Hobby hosting, personal accounts, or platforms that will not sign an agreement
Identity	The health system's identity provider, or an established identity service	Authentication code written by the assistant
Database	Managed database with encryption, private networking, and automated backup	A database with a public address, or row-level security left off
Health data storage	A managed FHIR service where one fits	A custom schema that must later be mapped to FHIR
Secrets	The cloud's secrets manager	Environment files in the repository
Model access	Enterprise endpoints through a gateway you control	Direct calls from the browser to a model API
Front end	Server-rendered or a thin client that holds no secrets and no bulk data	Business logic and access rules in client code
Observability	Centralized logging with PHI filtered out	Third-party analytics or session replay on pages that show patient data

Documents the architecture must produce

A data flow diagram showing every place patient data rests or moves, a list of every vendor that touches it, a threat model, and a record of which control answers which threat. These four documents serve the

HIPAA risk analysis, the DPIA, the health system's security review, and the AI Act's technical documentation alike.

9. Security control baseline

These controls are the minimum for any tool at stage 2 or above. They draw on the HIPAA Security Rule and its proposed update, NIS2, the NIST Cybersecurity Framework 2.0, and the OWASP Application Security Verification Standard.

Area	Control	How to verify it
Identity	Single sign-on through the institution's identity provider. Multi-factor authentication for every user and administrator. No shared accounts.	Attempt login without the second factor. It must fail.
Authorization	Role-based access enforced on the server for every request. Each record checked against the requesting user. Deny by default.	Log in as user A and request user B's record by changing the ID in the URL. It must fail.
Sessions	Short-lived tokens. Automatic timeout. Secure, HttpOnly, SameSite cookies.	Inspect cookies and token lifetimes.
Encryption in transit	TLS 1.2 or higher on every connection, internal and external.	Scan endpoints. Reject plain HTTP.
Encryption at rest	AES-256 or equivalent on databases, file stores, backups, and device storage. Keys held in a managed key service, separate from the data.	Review storage configuration. Confirm keys are not in code.
Secrets	No credentials, API keys, or tokens in source code, prompts, or client-side code. Use a secrets manager. Rotate on any suspicion.	Run a secret scanner on the full repository history.
Input and output	Validate all input on the server. Use parameterized queries. Encode output. Limit file types and sizes.	Static analysis plus injection tests.

Area	Control	How to verify it
Audit logging	Record every access to patient data: who, what, when, from where, and why where available. Logs are append-only and retained per policy, commonly six years in the US.	Open a record, then find the log entry. Confirm the log holds no PHI beyond identifiers.
Data minimisation	Collect and return only required fields. Mask by default. Set retention periods and delete on schedule.	Compare API responses to what the screen needs.
Network	Private subnets for data stores. No database exposed to the internet. Segmentation between environments. Web application firewall and rate limiting.	External scan. Confirm storage buckets are not public.
Dependencies	Lock versions. Scan for known vulnerabilities. Maintain an SBOM. Verify that every package exists and is reputable.	Software composition analysis in the build pipeline.
Environments	Separate development, test, and production. No production data in lower environments.	Review access lists and data in each.
Backup and recovery	Encrypted, tested backups. Documented recovery time. The proposed HIPAA rule sets a 72-hour restoration target for critical systems.	Restore from backup on a schedule.
Vulnerability management	Scanning at least every six months. Penetration test at least annually and before pilot. Timely patching.	Dated reports with findings closed.
Incident response	Written plan with named roles. Contract terms for notifying the health system promptly, commonly within 24 to 72 hours.	Run a tabletop exercise.

Area	Control	How to verify it
Error handling	Fail closed. Generic messages to users. No stack traces, record contents, or keys in errors or logs.	Trigger errors and read what comes back.

Two controls deserve particular attention in AI-built applications. Server-side authorization is the control most often missing, because generated code tends to check that a user is logged in without checking what that user may see. Backend-as-a-service platforms are the most common source of exposure, because their row-level security is frequently left off or misconfigured, and their keys end up in browser code. Test both by hand.

10. AI-assisted coding: specific risks and guardrails

AI coding tools produce software that works more reliably than they produce software that is safe. The evidence is consistent across independent studies.

- Veracode tested more than 100 language models and found that 45 percent of generated code samples introduced a vulnerability from the OWASP Top 10, a rate that did not improve through early 2026.
- A scan of applications built on popular vibe coding platforms found more than 2,000 critical vulnerabilities, over 400 leaked secrets, and 175 exposures of personal information, including medical records.
- A May 2026 scan of more than 5,000 publicly deployed AI-built apps found that about 40 percent exposed sensitive data, medical information among it.
- Georgia Tech's Vibe Security Radar attributed 6 published vulnerabilities to AI-generated code in January 2026 and 35 in March 2026.

The lesson is not to stop. It is to match the speed of creation with an equal discipline of verification.

The risks

Risk	What happens	Guardrail
Data leaves through the tool	PHI pasted into a prompt, or present in a file the assistant reads, is sent to a vendor with no BAA or DPA.	Synthetic data in development. Enterprise tools with zero retention and a signed agreement before any real data is near the workspace. Exclude data folders from the assistant's reach.
Missing authorization	The app checks login but not ownership. Any user can read any record.	Write the access rules into the specification. Test by hand with two accounts.

Risk	What happens	Guardrail
Exposed secrets	Keys are written into code or shipped to the browser.	Secrets manager. Pre-commit secret scanning. Rotate anything that was ever committed.
Open backend	Database or storage bucket reachable from the internet. Row-level security off.	Infrastructure reviewed by an engineer. External scan before every release.
Invented or malicious packages	The model names a library that does not exist. An attacker registers that name with harmful code.	Verify every new dependency. Lock versions. Scan in the pipeline.
Outdated patterns	Weak hashing, old TLS settings, deprecated libraries.	Static analysis and dependency scanning on every change.
Security erodes with iteration	Repeated prompting to fix or extend code removes earlier protections.	Tests for each security control, run automatically, so removal is caught.
Prompt injection of the coding agent	Instructions hidden in a web page, document, issue, or package cause the agent to run commands or leak files.	Run agents in a sandbox with least privilege. No production credentials. Human approval for commands, network access, and merges.
Agent with excess authority	An agent with broad tool access deletes data or changes infrastructure.	Read-only by default. Scoped, short-lived credentials. Separate accounts for agents.
No one understands the system	The builder cannot explain how the code handles data, so cannot complete a risk analysis or answer an auditor.	Require the assistant to produce architecture notes and data flow descriptions. Have a qualified person read the code before pilot.
Logging that leaks	Generated code writes full records to logs or error trackers.	Review log statements. Route logs to a protected store.
Unclear provenance	No record of which code was generated, by which tool, from which instruction.	Keep prompts, specifications, and review notes in the repository.

Working practices

1. **Specify before you generate.** Write the intended use, the data classes, the user roles, and the access rules first. Give them to the assistant as requirements.
2. **Load standing instructions.** Place a rules file in the repository so every session begins with the security requirements. The companion Markdown file is written for this purpose.
3. **Choose boring, proven components.** Use an established identity provider, a managed database with encryption, and a mature web framework. Do not let the assistant write its own authentication or cryptography.
4. **Keep changes small.** Review each change before accepting the next. Large generated changes cannot be meaningfully reviewed.
5. **Automate the checks.** Static analysis, dependency scanning, secret scanning, and infrastructure scanning run on every change and block merges on failure.
6. **Ask the assistant to attack its own work.** Request a threat model, a list of abuse cases, and tests for each. Then have a different tool or person review the result.
7. **Put a qualified human at the gate.** Before real data, someone who can read the code reviews authentication, authorization, data handling, and infrastructure.
8. **Record your decisions.** Keep a short log of what was decided and why. It becomes the evidence for your risk analysis and DPIA.

11. Language models and agents inside the product

When the product itself calls a model, the model becomes part of the data path and part of the clinical risk. Treat it as an untrusted component that happens to be very capable.

Choosing how the model is deployed

Tier	Description	Suitable for
Consumer AI service	Personal or free accounts. Data may be retained or used for training. No BAA or DPA.	Stages 0 and 1 only. Never real patient data.
Enterprise API or cloud platform	Contracted service with BAA or DPA, zero data retention, no training on your data, regional hosting, and access logging.	Stages 3 and 4, once the agreement is signed and the configuration is verified.
Private deployment	Open-weight model run inside the institution's own cloud tenancy or on its premises.	Highest sensitivity, strict residency requirements, or offline settings.

Confirm five points in writing before real data flows: the agreement covers the specific model and feature in use, retention is zero or defined, data is not used for training, processing stays in the required region, and sub-processors are listed.

Threats specific to models

The OWASP Top 10 for Large Language Model Applications and its companion work on agentic systems give the standard catalogue. Those that matter most in health:

- **Prompt injection.** Text inside a patient note, uploaded document, email, or web page instructs the model to do something else. Any content the model reads is a possible instruction. Separate trusted instructions from untrusted content, restrict what the model can do, and never let model output trigger a consequential action without a check.
- **Sensitive information disclosure.** The model reveals data from its context, memory, or retrieval store to the wrong user. Enforce the user's own permissions at the retrieval layer, before content reaches the model.
- **Excessive agency.** An agent with write access to orders, messages, or records acts on a mistaken or manipulated instruction. Grant the least capability needed, prefer read-only, and require human confirmation for anything that changes a record or reaches a patient.
- **Improper output handling.** Model output is inserted into a page, a query, or a command without validation. Treat output as untrusted input.
- **Misinformation.** The model states something false with confidence. In clinical use this is a patient safety hazard, not a quality defect.
- **Supply chain and poisoning.** A model, adapter, dataset, or tool server of unknown origin carries hidden behavior. Use sources you can verify.

Design requirements

1. **Minimise what the model sees.** Send only the fields needed. Remove direct identifiers where the task allows.
2. **Ground the output.** Use retrieval from approved sources and show citations so a clinician can check the basis. This also supports the fourth CDS criterion.
3. **Keep a human in authority.** For any output that informs care, a qualified person reviews and decides. Design the screen so that review is real, with the source material visible and disagreement easy.
4. **Tell people.** Disclose that AI is in use, to clinicians and to patients. The AI Act requires it in Europe, several US states require it for patient communication, and trust requires it everywhere.
5. **Log every exchange.** Record the input, retrieved context, output, model version, and the human decision, in a store protected as carefully as the record itself.
6. **Evaluate before and after release.** Test on representative cases, including subgroups by age, sex, language, and ethnicity. Monitor for drift. Give users a way to report errors, and act on the reports.
7. **Pin and control versions.** A change of model is a change to the product. Re-run evaluations before adopting it.
8. **Constrain tools and connectors.** For agents that use tools or the Model Context Protocol, allowlist servers, scope each credential narrowly, and review tool descriptions as you would review code.
9. **Plan for failure.** Define what the system does when the model is unavailable, slow, or wrong. Care must be able to continue without it.

Frameworks to draw on

NIST AI Risk Management Framework and its generative AI profile, ISO/IEC 42001 for AI management systems, the Coalition for Health AI assurance guidance, the FDA's Good Machine Learning Practice principles, and the WHO guidance on ethics and governance of large multimodal models.

12. Interoperability and connecting to the health record

Connect through the standards the health system already supports. A tool that reaches the record by any other route, such as screen scraping, shared credentials, or a direct database link, will not pass review and should not.

The standards

Standard	What it does	Analogy
HL7 FHIR (R4 is the deployed baseline)	Defines health data as resources such as Patient, Observation, and MedicationRequest, exchanged over a web API.	A shared grammar. Every system agrees on what a sentence about a patient looks like.
US Core and USCDI	The US profiles and data classes that certified EHRs must support.	The regional dialect and required vocabulary.
European EHR exchange format (EEHRxF)	The FHIR-based format required under the EHDS. Builds on the International Patient Summary.	The European dialect, now being written into law.
SMART on FHIR	OAuth 2.0 and OpenID Connect profiles for launching apps and granting scoped access.	A visitor badge. It names who you are, which rooms you may enter, and when it expires.
SMART Backend Services	System-to-system access with signed assertions, no user present.	A service entrance with its own key and its own log.
CDS Hooks	Lets the EHR call your service at decision points and show the result as a card in the workflow.	A colleague who speaks only when the moment calls for it.
Bulk Data Access	Exports large populations for analytics.	A scheduled freight delivery, as opposed to a courier.
HL7 v2, C-CDA, DICOM, IHE profiles	Older and specialty standards for messages, documents, and imaging, still widely used.	The existing wiring of the building.

Standard	What it does	Analogy
SNOMED CT, LOINC, RxNorm, ICD-10	Coding systems for clinical terms, laboratory tests, medicines, and diagnoses.	The dictionary.

Rules for integration

1. **Start in a sandbox.** The major EHR vendors and SMART Health IT offer public test environments with synthetic patients. Build and demonstrate there before approaching a health system.
2. **Request the narrowest scopes.** Ask for `patient/Observation.rs`, not `user/*.*`. Reviewers read scope requests closely, and broad requests are refused.
3. **Use the authorization flow as designed.** Authorization code with PKCE for user-facing apps. Asymmetric keys for backend services. Never embed a client secret in a browser or mobile app.
4. **Keep tokens short-lived and protected.** Store them server-side. Do not log them.
5. **Respect the source of truth.** Read from the record when needed and avoid building a shadow copy. Where you must store data, store the minimum and define its retention.
6. **Write back with care.** Writing to the record carries clinical safety consequences. Expect additional review, and attribute every write to a person.
7. **Follow the vendor's program.** Each EHR vendor has an app registration and review process. Each health system adds its own security and privacy review on top.
8. **Plan for variation.** Implementations of the same standard differ between vendors and sites. Test against each one you intend to support.

13. Guidance by builder

The principles are the same for everyone. The first mistake each group tends to make is different.

Students and first-time builders

Your work belongs at stages 0 and 1, and that is a strength, not a limit.

- Use synthetic data for everything. A well-made prototype on synthetic data is more persuasive than a risky one on real data.
- Never use data from a clinical rotation, a job, or a family member, even with names removed.
- Learn one sanctioned pattern well: a SMART on FHIR app against a public sandbox.
- Write your intended use sentence and your answers to the six triage questions. Bring them to your instructor.
- Treat the security baseline as part of the assignment. Adding login, authorization, and logging to a prototype is how the skill is learned.

Clinicians building inside a health system

You understand the problem better than anyone. The risk is that familiarity with the data makes its sensitivity feel ordinary.

- Your access to records for care does not extend to building software with them. A new purpose needs new approval.

- Do not paste patient information into any AI tool your institution has not approved for PHI. This includes coding assistants.
- Bring your idea early to the people whose role is to help: the innovation office, the privacy officer, the security team, and the AI governance committee. They are far more helpful at the beginning than at the end.
- A tool that runs on your own machine and touches patient data is unsanctioned software. Ask for an approved environment.
- If your tool offers a recommendation about a patient, raise the device question before you pilot.
- Decide who will maintain the tool when you are on service, on leave, or have moved on.

Founders selling to health systems

Your customer's trust is your product. Security and compliance evidence shortens the sales cycle more than any feature.

- Sign a BAA or DPA with every vendor before the first customer conversation that involves data. This includes cloud, model, analytics, logging, and support tools.
- Begin SOC 2 or ISO 27001 early. Most enterprise buyers require one. HITRUST helps in the US.
- Hire or contract a qualified security engineer before the pilot. AI-generated code reviewed only by AI will not satisfy a hospital security team.
- Settle your regulatory position in writing: device or not, and under the AI Act, provider or deployer, high-risk or not. Buyers will ask.
- Keep an SBOM, a penetration test report, an architecture diagram, and a data flow diagram ready to share.
- Build for both markets from the start if you intend to serve both. Regional hosting and GDPR-grade data rights are costly to add later.
- Price in the cost of compliance. It is a permanent operating cost, not a launch expense.

Hackathon teams working across countries

You have little time and members under different laws. Simplicity protects you.

- Synthetic or openly licensed data only. Organizers should state this as a rule and provide the datasets.
- If a host institution offers real data, it must come de-identified, under a written agreement, inside an environment the host controls. Data does not leave that environment and does not cross borders.
- Do not publish secrets. Public repositories from hackathons are a frequent source of leaked keys. Scan before you push.
- Record your intended use, your triage answers, and the open risks in the project README. Judges and future partners will value it.
- Treat the event's end as stage 1. The path to a pilot begins afterward, through section 7.
- When team members sit in several countries, apply the strictest rule among them to shared work.

For teams in countries outside the US and Europe, the same structure usually holds. Many national laws follow the GDPR model, including those of Brazil, the UK, Kenya, Nigeria, and South Africa, and some add

data localisation requirements. Identify the national data protection authority and the health ministry's digital health rules at the outset.

14. Checklists and sources

Before any code

- ☐ Intended use written in one sentence
- ☐ Six triage questions answered
- ☐ Jurisdictions of users and data identified
- ☐ Data classes named: none, synthetic, de-identified, or identifiable
- ☐ Current stage identified, and the next gate understood

Before leaving stage 1

- ☐ Only synthetic data has been used
- ☐ No secrets in the repository, confirmed by a scan of its history
- ☐ Authentication handled by an established provider
- ☐ Authorization rules written down and tested with two accounts
- ☐ Dependencies verified and locked
- ☐ Architecture and data flow described in the repository

Before real patient data

- ☐ BAA or DPA signed with every vendor in the data path, including AI tools
- ☐ Security risk analysis and, in Europe, DPIA completed
- ☐ Lawful basis or permitted use documented
- ☐ Encryption in transit and at rest verified
- ☐ Multi-factor authentication enforced
- ☐ Audit logging working and reviewed
- ☐ Code reviewed by a qualified engineer
- ☐ Static, dependency, secret, and infrastructure scans passing
- ☐ Penetration test completed and findings closed
- ☐ Device classification and AI Act role confirmed in writing
- ☐ Approval from privacy, security, and AI governance
- ☐ Ethics or IRB approval if the work is research
- ☐ Incident response plan with named contacts

Before production

- ☐ Certifications held or scheduled
- ☐ Backup restoration tested
- ☐ Monitoring and alerting live, including model performance
- ☐ SBOM generated for each release
- ☐ Vulnerability reporting channel published
- ☐ Named owner for maintenance and a plan for retirement
- ☐ Users and patients informed of AI use

- ☐ Process to honor individual rights requests

Sources

Time-sensitive facts in this guide were checked on 5 October 2026 against the following.

- HIPAA Journal: HIPAA updates and changes in 2026 (Security Rule proposal status)
- King & Spalding: FDA updates general wellness and CDS guidance
- Drummond Group: how the HTI-5 proposed rule affects certification
- Freshfields: the final Digital Omnibus on AI
- HIQA: EHDS, what a health data holder needs to know
- Matheson: Cyber Resilience Act reporting obligations
- Cloud Security Alliance: AI-generated code vulnerability surge
- NordLayer: vibe coding security

Primary texts to consult directly: 45 CFR Parts 160 and 164 (HIPAA); 42 CFR Part 2; 16 CFR Part 318 (Health Breach Notification Rule); FDA guidance on Clinical Decision Support Software and on Cybersecurity in Medical Devices; 45 CFR Parts 170 and 171 (certification and information blocking); Regulation (EU) 2016/679 (GDPR); Regulation (EU) 2024/1689 (AI Act) as amended by Regulation (EU) 2026/1744; Regulation (EU) 2017/745 (MDR); Regulation (EU) 2025/327 (EHDS); Directive (EU) 2022/2555 (NIS2); Regulation (EU) 2024/2847 (CRA); NIST SP 800-66 Rev. 2, Cybersecurity Framework 2.0, and AI RMF; OWASP ASVS and the OWASP Top 10 for LLM Applications.

Review cycle. Several rules cited here are in motion: the HIPAA Security Rule update, HTI-5, the MDR reform, and EHDS implementing acts. Review this guide every six months, and whenever one of them is finalized.