SAMPLE SECURITY FRAMEWORK

Medical research AI Data Risk Controls

Focus areas: 12 Controls, 6 Domains (D)

Purpose

Structured, actionable security controls for CISOs or IT teams governing AI tool usage in medical research environments. Each control includes priority, ownership, implementation steps, and testing protocols.

Scope: All AI/LLM tools used in proximity to genomic, phenotypic, or patient derived datasets.

CONTROLS SUMMARY

Control	Title	Priority	Owner	Domain
01	AI Tool Approval Gate	Critical	CISO / IT	Governance
02	Contractual Data Safeguards	Critical	Legal / CISO	Governance
03	Data Classification Matrix	Critical	Data Gov	Data Handling
04	De identification Standards	Critical	Data Gov / IT	Data Handling
05	Access Control & MFA	Critical	CISO / IT	Access Control
06	API Key Management	High	Security Eng	Access Control
07	Data Loss Prevention	Critical	Security Eng	Monitoring
08	Audit Logging & SIEM	High	Security Eng	Monitoring
09	Prompt Injection Prevention	High	Security Eng	Model Abuse
10	Output Validation	High	Research IT	Model Abuse
11	AI Security Training	High	CISO / HR	Training
12	AI Incident Response	High	CISO / SecOps	Incident Resp.

D1 -AI TOOL GOVERNANCE & PROCUREMENT

Evaluate, approve, and monitor all AI tools before they interact with research data.

01 AI Tool Approval Gate

Critical | CISO / IT

Implementation Steps

1. Maintain a central AI Tool Register: vendor, data location, retention policy, certifications.
2. Require a Security Assessment Questionnaire (SAQ) before any AI tool touches research data.
3. An AI Approval Board (CISO, DPO, Research IT) issues Approved / Conditional / Rejected decisions.
4. Publish the approved list quarterly; provide a shadow AI reporting channel.

Testing & Validation

T1 Attempt to onboard a vendor without SAQ confirm procurement blocks it.

T2 Survey 10 staff: can they locate the approved tool list?

T3 Review, register for tools unused 90+ days; remove or re assess.

02 Contractual Data Safeguards

Critical | Legal / CISO

Implementation Steps

1. Execute a UK GDPR Art. 28 DPA with every AI vendor that may process personal data.
2. Insert a model training prohibition clause: no use of submitted data to train/fine-tune models.
3. Mandate a 30 day data deletion guarantee on contract termination, with written confirmation.
4. Include a right to audit clause and sub processor notification obligation (30 days' notice).

Testing & Validation

T1 Pull 3 active contracts confirm each contains a training prohibition clause.

T2 Request sub processor lists from 2 vendors and validate against contract terms.

T3 Simulate contract termination; confirm data deletion within the SLA.

D2-DATA CLASSIFICATION & DE-IDENTIFICATION

Prevent sensitive biobank data from entering AI systems in an identifiable form.

03 Data Classification & AI Restriction Matrix

Critical | Data Gov

Implementation Steps

1. Define 4 tiers: T1 Identifiable (prohibited) → T2 Pseudonymised (approved tools + DPA) → T3 Aggregated (approved tools) → T4 Public (unrestricted).
2. Apply classification tags in the data catalogue and embed in all dataset metadata.
3. Train all data custodians annually; include a short quiz to confirm understanding.

Testing & Validation

T1 Sample 20 datasets verify each has a classification tag.

T2 Ask a researcher to classify an unlabelled dataset; measure accuracy.

T3 Attempt to submit a T1 dataset to an AI tool confirm DLP blocks it.

04 De-identification Standards

Critical | Data Gov / IT

Implementation Steps

1. Adopt Safe Harbour as minimum baseline: remove all 18 HIPAA/UK equivalent identifiers.
2. For genomic data: suppress rare variants (MAF < 1%), restrict full sequence to T1 handling.
3. Store pseudonymisation key mapping in an HSM protected vault, access-controlled separately.
4. Run a re-identification risk assessment (k-anonymity ≥ 5) before any dataset reaches an AI tool.

Testing & Validation

T1 Run ARX or sdcMicro on a sample dataset; confirm k-anonymity ≥ 5.

T2 Cross reference a pseudonymised cohort with a public dataset; document re ID feasibility.

T3 Confirm key vault has no network path from the AI tool environment.

D3-ACCESS CONTROL & IDENTITY MANAGEMENT

Enforce least-privilege access and ensure all AI interactions are attributable.

05 Access Control & MFA for AI Tools

Critical | CISO / IT

Implementation Steps

1. Define roles: Viewer, Operator, Data-Operator (requires manager + DPO approval), Admin.
2. Route all AI tool access through SSO/SAML; enforce MF reject password only auth.
3. Set 12-month auto expiry on Data Operator roles; conduct quarterly access reviews.
4. Revoke access within 24 hours of a leaver integrate with offboarding workflow.

Testing & Validation

T1 Pull all Data-Operator role holders; confirm each has an approved access request on file.

T2 Simulate a leaver verify AI access is revoked within the SLA.

T3 Attempt AI tool access without MFA confirm rejection.

06 API Key & Credential Management

High | Security Eng

Implementation Steps

1. Store all API keys in a secrets vault (HashiCorp Vault / AWS Secrets Manager) never in code repos.
2. Enforce 90-day rotation; use separate keys for prod, staging, and dev environments.
3. Enable vault audit logging: capture every key read with identity and timestamp.
4. Maintain a tested key revocation runbook: invalidate a compromised key within 15 minutes.

Testing & Validation

T1 Run truffleHog / GitGuardian across all repos confirm no plaintext AI keys.

T2 Verify all keys rotated within 90 days; escalate overdue ones.

T3 Execute the revocation runbook in test; measure against the 15-minute SLA.

D4-MONITORING, LOGGING & THREAT DETECTION

Establish visibility into AI interactions to detect exfiltration and anomalous behaviour.

07 Data Loss Prevention (DLP)

Critical | Security Eng

Implementation Steps

1. Deploy DLP on web proxy, API gateways, and endpoints covering all AI tool channels.
2. Define biobank-specific rules: genomic string patterns (FASTA/VCF), NHS number regex, DOB proximity.
3. Block T1 data; alert-and-log T2; log only T3. Route P1 alerts to SIEM (ACK within 30 min).
4. Conduct monthly rule-effectiveness reviews; maintain a CISO/DPO-approved exceptions process.

Testing & Validation

T1 Submit synthetic NHS number and mock genomic string to each approved tool confirm alert within 5 min.

T2 Review 30 days of alerts for any non test T1 detections; investigate findings.

T3 Attempt to submit a T1 file via web proxy confirm it is blocked and user notified.

08 Audit Logging & SIEM Integration

High | Security Eng

Implementation Steps

1. Mandate AI tool logs: user identity, timestamps, prompt hash, output hash, data volume.
2. Ingest to SIEM within 5 minutes; retain for 12 months minimum.
3. Activate detection rules: large prompts (>10k tokens), high frequency queries, off hours access, unusual geo.
4. Produce a monthly AI Usage Report for CISO and DPO.

Testing & Validation

T1 Submit an oversized prompt from a test account confirm SIEM alert fires within 10 min.

T2 Cross-reference session records vs SIEM confirm 100% log capture.

T3 Verify logs from 13 months ago are present and quarriable.

D5-PROMPT INJECTION & MODEL ABUSE

Defend against adversarial prompt techniques and ensure output integrity.

09 Prompt Injection Prevention

High | Security Eng

Implementation Steps

1. Apply input sanitisation layer: strip role-switching, jailbreak patterns, nested instruction blocks.
2. Prepend a hardened system prompt to every API call defining role, data constraints, and refusals.
3. For document upload tools, sanitise all file content before passing to the LLM.
4. Maintain a Prompt Injection Playbook: preserve evidence, revoke session, notify CISO.

Testing & Validation

T1 Submit 10 OWASP LLM Top-10 injection strings to each hosted tool document pass/fail.

T2 Embed a malicious instruction in a mock document and upload confirm it is not executed.

T3 Review output logs (30 days) for any system-prompt fragments in responses.

10 Output Validation & Scientific Integrity

High | Research IT

Implementation Steps

1. Mandate a human review gate: no AI output used in publications or regulatory submissions without expert sign off.
2. Require researchers to log: prompt, model version, temperature, and output stored in the research audit trail.
3. Define prohibited uses: AI must not be the sole basis for variant pathogenicity calls or clinical eligibility decisions.
4. Conduct quarterly output audits: sample 10 AI-assisted analyses for accuracy and process compliance.

Testing & Validation

T1 Submit a hallucination-inducing prompt — verify the review process would catch it.

T2 Check 3 recent AI-assisted analyses: confirm prompts, model versions, and sign offs are recorded.

T3 Survey 5 researchers on the Human Review Gate and prohibited use cases.

D6 — TRAINING & INCIDENT RESPONSE

Equip staff and security teams to prevent and respond to AI-related security incidents.

11 AI Security Awareness Training

High | CISO / HR

Implementation Steps

1. Deliver a mandatory 30 minute AI Security module: approved tools, data tiers, prompt injection, reporting.
2. Require completion within 30 days of onboarding; annually thereafter track via LMS.
3. Run quarterly phishing style AI Risk Simulations; measure and publish reporting rates.
4. Publish a monthly one page newsletter or a game to all staff

Testing & Validation

T1 Confirm >90% LMS completion for staff with data access.

T2 Measure simulation reporting rate target >30% report vs clicks.

T3 Interview 3 researchers and 3 IT staff to assess knowledge retention.

12 AI Incident Response Plan

High | CISO / SecOps

Implementation Steps

1. Create an AI IRP annex covering: AI facilitated breach, prompt injection compromise, vendor breach, shadow AI discovery.
2. Define escalation triggers: DPO notified within 1 hour of suspected personal data breach.
3. Appoint an AI Incident Response Lead; conduct an annual tabletop exercise.
4. Require post-incident root cause analysis and lessons-learned report within 30 days of resolution.

Testing & Validation

T1 Run the annual tabletop — measure time-to-notify DPO and time-to-isolate the affected tool.

T2 Review last 3 vendor security notifications; confirm acted on within contractual SLA.

T3 Confirm the AI IRP annex has been reviewed and updated within the last 12 months.