Week 7 — Training, Release, Monitoring & Maintenance

HEALTHCARE TRACK

MAYO CLINIC CLINICAL AI GOVERNANCE

AI Governance & Risk Management – Week 7 explores the **last mile and long haul of AI safety.** From the moment a model goes live, governance transforms from a testing exercise into continuous, legally accountable oversight.

The Last Mile & Long Haul of AI Safety

CASE STUDY: MAYO CLINIC

This week centers on a multilingual clinical decision AI deployed at scale. Three fundamental questions frame every governance decision:

Before Deployment

What must be provably true about training data, performance thresholds, and bias mitigation before any patient sees the model's output?

During Release

Which deployment strategy minimizes risk exposure? How do canary rollouts, kill switches, and rollback protocols protect patient safety?

After Go-Live

How is continuous accountability maintained? What metrics tie directly to patient harm and trigger mandatory escalation?

This is where legal liability begins. Governance frameworks shift from internal quality assurance to externally auditable, legally defensible accountability.

Why Accountability Begins at Release

RELEASE = LEGAL LIABILITY THRESHOLD

The moment an AI system transitions from testing to live clinical use, the governance calculus changes entirely. What was once a model error becomes a patient harm exposure.

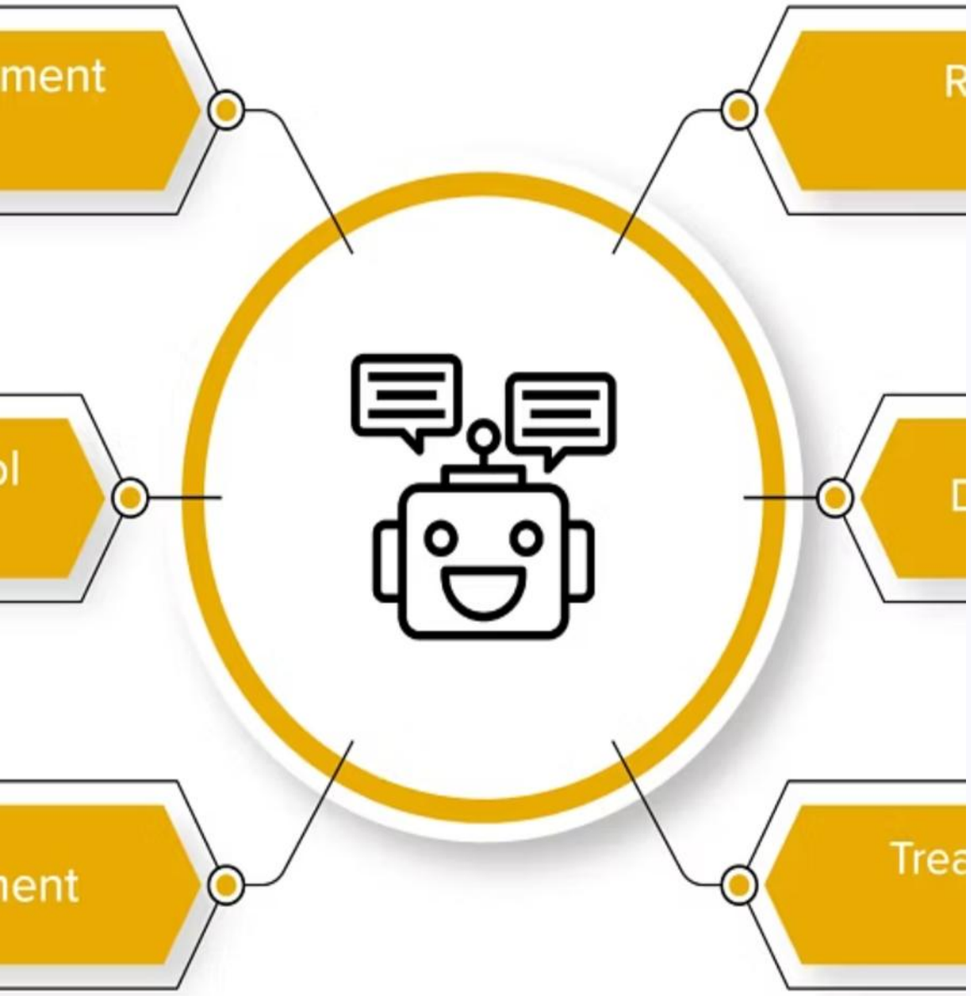
Once AI Goes Live

- Real users making real clinical decisions
- Real patient impact – including urgent triage
- Real regulatory exposure under FDA and CMS rules
- Real malpractice risk for the institution

Healthcare Example

If Mayo Clinic's AI triage system mis-ranks a critical patient – assigning low urgency to someone experiencing an acute cardiac event – that is no longer classified as a "model error." It becomes **patient harm exposure**, subject to incident reporting, regulatory review, and potential litigation.

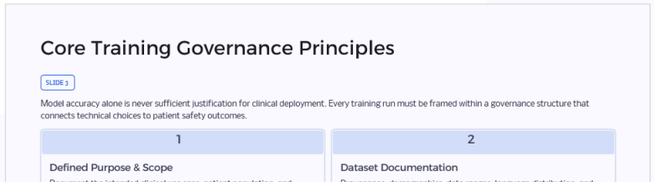
Governance shifts from **testing → continuous accountability.** Every post-deployment metric carries legal and ethical weight.



Part 1 — Training Governance

FOUNDATION

Get the training right — everything downstream depends on it.



Core Training Governance Principles

SLIDE 3

Model accuracy alone is never sufficient justification for clinical deployment. Every training run must be framed within a governance structure that connects technical choices to patient safety outcomes.

1

Defined Purpose & Scope

Document the intended clinical use case, patient population, and decision support role before any training begins.

2

Dataset Documentation

Provenance, demographics, date ranges, language distribution, and known gaps must all be formally recorded.

3

Clinical Evaluation Framework

Define clinically meaningful metrics – not just accuracy – before training begins, not after results are seen.

4

Harm-Based Threshold Alignment

Set acceptable false negative rates, override rates, and language parity gaps with clinical stakeholders, not engineers alone.

✓ Training Governance Checklist

✓ Dataset Documentation Sheet

Reproducible Training Environments

SLIDE 4

In healthcare AI, reproducibility is not a nice-to-have – it is a legal requirement. If Mayo Clinic is ever challenged in court or by a regulator, the institution must be able to recreate the exact model that made a clinical recommendation.

Reproducibility Controls Required

- Pinned software and library versions
- Fixed random seeds logged per run
- Containerized environments (e.g., Docker)
- Immutable dataset snapshots with version IDs
- Cryptographic hashes for all training artifacts

If Sued, Mayo Must Reproduce Exactly:

- Which model version was active
- Which dataset version was used
- Which hyperparameters were applied
- Which compute environment ran the job

Failure to reproduce = inability to defend the system in a legal or regulatory proceeding.

✓ Reproducibility Control Matrix

Data Leakage & Test Set Integrity

SLIDE 5

Data leakage – when information from the test set contaminates the training set – is one of the most dangerous and invisible failure modes in clinical AI. It produces artificially inflated performance metrics that collapse in production.

Strict Train/Test Separation

No data point in the test set may appear in any form during training. This includes augmented, anonymized, or derived versions of records.

Sealed Test Slices

Hold-out sets must be sealed and inaccessible to engineers until formal evaluation. Access logs must be maintained.

Rare-Case Challenge Sets

Critical edge cases – rare diagnoses, pediatric populations, non-English speakers – must be deliberately included in evaluation sets.

Healthcare Impact: Leakage produces inflated recall scores → missed critical cases in production → direct regulatory exposure and patient harm liability.

✓ Train/Test Separation Validation Report

Part 2 — Drift & Performance Governance

ONGOING VIGILANCE

A model that was safe on launch day may not be safe next quarter.

Types of Drift to Monitor

SLIDE 6

Clinical AI systems are deployed into dynamic environments. Policies change, patient populations shift, clinical language evolves, and knowledge bases become outdated. Each creates a distinct and measurable form of drift that governance teams must detect before it causes harm.

Policy Drift

CMS coverage rules or clinical guidelines update, but the model's recommendations do not reflect the change. Example: a new covered screening protocol goes unrecognized.

Language Drift

Shifts in how patients or clinicians phrase symptoms in Spanish, Somali, or Hmong reduce accuracy for non-English speakers over time.

Retrieval Drift

The RAG corpus becomes outdated. New clinical evidence exists but the model continues citing older, less accurate sources.

Distribution Drift

Flu season surges or pandemic conditions alter the patient mix, pushing the model outside its validated operating range.

✓ Drift Monitoring Plan

Performance Thresholds That Matter

SLIDE 7

These are not abstract technical targets. Each threshold is calibrated to a specific patient safety outcome. Falling below any one of them triggers mandatory governance review.

≥0.88

Recall

Minimum sensitivity required to avoid missing critical clinical findings across all patient cohorts.

≥0.90

Grounding

Proportion of recommendations traceable to verified clinical evidence. Below this, citation integrity is compromised.

≤5pp

Language Parity Gap

Maximum allowable performance difference between English and any supported non-English language. Exceeding this signals discriminatory impact.

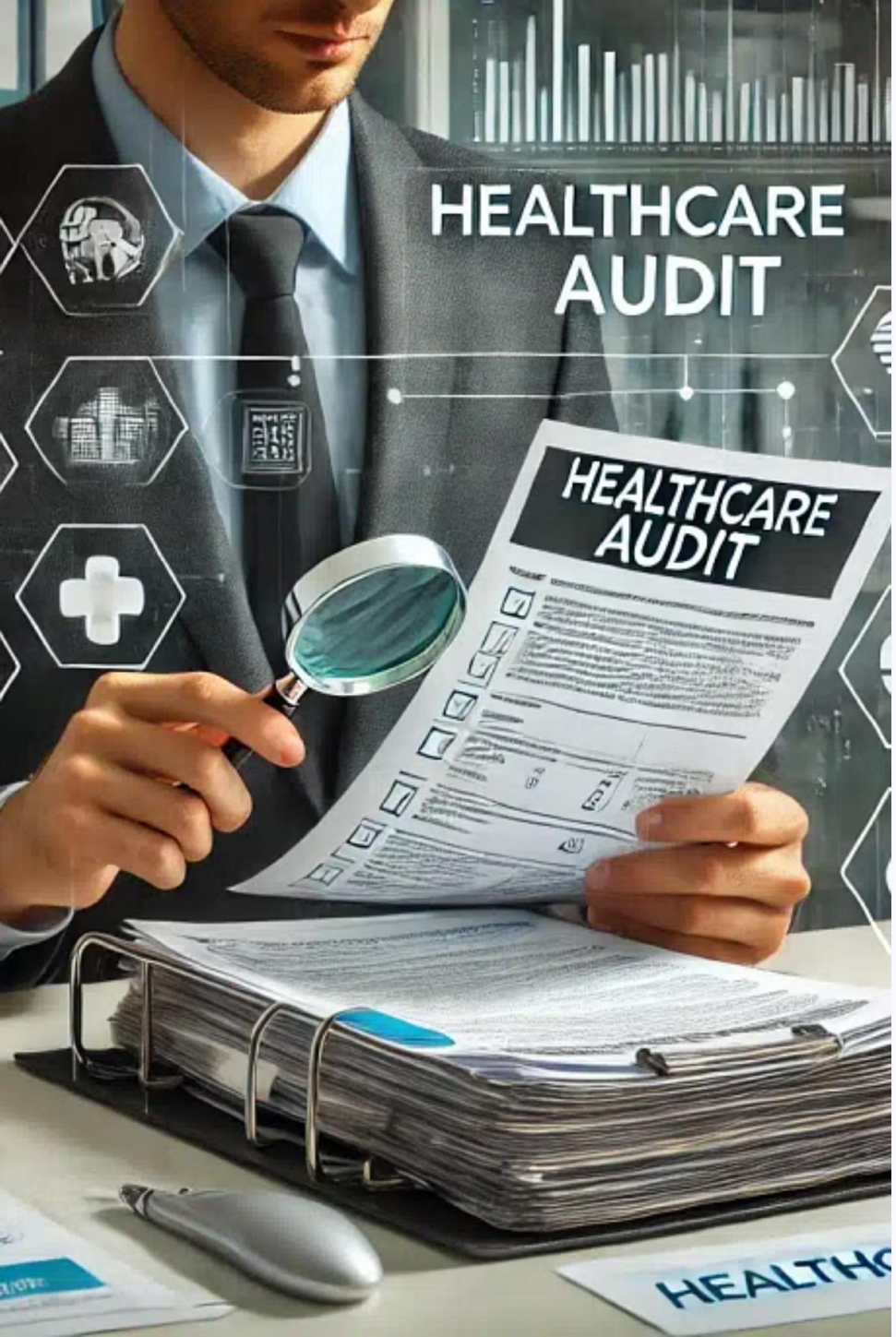
≤9%

Override Rate

If clinicians override the AI more than 9% of the time, the system is generating alert fatigue or untrustworthy recommendations.

These thresholds map directly to: **missed urgent cases, care denial, discriminatory impact, and clinician burden.** Each must be justified in writing before deployment.

✓ Threshold Justification Memo



Part 3 — Evidence Package Before Go-Live

PRE-DEPLOYMENT GATE

No evidence package. No deployment.

The Mandatory Evidence Package

SLIDE 8

Before any clinical AI system goes live, a complete, reviewable evidence package must be assembled, reviewed, and approved by the governance committee. Incomplete documentation is not a minor gap – it is a governance failure that voids deployment authorization.

Cohort Performance Metrics

Disaggregated results by language, age group, care setting, and diagnosis category – not just aggregate accuracy scores.

Red-Team Testing Results

Documentation of adversarial testing: edge cases, prompt injection attempts, worst-case failure scenarios, and clinician stress tests.

Bias Mitigation Documentation

Evidence that identified disparities were addressed, not simply acknowledged. Includes before/after metric comparisons.

Rollback Plan & Operator Playbook

Step-by-step procedures for reverting to the prior system version and day-one operator guidance for clinical staff.

Without this package, deployment equals governance failure – exposing Mayo Clinic to regulatory sanctions, accreditation risk, and legal liability.

✓ AI Go-Live Evidence Binder

Part 4 — Deployment Strategy

CONTROLLED RELEASE

How you release is as important as what you release.

Strategic Deployment Patterns

SLIDE 9

Healthcare AI does not tolerate big-bang deployments. Staged release strategies reduce patient risk exposure and provide real-world governance data before full activation. For high-risk clinical AI, canary deployment is not optional.

1

Blue-Green

Two identical environments run in parallel. Traffic switches instantly between versions, enabling zero-downtime rollback if the new version fails.

2

Canary

New version routes to a small, monitored cohort first – e.g., 10% of ICU cases. Drift and harm signals are validated before wider expansion.

3

Shadow Mode

Model runs in parallel with existing workflow, generating recommendations that are logged but not shown to clinicians. Safe for validation.

4

A/B Testing

Two model variants serve different cohorts simultaneously, enabling controlled comparison of clinical decision quality under real conditions.

Mayo Clinic Protocol: ICU deployment begins at 10% → monitor for drift signals → expand to 25% → 50% → full activation only after each stage clears governance review.

✓ Deployment Strategy Plan

Kill Switch & Rollback Protocols

SLIDE 10

Every clinical AI system must have a pre-defined, tested, and documented path to immediate deactivation. Rollback is not an admission of failure – it is the governance mechanism that protects patients when conditions exceed the model's validated operating range.

Automatic Kill Switch Triggers

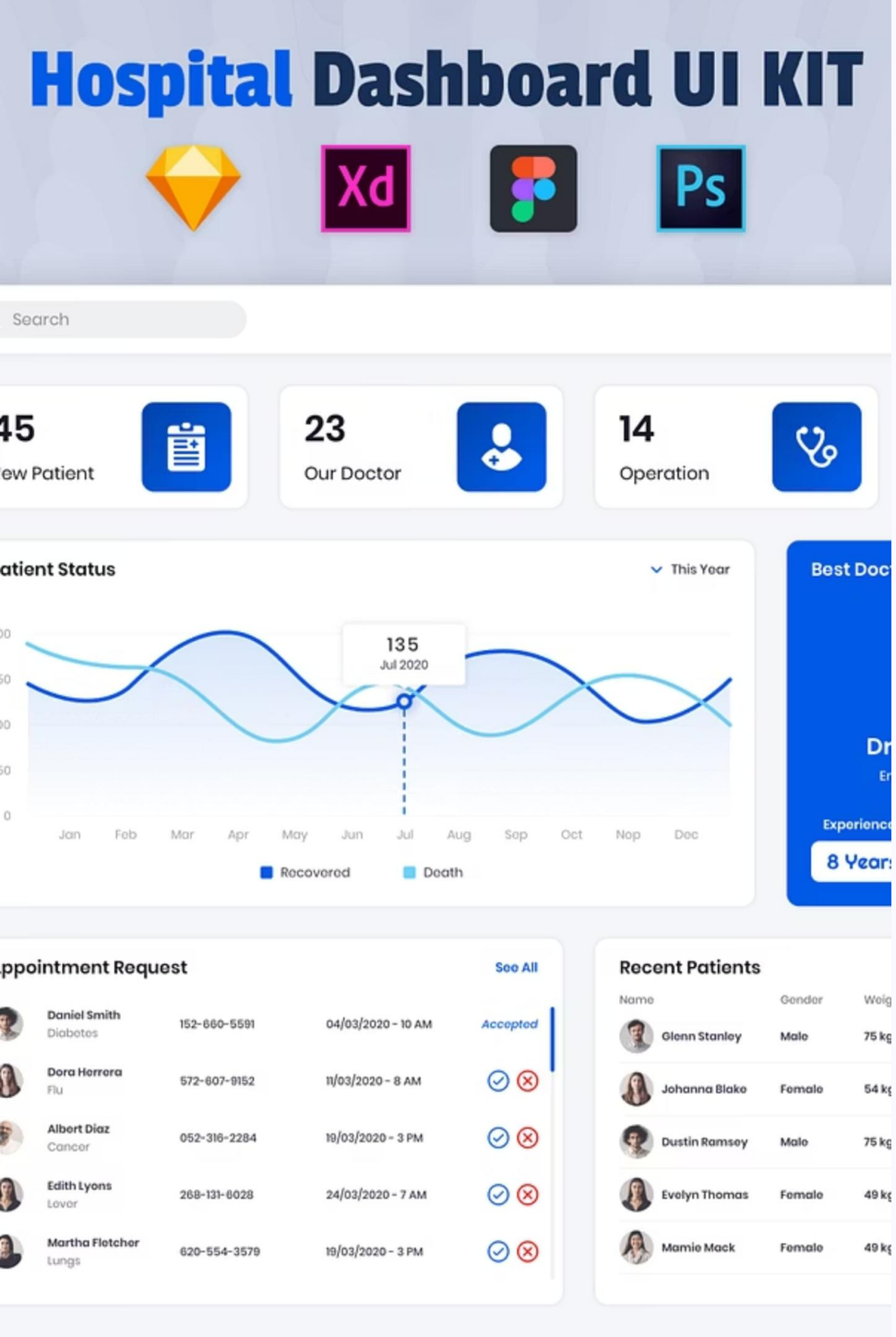
- Harmful output rate exceeds **0.2%**
- Response latency exceeds **2.5 seconds**
- Language parity violation exceeds **10 percentage points**
- Grounding score falls below defined threshold
- Urgent mis-ranking rate spikes above baseline

Rollback Is Responsible Governance

Governance teams that hesitate to roll back a failing system – out of concern for operational disruption or reputational risk – are prioritizing institutional comfort over patient safety.

The rollback SOP must specify: who can authorize, how long rollback takes, how clinicians are notified, and how incidents are logged for post-mortem review.

✓ Kill Switch & Rollback SOP



Part 5 — Monitoring & Alerting

CONTINUOUS ACCOUNTABILITY

If a metric doesn't tie to patient harm, it's vanity.

Monitoring That Ties to Harm

SLIDE 11

Effective post-deployment monitoring is not about collecting every available metric. It is about tracking only those signals with a direct, documented link to patient safety outcomes. Vanity metrics – those that look good in reports but don't detect harm – waste governance capacity and create false confidence.

Recall@k Per Language

Tracks whether non-English speakers receive equivalent quality of clinical decision support. A recall gap is a health equity violation.

Harmful Output & Override Trends

Rising override rates signal clinician distrust or recommendation failure. Harmful output trends are escalation triggers, not quality improvement inputs.

Urgent Mis-Ranking & Denial Reversals

If critical patients are being under-triaged, or if prior authorization denials are being reversed on appeal at high rates, the model is causing active harm.

✓ Harm-Correlated Monitoring Dashboard

Alerting With Clear Ownership

SLIDE 12

An alert without an owner is noise. Every monitoring threshold that fires must immediately activate a named individual on a defined timeline via a specified channel. Ambiguous escalation paths are a governance gap – not a technical inconvenience.

WHO

Named individual or role – not a team inbox. Accountability requires a human name attached to every alert category.

WHEN

Critical alerts: **15-minute** response SLA. Standard alerts: defined within the runbook by severity tier.

HOW

Paging system, secure clinical messaging, email, or direct call – channel defined per alert type, not left to individual discretion.

ESCALATION PATH

If primary owner does not respond within the SLA, the alert automatically routes to their supervisor and the governance committee chair.

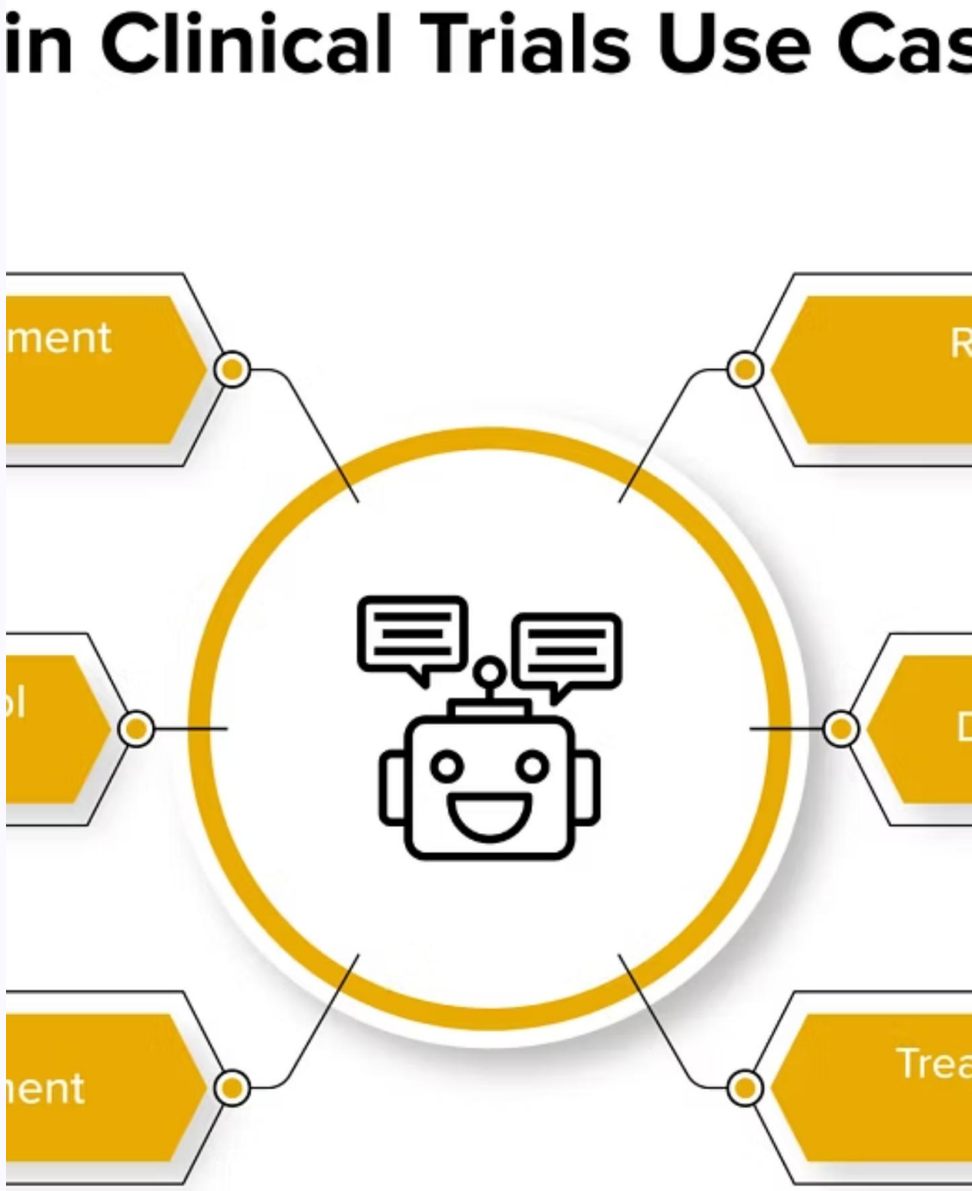
✓ Alert Runbook

✓ Escalation Matrix

Part 6 — Retraining Governance

CLOSING THE LOOP

Retraining is not a reset. It is a new deployment.



Retraining Triggers & Governance Requirements

SLIDE 13

When monitoring detects drift, bias, or degraded performance, retraining must be initiated – but it must follow the same rigorous governance path as the original training. Urgency does not justify shortcuts. A rushed retrain that introduces new bias is worse than a controlled delay.

1 Parity Violation

Language parity gap exceeds the 5pp threshold. Immediate retraining trigger with bias audit required before redeployment.

2 Policy Update

CMS, FDA, or institutional policy change invalidates prior training assumptions. Corpus and labeling schema must be updated.

3 Seasonal Shift

Flu season, pandemic surge, or population shift pushes distribution beyond validated range. Scheduled retraining with new cohort data.

4 Performance Degradation

Recall or grounding scores fall below minimum thresholds over a rolling 30-day window. Triggered retraining with root cause analysis.

Governance principle: Retraining must follow the same checklist, documentation standards, evidence package requirements, and deployment strategy as the original model. No exceptions under operational pressure.

✓ Retraining Governance SOP