From 0→1 QA in AI Healthcare

A Blueprint for Safe, Reliable, and Scalable Quality Assurance

Introduction

Healthcare demands the highest standard of reliability. When AI systems are deployed in clinical contexts, a single regression can translate into patient harm, regulatory risk, and lost trust. Traditional QA approaches—reactive test scripts, siloed ownership, slow manual validation—cannot keep pace with the speed of AI product cycles.

This paper outlines a practical 0→1 QA blueprint designed for AI healthcare startups. The approach balances clinical safety, regulatory compliance, and engineering velocity, while leveraging modern automation and intelligent test tooling to build reliability from day one.

Principles of Modern QA in Healthcare

Risk-Weighted Coverage

Not every flow is equal. Clinical decision support, PHI handling, and billing are P0 flows—they must be airtight before release. Marketing copy alignment is not. By tagging tests (risk:high, risk:medium, risk:low), the QA strategy ensures that critical patient-facing functions are never compromised.

Synthetic PHI by Default

HIPAA and SOC-2 compliance demand strict test data governance. All test automation must operate on synthetic or anonymized PHI datasets. Logs are redacted at source, and build artifacts expire on strict retention policies to avoid compliance breaches.

LLM as a Test Surface

Testing AI is not like testing CRUD. Prompt-based models must be validated with golden sets, adversarial probes, and refusal scoring. QA must treat hallucinations, prompt injections, and inconsistent refusals as first-class Sev-1 defects. For AI medical agents, prompt stability and guardrail compliance are as critical as functional correctness. A hallucination is a Sev-1 defect

CI/CD Evidence & Auditability

Every QA run should produce immutable, signed manifests containing commit SHAs, environment details, and test results. These artifacts double as SOC-2 evidence, ensuring audit readiness without extra overhead.

Blueprint: Building QA from 0→1

Phase 1: Baseline (Days 0–30)

- Stand up a Playwright framework for core API and UI paths.

- Define risk taxonomy (clinical, PHI-handling, billing, auth).
- Implement CI gating in GitHub Actions; block merges if P0 flows fail.
- Seed an LLM golden set to validate baseline AI behavior.
- Enforce synthetic PHI only; no real patient data in test runs.

Phase 2: Reliability (Days 31–60)

- Expand automation to ~30–40% of critical flows.
- Add flake quarantine and failure bucketing to keep builds green.
- Integrate SOC-2 controls: immutable manifests, log redaction, artifact retention.
- Scale LLM evaluation with hallucination traps, refusal scoring, jailbreak detection.
- Establish build health SLOs (e.g., <24h to green).

Phase 3: Scale (Days 61–90)

- Parallelize test execution on Kubernetes runners for speed.
- Introduce visual regression checks on stable UI surfaces.
- Add performance baselines with k6 or Locust for high-volume APIs.
- Formalize a QA guild: embed QA engineers into squads, run weekly risk reviews.
- Track and report defect escape rates, MTTD, MTTR, and AI eval score trends.

Innovations from QA Brain

This approach builds on the principles behind QA Brain, a local-first QA intelligence toolkit. While the full system remains private, its architectural ideas inform the blueprint:

- Replay-Driven Test Generation: Convert production logs and bug reports into reproducible test cases.
- Origin-Aware Quality Grading: Score tests differently if they’re manual, AI-generated, or exploration-driven.
- Self-Healing Selectors: Automatically repair brittle locators to maintain stability at scale.
- Heuristics Library: Domain-specific test recipes for common risks (auth, payments, PHI).
- Testrunner + Memory: Local execution engine with results stored as an evolving knowledge base.

Conclusion

For AI healthcare, quality cannot be bolted on—it must be architected in from the start. By focusing on risk-weighted coverage, synthetic PHI safety, LLM evaluation, and auditable CI pipelines, startups can achieve clinical-grade reliability without slowing velocity.

The 0→1 QA blueprint is not just about preventing regressions—it’s about enabling trust. In healthcare, trust is everything.