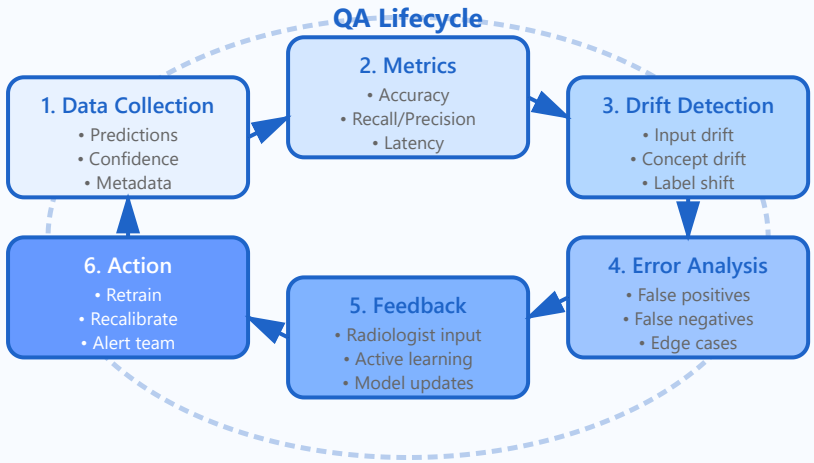
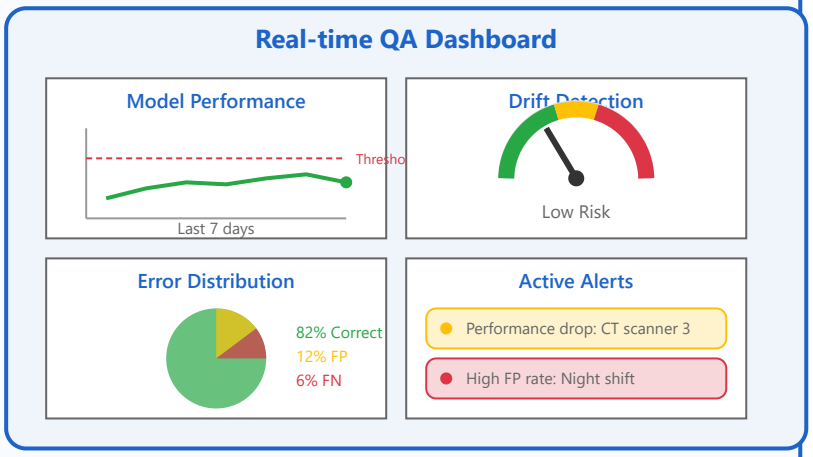
Quality Assurance

Continuous Quality Assurance Framework





Performance Monitoring

Track accuracy, precision, recall over time. Automated dashboards

Drift Detection

Identify distribution shifts. Input drift (scanner changes) vs concept drift (disease patterns)

Error Analysis

Systematic review of failures. Identify error patterns and edge cases

Feedback Loops

Radiologist corrections. Active learning to improve model from production data

Continuous Improvement

Iterative model updates. A/B testing of model versions

Detailed Analysis of Quality Assurance Components

1. Performance Monitoring - Deep Dive

What is Performance Monitoring?

Performance monitoring is the continuous tracking of AI model behavior in production. It involves measuring key metrics over time, comparing against baselines, and detecting degradation before it impacts patient care. Unlike one-time validation, monitoring provides ongoing surveillance to ensure consistent quality.

💡 Real-World Example: Chest X-Ray Pneumonia Detection

Scenario: Hospital deploys pneumonia detection AI with 96% validation accuracy

Week 1-8: Stable at 95.5-96.2% accuracy ✓
Week 9-10: Drops to 93.8% - warning alert triggered ⚠️
Week 11: Falls to 91.2% - critical alert 🚨

Investigation revealed: New CT scanner produced images with different contrast
Action taken: Recalibrated preprocessing, performance recovered to 95.8%

Impact: Issue caught before affecting patient care significantly

🔑 Key Metrics to Monitor

- **Accuracy:** Overall correctness - primary indicator of model health
- **Sensitivity (Recall):** Ability to catch positive cases - critical for patient safety
- **Specificity:** Ability to correctly identify negatives - reduces false alarms
- **Precision (PPV):** Accuracy of positive predictions - impacts clinician trust
- **AUC-ROC:** Overall discriminative ability across all thresholds
- **Inference Latency:** Prediction time - impacts clinical workflow
- **Confidence Calibration:** Whether predicted probabilities match actual outcomes

2. Drift Detection - Understanding Changes

Three Types of Drift

📊 Input Drift

What: Distribution of input features changes
Example: New

🔄 Concept Drift

What: Relationship between features and

📈 Label Drift

What: Distribution of outcomes changes
Example: Flu

3. Error Analysis - Learning from Failures

Systematic Approach to Understanding Model Failures

🎯 Common Error Patterns

- **Image Quality Issues:** Poor contrast, artifacts, motion blur, noise

scanner produces brighter images
Impact: Model sees different input than training

labels changes
Example: New disease variant shows different patterns
Impact: What model learned no longer applies

season increases pneumonia prevalence 67% → 85%
Impact: Class imbalance affects predictions

📋 Case Study: Multi-Site Deployment

Scenario: Lung nodule detection deployed across 15 hospitals

Site A (GE Scanner):
• Detected: Input drift - slice thickness changed 1mm → 2.5mm
• Action: Recalibrated preprocessing pipeline
• Result: Performance recovered from 87% → 94%

Site B (Teaching Hospital):
• Detected: Concept drift - COVID increased ground-glass opacities
• Action: Collected new training data, retrained model
• Result: Improved detection of evolving pathology patterns

Site C (Cancer Center):
• Detected: Label drift - 85% positive cases vs. 15% in training
• Action: Adjusted decision threshold for prevalence
• Result: Reduced false positive rate by 30%

- **Rare Presentations:** Atypical disease manifestations not well-represented in training
- **Boundary Cases:** Subtle findings near decision threshold
- **Overlapping Pathologies:** Multiple conditions present simultaneously
- **Patient Demographics:** Pediatric, elderly, or obese patients with different anatomy
- **Technical Factors:** Equipment differences, patient positioning, scan timing
- **Early-Stage Disease:** Minimal visible changes difficult even for experts

📊 Deep Dive: Diabetic Retinopathy Error Analysis

Initial Performance: 92% sensitivity, 95% specificity
500 errors reviewed by ophthalmologists:

False Negatives (120 cases) - Most Critical:
• 45% early-stage mild cases with minimal changes
• 30% image quality issues (blur, poor illumination)
• 15% patients with dark fundus pigmentation
• 10% concurrent conditions obscuring view

False Positives (380 cases):
• 50% age-related changes misclassified as diabetic changes
• 25% image artifacts mistaken for lesions
• 15% borderline cases near decision threshold
• 10% other retinal conditions with similar appearance

Improvements Implemented:
1. Added 2,000 early-stage images to training
2. Implemented image quality pre-screening
3. Augmented with dark fundus images
4. Added patient age as model feature
5. Confidence-based referral for borderline cases

Result: 96% sensitivity, 97% specificity | 60% reduction in false negatives ✓

4. Feedback Loops - Continuous Learning

Leveraging Production Data for Model Improvement

🔄 Active Learning Strategies

- **Uncertainty Sampling:** Prioritize cases with 45-55% confidence (near decision boundary)
- **Diversity Sampling:** Select cases from underrepresented regions of feature space
- **Error-Based Selection:** All false negatives (highest priority), false positives (medium priority)
- **Representative Sampling:** Include correct predictions to avoid dataset bias
- **Clinical Flagging:** Cases marked as challenging by radiologists

🏥 Implementation: Breast Cancer Mammography

System Setup:
• 4 radiologists review 200 AI-flagged cases/week
• Monthly model updates with accumulated feedback
• Active learning prioritizes high-value cases

6-Month Progress:

5. Continuous Improvement - Version Management

Systematic Process for Model Evolution

🚀 Deployment Pipeline (Safety-First Approach)

Stage 1: Development → Experiment with architectures, train models
Stage 2: Validation → Test on held-out data, cross-site validation
Stage 3: Shadow Mode → Run in parallel (no clinical impact), compare to current model
Stage 4: A/B Testing → 10% traffic, statistical comparison, safety monitoring
Stage 5: Gradual Rollout → 50% → 100% with automated rollback capability
Stage 6: Full Deployment → All traffic, continuous monitoring

📈 Evolution Example: CT Pulmonary Embolism Detection

v1.0 (Jan 2023): 89% sens., 92% spec. - Initial deployment
• Known issues: Subsegmental PE, motion artifacts

- **Month 1:** 92% sens., 88% spec. (baseline)
- **Month 2:** 93.5% sens., 89.2% spec. (+1.5% / +1.2%)
- **Month 3:** 94.2% sens., 90.5% spec. (+0.7% / +1.3%)
- **Month 6:** 96.1% sens., 92.8% spec. (+4.1% / +4.8% total)

Impact:
- 4,200 new labeled cases incorporated
- False negative rate reduced 65%
- Review time: 2.0min → 1.2min per case
- Radiologists report seeing direct improvements from their input
- Estimated 15-20 additional cancers detected per year

v1.1 (Apr 2023): 91% sens., 93% spec. - First update
- Added 3,000 subsegmental PE cases
- A/B tested with 20% traffic for 2 weeks
- Result: 15% reduction in false negatives ✓

v1.2 (Jul 2023): 91.5% sens., 94% spec.
- Motion artifact detection preprocessing
- Result: 25% reduction in motion-related false positives

v2.0 (Oct 2023): 94% sens., 95% spec. - Architecture upgrade
- Switched to 3D transformer
- Extended testing: 4 weeks shadow, 6 weeks A/B
- Significant improvement, became new standard

v2.1 (Jan 2024): 94.5% sens., 96% spec.
- Multi-site domain adaptation
- Consistent performance across 12 hospitals

Total Improvement: +5.5% sensitivity, +4% specificity over 12 months
Clinical Impact: ~50 additional PE cases detected annually

✅ Regression Testing Checklist

- **Performance Tests:** Overall metrics ≥ baseline, subgroup performance maintained
- **Consistency Tests:** No prediction flip-flops on stable cases
- **Safety Tests:** Zero increase in critical false negatives
- **Technical Tests:** Inference latency < threshold, memory requirements

- **Integration Tests:** API compatibility, deployment infrastructure

Summary: Building Robust QA Systems

🎯 Key Takeaways for Implementation

💡 Start Simple, Scale Gradually

Begin with basic monitoring and error tracking. Add sophisticated drift detection and active learning as system matures.

🏗️ Invest in Infrastructure

Automated monitoring and dashboards pay for themselves through early problem detection and reduced downtime.

🤝 Engage Clinicians

Treat clinical staff as partners. Their feedback and domain expertise are invaluable for improvement.

📊 Measure What Matters

Focus on metrics aligned with patient outcomes and clinical workflow, not just technical performance.

🛡️ Build for Safety

In medical AI, false negatives are often more

📝 Document Everything

Maintain detailed records of versions, changes, and

critical than false positives. Design QA accordingly.

impacts. Critical for compliance and learning.

The Five Pillars Work Together

Performance Monitoring detects issues → **Drift Detection** identifies root causes → **Error Analysis** reveals patterns → **Feedback Loops** enable learning → **Continuous Improvement** systematically enhances the model