

AI and Quality Improvement

What changes for the quality professional, and what does not

Razaq Badamosi, MD, FCCP

Vice President, Network Quality and Chief Quality Officer
Associate Professor of Medicine Burnett School of Medicine

A Quick Question for the Chat

Where does your Quality team spend the most time today?

1

Finding the data

locating cases, reports, source fields

2

Abstracting / Validating

manual review, reconciliation, submission

3

Explaining the measure

why the rate moved or failed

4

Improving the system

testing, redesigning, sustaining change



The promise of AI is not simply faster reporting. It is to move Quality work to the right.

Less time finding the signal. More time understanding it and changing care.

Session Objectives

1

Describe

The evolving AI Landscape.

2

Explore

The role of the Quality and Patient Safety Professional.

3

Discuss

AI Governance and changing external requirements.

AI Is Not a Single Thing



TYPE	WHAT IT DOES	CLINICAL USE CASES	QUALITY AND OPERATIONS USE CASES
Predictive models	Scores the probability of a future event from structured data	Sepsis and deterioration scores, readmission, no-show, fall and delirium risk	PSI and HAC case triage, readmission outreach lists, LOS forecasting
Computer vision	Interprets images and video	Radiology triage (bleed, PE, LVO), retinal and pathology screening	Wound and pressure injury imaging, room-based fall detection, hand hygiene
Language classification (NLP)	Reads free text and assigns codes, categories or flags	Problem list and coding suggestions, registry and abstraction support	Safety event categorization, HAI and PSI chart screening, complaint themes
Generative AI (LLMs)	Produces new text, summaries and answers	Ambient documentation, discharge summaries, patient message drafts, chart Q&A	Policy crosswalks, RCA and PSI narrative drafting, survey readiness search
Rules and automation	Executes fixed logic and routes work (RPA, workflow bots)	Order sets, interruptive alerts, prior authorization, scheduling, coding edits	eCQM logic, worklists, escalation routing, report distribution
Agentic AI	Chains tools and actions toward a goal with limited supervision	Referral and scheduling agents, care-gap outreach (emerging)	Automated measure investigation, tracer evidence gathering (emerging)

Most organizations already run rows 1, 3 and 5 without calling them AI. An honest inventory counts all six.

SO THE EVIDENCE YOU DEMAND AND THE WAY YOU MONITOR MUST MATCH THE FUNCTION

Each Type Fails Differently



TYPE	HOW IT FAILS	EVIDENCE BEFORE GO-LIVE	WHAT TO MONITOR AFTER
Predictive models	Drift, miscalibration, subgroup degradation	Local sensitivity, PPV and calibration at our prevalence	Flag rate, PPV, subgroup performance, version changes
Computer vision	Poor transfer to our scanners, population and edge cases	Local validation on our images; FDA status and intended use	Discordance with clinician read, missed findings, turnaround
Language classification (NLP)	Misclassification, terminology mismatch, missed negation	Precision and recall against a human abstraction sample	Ongoing adjudication sample, override and correction rate
Generative AI (LLMs)	Hallucination, omission, subtle distortion, confident errors	Structured review against the source chart; defined intended use	Edit rate, error taxonomy, patient-facing complaints
Rules and automation	Silent logic change, stale value sets, errors at scale	Test cases, value-set version control, change control	Logic change log, exception queue, downstream harm signals
Agentic AI	Compounding errors across steps, acting outside scope	Scope limits, human checkpoints, a kill switch	Action logs, escalation rate, stop-rule triggers

One governance process. Different evidence, different monitoring, matched to the function.

PART ONE

The tools are not obsolete. The assumptions are changing.

Our tools will not suddenly stop working. Their assumptions become easier to violate.



JPS Health Network
Fort Worth, Texas

What Is Quality Improvement?

Quality Improvement is the systematic, data-guided work of **changing** care processes so outcomes become reliably better, and stay better.

AI enters QI through the measurement system, the analysis system and the work system.

It can change what we see, how quickly we see it, and sometimes the measure itself.

QI REQUIRES FOUR THINGS

AIM

What are we trying to accomplish?

MEASURE

How will we know a change is an improvement?

CHANGE

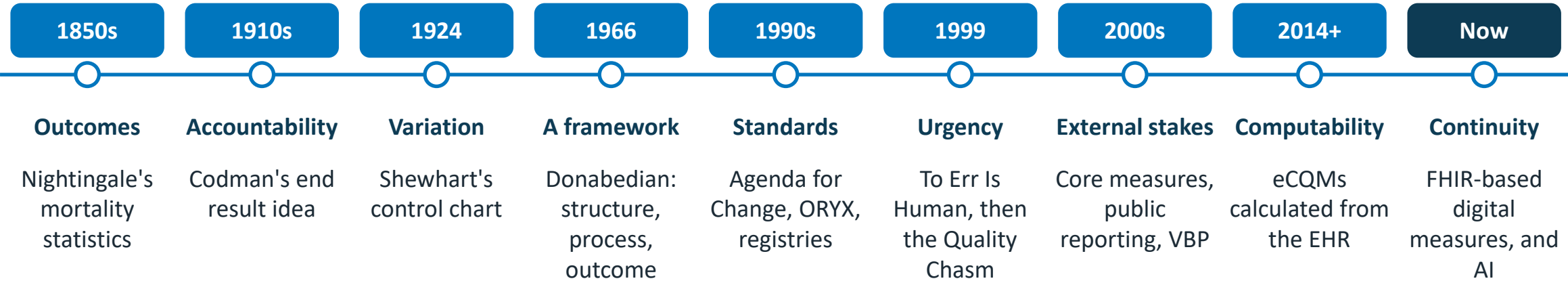
What can we alter in the system?

LEARNING

What did the test teach us?

EACH ERA MEASURED WHAT IT COULD OBTAIN

How Healthcare Learned to Measure Quality



The method was never chosen purely on science. It was chosen on what data could be obtained, at what cost, at that moment.

THE 1966 FRAMEWORK IS STILL THE CLASSIFICATION BEHIND MOST MEASURE SETS

Structure The setting and capacity of care: staffing, equipment, policies, governance	Process What is actually done for the patient: antibiotic within the window, screening completed	Outcome What happened to the patient: mortality, harm, readmission, function, experience
---	--	--

The Quality Professional's Toolbox

Measurement

- Chart abstraction
- Sampling
- Run charts
- SPC and control charts
- Risk adjustment and O/E
- Benchmarking

Improvement

- PDSA and the Model for Improvement
- DMAIC and Lean
- Driver diagrams
- Pareto analysis
- Standard work
- Spread and sustainability

Safety and investigation

- RCA and RCA squared
- 5 Whys and fishbone
- FMEA and HFMEA
- Event reporting taxonomy
- SEIPS and human factors
- STPA and CAST

The question is not “Which tools die?” It is “Which assumptions underneath them change?”

Did Care Change, or Did the Measurement Change?

1

AI / ambient documentation

More complete capture of comorbidities

2

Coding / data capture

Case-mix severity becomes more complete

3

Expected outcome

Expected mortality or complication risk rises

4

Risk-adjusted result

O/E improves even if observed outcomes are unchanged

The control chart may be faithfully showing a real shift, in documentation or case-mix capture, not in clinical care.

A signal appeared. Ask both sets of questions.

ASK FIRST, AS ALWAYS

What changed clinically?

Who, where, and when?

Is it sustained or a blip?

Does the frontline recognize it?

NOW ASK AS WELL

Did a documentation tool go live?

Did a model version or threshold change?

Did an Epic field or mapping change?

Did the measure specification change?

Practical answer: annotate the chart with model, documentation and specification versions, the same way you already annotate an intervention. Keep a measurement-system change log next to the improvement change log.

PDSA Still Works. Now Version the Intervention.

Plan

Define the change, the population, the measure, and the version of every system involved.

Do

Deploy, and record actual use. Overrides and workarounds are data, not noise.

Study

Look at outcome, workflow response, and whether the measurement system held still.

Act

Adopt, adapt, abandon, and record what version you are carrying forward.

IF ANY OF THESE CHANGED BETWEEN DO AND STUDY, YOU ARE NO LONGER STUDYING THE SAME INTERVENTION

Model version

Prompt or template

Field mapping

Alert suppression

Documentation AI

This is not a flaw in PDSA. It is a measurement and attribution problem that PDSA can absorb, if you write the versions down.

How AI Changes the QI Toolbox



QI TOOL	TRADITIONAL ASSUMPTION	WHAT AI CHANGES	THE QUALITY RESPONSE
Chart abstraction	Humans review a sample	Near-complete screening becomes feasible	Human adjudication becomes the validation layer
SPC and control charts	The measurement system is stable	Documentation or pipeline change moves the measure without changing care	Annotate model and data versions; re-baseline when justified
Risk-adjusted O/E	Case-mix capture is consistent	AI documentation and coding shift comorbidity capture	Track case-mix drift alongside the clinical outcome
Pareto analysis	Categories are predefined	Language models surface clusters outside the taxonomy	Pair the fixed taxonomy with discovered clusters
PDSA	The intervention holds still during the test	Model, prompt, mapping or workflow may change mid-cycle	Version-control the intervention and the measurement system
RCA, 5 Whys, fishbone	The causal chain can be reconstructed	Model, data, workflow and human behavior interact	Add data lineage and model version; use STPA or CAST when warranted
FMEA	Failures are discrete, foreseeable steps	Drift and subgroup degradation unfold over time	Add continuous model and data assurance

None of these tools is retired. Each needs one additional question asked of it.

PART TWO

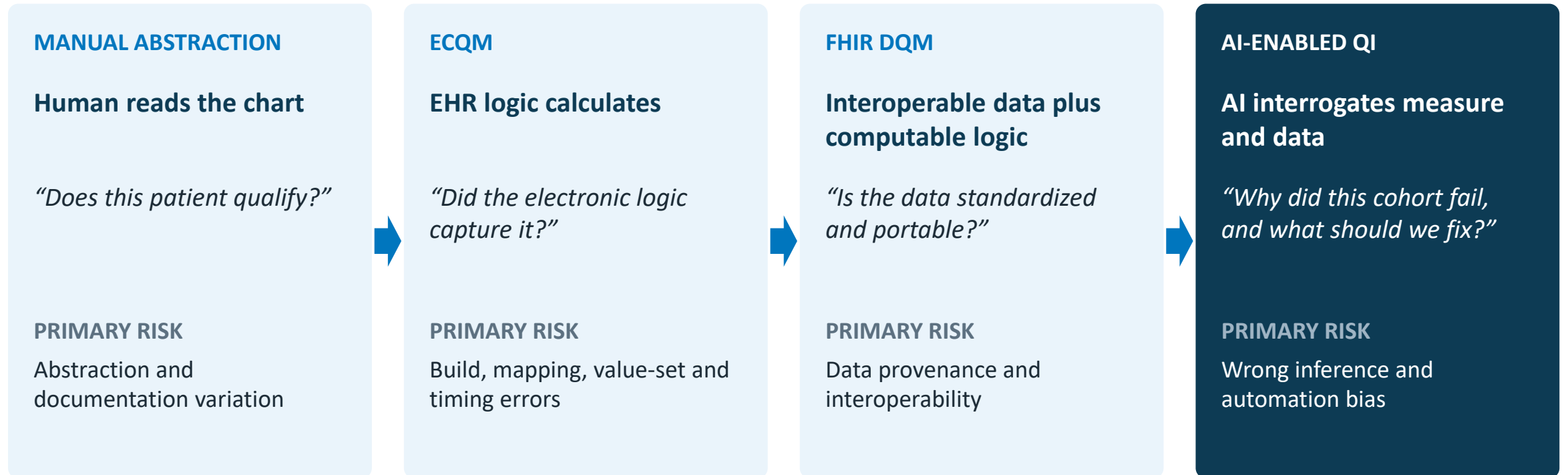
What happens to our day job?

From chart abstraction, to eCQM, to digital quality measurement



JPS Health Network
Fort Worth, Texas

Chart Abstraction → eCQM → Digital Quality → AI-enabled QI



The value of the Quality professional moves from collecting data, to validating meaning, to designing improvement.

The Chain Between a Patient and a Percentage



The measure result can fail at any one of these six points, and only the first one is about care.

What was found preparing for one measure

- Required elements documented in narrative text but not in the structured location the logic reads
- Local terminology that did not match the expected value set
- Inconsistent present-on-admission and timing capture
- Conflicting information across two documentation sources

The uncomfortable implication

A prevention program can be working while the measure says it is not.

And a measure can look acceptable while care is not, if the capture is incomplete in a convenient direction.

Neither error is visible from the dashboard alone.

Four Very Different Answers

Care gap	Capture gap	Mapping gap	Specification gap
The care genuinely did not happen. THE FIX Clinical improvement work. This is the one we are trained for.	The care happened, but not where the logic looks. THE FIX Documentation redesign and clinician workflow.	The data exists, but the local field or code does not map to the measure. THE FIX Build and terminology work with IT.	The measure changed and our implementation did not change with it. THE FIX Annual measure-year review and version control.

Same red measure. Four different owners. Diagnosis first, then improvement.

From Abstractor to Improvement Designer

YESTERDAY

Find

Read

Abstract

Submit



TODAY AND FUTURE STATE

Validate

Interrogate

Explain

Intervene

Monitor

What becomes scarce

Not the ability to pull a report. Software does that.

Not the ability to abstract a chart. That becomes a validation sample.

What becomes scarce is the person who can look at a red measure and say whether it is a care problem, a documentation problem, or a build problem, and then fix the right one.

That skill is not automatable.

PART THREE

The Regulatory Landscape: The rules have changed. Have you?

What changes for regulatory, accreditation and risk work



JPS Health Network
Fort Worth, Texas

Accountability Flips to the Deployer



Federal: device

FDA lifecycle guidance and post-market monitoring. Most generative AI documentation and decision support is non-device, so the deployer owns safety.



Federal: transparency

The HTI-1 rule requires a predictive decision-support “nutrition label”: source attributes for training data, intended use, and validity and fairness on local data.



Accreditation

Joint Commission Responsible Use of AI guidance and certification pathway, plus the Coalition for Health AI playbooks, are the readiness target.

Five obligations recur across every framework



Documented risk and impact assessments



Transparency on training data and local performance



Mandated post-deployment monitoring



Algorithmic-discrimination and equity duties



Disclosure and consent for patient-facing AI

Start with a registry of every deployed AI tool. You cannot govern, monitor or investigate what you have not inventoried.

Key AI Regulations at a Glance



TRAIGA

Texas Responsible AI Governance Act

- Effective January 1, 2026; intent-based, enforced by the Attorney General
- Providers must disclose AI use in diagnosis or treatment
- Penalties up to \$200,000 per violation, with a 60-day cure period



SB 1188

Texas Senate Bill 1188 (in effect September 2025)

- Physician review of AI-generated diagnostic records
- US-based storage of electronic health records
- Patient disclosure duties for clinical AI



HTI-1

ONC/ASTP algorithm transparency rule

- Predictive decision-support “nutrition label” of source attributes
- Fair, Appropriate, Valid, Effective, Safe (FAVES) evaluation lens
- Applies through your certified EHR developer



NIST safe harbor

NIST AI Risk Management Framework

- Documenting your program to the NIST AI RMF is an affirmative defense under TRAIGA
- Map controls to it once to satisfy compliance and defense together

Seven Elements of Responsible Use of AI in Healthcare



Roadmap: High-level guidance → Responsible-Use Playbooks (baseline controls) → Voluntary Joint Commission certification

Already Happening. Be Precise About What Is Proven.

WHAT EXISTS TODAY

- In May 2025, The Joint Commission announced a long-term partnership with Palantir to apply AI and analytics to accreditation and certification, with the Joint Commission retaining data ownership.
- Health systems named in that announcement report operational gains (placement time, PACU holds, MRI turnaround) and clinical gains (sepsis length of stay).
- Machine-readable measure specifications are published openly.
- Platform capabilities for ontology, lineage and data health exist.
- CMS validation now audits eCQM output against the chart.

WHAT I AM INFERRING

I have not seen public evidence that any organization is currently feeding measure logic and local mapping into one of these platforms to diagnose capture and configuration failures at scale.

What I am arguing is that the pieces now exist, and that this is the obvious next use.

The reported vendor results are the organizations' own numbers. Apply the same evidence standard to them that you would apply to your own.

Naming your own inference is the behavior we are asking of vendors.

Regulatory Readiness Becomes Evidence Management

TODAY

Standard

Manual crosswalk

Policy search

Spreadsheet

Document request

Mock tracer

Corrective action



THE QUESTIONS AI CAN ASK OF LINKED EVIDENCE

- 1 Which requirements have no supporting policy?
- 2 Where does the policy say one thing and the workflow do another?
- 3 Which requirements changed since the last version?
- 4 Which areas have weak or stale evidence?
- 5 Where should tomorrow's tracers be targeted?
- 6 Are corrective actions still holding six months later?

Know your own evidence before anyone asks for it.

PART FOUR

AI governance: Does it work here? Will it keep working? How would we know?

Case study: local validation of a predictive sepsis model



JPS Health Network
Fort Worth, Texas

The Tool

Score every eligible patient, continuously, on the probability of meeting sepsis criteria within a defined future window, so the team can intervene earlier.

- 1 Continuous** Rescored every 30 minutes, around the clock
- 2 Population-wide** Every eligible inpatient, ED and observation encounter
- 3 Probabilistic** It produces a score, not a diagnosis
- 4 Opaque** Dozens of variables, no clinically inspectable chain

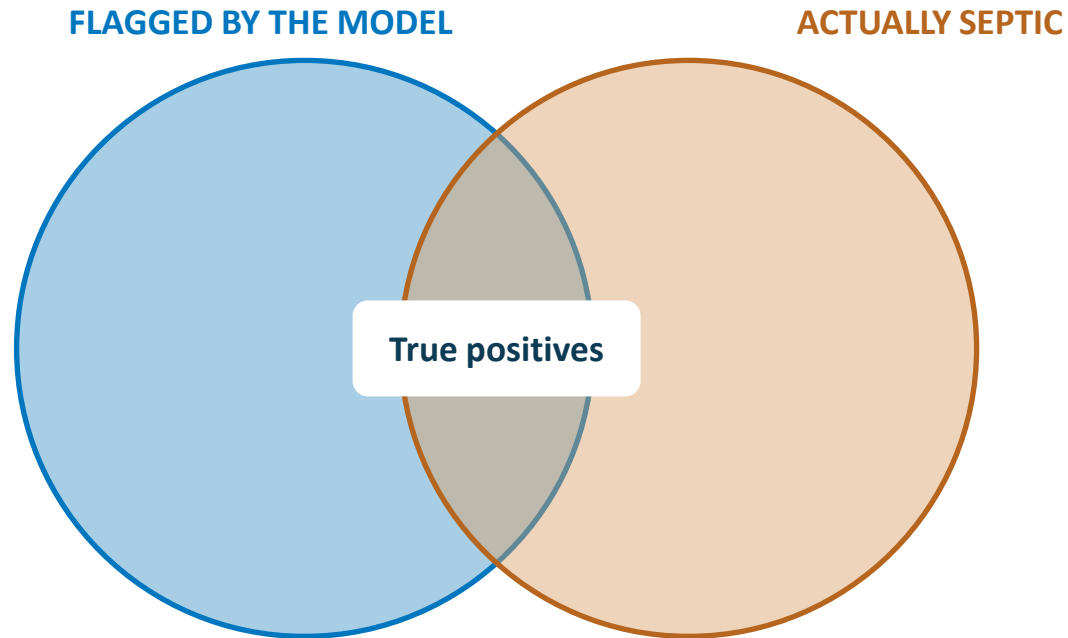
Why this one?

Sepsis is time critical, high volume, and high mortality.

The model is deployed at hundreds of US hospitals, so many of you already have it running.

And it is the clearest possible demonstration that vendor performance and local performance are not the same number.

The Fishing Net



Widen the net and you catch more fish, but less of the catch is fish. That trade is a governance decision, not a technical one.

Flag rate

How big is the net?

The share of all patients the model alerts on. This is your workload.

Sensitivity

How many fish did we catch?

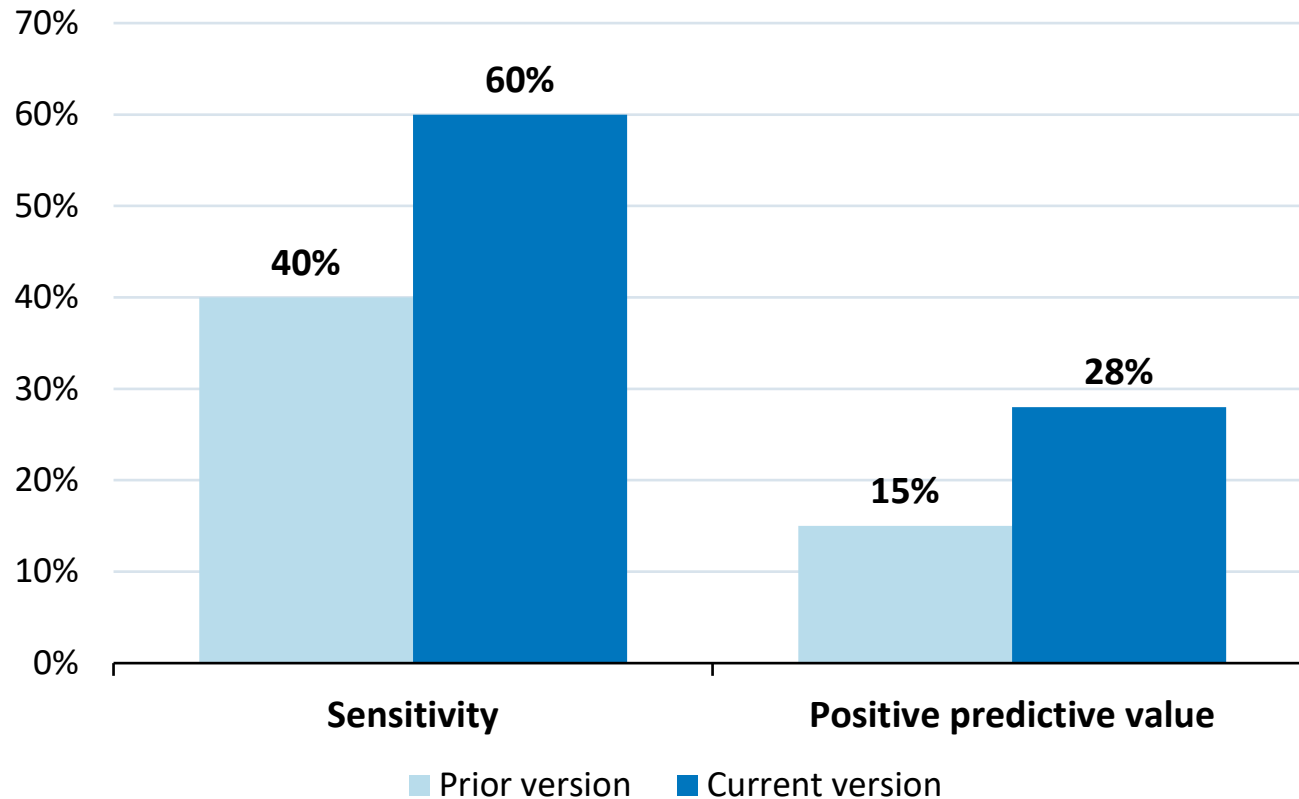
The share of truly septic patients the model flagged.

PPV

How much of the catch is fish?

The share of alerts that are real. This is your credibility.

What Local Validation Showed Us



DISCRIMINATION

0.74 → 0.89

C-statistic

ALERTS PER TRUE CASE

6 → 3.5

Alert burden

Directionally better on every measure. Still our threshold to set.

Values rounded and de-identified. The two versions were validated on different cohorts, so these are directional, not a controlled head-to-head comparison.

The Threshold Is a Governance Decision

Earlier detection

versus

Alert burden

Fewer missed cases

versus

More false alerts

So who gets to decide where the threshold sits?

Data scientist

Vendor

CMIO

Sepsis physician lead

Nursing leadership

**Quality and patient
safety**

The honest answer

None of them alone.

Threshold selection is a multidisciplinary clinical policy decision informed by data, not a setting configured by whoever built the model.

Document who decided, on what evidence, and when it will be revisited. That record is what you will want when someone asks why the alert did or did not fire.

And Performance Really Does Vary by Hospital

227,091

inpatient encounters

across 4 major US health systems, prospectively validating the same commercially available sepsis model.

- Discrimination ranged from 0.82 to 0.92 across sites
- High institutional variability in performance
- Low positive predictive value and high alert burden

The authors' own conclusion: you are responsible

Institutions implementing this model should conduct local validation studies to verify performance, design workflows to manage false positives, and use alert silencing strategies to reduce burden.

Vendor: *"This model works."*

Quality: **"Does it work here?"**

Model Performance Is Not Clinical Performance



Any term can be near zero. A technically excellent model that nobody acts upon is a poor clinical intervention.

Automation bias

Accepting the output because the machine said so

Automation complacency

Reduced vigilance when a tool is usually right

Alert fatigue

The rational response to low positive predictive value

Deskilling

Erosion of the judgment the tool was meant to support

Human factors, automation bias and alert fatigue are patient safety work. Put the human-AI interaction on the FMEA.

Monitor the Model AND the System

Technical

- Discrimination
- Calibration
- Pipeline integrity

Clinical

- Sensitivity and PPV
- Time to treatment
- Outcomes

Workflow

- Response rate
- Override rate
- Acceptance

Equity

- By age and race
- By language and payer
- By site

Safety

- AI-involved events
- Near misses
- Reported concerns

Most organizations monitor the first dial, sometimes the second, and almost never the last three. The last three are where quiet failure lives.

Know When to Stop

1

Performance falls below the pre-agreed floor

2

A major model or version change is pushed by the vendor

3

An unsafe signal appears in any monitored subgroup

4

Alert burden becomes clinically unsustainable

5

Frontline abandonment or systematic workaround is detected

6

A serious safety signal remains unresolved

Write the stop rules before go-live, while everyone is optimistic. Nobody writes a good stop rule during a crisis.

This is classic quality thinking. We do it for medications and devices. We rarely do it for algorithms.

PART FIVE

AI and Patient Safety

Seeing AI when it causes or contributes to harm



JPS Health Network
Fort Worth, Texas

Another Question for the Chat

How many AI-enabled tools are currently operating in your organization?

0 to 5

6 to 20

More than 20

I honestly do not know

You cannot govern, monitor or investigate what you have not inventoried. Every recommendation in this presentation depends on knowing what is running.

Count all six types from the taxonomy: predictive, vision, language classification, generative, rules and automation, agentic. Most inventories miss the rules and automation layer entirely.

The Invisible Event Problem

Our incident reporting system asks whether a medication, a device, or a fall was involved. Does it ask whether AI was involved?

No field leads to **No data**

If the system never asks, the answer is never recorded.

No data leads to **No signal**

AI-involved harm cannot trend, because it cannot be counted.

No signal leads to **No learning**

Governance reviews performance dashboards and never sees the events.

A Pareto chart of safety events is only as good as its categories. Today, for most of us, AI is not one of them.

A Minimum AI Safety Event Dataset

1 Was AI involved in this event?

2 Which system, and which version?

3 What did the AI actually do or output?

4 How did the human respond: accepted, overridden, ignored?

5 What failure mode occurred?

6 Did it reach the patient, and did it cause harm?

7 Does it require a regulatory reportability assessment?

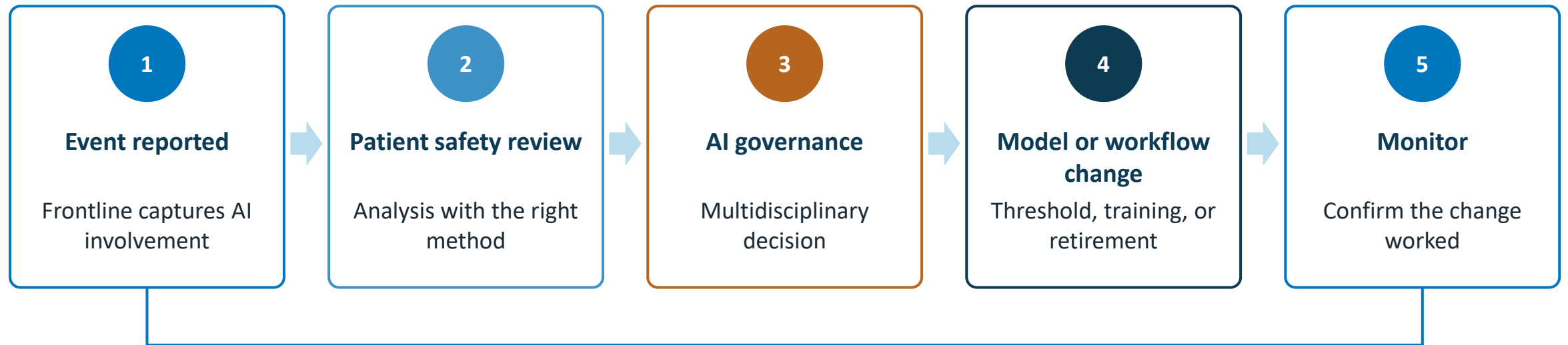
Where we are

We have designed an AI event type for our reporting platform and it is in development with the vendor: a gate question, system and version, failure mode, human review status, event reach and a reportability flag.

Why it matters now

The Joint Commission's responsible AI framework describes internal and external reporting processes for AI safety events. The field is coming whether or not we build it.

Close the Learning Loop



and the monitoring result returns to the top of the loop

The loop is what separates AI governance from AI paperwork. If an AI-involved event cannot change a threshold, a workflow, or a decision to keep using the tool, the committee is decorative.

PART SIX

The AI-enabled Quality Department

Where the work goes, and the guardrail that comes with it



JPS Health Network
Fort Worth, Texas

AI Assists. Humans Adjudicate.



AI

- Finds
- Organizes
- Compares
- Proposes

PEOPLE

- Adjudicate
- Understand context
- Decide
- Lead the change

And govern the tools you use on yourselves

Every principle in this talk applies to the AI your own department adopts. Inventory it, name an owner, validate it locally before you trust its output, and be able to say how you would know if it started getting things wrong.

LOWER RISK

Summarize

Categorize

Draft

Lighter controls are reasonable

Human reviews the output before use · Sampling-based quality checks · Documented intended use

HIGHER RISK

Find patterns

Recommend

Decide

Stronger controls are required

Local validation before deployment · Subgroup performance audit · Named clinical owner and stop rules

Three Things to Do Monday Morning

1



Inventory every AI tool

All six types, including the rules and automation layer nobody calls AI. Name an owner for each. You cannot govern what you have not counted.

PART THREE

2



Annotate the measurement system

Add model versions, documentation tools, mapping and specification changes to your control charts and change logs, the same way you annotate an intervention.

PART ONE

3



Add the gate question

Put “Was AI involved?” on the event report. No field, no data, no signal, no learning. Then close the loop so the answer can change something.

PART FIVE

None of these needs a new budget. All three need a decision.

The Work Moves Upstream

FROM

Find the data

Abstract the chart

Calculate the measure

Report what happened months ago



TO

Validate the data

Understand the measure

Explain the variation

Find the system weakness

Design the intervention

Prove patients actually got better

AI should not remove quality professionals from improvement. It should remove the work that keeps us from improving quality.

What AI Cannot Replicate

E	Empathy	Understand what patients and frontline clinicians actually experienced
P	Presence	Build the psychological safety and trust that make people report and speak up
O	Opinion	Judge acceptable residual risk when the evidence is incomplete
C	Creativity	Redesign the system when the familiar approaches have already failed
H	Hope	Set a safety vision and help teams believe improvement is still possible

AI can accelerate the analysis. Human leadership decides what it means and what happens next.

Acknowledgements

Brooke Ortiz, Executive Assistant

Keeps me organized, my calendar and myself protected, which is the only reason work like this gets finished. Thank you, Brooke.

The JPS Network Quality Team

The best quality team in the world.

Every example in this hour came from a team doing the work, not from a slide.

THANK YOU

Questions



Razaq Badamosi, MD, FCCP

Vice President, Network Quality
Chief Quality Officer

Associate Professor of Medicine Burnett School of Medicine

Selected References and Sources

- Donabedian A. Evaluating the quality of medical care. 1966. Structure, process and outcome classification, still used by AHRQ.
- Institute of Medicine. To Err Is Human, 1999; Crossing the Quality Chasm, 2001.
- Shewhart WA. Control chart, Bell Laboratories, 1924.
- CMS. FY2025 IPPS/LTCH final rule: eCQM expansion and the start of eCQM data validation scoring with CY2025 discharges.
- CMS and eCQI Resource Center. Digital quality measures, FHIR quality measure implementation guide, QI-Core, CQL, and MADiE.
- NCQA. Proposed timeline for retiring and replacing HEDIS hybrid measures.
- The Joint Commission. Accreditation 360 and the May 2025 Palantir partnership announcement.
- Wong A, et al. Multicenter prospective validation of an updated proprietary sepsis prediction model. JAMA Netw Open. 2026.
- Sittig DF, Singh H. Recommendations to ensure safety of AI in real-world clinical care. JAMA. 2025.