

White Paper

AI Safety in Health Systems: Building Infrastructure and Strengthening Risk Management Practices



March 26, 2026

Authors

Cameron Joyce, MPA,

Duke-Margolis Institute for Health Policy

Nicoleta J Economou, PhD,

Duke Health AI Evaluation and Governance

Christina Silcox, PhD,

Duke-Margolis Institute for Health Policy

Acknowledgments

The authors would like to thank several individuals for their contributions to this white paper. First, we thank the participants of our expert workshop, who are listed at the end of the paper, for sharing their expertise and experiences, as well as the multiple other health system representatives and policy influencers that held individual informational calls with us. We would also like to thank Hannah Vitello, Luke Durocher, and Michelle Langlois for their help with this paper and the associated meetings, and Laura Hughes for design support. Any opinions expressed in this paper are solely those of the authors and do not necessarily represent the views or policies of any other person or organization external to Duke-Margolis. This work was funded by the Duke Endowment.

About the Duke-Margolis Institute for Health Policy

The Robert J. Margolis, MD, Institute for Health Policy at Duke University is directed by Mark McClellan, and brings together expertise from the Washington, DC, policy community, Duke University, and Duke Health to address the most pressing issues in health policy. The mission of Duke-Margolis is to improve health, health equity, and the value of health care through practical, innovative, and evidence-based policy solutions. Duke-Margolis catalyzes Duke University's leading capabilities, including interdisciplinary academic research and capacity for education and engagement, to inform policymaking and implementation for better health and health care. For more information, visit healthpolicy.duke.edu.

Duke Health AI Evaluation & Governance

Duke Health is an academic health system advancing care through clinical excellence, education, and research. Its network of hospitals, clinics, and specialty centers serves diverse populations across North Carolina and beyond. Duke Health emphasizes innovation, fairness, and continuous improvement, with a focus on precision medicine, AI-enabled care, and population health. Through the Duke Health AI Evaluation and Governance Program, Duke Health works to promote the responsible and trustworthy use of AI within clinical care and beyond. The program integrates evaluation, oversight, and ongoing monitoring into health system operations to support safe and effective use of AI in practice. It also contributes to broader efforts to advance best practices, raise awareness, and improve how health systems implement and oversee AI. This work is grounded in a shared goal of ensuring that AI solutions are safe, trusted, fair, and meaningfully integrated into patient care to improve health outcomes.

OVERVIEW

As clinical artificial intelligence (AI) tools are [deployed at increasing scale](#) across health care settings, health systems and regulators face growing challenges to understand and manage the potential patient safety risks. Emerging AI safety risks are difficult to detect through existing patient safety monitoring systems. Traditionally, safety events are often identified by clinicians, patients, or family members, with patients and families typically serving as early detectors of harm. However, clinical AI tools used in health systems are often largely invisible to patients, which makes patient reporting of AI safety events more challenging. As such, [effective system-level proactive risk management](#), including but not limited to approaches to identify, track, and respond to patient safety events, are needed to support the [responsible use of these technologies](#). This white paper summarizes key insights from meetings and discussions with leaders from health systems, AI tool developers, industry organizations, policy experts and regulators. It discusses emerging practices and gaps in AI tool tracking and monitoring and outlines policymakers' considerations that could strengthen the consistent prevention, identification, documentation, and oversight of patient safety associated with the use of clinical AI tools.

A **patient safety** event is commonly defined as any process, action, or omission that causes or has the potential to cause patient harm, meaning unintended physical or emotional injury resulting from or contributed to by medical care that requires additional monitoring, treatment, hospitalization, or results in death. The [World Health Organization](#) and [American Medical Association](#) note that patient safety events should include errors, system failures, equipment issues, or communication breakdowns. In [the 2025 CHAI and TJC Responsible Use of AI in Healthcare Guidance](#), patient safety events are also described as systemic errors that involve unsafe recommendations, performance failures/degradation, or near-misses.

INTRODUCTION TO LIFECYCLE-BASED RISK MANAGEMENT

While AI tools can provide value to clinicians and patients, the complexity of AI technology introduces additional risks compared to traditional clinical tools, including the risk of performance drift and [bias](#). In addition, clinical tools that rely on large language models (LLMs) also introduce the risk of hallucinations and overreliance by the user, often due to the authority with which outputs are presented regardless of the uncertainty in the calculation. The rapid acceleration of AI capabilities and widespread push to deploy the technologies across clinical environments creates an urgent need for health systems to develop AI risk management plans processes on how to prevent, identify, track, and monitor patient safety events associated with these tools.

Health systems have ethical and legal obligations to identify and manage risks that threaten patient safety. In addition to traditional safety reporting/issue management and documentation, proactive risk management of clinical AI should include workflow integration, training, and monitoring that takes individual model performance, fairness, and usability into consideration. Health systems need clear governance structures that define ownership and accountability for managing clinical AI tool-related risks across the tool's entire lifecycle. These processes help anticipate and mitigate potential failures early and adapt mitigation strategies as new risks emerge during real-world use.

While there is much written on AI governance and general risk management, there are no widely adopted organizational practices and procedures to ensure patient safety when clinical AI tools are used. Also, less explicit discussion exists for patient safety-specific risk mitigation, monitoring, and event reporting involving clinical AI tools. At this time, many health systems lack sufficient expertise and information to fully anticipate the range of patient safety risks introduced by clinical AI tools. This is due in part to the lack of clear standardized approaches for monitoring clinical AI tool safety performance or clearly established accountability for performance and monitoring within health organizations and vendor partners.

The current regulatory framework for clinical AI tools is fragmented and often confusing. Oversight is distributed across multiple federal agencies rather than centralized in a single authority. Depending on a tool's function and implementation, it may fall under the jurisdiction of the U.S. Food and Drug Administration (FDA), the U.S. Department of Health and Human Services' (HHS) Office for Civil Rights, or the Office of the National Coordinator

for Health IT. For example, the FDA regulates AI that qualifies as software as a medical device (SaMD), including certain clinical decision support (CDS) tools, with oversight spanning pre-market review and select post-market changes. However, many tools like administrative or documentation-focused AI systems fall outside of FDA authority and even any federal regulation. Adding further complexity, states are increasingly enacting their own AI health care laws.

Without intentional planning and investments, resource constraints and limited access to staff with AI expertise risk widening existing resource disparities between health systems and exacerbating inequities in patient experience and clinical outcomes.

Risk Management Frameworks

While not specific to health, the NIST Artificial Intelligence Risk Management [Framework](#) and [Playbook](#) states that AI trustworthiness includes validity, reliability, security, resilience, accountability, transparency, explainability, integration, privacy protection, and fairness with bias mitigation. It defines risk as both technical and socio-technical, emphasizing context. Systems should establish governance, define the use case, assess risk, evaluate trustworthiness characteristics, and implement controls, monitoring, and incident response.

The IHI Lucian Leape Institute convened an [expert panel](#) in early 2024 to discuss the safety implications of generative AI in healthcare, with a focus on patient safety, clinician impact, and workflow integration. Their findings were published in the report [Patient Safety and Artificial Intelligence: Opportunities and Challenges for Care Delivery](#) outlines potential AI harms and mitigation strategies, with recommendations to manage risk and maximize benefit: prioritize ethical, patient-centered design; rigorously validate tools for efficacy and bias; engage clinicians in deployment; establish strong organizational and federal governance; ensure clinical relevance; and support continuous evaluation and cross-system learning.

The Coalition for Health AI (CHAI) developed a consensus-based framework and actionable guidance to drive the responsible, safe, and trustworthy adoption of AI in healthcare through principles, checklists, and lifecycle risk management. Their work emphasizes patient safety, bias mitigation, transparency, governance, and continuous monitoring. Their [Responsible AI Guide](#), operationalize core principles that explicitly embed safety considerations into the early planning, design, deployment, and monitoring phases of an AI system's lifecycle.

Patient Safety Reporting

One of the most important elements of traditional risk management is a robust system for tracking and monitoring safety events, which enables health systems to capture both adverse events and near misses across the full continuum of care. Traditional safety reporting processes are typically based on industry-wide standards and best practices, which are developed and shared by professional organizations and learning networks. Regulatory frameworks and financial incentives further reinforce these monitoring systems. These include [Medicare Conditions of Participation](#), Joint Commission (JC) accreditation, the [Patient Safety and Quality Improvement Act](#), or payment programs administered by Center for Medicare & Medicaid Services (CMS) to help incentivize reporting and performance on safety outcomes through pay-for-reporting, value-based purchasing, and penalty-based programs tied to preventable harm.

Health systems face multiple challenges in tracking and reporting AI-related safety events, including consistent errors introduced by AI that may not be identified by an

organization or captured within traditional patient safety event reporting systems. This may be exacerbated by frequent software updates that potentially alter risks and inconsistent user awareness of AI tools that may be part of clinical workflows. Gaps in education, governance, and communication, along with varying clinician trust in AI tools, can lead to under-reporting or misattribution of patient safety events. Since AI tools used in health systems are largely invisible to patients (especially in the lower risk categories of AI tool), safety events may be less likely to be reported by patients and caregivers, or the report may not make mention of an AI tool being a contributor to an event. Clear communication and accessible reporting mechanisms will be essential, particularly towards improving capture of “near misses” as well as events. Near misses often occur more frequently than adverse events and provide critical opportunities to identify system weaknesses and prevent future harm. Since individual near misses and events may vary in severity, systematic tracking supports the identification of broader patterns and informs targeted interventions to reduce risk to patients and the organization.

Operationalizing AI Tool Safety Across the Tool Lifecycle

The following section is intended to help operationalize AI-related patient safety risk management principles across the AI tool lifecycle. This begins with the development and maintenance of frameworks, governance tools, and processes by health systems’ governance body responsible for overall oversight of AI tools. These processes will then be operationalized by deployment and governance teams, which consist of a subset of the larger governance body or other designated individuals, with responsibilities in both pre-deployment and post-deployment. The structure of these teams will vary between health systems and potentially by the assessed risk of the clinical AI tool. Deployment teams should include [personnel with significant insight](#) into the clinical workflows affected by the tool (commonly known as the *business owner or champion*), as well as into the tool itself (commonly known as the *technology owner*). If the tool is developed in-house, it would be common for one of the developers to be the technology owner. For vendor-supplied tools, an internal technology owner is still needed. Some health systems reported that they designate someone from the IT staff to serve in that role. While some health systems may include governance staff

on the deployment team, there needs to be a separate governance team tasked with oversight of the deployment that should avoid including overlapping individuals with the deployment team, if possible.

AI Governance

[The AI governance system](#) provides the health system with the foundation for operationalizing clinical AI tool safety by offering structure, tools, policies, and oversight across the organization and works in compliance with applicable laws and regulatory requirements. AI governance processes should align with established patient safety standards and relevant regulatory guidance, including FDA medical device and software oversight frameworks, Office of the National Coordinator for Health Information Technology (ONC) health IT certification and transparency requirements, The HHS Office for Civil Rights (OCR) nondiscrimination and privacy protections, and applicable accreditation and quality standards (e.g., Joint Commission), as well as state and federal policies governing clinical risk management, reporting, and quality improvement.

Clear accountability at the clinical AI tool level is essential and is often lacking in practice. This starts with developing

Bias presents an important patient safety concern for traditional AI tools. [Design and implementation limitations](#) can result in systematic errors, such as consistently higher false-positive or false-negative rates for specific sociodemographic groups, leading to inaccurate or unreliable model outputs. [Empirical studies](#) have demonstrated that such biases are not theoretical, and they have been documented across a range of clinical algorithms, including risk prediction, disease detection, and resource allocation tools. While most research in this space is focused on [sociodemographic bias](#), other types of data differences between subpopulations can also lead to bias (rural-urban, insurance type, etc.). When applied in high-stakes clinical contexts, such as sepsis prediction, these biases can have serious patient safety implications. For example, prior analyses of sepsis and risk-stratification algorithms have shown that models trained on historically skewed data may underestimate risk in non-white patients, leading to fewer alerts and delayed clinical escalation. As such, algorithmic bias should be treated as a patient safety risk rather than solely a technical performance issue.

processes for governance body assessments of proposed clinical AI solutions. Assessments should include structured, risk-based reviews based on the potential severity and likelihood of a patient safety event. The governance body will also need to create an accountability framework that assigns ownership throughout a tool's lifecycle. This includes establishing a shared understanding of who is responsible for implementing and enforcing mitigation strategies, reviewing monitoring results and safety events, and managing version updates, as well as who is accountable for final safety-related decisions. Roles, escalations pathways and decision-making authority should be clearly articulated to avoid ambiguity. These responsibilities may include individuals in either the deployment team, the governance team, or senior members of the governance body.

Governance bodies can create tools to facilitate their work. This could include publishing pre-assessment questionnaires that clearly set out expectations for developers and the potential deployment team, such as utilizing best practices on performance testing or ensuring human-informed implementation design. This should also set out the general elements required in a deployment team's risk management plan including monitoring plans and recommended mitigations such as the level of human oversight.

Another essential governance tool is a centralized, searchable inventory of all clinical AI tools used within the organization. Requiring identification and registration of all AI tools into the inventory will reduce the risk of unknown tools being introduced into clinical practice. The inventory can support the appropriate use of the tool through documentation on the intended users, conditions of use, and purpose of the tool, and it should also document who within the organization is accountable for the tool. The

inventory information could also support the creation of workforce and patient information resources on the AI tools in use in the health system.

Governance bodies should also establish clear procedures and ownership for clinical AI-related safety event management through formal AI governance structures, including responsibility for event triage, technical and clinical assessment, documentation, and external reporting, where appropriate. Some health systems require that any patient safety report potentially involving a clinical AI tool be reviewed by governance staff or an AI subject-matter expert who is embedded within the patient safety workflow. Embedding AI-specific flags within existing enterprise patient safety systems and reporting workflows can facilitate faster review of reported events by a clinical expert member of the AI governance body. This approach could also surface concerns or irregularities that may not meet the threshold of formal safety events but warrant review. Simple workflow-based monitoring helps avoid duplicative processes and supports learning across tools and clinical domains. AI governance bodies may also want to consider creating centralized, low-burden user feedback mechanisms for identifying outputs that seem inaccurate or strange (such a "thumbs up/thumps down" button) to quickly surface concerns and serve as early warning signals.

Finally, escalation pathways must be established so that significant AI-related safety events are elevated to the AI governance body and to senior leadership when warranted. These processes should align with existing quality, compliance, and auditing functions to ensure accountability, reduce redundancy, and support enterprise-wide risk management rather than fragmented, tool-by-tool oversight. This should include pause, refinement, or decommissioning plans in an emergency situation.

Bryan Health System integrates AI risk assessment governance into its existing IT processes through the establishment of a structured framework that leverages interdisciplinary expertise to evaluate AI tools prior to implementation. This model centers on a Data and AI Governance Workgroup composed of leaders representing each sub workgroup. Their charge is to triage proposed AI solutions to determine whether they require full review across all domains, an approach designed to ensure thorough oversight without impeding innovation or deployment timelines. Sub workgroups include Clinical, Information Technology and Security, Legal and Compliance, Quality, and Data Analytics, each of which has developed a 12 question assessment reflecting its specific risk lens. Through this multidisciplinary structure, the organization can efficiently identify concerns related to workflow integration, security vulnerabilities, data integrity, operational ownership, bias, and other critical factors. Following review, the Governance Workgroup synthesizes the findings and returns a consolidated risk profile that enables implementation teams to anticipate and mitigate issues early. Work is also underway to define criteria for ongoing system oversight of AI tools, including post deployment monitoring for value and return on investment. As AI capabilities expand at a pace that challenges traditional governance models, aligning evaluation processes with organizational strategy becomes essential. At Bryan Health, the AI and Data Governance framework ensures that innovation advances responsibly, safeguarding both clinical excellence and patient outcomes.

Human-in-the-loop (HITL) review is a common mitigation strategy that requires trained professionals to evaluate AI outputs. This may not be required for every output, such as agentic solutions where rigorous risk thresholds trigger mandatory human intervention. Implementing HITL in a risk-based fashion proactively mitigates safety events while facilitating human-approved algorithmic improvements. However, HITL is not foolproof, as [automation bias can lead users to over-rely on AI outputs](#) or [fail to critically assess them](#). Effective strategies must move beyond passive acceptance, utilizing safeguards that reinforce clinician engagement and promote appropriate skepticism of AI-generated results.

Pre-Deployment Tasks

These next two sections lay out the pre-deployment and post-deployment tasks for both the governance team and the deployment team, in accordance with the governance processes discussed above. In the pre-deployment phase, known risks identified during development, testing, and local validation should be systematically documented by the deployment team and assessed based on both their potential severity and likelihood of occurrence. The deployment team will submit a tool for review that is supported by structured information regarding technical performance, workflow considerations, and a comprehensive risk management and monitoring plan. Using established AI governance processes and the information provided by the proposed deployment team, the governance team will identify the specific questions needed to evaluate safety, effectiveness, and usability of the specific tool in real-world clinical settings. The governance team will then review the information provided and may work with the deployment team to

gather more information or, in certain cases, propose changes to the risk management plan. Finally, the governance team will report their findings and any recommendations to their AI governance body.

Below is a high-level list of tasks involved in this stage. Except where specified, these tasks should be executed by the deployment team, then reviewed and accepted or modified by the governance team in alignment with processes established by their AI governance body.

Identify Purpose: Confirm or clarify the tool's intended use or range of uses as well as the clinical context of use.

- In cases of broad-purpose AI (some LLMs or general-purpose tools), the tool may have a broad range of uses, and those will be defined by user in each instance of use. The deployment team is responsible for defining the tool's range of use and performing safety and performance assessments for that range.

Health systems should treat clinical and clinical-adjacent large language model tools as a source of meaningful risk, since they introduce distinct safety considerations due to the risk of hallucinations, variability in outputs, and overreliance by the user due to authoritative language regardless of uncertainty.

Create a risk management plan: Conduct and supply an initial safety assessment and document the known limitations of the tool while considering key patient risk dimensions, their impact on patient safety, their severity, the likelihood of their occurrence, and mitigation strategies.

- Propose specific mitigations for known or suspected risks of the tool.
 - These should be designed to integrate into existing clinical workflows as much as possible and may include strategies such as human-in-the-loop review, secondary checks, alternative actions within the workflow, temporary use restrictions, additional end user training, or adjustments to tool configuration.
 - These mitigation measures should be practical and not create undue burden.
- Adjust user training content and competency expectations (such as credentialing for high-risk clinical tools) in line with the assessed risks and mitigation strategies in the risk management plan.

Provide validation: Provide local validation or other relevant testing results for the clinical AI tool and outline the anticipated benefits and risks to the health system's patient population. Ideally, testing is performed with the proposed workflows and within the patient population, but this may not always be possible.

- When applicable, validation testing should include assessments on alignment with clinical judgment to compare with current human decision-making and outcomes.

Propose a monitoring plan: Safety metrics and thresholds should be defined for performance drift, errors, and safety-related events once deployed in real-world settings.

- Establish plans of action for pausing or disabling AI tools if safety risks are identified.

- Create a decommission or substitute workflow plan to be implemented if a clinical AI tool is taken offline, to avoid disruption to clinical care.
- Establish methods to efficiently collect and analyze feedback, and process the data to make post-deployment patient safety sustainable at scale.
 - Potential approaches range from user feedback/focus groups, usage analytics, and adversarial monitoring (AI checking AI) to detect drift or anomalies.
- Monitoring strategies can be risk-based. Lower-risk tools may be able to rely on low-burden user feedback mechanisms and periodic qualitative reviews, while higher-risk tools may warrant more robust oversight.

Establish accountability: Identify the specific individuals who are responsible for the activities laid out by the governance body, such as implementing mitigation strategies, reviewing monitoring results and safety events, and managing version updates, as well as who is accountable for final safety-related decisions.

Vendor Relations

Clear and reliable bidirectional communication pathways between the health system and the AI tool vendors are crucial to effective risk management and patient safety. When selecting AI tools, health systems should prioritize vendors that demonstrate alignment with their organizational safety goals and are willing to meet the transparency expectations related to known safety risks, limitations, and safety issues. A transparency-focused partnership supports effective and scalable safety monitoring, improved workflow integration, earlier risk identification, and collaborative problem-solving throughout the tool's lifecycle. Developers should also systematically document and assess known risks based on both their potential severity and likelihood of occurrence.

Health systems should clearly describe what is needed for their safety evaluations and request that vendors disclose product updates, particularly for tools that are newly modified with AI functionality. The deployers should require that vendors clearly communicate and highlight version changes that may affect patient safety. Vendors should engage with the health system deployment teams beyond baseline disclosures and discuss tool design, testing, validation, and post-deployment monitoring practices. For more complex tools, like LLMs, vendors should provide the necessary information to support effective evaluation and monitoring before deployment and as part of ongoing operational oversight.

There must also be a defined process for reporting and addressing patient safety concerns. These communication channels should engage business owner/champion within the health system and enable direct collaboration with vendors to investigate and mitigate identified safety issues.

Post-Deployment Tasks

Post-deployment risk management utilizes the plans developed in the pre-implementation stage and safety reporting mentioned above to monitor, detect, and respond to safety issues once a tool is in active use within workflows. Responsibilities include tracking post-deployment metrics, monitoring drift, triggering reviews as needed based on the pre-defined thresholds, and investigating safety events. Different health systems may differ on whether they assign these tasks to governance or deployment teams, but responsibility should be clearly defined.

Review and track patient safety events: When a safety event is reported, follow the established governance protocol for event triage, technical and clinical assessment, documentation, and any needed external reporting.

- Record all safety events and close calls to quantify their occurrence and severity.
- Use safety event information to improve the risk management plan and inform future iterations of the tool.

Monitor tool performance: Ensure that tools are performing as expected using the monitoring plan designed and approved in the pre-deployment stage.

- If pre-defined safety thresholds are hit, implement the appropriate action plan and contact relevant personnel based on the accountability framework.

Evaluate mitigation effectiveness: Review real-world implementation of planned mitigations to determine if they are working as expected and are not disrupting care or hindering clinician adoption.

- Deployment teams should maintain transparent communication with end-users throughout the tool's active use.
- Ensure AI tools are being used appropriately by tracking usage. Users should be retrained as needed, or have access removed for serious misuse.
- Update training and educational materials as new risks and mitigation strategies emerge, with clear guidance on severity, reporting expectations, and escalation pathways to AI governance and senior leadership when warranted.

Implement change-management: Changes to the tool or to the risk management plan may necessitate a re-review process.

- Monitor vendor updates carefully and identify potential changes to patient safety risks.
- As personnel change roles or employment, ensure that accountability is re-assigned appropriately.

Duke Health's AI Safety Framework is an example of a health system's approach to strengthen their reporting systems by integrating AI tool safety monitoring into the existing safety reporting systems rather than duplicating systems. It triangulates governance, patient safety, and structured reporting. An important component is the Safety Reporting System (SRS), which integrates an AI flag into the typical patient safety event reporting procedures. This enables faster identification of events involving AI tools without adding burden to frontline reporters, who are typically end-users of AI solutions or patient relations professionals.

Duke Health's AI safety framework recognizes the importance of risk management across the AI lifecycle by systematically identifying, assessing, and mitigating risks from design and development through deployment and monitoring. It is important for end users to receive training on known AI-related risks and reporting avenues before use, reinforced through ongoing updates on known risks post-deployment.

The system process then emphasizes structured governance and accountability after a flagged event through the following actions:

- Each AI solution has designated clinical and model owners, accountable for deployment as established by Duke's AI governance processes.
- When an event is flagged for AI involvement, an AI clinical reviewer and a central monitor from the AI governance team participate in health system's established process for patient safety event review, working alongside clinical and model owners (those accountable for AI tool) to incorporate technical and clinical perspectives into root cause analysis and mitigation.
- The clinical reviewer documents event details, severity assessment, root cause, mitigation steps, and potential for harm (including cases where harm was mitigated or did not occur) in SRS and issue management repository, and clinical owners and model owners implement mitigation strategy.
- The health system-wide AI Oversight policy states how AI-related safety events should be reported.
- High severity events are escalated, as needed, to AI governance leadership in a timely fashion.

A centralized issue management repository captures non-safety event AI-related concerns (such as usability, fairness, or workflow misalignments) providing a holistic view of AI's impact on patient safety.

Duke Health designed quantitative surveillance dashboards to track trends in frequency, severity, and type of AI-related issues. The dashboards enable leadership transparency and intervention on issues and recurring events and are a vehicle for communication to AI governance, patient safety, and quality oversight committees. Metrics include normalized counts of issues, safety events, and high-severity safety events, categorized by severity (e.g., harm, mitigated harm, near miss), AI solution, and deployment location.

Cross-System Learning Networks

Cross-system learning networks strengthen patient safety by fostering collaboration among health systems, regulators, and developers through protected learning environments and regular convenings that promote transparency, oversight, and continuous improvement. Traditionally, learning networks focused on quality improvement goals (including patient safety) or on disease-specific outcomes within and across health care institutions. Effective AI governance approaches are still emerging, so cross-system communication and shared learning are essential methods to quickly identify emerging practices and novel risks.

Existing models of shared clinical AI tool learning include vendor-led user forums, such as Epic's User Groups, and multi-stakeholder initiatives like the [Health AI Partnership](#), [Project ECHO's Artificial Intelligence in Medicine program](#), and the [Trustworthy and Responsible AI Network \(TRAIN\)](#). When appropriately structured, these collaborative forums can accelerate collective learning, reduce duplicative efforts to address risk, and strengthen the safety of AI tools across the health care ecosystem. Systems should maintain a culture that promotes shared learning within and outside of their system, and work together to share their implementation best practices that include their context for intended use, outcome measures, risks, and safety events.

Many health systems already maintain internal AI inventories that include model cards and system-specific analyses (capturing details such as intended use, vendor or developer, data inputs, model version, approval status, ownership, known risks, and monitoring practices). These health systems often use these inventories to support internal information sharing through role-based access for clinicians, safety teams, and operational leaders. Health systems could consider leveraging these inventories to facilitate sharing selected governance and safety data about clinical AI tools. This would likely require structured information-exchange platforms designed to be interoperable and allow organizations to indicate what data can be shared with specific external organizations. But automated curation and sharing of information could reduce staff

effort needed for voluntary sharing, advancing the goal of reducing redundant effort while enhancing collective learning.

Another method to facilitate cross-community learning could be the establishment of outcome registries that collate and analyze AI tool implementation performance and safety. An early example of this is the National Radiology Data Registry hosted by the American College of Radiology (ACR). Through an automated process, ACR aggregates outcome data from certain types of AI tools sent by participating health organizations. They then report back with organizational level insights and can identify broader trends in performance faster than could be identified by individual organizations that have access to less data. While currently tailored to specific domains, this registry demonstrates how shared infrastructure can protect organizational privacy and illustrates the value of centralized data systems in scaling the risk management and reporting essential for clinical AI.

While there are multiple avenues to explore and there will be concerns that need to be addressed, patient safety will benefit from health systems' working together to share effective risk assessment and management practices that improve safety outcomes.

The Role of Policymakers

Despite recent AI governance efforts, there is disparity in the capacity to implement and monitor clinical AI tools across health systems. The rapid evolution and new risks from AI underscore the need for flexible and adaptive governance and monitoring frameworks. Industry-led organizations and multi-stakeholder initiatives (e.g., The Joint Commission, CHAI, the Health AI Partnership) play a role in convening stakeholders, aggregating real-world experience, and developing shared best practices and common frameworks for clinical AI governance and risk management.

Policymakers should strengthen and clarify the regulatory framework for clinical AI tools and particularly where uncertainty exists. Priority areas include addressing regulatory ambiguity between homegrown and commercial tools, clarifying oversight and safety expectations for large language models and increasingly autonomous systems, and translating requirements appropriately across clinical and operational use cases.

Policymakers should also consider the use of safe harbors or other liability protections to encourage information sharing, support learning around AI safety risks and outcomes across institutions, and reduce legal disincentives to transparent reporting of AI-related safety events. Policymakers could also incentivize health systems to implement safety and risk management infrastructure and procedures described above through CMS quality measures. Duke-Margolis and the Duke Health AI Evaluation & Governance program will also publish an issue brief that goes into further detail about how policymakers can support AI risk management. The *upcoming issue brief, scheduled for release in April 2026, [will be available here](#).*

CONCLUSION

The integration of AI into clinical care could be transformative with regards to outcomes, costs, and access, but success is dependent on proactive risk management and responsible implementation. To ensure that these tools enhance rather than compromise patient safety, health systems need AI governance bodies to create systems and processes to identify and mitigate risks pre-deployment, and enforce accountability for monitoring and mitigation responsibilities post-deployment. Low burden tracking mechanisms, such as digital flagging, could be useful to enable visibility into real-world AI performance in the clinical workflow without burdening the staff.

The complexity of AI-driven risk and the speed of innovation necessitate a collaborative approach to safety. With learning networks, health systems could share critical patient safety insights and best practices. The role of policymakers remains central, and a key priority is ensuring that technical support and infrastructure is available to less-resourced systems to implement these tools safely and equitably, in compliance with relevant laws and regulations. In the absence of this, we will see an increasing “AI divide,” with some communities being left behind. Ultimately, the goal is a health care environment where AI tool deployment and implementation workflows are carefully planned, monitored and refined as needed, ensuring that innovation always remains tethered to the fundamental principle to do no harm.

Expert Workshop Participant List

Advancing Responsible AI in Clinical Care: Safety Management

September 30, 2025

The authors would like to thank these individuals for sharing their experiences and expertise with us. The opinions expressed in this paper are solely those of the authors and do not necessarily represent the views or policies of any other person or organization. Affiliations are as of the time of the workshop.

Ada Tsoi, PhD

Senior Data Scientist
UNC Health

Alex Treacher, PhD

Principal Data and Applied Scientist
Parkland Healthcare

Bakul Patel

Senior Director
Digital Health Regulatory Strategy
Google

Cora Han

Chief Health Data Officer
University of California Health

Corey Miller

Vice President, R&D
Epic

Daniel A. Carnegie, MD, MPH, MBA

Chief Data Officer
NC DHHS

Daniel Shieh

Senior Advisor
HHS Office for Civil Rights

Dave Galich

Chief Product Officer
Accenture / Avanade's SAIGE

Eric Poon, MD, MPH, FACMI

Chief Health Information Officer
Duke Health

Eric Rosenthal

Medical Director
Mass General Brigham

Jeffery Smith

Deputy Director, Certification
& Testing Division, HHS Assistant
Secretary for Technology Policy
ASTP

Jennifer Stoll

Chief External Affairs Officer
OCHIN

Jig Patel

Vice President, OS
Tempus AI

Jodi Daniel

Partner
Wilson Sonsini

Maya Sandalow

Associate Director
Bipartisan Policy Center

Kev Coleman

Research Fellow
Paragon Health Institute

Lacy A. Knight, MD

CHIO
Piedmont Healthcare

Laura P Coombs, PhD

Vice President, Data Science
and Informatics
American College of Radiology

Lee A. Fleisher, MD, ML

CEO
Rubrum Advising

Matthew Diamond, MD, PhD

Chief Medical Officer
Digital Health Center of Excellence
Center for Devices and Radiological
Health
US FDA

Merage Ghane

Director of Responsible AI
Coalition for Health AI (CHAI)

Michael Pencina, PhD

Chief Data Scientist
Duke Health

W.B. "Mitch" Mitchell

Chief Growth Officer
Government Sector
Hippocratic AI

**Patricia McGaffigan, MS,
RN, CPPS, CPHFH**

Senior Advisor, Safety;
President, Certification Board
for Professionals in Patient Safety
Institute for Healthcare Improvement

Shannon Robinson, DNP, MSN, RN

System Clinical Innovation Director
Bryan Health

Susan Harkness Regli, PhD

Human Factors Scientist
University of Pennsylvania Health
System

Tara Montgomery

Public Interest Advocate and
Founder & Principal
Civic Health Partners

Walter Wiggins, MD, PhD

Director, Clinical AI
Radiology Partner