

# Patient Safety in the Era of AI: Draft Conditions for the Design of Safe Care

**Laura Adams, MS**, National Academy of Medicine; **Kirsten Austad, MD, MPH**, Boston Medical Center; **Elaine Fontaine**, National Academy of Medicine; **Michael Matheny, MD, MS, MPH**, Vanderbilt University Medical Center; **Sanjiv Mehta, MD, MSCE**, Children's Hospital of Philadelphia; and **Julianna Yeung, MPH, MEng**, Carers Hong Kong

*This commentary was developed under the auspices of the National Academy of Medicine (NAM)'s Steering Group on Patient Safety in the Era of AI.*

July 13, 2026

## Introduction

In the quarter-century since the publication of *To Err Is Human*, both the awareness and the context for patient safety have changed markedly (IOM, 2000). Treatment and diagnostic modalities have increased considerably, care systems have grown substantially more complex, and the medical knowledge base has expanded in both depth and volume. Digital infrastructure and artificial intelligence (AI) are reshaping the capabilities, expectations, and responsibilities of patients and clinicians. AI offers an opportunity to transform health care writ large by enabling care that is inherently safe, highly personalized to each individual, and continuously monitored and adapted to change. Advances in AI now make it feasible to integrate diverse clinical, behavioral, environmental, and experiential information and act on it in real time. AI also offers novel capabilities to address historic barriers to patient safety, including accurate measurement, proactive identification of harm risk, and real-time intervention, and to advance the shared commitments of a national learning health system (LHS) (Madara et al., 2025).

Building on the NAM AI Code of Conduct, the NAM's initiative on Patient Safety in the Era of AI aims to develop approaches for a national strategy

that reaffirms safety as an inviolable precondition of care and applies AI to help identify and dismantle persistent barriers to patient safety (NAM, 2025). Although safety in the application of AI is an important focus, the initiative extends beyond that goal to using AI to make care safer for all patients, wherever and whenever they seek care. The strategy will be deliberately focused on a limited set of high-leverage actions that can unlock cascading opportunities across the broader health ecosystem—that is, the interdependent network of key stakeholders and the environmental context that shapes health and care delivery. This commentary introduces the draft essential system conditions for safe care, describes approaches to the highest-leverage priority actions, and posits a roadmap toward AI-enabled care that is not only of unprecedented safety, but also establishes the foundation for care of unprecedented effectiveness.

## Background

Increasingly sophisticated efforts to improve patient safety have been deployed over the past several decades across clinical, research, and policy domains. Health care improvement science brings rigor to systems management, workflow design, human factors, and variation reduction.

Recognition of health care as a complex adaptive system—with multiple care networks, settings, agents, and unpredictable interactions—has resulted in the application of complexity science to safety. Safety practitioners have therefore expanded the aperture of assessment from a focus primarily on adverse events to learning from successful processes and acknowledging humans as adaptive agents capable of dynamic response processes (Hollnagel et al., 2015).

While patient safety improvement initiatives have yielded meaningful successes in harm reduction, including the introduction of surgical controls and reduction in health care-associated infections, gains have remained uneven, concentrated in pockets of excellence rather than spread system-wide, and difficult to sustain and scale. The overall rate of patient harm has been resistant to change and requires a systematically different approach to pursue wider improvements in patient safety (Office of Inspector General, 2022).

Both improvement science and various policies targeting improving safety, such as pay-for-performance and “never events” depend on access to data and systematic measurement of processes and outcomes (CMS, 2008). With digitization of health care, growing interoperability, and patient access to computable health care information, data availability is expanding logarithmically, challenging traditional mechanisms of data management and inference (Gebler et al., 2025).

Contemporaneous to the digitization of health care, AI has also undergone transformation in the capacity and scale at which it can synthesize information to generate insights and inference. AI has tremendous ability to integrate into patient safety. Health delivery systems show increasing awareness and receptivity to applying AI within routine clinical processes, ranging from cognitive and clinical workflow support to the use of autonomous AI agents for managing increasingly complex processes. And, as health systems adopt AI to improve both clinical care and operations, patients and caregivers are increasingly turning to AI in response to unmet needs, including but not limited to access to care. While barriers to maturity,

governance, and system integration remain, in the coming decade AI is poised to address fundamental issues in supporting patient safety across the health ecosystem.

### **National and Multinational Approaches to Patient Safety**

The authors reviewed national strategies and multinational patient safety frameworks published between 2000 and 2025. Frameworks from selected high-income countries that comprehensively addressed patient-safety components were selected for further consideration. Importantly, selected approaches were required to address patient safety at the macro-system level rather than an institutional level. Components of the multinational frameworks from the Organisation for Economic Co-operation and Development (OECD) and World Health Organization (WHO) and national strategies from the UK, Australia, Canada, the Netherlands, Ireland, and Norway were evaluated for design approach, component inclusion, emphasis and overlap, as well as applicability to the US health system (Busse et al., 2019; WHO, 2021).

The authors agree that OECD and WHO frameworks consistently emphasized leadership and culture, policy and governance, data and transparency, patient and stakeholder engagement, teamwork, and capacity building. However, these components were presented in reinforcing but largely linear models that did not reflect the dynamic interactions among them. Review of national strategies also showed that each country’s approach was closely shaped by its underlying health system, limiting direct applicability to other contexts. Nevertheless, the authors identified several key commonalities. All countries had an independent national patient-safety agency. Regulation and accreditation were the primary policy levers, rather than financial incentives. Patient engagement, while consistently promoted, was deemed difficult to sustain. System-based continuous learning and just culture were emphasized but not implemented consistently or at scale. In addition, all strategies were developed before the widespread adoption of generative AI

and therefore did not address its implications for patient safety.

Importantly, the reviewed frameworks and strategies did not incorporate AI as a mechanism to advance patient safety at scale, nor did they reflect the dynamic interactions among core patient safety components, making them a useful reference point, but not a blueprint for a national strategy for the US.

### **Transformational Opportunities in Patient Safety**

Safety efforts to date have yielded insufficient progress; there is a need and an opportunity for a novel approach: a national strategy for patient safety that addresses the gaps identified in prior national and multinational approaches, while also setting the stage for broader advances in the system-wide precision and effectiveness of health care.

Patient safety provides an urgent and critical starting point, and AI has transformational power for its improvement. AI is currently being used to detect adverse events, predict medication errors, assess fall risks, and prevent pressure injuries (De Micco et al., 2025). However, these examples have largely been applied within individual health systems, and do not adequately highlight the transformational opportunity that AI makes possible. Patient safety could be markedly improved by employing AI to identify and close safety gaps on a national level; for example, a standard national measurement system could detect safety patterns across various health systems in the US, supporting shared learning and improvement.

AI could also offer a way to connect currently fragmented spaces between patients, their caregivers, and the broader health system, improving transparency, communication, and downstream outcomes consistent with patient goals. Patients can now use commercially available AI tools to query medical diagnosis and treatment information translated to their desired health literacy and do so in conjunction with aggregating their medical records and wearables data to receive personalized health advice. This capability is transforming

the way patients can direct their health care, but it also presents new safety risks and requires providers in the traditional health care context to engage, communicate, and partner with patients in novel ways.

Existing patient safety strategies have inadequately accounted for the interactions among system components that drive behaviors and outcomes. Patient safety emerges from the dynamic interplay among multiple actors, incentives, and contexts, creating feedback loops that can amplify improvements or trigger unintended consequences. Systems science tools, such as causal loop diagrams, make these dynamics visible by mapping how components influence one another over time, enabling both anticipation of problems and ongoing monitoring. AI further advances this approach by autonomously discovering complex interactions within observed data, monitoring feedback loops in real time, and facilitating adaptive interventions at scale. This represents the essence of an LHS paradigm, proactively managing complexity to deliver scalable, sustainable safety improvements.

### **Essential System Conditions for AI-Enabled Patient Safety at Scale**

The authors' review of existing system components and patient safety strategies revealed that these elements have largely been treated as independent rather than dynamically interacting. As agentic AI increasingly couples these components, accelerating feedback loops and creating new interdependencies, a framework that accounts for these interactions becomes essential.

In looking ahead to system performance that ensures patient safety, attention is best focused not on enumerating the system components but on the context and nature of the interacting conditions under which safety reliably emerges for patients, regardless of care setting. The anchor focus is on patients and caregivers, whose experiences and active co-production are determinants of system performance, and on using AI to ensure that optimal information and perspectives are in play throughout the care process. *Box 1* presents

**BOX 1 | Proposed Essential System Conditions for AI-Enabled Patient Safety at Scale**

**Aligned Governance:** Because health care is complex in both delivery and governance, decision making, accountability, and authority processes must be aligned across multiple levels and actors in the health ecosystem driven by the needs of patients and caregivers.

**Adaptive Capacity:** To ensure safety, the availability of robust and adaptive resources is essential; this includes people (patients and health care workforce), infrastructure (interoperable AI-ready data), technology (along with sociotechnical integration), and investments needed to meet demand, continuously learn, and respond to emerging opportunities and threats.

**Co-created Solutions:** Close mutual and collaborative engagement, AI-facilitated as indicated, among patients, caregivers, communities, and frontline health care workers is critical in identifying concerns and shaping solutions to improve patient safety.

**Automated Visibility:** Automated, standardized measurement and transparent reporting of established and latent safety performance indicators is imperative for learning, improvement, and accountability across the health ecosystem.

**Cohesive Signaling:** In complex systems, system signals influencing the delivery of care (e.g., financial and economic, reputational risk, market share, liability) must reinforce the preeminent prioritization of patient safety across the health ecosystem.

**SOURCE:** Created by the authors.

these dynamics as Essential System Conditions—the structural and foundational requirements to achieve transformational change in patient safety.

Together, these conditions describe an interconnected LHS organized around patients as both the orienting force and the ultimate measure of success. Each condition functions simultaneously as a domain of focus and a target state. For example, Adaptive Capacity defines what must be present (workforce, technology, data, infrastructure, investment) and the system property those resources must collectively enable (the ability to learn, evolve, and respond). This dual framing reflects the LHS paradigm, in which safety is not a state to be achieved and preserved but a capacity to be continuously generated in the face of evolving complexity and emergent risk.

Achieving these conditions at scale depends on two elements that pervade the framework without constituting discrete conditions. Patients and communities are central to how these conditions interact—their goals orienting the system, their outcomes defining its performance, and

their participation enabling its adaptation (see *Figure 1*). AI is similarly pervasive but occupies a distinct structural role as the evolving infrastructure embedded within each condition that makes real-time measurement, adaptive learning, system-wide coordination, and scaled patient participation operationally feasible. Because a durable framework must define system properties rather than the technologies enabling them, AI is understood as the computational capacity that connects and amplifies each condition rather than standing alongside them.

Framing patient safety through interacting system conditions offers policy makers a shared evaluative lens for assessing strategic actions. Rather than asking whether an intervention addresses a single safety domain, leaders can trace how a proposed action synergistically propagates across conditions, and where it may introduce tension or unintended effects. Crucially, well-defined system dynamics build accountability by clarifying which actors own which levers and making deviations from anticipated system behaviors



**FIGURE 1 |** Centered Around the Patient: Essential System Conditions for an AI-Enabled National Strategy for Patient Safety

**SOURCE:** Created by the authors.

easier to address. This approach also provides the structural foundation for computational policy simulation, enabling AI tools to test how candidate strategies interact within and across conditions before adoption.

### **Approach to Developing an AI-Enabled National Strategy for Patient Safety**

The development of a novel national strategy for patient safety in the AI era begins with an aspirational vision of inherently safe, highly personalized care for every patient, everywhere. Achieving this vision requires a fundamental, AI-enabled shift in focus—from a system designed around institutional imperatives to one focused on and accountable to the person receiving care; from one primarily dependent on human intervention to one supported by an agentic AI safety infrastructure.

AI can be an effective tool to ameliorate persistent barriers to safety, attenuate system constraints, and make highly personalized health care possible, moving well beyond the precision medicine paradigm. It can also serve as a tool for equalizing

knowledge and power, making care more responsive by ensuring that every patient understands their condition(s) and treatment options, can participate in decisions, and is heard. Thus, it will be increasingly important that safety strategies incorporate patient co-designed solutions.

Today's care delivery system falls short of this vision and has proved resistant to change. An effective AI-enabled national strategy for patient safety must incorporate lessons from prior efforts, identifying and generalizing system conditions that enabled successes or impeded progress. It must also anticipate downstream, unintended consequences and acknowledge that the gap between the current and future states may require temporary bridging solutions.

Avoiding diffusion of effort that undermines impact and effectiveness will require focus and disciplined prioritization to enable cascade effects that unlock downstream opportunities. Much will remain undone as this initiative reflects a compelling initial approach in the era of AI-enabled patient safety. Finally, to ensure sustainability, this strategy



must be both grounded in immutable principles and adaptable to rapid change.

### Next Steps

1. Solicit and integrate key stakeholder feedback on the essential system conditions for a broad AI-enabled strategy for patient safety.
2. Identify highest-leverage priority actions with associated evaluation criteria, as well as a possible roadmap for implementation to advance patient safety, while addressing the interactions and unanticipated consequences of change across various system conditions.
3. Describe how patient safety strategy that is highly individualized to each patient's goals, preferences, circumstances, and priorities serves as a foundation for system-wide advances in the effectiveness of patient care more broadly.
4. Solicit and integrate key stakeholder feedback on draft priority actions to advance patient safety.
5. Release an NAM publication presenting the approach to an AI-enabled national strategy for patient safety at scale and including the essential system conditions, priority actions, and roadmap for implementation.

### Key Themes

Over the past quarter-century, genuine progress in patient safety has coincided with the broad digitization of the health enterprise and the application of complexity science to a system that has long demanded it. But progress has been uneven and inadequate in an ever more complex environment, and patients have too often been absent from the design of care, measurement of safety, and governance structures. The emergence of large language models and agentic AI marks a watershed moment. AI tools now exist that allow continuous learning at scale, identification of risk before harm occurs, and inclusion of the patient voice in ways previously unimagined.

Based on a review of national strategies and multinational frameworks for patient safety, guidance from the Steering Group, and ongoing dialogue with national thought leaders, this commentary proposes a draft set of essential system conditions for safe care and outlines an approach to developing a set of priority actions in the context of the dynamic system interplay. Engagement of all key stakeholders in the co-creation of a broad strategy for Patient Safety in the Era of AI is essential to ensure that the pivotal opportunity to use AI to substantially improve patient safety is fully realized throughout the nation.

### References

- Busse, R., N. Klazinga, D. Panteli, and W. Quentin, eds. 2019. *Improving health care quality in Europe: Characteristics, effectiveness, and implementation of different strategies*. Copenhagen (Denmark): European Observatory on Health Systems and Policies.
- CMS (Centers for Medicare & Medicaid Services). 2008. *CMS improves patient safety for Medicare and Medicaid by addressing never events*. Available at: <https://www.cms.gov/newsroom/fact-sheets/cms-improves-patient-safety-medicare-and-medicaid-addressing-never-events> (accessed June 25, 2026).
- De Micco, F., G. Di Palma, D. Ferorelli, A. De Benedictis, L. Tomassini, V. Tambone, M. Cingolani, and R. Scendoni. 2025. Artificial intelligence in health care: Transforming patient safety with intelligent systems—A systematic review. *Frontiers in Medicine* 11:1522554. <https://doi.org/10.3389/fmed.2024.1522554>.
- Gebler, R., I. Reinecke, M. Sedlmayr, and M. Goldammer. 2025. Enhancing clinical data infrastructure for AI research: Comparative evaluation of data management architectures. *Journal of Medical Internet Research* 27:e74976. <https://doi.org/10.2196/74976>.
- Hollnagel, E., R. L. Wears, and J. Braithwaite. 2015. *From Safety-I to Safety-II: A white paper*.

Middelfart (Denmark): Resilient Health Care Net. IOM (Institute of Medicine). 2000. *To err is human: Building a safer health system*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/9728>.

Madara, J. L., S. Miyamoto, S. C. Dowdy, S. M. Greene, J. Tarrant, and P. A. Margolis. 2025. Transforming health care: Shared commitments for a learning health system. *New England Journal of Medicine* 393:192-197. <https://doi.org/10.1056/NEJMSb2507600>.

NAM (National Academy of Medicine). 2025. *An artificial intelligence code of conduct for health and medicine: Essential guidance for aligned action*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/29087>.

Office of Inspector General. 2022. *Adverse events in hospitals: A quarter of Medicare patients experienced harm in October 2018*. OEI-06-18-00400. Washington, DC: US Department of Health and Human Services. Available at: <https://oig.hhs.gov/reports/all/2022/adverse-events-in-hospitals-a-quarter-of-medicare-patients-experienced-harm-in-october-2018> (accessed June 25, 2026).

WHO (World Health Organization). 2021. *Global patient safety action plan 2021-2030: Towards eliminating avoidable harm in health care*. Geneva (Switzerland): World Health Organization.

## DOI

<https://doi.org/10.31478/202607a>

## Suggested Citation

Adams, L., K. Austad, E. Fontaine, M. Matheny, S. Mehta, and J. Yeung. 2026. Patient Safety in the Era of AI: Draft Conditions for the Design of Safe Care. *NAM Perspectives*. Commentary, National

Academy of Medicine, Washington, DC. <https://doi.org/10.31478/202607a>.

## Author Information

**Laura Adams, MS**, is Senior Advisor, National Academy of Medicine. **Kirsten Austad, MD, MPH**, is Assistant Professor of Family Medicine, Boston Medical Center. **Elaine Fontaine** is Special Advisor, National Academy of Medicine. **Michael Matheny, MD, MS, MPH**, is Professor of Biomedical Informatics, Biostatistics, and Medicine, Vanderbilt University Medical Center. **Sanjiv Mehta, MD, MSCE**, is Assistant Professor of Critical Care Medicine, Children's Hospital of Philadelphia. **Julianna Yeung, MPH, MEng**, is Non-Executive Director, Carers Hong Kong.

## Acknowledgments

**Sunita Krishnan**, Senior Program Officer at the National Academy of Medicine, **Audrey Elliott**, Associate Program Officer at the National Academy of Medicine, and **Ravi Parikh**, Associate Professor at Emory University School of Medicine, provided valuable support for this paper.

## Conflict-of-Interest Disclosures

None to disclose.

## Disclaimer

The views expressed in this paper are those of the authors and not necessarily of the authors' organizations, the National Academy of Medicine (NAM), or the National Academies of Sciences, Engineering, and Medicine (the National Academies). The paper is intended to help inform and stimulate discussion. It is not a report of the NAM or the National Academies. Copyright by the National Academy of Sciences. All rights reserved.