Institute *for* Healthcare Improvement

Patient Safety and Artificial Intelligence

Opportunities and Challenges for Care Delivery

Content provided by:

Lucian Leape Institute, an initiative of the Institute for Healthcare Improvement, guiding the global patient safety community.

ihi.org

How to Cite This Document: Lucian Leape Institute. *Patient Safety and Artificial Intelligence: Opportunities and Challenges for Care Delivery.* Boston: Institute for Healthcare Improvement; 2024. (Available at ihi.org)

Institute for Healthcare Improvement

For more than 30 years, the Institute for Healthcare Improvement (IHI) has used improvement science to advance and sustain better outcomes in health and health systems across the world. We bring awareness of safety and quality to millions, accelerate learning and the systematic improvement of care, develop solutions to previously intractable challenges, and mobilize health systems, communities, regions, and nations to reduce harm and deaths. We work in collaboration with the growing IHI community to spark bold, inventive ways to improve the health of individuals and populations. We generate optimism, harvest fresh ideas, and support anyone, anywhere who wants to profoundly change health and health care for the better. Learn more at ihi.org.

Acknowledgments

The Institute for Healthcare Improvement (IHI) Lucian Leape Institute gratefully acknowledges the experts that contributed to this work. Experts with asterisk (*) are members of the IHI Lucian Leape Institute. We are also thankful to IHI team members Kate Feske-Kirby, Britney Pierre, Gabriel Cenizal, and Val Weber, who provided editorial and logistical support in the creation of this report.

Project Co-Chairs	
• Robert Wachter, MD,* Professor and Chair, Department of Medicine, University of California, San Francisco	• Kaveh Shojania, MD, Professor and Vice Chair (Quality and Innovation), Department of Medicine, University of Toronto
Expert Participants	
• Nasim Afsar, MD, MBA, MHM, Chief Health Officer, Oracle • David Bates, MD, MS, Chief of General Internal Medicine, Brigham and Women's Hospital; Professor of Medicine, Harvard Medical School • Leah Binder, MA, MGA, President and CEO, The Leapfrog Group • David Classen, MD, MS, Professor of Medicine, University of Utah School of Medicine; Chief Medical Information Officer, Pascal Metrics • Pamela Cipriano, PhD, RN, NEA-BC, FAAN, President, International Council of Nurses; Professor, University of Virginia School of Nursing • Patricia Folcarelli, PhD, MA, RN, Senior Vice President, Patient Care Services and Chief Nursing Officer, Beth Israel Deaconess Medical Center • Tejal Gandhi, MD, MPH, CPPS, Chief Safety and Transformation Officer, Press Ganey • Eric Horvitz, MD, PhD, Chief Scientific Officer, Microsoft • Gary Kaplan, MD, FACMPE,* Former CEO, Virginia Mason Franciscan Health; Chair, Lucian Leape Institute	• Della Lin, MS, MD, FASA, Consultant; Executive Officer, Anesthesia Patient Safety Foundation; Physician Lead, Hawaii Safer Care • Kedar Mate, MD,* President and CEO, Institute for Healthcare Improvement • Muhammad Mamdani, MPH, MA, PharmD, Vice President, Data Science and Advanced Analytics, Unity Health Toronto • Patricia McGaffigan, MS, CPPS, Vice President, Patient Safety, Institute for Healthcare Improvement • Genevieve Melton-Meaux, MD, PhD, Chief Analytics and Care Innovation Officer, M Health Fairview • Eric Poon, MD, MPH, Chief Health Information Officer, Duke Health • Vardit Ravitsky, PhD, President and CEO, The Hastings Center • Tina Shah, MD, MPH, Chief Clinical Officer, Abridge • Rod Tarrago, MD, Lead of Clinical Informatics, Amazon Web Services, Academic Medicine • Beth Daley Ullem, MBA, Co-Founder, Patients for Patient Safety; Co-Chair, SPS AI Steering Committee; Board of Directors, Institute for Healthcare Improvement

Contents

Executive Summary

The introduction of OpenAI's ChatGPT in late 2022 marked a significant milestone in the evolution of generative artificial intelligence (genAI), garnering attention across all sectors of the economy. In health care, there is hope that genAI may improve care safety and quality, lower costs, and enhance both patient and clinician experiences. However, there are also concerns that genAI may introduce new threats to patient safety.

In January 2024, the Institute for Healthcare Improvement Lucian Leape Institute convened an expert panel to further explore genAI's promise and potential risks for patient safety. The panel reviewed the literature on AI and patient safety and engaged in a robust in-person discussion that focused on three likely use cases for genAI in health care: documentation support, clinical decision support, and patient-facing chatbots.

Panel members are enthusiastic about the potential for genAI tools to reduce clinician burnout and cognitive load, facilitate the provision of evidence-based practices, improve diagnostic accuracy, and potentially reduce cost. Well-designed AI also has the potential to identify care gaps and error-prone situations; doing so in real time may offer opportunities for timely intervention that can prevent some mistakes.

The expert panel also identified various risks, including the depersonalization of care, the possibility of genAI producing inaccurate (or even fabricated) predictions and recommendations, the weakness of human oversight of generally accurate technological outputs, the tendency of health care organizations to turn any AI-generated efficiencies into increased productivity expectations, and the challenges of integrating AI into existing workflows for clinicians and patients.

In addition, the potential for biased AI outputs and clinical deskilling deserves special attention. It is possible that bias can be mitigated by ensuring broad representation in AI datasets, full transparency regarding conflicts of interest that may influence results, and through novel computational methods. Deskilling is particularly vexing since newer clinicians, coming of age after widespread AI implementation, may never acquire the skills of more senior clinicians. This is an area that will require active research and experimentation.

Despite these serious concerns, panelists remain generally enthusiastic about the promise of genAI for improving safety and quality. In pursuing the ongoing development of genAI tools and their integration into clinical care delivery, the expert panel has recommendations:

- Serve and safeguard the patient.
- Learn with, engage, and listen to clinicians.
- Evaluate and ensure AI efficacy and freedom from bias.
- Establish strict AI governance, oversight, and guidance both for individual health delivery systems and the federal government.
- Be intentional with the design, implementation, and ongoing evaluation of AI tools.
- Engage in collaborative learning across health care systems.

Along with these recommendations, further guardrails must be implemented based on the following concerns highlighted by the panel:

- Relying on clinicians alone to double-check the accuracy of AI results and recommendations is an unreliable safety strategy.
- The risk of deskilling is high and will require proactive mitigation strategies.
- AI-driven efficiencies will simply result in more duties assigned to clinicians, with no relief from their current workload and cognitive burden.

In addition to addressing the uses of genAI in health care practice, panelists also discussed the potential for genAI to improve the practice of patient safety itself. GenAI could help streamline cumbersome tasks associated with incident reporting and leverage care episode and patient data to better identify hazardous conditions and suggest fixes.

Based on the IHI Lucian Leape Institute expert panel's review and discussion of AI implications for patient safety, this report summarizes:

- For the three use cases, the potential benefits, risks, and challenges of genAI implementation in clinical care;
- A detailed review of mitigation and monitoring strategies and expert panel recommendations; and
- An appraisal of the implications of genAI for the patient safety field.

For more information about the expert panel, the report, and related content, please visit ihi.org/LLISafetyAI.

Introduction

With the public release of OpenAI's ChatGPT in 2022, generative artificial intelligence (genAI) tools quickly captured widespread interest from every segment of the economy. Given the importance, complexity, and costs of health care, the potential for genAI to transform this field is particularly exciting. While genAI may improve care quality, lower costs, and enhance both patient and clinician experiences, these tools also have flaws that may compromise patient safety, including hallucinations (situations in which genAI fabricates results), bias, and the potential for clinical deskilling (the reduction of skill level necessary to complete a job due to the introduction of new technology).

The need for new approaches to patient safety is clear. A quarter-century after the release of the Institute of Medicine's seminal report, *To Err Is Human*, errors and harm in health care persist.[1] One study found that, in a random sample of Massachusetts hospital admissions in 2018, an adverse event occurred in nearly one in four admissions; approximately 23 percent of these events were judged to be preventable.[2] Moreover, the US Office of Inspector General found that 25 percent of Medicare patients hospitalized in October 2018 experienced harm during their hospital stay, and nearly half (43 percent) of these events were found to be preventable.[3]

The Institute for Healthcare Improvement (IHI) Lucian Leape Institute (LLI) convened an expert panel in January 2024 to assess genAI's impact on patient safety. The panel's goal was to identify areas where genAI could enhance safety, name potential threats, and suggest ways to maximize benefits and minimize harm. Out-of-scope topics included the impact of genAI on other elements of value (e.g., equity, access, patient and provider satisfaction), data security or privacy, revenue cycle and operations, and health care profession education. The goal of this pre-work was to allow the expert group to focus on a series of use cases, which

What Is Generative AI?

Broadly speaking, generative AI (genAI) refers to digital systems capable of generating text, images, code, or other types of content, often in response to a prompt or question entered by a user through a chat interface.

GenAI systems use advanced algorithms like deep learning to analyze and understand patterns in training data sets (which may include everything from individually curated databases to the entire Internet), allowing for the creation of models that generate output with similar characteristics to that found in the training data.

It is important to note that content produced by genAI is influenced by patterns present in the training data, and these patterns can include biases

highlighted areas in which genAI could significantly impact patient safety. The expert panel also discussed the broader implications of genAI for the field of patient safety and the work of safety professionals.

From among the hundreds of potential uses for genAI in health care, three use cases were selected for review by the IHI Lucian Leape Institute expert panel as being broadly representative of anticipated clinical uses of AI in the next several years:

1. Clinical documentation support (e.g., creating clinical notes directly from verbal interactions between clinicians and patients, chart summarization);

2. Clinical decision support (e.g., suggesting diagnoses or treatments to clinicians); and

3. Chatbots that provide patient support (e.g., tools that leverage genAI to answer health and health care questions from patients).

This report presents:

- For the three use cases, potential benefits, risks, and challenges of genAI implementation in clinical care;
- A detailed review of mitigation and monitoring strategies and expert panel recommendations; and
- An appraisal of the implications of genAI for the patient safety field.

While the panel's primary focus was on the American health care system, most findings and recommendations are relevant to other systems in which genAI is being implemented.

The following content is included in the appendices:

- Appendix A: GenAI Use Case Summaries
- Appendix B: GenAI Use Cases
- Additional Resources: IHI Innovation Report and Considerations for Key Groups

Examination of Use Cases: Benefits and Challenges

For each of the three use cases, the IHI Lucian Leape Institute expert panel reviewed potential benefits, risks, and challenges of genAI implementation in clinical care.

GenAI Use Case #1: Documentation Support

Uses of generative AI for documentation support includes developing patient history summaries, supporting patient record reconciliation (including medication reconciliation), ambient recording of patient-clinician conversations, and drafting documentation.

As clinicians grapple with complex cases and heavy patient workloads, the burden of documentation, including writing notes and responding to electronic health record (EHR) inbox messages, can lead to burnout and cognitive strain. The repetitive nature of these activities can also adversely impact clinicians' ability to connect with patients empathically and productively.[4,5] Studies have demonstrated that AI-assisted documentation can reduce overall documentation burden, freeing up clinicians' workloads and lowering cognitive load.[6,7] Moreover, if designed and implemented correctly, genAI can detect opportunities to resolve inaccuracies within the EHR, such as code status, and to standardize common tasks like medication reconciliation.

The expert panel was enthusiastic about these uses of genAI and their impact on clinicians. Panelists were also enthusiastic about potential patient-facing benefits. For example, genAI tools can provide patients and caregivers with accessible documentation that clarifies or simplifies medical jargon, tailored to the individual's reading or health literacy level and in their preferred language. While all of these functions seek to enhance both safety and equity, steps must be taken to also ensure accuracy.

Even as genAI tools promise to free up clinician time by managing certain clerical tasks and flagging discrepancies in the medical record, the tools may create more work for clinicians if reconciliation requires manual review. Moreover, panelists expressed concern that increases in clinicians' time for other work that result from AI-generated efficiencies may be repurposed by health care systems into expectations of commensurate increases in throughput (e.g., seeing more patients, reading more x-rays or pathology slides), leaving clinicians with no relief from their current workload and cognitive burden.

The need for human oversight of genAI clinical output will raise a major challenge to patient safety. Panelists anticipate that many safety systems will involve a first pass from a genAI tool (for example, producing a chart summary or a clinical note, or even a radiology reading). Since these AI-generated outputs will be accurate much of the time but not perfect, the final signoff

(and the hoped-for guarantee of safety) will be performed by a human operator, a system sometimes called "human in the loop."

While this AI/human dyad sounds like a robust safety system, the "human in the loop" double check may be a surprisingly weak layer of protection for several reasons.

- Humans are not particularly good at vigilance, especially when they are passive participants in a process.
- The risk is further complicated by concerns about the potential for clinical deskilling and automation bias — an overreliance on technological tools and support that may lead to unnoticed mistakes and inadequate oversight.
- Clinician review of AI-generated outputs may be compromised if any productivity gains from genAI are translated into higher productivity expectations, leaving clinicians with no dedicated time for thoughtful review.

Whatever the genAI output in question (e.g., documentation, differential diagnoses, treatment recommendations), unless the AI is perfect (which is unlikely) most systems will depend on this technology/human dyad to ensure safety. It will be critical to study this issue and test solutions that increase the probability that such systems are operating as intended and do, in fact, produce safe results. A number of potential solutions were raised by the expert panel and in a recent article on the topic:[8]

- Program AI systems to report the level of confidence in a given output, perhaps by color-coding.
- Employ strategies to ensure that the humans whose responsibility it is to verify AI-generated outputs remain alert and identify gaps in the system. (As one example, TSA agents are periodically shown images of guns or bombs in luggage, both to test their level of attention and to promote vigilance.[9,10])
- Audit the frequency with which clinicians change the AI-generated content (and potentially "nudge" clinicians when they rarely edit the content).
- Design the AI system so that the clinician is asked to weigh in first (akin to a spellchecker reviewing a document after it's already been created).

Creating safe AI/human dyad systems (sometimes referred to as co-pilots) will require substantial research since the solution may lie in new approaches, or in combinations of approaches. For now, health systems need to test the results of their AI/human dyads to ensure that they are producing safe outcomes. In addition, these assessment strategies also need to monitor for algorithm validity since what begins as a safe system may drift over time as patient populations change or new research renders old, previously safe, practices no longer appropriate.

While this may not fall entirely within the realm of patient safety, panelists raised questions about the potential for genAI tools to produce misleading documentation, or documentation that depersonalizes the patient. Many ambient listening tools (e.g., digital scribes) are programmed to omit social conversation (e.g., "How are the grandchildren?" or "How's your

tennis game?") from the note. Yet sometimes it is precisely these snippets that personalize the documentation, thereby creating a connection between the patient and clinician and bolstering person-centered care. Further, current digital scribe models cannot recognize body language or appreciate clinical clues that may come from a patient's tone of voice, which may further compromise the accuracy of AI-generated documentation and recommendations.

Importantly, and also only peripherally related to patient safety, panel members expressed concerns regarding patient autonomy — in particular, whether the general public can provide truly informed consent for the use of AI in their care. While patients must be provided with the appropriate information, in the language and health literacy level they require, to consent to or refuse the use of genAI in their care, it could be difficult for anyone to fully grasp the implications of their choices, including whether their data are being shared or sold.

GenAI Use Case #2: Clinical Decision Support

Uses of generative AI for clinical decision support include providing diagnostic support and recommendations, offering early detection or warning on changes to patient condition, and developing potential treatment plans.

While computerized clinical decision support systems (CDSS) have existed for decades, genAI introduces exciting new capabilities. At the point of care, genAI can process and analyze substantial amounts of patient data to assist clinicians in making better decisions about care. Importantly, AI-based CDSS, like all CDSS, is intended as a "support," or a co-pilot, and not the final decision-maker. As a support, AI-based CDSS can help clinicians assess patient cases by analyzing reported symptoms, patient history, and test results, and then suggest potential diagnoses and next steps.

While diagnostic decision support for physicians has understandably generated substantial interest, many other potential applications could facilitate care by nurses, pharmacists, therapists, and other health professionals. Such genAI-based decision support applications might include the following:

- Identifying relevant guidelines or standards of care based on data from patient assessments by nurses;
- Flagging patients at risk for common in-hospital complications such as falls and pressure ulcers; and
- Identifying patients likely to benefit from medication reviews by clinical pharmacists.

Importantly, genAI CDSS tools can, in theory, generate recommendations for a given patient that evolve over time, based on data in the EHR, as opposed to traditional decision support tools, which usually rely on a few fixed characteristics. Further, CDSS has the potential to address basic safety challenges such as identifying tests that have not been followed up on or incomplete medication reconciliation. Each of these potential functions could improve

diagnostic accuracy, save clinicians time, and reduce costs by guiding clinicians to the most cost-effective workups.

Despite this promise, rigorous evaluations of CDSS – most conducted before the advent of genAI – have yielded disappointingly small improvements in the desired and recommended clinician behaviors.[11] There have been examples of more successful AI-based CDSS, yet to date these successes have involved supervised machine learning such as early detection of clinical deterioration for hospitalized patients or the use of AI-based computer aided detection (CAD) in radiology, which is now routinely used in mammography.[12,13] Enthusiasm about these successes, though, has been tempered by evidence that CDSS can contribute to alert fatigue and clinicians' frustrations with EHRs.[14]

Given the limited evidence supporting AI-based CDSS, many clinicians lack confidence that genAI tools will produce accurate diagnoses. Some of these flaws will be due to weaknesses in the AI itself, while others will be due to inaccuracies in the medical record itself ("garbage in– garbage out").[15] Whatever the source of the untrustworthy output, important questions arise: Will clinicians tolerate incorrect or even inappropriate suggestions if accompanied by valuable ones? Or will they simply reject all AI-generated diagnoses or suggestions after seeing a few that are clearly wrong? Interestingly, this was precisely the problem that doomed many of the early diagnostic AI programs in the 1970s and 1980s.[15,16] Previous failures paired with existing concerns with medical record documentation accuracy — that is, ongoing errors in the medical record due to incorrect, dated, or missing information — could decrease the accuracy of genAI outputs.

The trustworthiness of AI-based CDSS is further compromised by the lack of transparency regarding how AI-based tools work and what their recommendations are based on — the "black box" problem. In addition, there are concerns regarding AI companies' use of proprietary data and designs, the dearth of regulation and oversight of AI in health care,[16] and the possibility of performance bias, particularly affecting patients from marginalized or underserved groups. The latter point arises because genAI tools are often trained on datasets that contain insufficient data from marginalized populations, leading to inaccurate or biased outputs. Finally, experts are concerned about the technology's limitations, such as its inability to conduct physical examinations or use human senses, and the challenges of integrating AI-based CDSS into existing workflows. Ultimately, while gen-AI-based CDSS may produce somewhat more reliable results than prior versions of CDSS, some of the core problems that have bedeviled the use of CDSS in actual clinical work are unlikely to be solved simply by employing better algorithms.

Each of these concerns lead to the more general challenge of how AI-based CDSS will be validated. It is likely that among genAI's greatest impacts will be providing generalized clinical decision support such as suggesting diagnoses to consider or recommending treatments. This presents a more complex challenge when it comes to validation. For example, consider a patient who presents with a new cough and fatigue. Whereas a sepsis predictor can be tested against a well-validated definition of sepsis, it is not as simple to identify all patients who have similar symptoms, such as cough and fatigue, in their records, nor to determine how well the AI-based CDSS performed in terms of correctly suggesting the various possible diagnoses while avoiding incorrect suggestions. This predicament stresses the necessity of assembling more complex cohorts of patients (e.g., all patients with symptoms that are similar enough to cough

and fatigue) and checking the performance of the genAI against all the possible diagnoses these symptoms could represent. Further, calibrating genAI-based CDSS to ensure the presentation of robust lists of potential diagnoses, offer diagnosis probabilities appropriately influenced by both population prevalence and individual patient characteristics, and flag "can't miss" diagnoses will require significant additional research and validation.

GenAI Use Case #3: Chatbots That Provide Patient Support

Uses of chatbots that provide patient support include acting as a data collector to support triage, interacting with patients and responding to their questions and concerns, and supporting care navigation.

Although chatbots are already frequently used in other industries, health care has yet to widely adopt them. Chatbots offer the opportunity to expand access to care and to credible health care information. These automated tools can answer medical questions, support prescription refills, and help patients locate services without the long wait times experienced with call centers, emergency departments, or primary care and specialist appointments. Chatbots implemented and supervised by health care systems could, in theory, provide more accurate and reliable data than that provided by unverified online resources and could further increase knowledge accessibility by offering information at appropriate health literacy levels and in various languages.

In addition, chatbots might be used to manage emails from patients in the EHR inbox, which has contributed significantly to clinician burnout. Health systems could, and likely will, employ genAI solutions to respond to patients' basic concerns or questions, particularly when access to primary care or specialist appointments is limited. Assuming the responses are accurate, chatbots could improve access and equity, while enabling clinicians to practice closer to the top of their licenses.

Panel members underscore the importance of health care systems disclosing to patients the use of an AI-based chatbot, to mitigate any confusion and instill trust between patients and the health system employing the technology. Patients may be uncomfortable with automation, so they should be given clear guidance regarding what the chatbot can and cannot do and be presented with alternative options for patient-clinician engagement or emergencies. There was not uniform agreement on this point, with some panelists noting the frequent use of chatbots in other industries, thereby potentially reducing the need for patient education and consent.

As with all genAI output, it will also be important to ensure ongoing accuracy, which may require that chatbots are finetuned with updates and iterative improvement cycles. For example, Google reported on the use of such a chatbot, the Articulate Medical Intelligence Explorer (AMIE), which was refined by listening to thousands of patient-clinician conversations gathered from a de-

identified dataset of medical conversation transcripts licensed from a research organization. In the view of both specialist-clinician observers and patient-actors blinded to whether they were conversing with a doctor or the chatbot, AMIE conversations were deemed to deliver more accurate diagnoses and to be more empathic than those with actual clinicians.[17]

Despite these reassuring results, the expert panelists noted several important considerations about the use of chatbots.

- Although recent studies have indicated that genAI models can give patients the sense of high levels of empathy, there is still the potential for erosion of trust in the patient-clinician relationship.[18]

- If not appropriately developed and monitored, AI-driven chatbots may engage in misleading or even harmful conversation. OpenAI, for one, has begun to develop guardrails that aim to ensure that their tools refuse to present harmful content such as encouragement of self-harm, demeaning and hateful responses, and inappropriately graphic materials.[19]

- Chatbots must prove their ability to manage complex medical cases, consider comorbidities, and reconcile conflicting information and interests.

- Information flow and triage are also critical concerns for chatbots. If the threshold to escalate care (for example, to recommend that a patient with fever, shortness of breath, and cough go to the emergency room or urgent care) is not properly calibrated, chatbots could pose a danger to patients and their ability to access care in a timely manner.

Recommendations and Mitigation Strategies

After the extensive literature review and discussion, the expert panel remained generally enthusiastic about the potential for genAI to improve patient safety. One relatively easy win is the use of genAI tools for documentation support, freeing up clinicians' time to focus on higher level cognitive tasks and help reduce burnout. Over time, the implementation of AI tools that can provide clinicians and patients with accurate information in a customized, accessible way could also be transformational for patient safety.

Yet, the panel was clear-eyed about the possibility that genAI may not live up to its promise and might even lead to harm. Probably more so than any prior technological advances in health care, the unique power of these tools requires that we temper our enthusiasm with both caution and some skepticism.

It will be critical to heed calls for trust, transparency, accuracy, and human-centered design and implementation. While some challenges with genAI tools are expected to be addressed by regulators and accreditors at the federal or national level, there will be thousands of use cases and tens of thousands of algorithms implemented in all facets of health care, including tools made available directly to patients. This means that whether genAI improves or harms safety will likely be determined more by what occurs within hospitals, health care systems, and patients' homes (via direct-to-consumer apps and other tools) than by the actions of federal, state, or non-governmental regulators and accreditors.

To maximize the benefits and minimize the risks of genAI to patients and patient safety, key groups must consider the expert panel's recommendations and mitigation strategies.

LLI Expert Panel Recommendations

- **Serve and safeguard the patient.** Disclose and explain the use of patient-facing AI-based tools to patients.

- **Learn with, engage, and listen to clinicians.** Equip clinicians with general knowledge on genAI and related ethical issues, as well as specific instruction on how to use available AI-based tools.

- **Evaluate and ensure AI efficacy and freedom from bias.** Establish an evidence base of rigorously tested and validated AI-based tools, including the results of their use in real-life clinical situations.

- **Establish strict AI governance, oversight, and guidance** both for individual health delivery systems and the federal government.

- **Be intentional with the design, implementation, and ongoing evaluation of AI tools.** Follow human-centered design principles, actively engage end users in all phases of design, and validate models and tools with small-scale tests of real-world clinical uses.

- **Engage in collaborative learning across health care systems.**

Serve and safeguard the patient.

- Disclose and explain the use of patient-facing AI-based tools to patients. Health care organizations must partner with patients and advocates to co-design educational and support resources for AI-based tools to ensure that they understand the use and impact of the tools being used in their care.

- Empower patients to prioritize their needs and preferences, including requiring informed consent for the use of patient-facing AI-based tools and use of patient data, and allow patients to refuse AI-based services without compromising their care.

- Engage patients and patient advocates in the development, implementation, governance, and monitoring of genAI tools to ensure that their needs and safety concerns are addressed.

Learn with, engage, and listen to clinicians.

- Equip clinicians with general knowledge on genAI and related ethical issues, as well as specific instruction on how to use available AI-based tools.

- Building awareness and educating clinicians is a shared responsibility of health care systems and their staff. Health systems must offer ongoing training and education opportunities and consistently revise educational content to ensure that knowledge is up to date. Ongoing research may help elucidate how best to deliver this education; it is likely that genAI itself could aid in creating effective new educational tools.

- Health care systems also need to further invest in the clinical workforce by ensuring that time made available by AI-generated efficiencies is partly repurposed into clinician well-being and provision of high-quality, safe care and not entirely allocated to cost saving or production pressures, including expectations of seeing more patients.

- Organizations need to establish feedback mechanisms to harness clinician input on AI tool performance that informs both health care delivery systems and AI tool developers to ensure appropriate use and enable improvement of such tools.

Evaluate and ensure AI efficacy and freedom from bias.

- Establish an evidence base of rigorously tested and validated AI-based tools, including the results of their use in real-life clinical situations. Much of the concern surrounding genAI use in health care can be traced back to the lack of algorithm transparency and paucity of evidence-based research and validation; thus many patients and clinicians remain wary of genAI use in clinical care.

- All key parties, including the government, AI developers, health care systems, and foundation and academic institutions, need to invest time and resources in appropriately acquiring adequate and unbiased datasets and building an impartial evidence base to

test and validate the safety of AI-based tools and whether the tools consistently and equitably improve clinical outcomes.

- While development for evaluation and validation is difficult, time-consuming, and expensive, it is critical to ensure safe use, and to establish trust with clinicians and patients.

Establish strict AI governance, oversight, and guidance both for individual health delivery systems and the federal government.

- Prior to implementation of genAI tools, panel members encourage health systems to develop governance structures and strict policies and procedures on the use of genAI. In its nascent stage, the panel encourages a conservative approach to genAI with clear and firm guardrails, particularly in high-risk areas such as diagnosis and treatment recommendations.

- Experts also favored the early engagement of patient safety and quality personnel as well as patient and family advocates in genAI governance, design, implementation, and monitoring.

- The panel favored granting patients and clinicians the choice to opt in to using AI-based tools, at least in the early days of genAI.[20] This may change over time if certain AI-based tools become the standard of care, although governance and oversight efforts should remain consistent in protecting patient's privacy with regard to data sharing and monetization.

- Governance structures within health care organizations must provide clear processes and requirements for genAI development, testing, and implementation. Governance committees must include patient and family advocates, patient safety professionals, and clinician end users.

- Health systems must establish standards that include acceptable thresholds for accuracy or care escalation, which may differ by AI tool, specialty, and patient population. Regulatory standards need to require ongoing feedback loops and audit processes with mixed methods and real-time assessments.

- Relatedly, oversight and guidance needs to align with broader ongoing efforts to standardize guidance and guardrails, including the National Academy of Medicine's Health Care Artificial Intelligence Code of Conduct[21,22] and the guidelines and guardrails being created by the Coalition for Health AI[23].

Be intentional with the design, implementation, and ongoing evaluation of AI tools.

- Follow human-centered design principles and actively engage end users (e.g., patients, clinicians) in all phases of design and validation of AI models and tools.
- Prior to AI technology implementation, it is important to study the real-world uses of the technology while ensuring that new technologies are not forced to fit existing models or workflows. This may be accomplished with small-scale (and closely monitored) pilot studies in the actual workplace, or by employing simulation and testbed strategies that replicate actual workflows.[24]
- Prior to care delivery, develop and distribute implementation plans as part of clinician education and training.
- Following implementation, AI tools must undergo ongoing evaluation to ensure appropriate application and function and to identify improvement opportunities.

Engage in collaborative learning across health care systems.

- With genAI's swift introduction into health care, to ensure that genAI is safe for clinical care health systems must engage in collaborative learning and prioritize patient and clinician safety and well-being.
- Efforts to foster a culture of shared learning might include hosting or participating in collaboratives or communities focused on AI in health care, partnering with Patient Safety Organizations (PSOs), developing quality assurance laboratories, publishing findings related to an AI-based tool's performance or outcomes, investing in research on genAI or human factors that impact its design or implementation, and developing and offering educational or training materials for health care professionals and the public.
- Examples of ongoing collaborative learning on AI in health care include the National Academy of Medicine's Health Care Artificial Intelligence Code of Conduct, the Coalition for Health AI, and The Health Management Academy's The AI Collaborative[25].

Impact of GenAI on Patient Safety

Beyond the safety considerations for clinical genAI applications in direct patient care, the IHI Lucian Leape Institute expert panel discussed the potential of genAI to impact the field of patient safety itself. For instance, genAI tools might support the following:[26]

- Aggregate and examine incident reports, care episodes, and root cause analysis data
- Leverage data for real-time identification and resolution of safety concerns
- Identify contributing factors that lead to safety issues over time
- Improve audit tracking
- Aid in redesigning workflow to be more efficient and intuitive
- Train safety and quality professionals
- Integrate patient experiences and feedback into safety efforts (e.g., connecting data from patient engagement surveys, patient complaints or incident reports, claims data)

Each of these functions has the potential to help health care professionals shift from reacting to errors and harm to upstream prediction. They might also enhance the ability of safety and quality professionals to respond in real time with data-informed solutions.

It is useful to think about the patient safety field as partly being about proactive ascertainment and partly about action and response. Up to this point, the dominant ascertainment methods — analyzing adverse events and unsafe conditions — have been through incident reports and root cause analysis. Many studies have shown that incident reports produce unsatisfying results, missing many problems and often capturing issues that are peripheral to the relevant patient safety problem.[27,28,29,30]

In terms of action plans, root cause analyses — detailed interdisciplinary reviews of adverse events — have been the primary means by which health systems have come to understand and address safety issues. The creative use of genAI could identify patterns of care and suggest solutions based on past events at a given health care organization (or others if their data is accessible). The AI tools could also then help monitor if these solutions are working over time. Moreover, by creating a mechanism to continuously review every patient's EHR for signals of adverse events and unsafe conditions, genAI has the potential to reshape the way that delivery systems learn about safety. Indeed, such real-time monitoring of AI systems have already been built into EHRs and have been found to detect more than 10 times as many incidents of harm, in real time, and predict harm up to 72 hours before it occurs.[31]

Finally, paralleling the problems of documentation burden for clinicians, the work of patient safety and quality professionals is currently characterized by an enormous amount of time (more than 50 percent in some estimates) spent in chart review, often for the purpose of completing documentation requirements for accreditors, insurers, registries, and others.[32,33] GenAI-based record review has the potential to free up much of this low-yield time. It is crucial that health care organizations repurpose some of this freed-up staff time into more meaningful safety and quality activities, rather thandownsizing the department in direct proportion to the liberated time.

How AI Can Support and Improve Patient Safety
- Aggregate and examine incident reports, care episodes, and root cause analysis data[26]   - Leverage data for real-time identification and resolution of safety concerns[26]   - Identify contributing factors that lead to safety issues over time[26]   - Improve audit tracking, aid in redesigning workflow to be more efficient and intuitive[26]   - Train safety and quality professionals[26]   - Integrate patient experiences and feedback into safety data and other efforts[26]   - Assisting with identifying diagnosis and facilitating early detection of common inpatient safety concerns, including hospital-associated infections, adverse drug events, venous thromboeblism, and surgical complications[34,35,36,37,38]   - Reduce workload of patient safety and quality improvement staff, allowing greater resource allocation to implementing patient safety efforts[39,40]

Conclusion

Generative artificial intelligence stands poised to revolutionize health care, including having positive effects on patient safety and quality of care. However, its adoption must be navigated with caution, acknowledging that new technologies in health care always introduce unanticipated consequences and that certain special characteristics of generative AI may create unique hazards. Key groups must work together with intention and discipline to implement genAI in ways that enhance patient safety.

IHI Lucian Leape Institute expert panelists are generally optimistic about the potential for genAI to improve many aspects of health care, including patient safety, and hope that this report will encourage a variety of audiences to take actions that will ensure the safe and responsible adoption and integration of genAI into the health care system.

Appendix A: GenAI Use Case Summaries

The tables below summarize the IHI Lucian Leape Institute expert panel's discussion of three use cases for implementing generative AI in clinical care.

GenAI Use Case #1: Documentation Support	
Function	• Develop patient history summaries • Support patient record reconciliation, including medication reconciliation • Create documentation of patient-clinician conversations via ambient listening tools (e.g., digital scribes) • Draft responses to patient messages, including EHR inbox messages
Benefits	• Reduce clinical documentation burden, thereby reducing clinician burnout and cognitive load • Identify and potentially resolve inaccuracies in the EHR • Standardize common tasks like medication reconciliation • Improve accessibility of documentation for patients and caregivers (e.g., define medical jargon; tailor for health literacy level and preferred language) • Strengthen trust and communication between patients and providers (i.e., clinicians can focus on the patient because they are less focused on typing into the computer)
Risks and Challenges	• Decenter the patient and their rights by failing to provide understandable information about risks and benefits of genAI tools used in their care, obtain informed consent, or provide quality alternatives for patients who choose not to consent • Potential to increase clinician's workload if manual review of flagged inaccuracies or other forms of AI double-checks are unduly burdensome • AI-generated efficiencies are used for cost savings instead of providing relief for clinicians (e.g., asking clinicians to see more patients vs. providing cognitive breaks or increasing encounter time with patients) • Concerns with the accuracy and transparency of AI-supported documentation and need for human oversight (and the weakness of human oversight of AI-generated outputs as a patient safety mechanism) • Depersonalize documentation due to the loss of body language and "small talk" that may not be entered into the medical record note when collected by ambient listening tools
Mitigation Strategies and Expert Panel Recommendations	• Ensure that patients are informed of genAI documentation support and given the opportunity to provide informed consent to or refuse use • Provide alternative, high-quality options for patients who refuse the use of genAI tools • Create a philosophy regarding how to repurpose AI-generated efficiencies, ensuring that some of the newly available clinician time is used to prevent cognitive overload, promote empathy, and allow for meaningful oversight

	• Develop strict human oversight of genAI tools in health systems, including governance structures, standard use guidelines, appropriate clinician double-checks, and feedback loops and audits
	• Monitor patient engagement with genAI tools using audit systems (e.g., opting in or out), use of overrides by clinicians, and whether clinicians are following or deviating from AI-generated recommendations

GenAI Use Case #2: Clinical Decision Support	
Function	• Provide diagnostic support and recommendations • Offer early detection or warning of changes to patient condition • Develop and suggest potential treatment plans
Benefits	• Function as an aide for clinicians ("co-pilot") by analyzing information and suggestion potential diagnoses and treatment plans, which could support more evidence-based decisions at the point of care • Could improve diagnostic accuracy, save clinicians time, and potentially reduce costs
Risks and Challenges	• Paucity of evidence validating AI-based clinical decision support systems (CDSS) tools for clinical use (note that past CDSS failures impacted clinicians' confidence in usefulness of CDSS tools, and thus new tools will require intentional trust-building) • Concerns over clinical overreliance, compliance, and automation bias • Lack of transparency, explainability, and validation (e.g., the "black box") • Prioritization of proprietary data and designs over patient safety and quality care • Limitations of existing training data sets, which could codify human bias (e.g., possibility that racism, sexism, and other biases will be baked into AI-generated outputs drawn from past practices) • Limitations of technology (e.g., inability to conduct physical exams or challenges of integrating genAI into existing workflow)
Mitigation Strategies and Expert Panel Recommendations	• When appropriate and feasible, ensure that patients are aware of AI-based CDSS use and that health systems support patient autonomy and decision making (e.g., consent or refusal to use of digital scribes) • Invest in building an evidence base to test and validate AI-based tools' performance and outcomes and to ensure the safe, equitable application of such tools in clinical care delivery • Devote resources to educate and train clinicians and health care staff on AI basics, associated ethics, and simulation training • Prior to implementation, verify that newly introduced genAI technology will function within the current workflow and identify challenges that could prevent appropriate and timely adoption • Promote the philosophy that clinical use of AI-supported CDSS is an aide to, not a substitute for, clinician decision-making • Work with developers to create transparency and confidence in CDSS outputs, for example, color-coding outputs to identify the level of confidence

GenAI Use Case #3: Chatbots That Provide Patient Support	
Function	• Function as a data collector to support patient triage • Interact with patients and respond to basic questions and concerns • Support care navigation (e.g., connect users to resources such as care center locations, operating times, and scheduling or modifying appointments)
Benefits	• Expand access to care • Democratize access to credible health care information • Provide more accurate and reliable data
Risks and Challenges	• Ethical concerns regarding technology that mimics humans and need for proper disclosure that genAI is being used to provide advice • Chatbot accuracy and the need for ongoing auditing, maintenance, and updates • Loss of human connection (e.g., chatbots cannot convey tone or emotion) and potential to erode trust between patients and clinicians • Information flow and triage (e.g., How is clinician oversight exercised? How to ensure that patients get the care they need if the genAI triages them into a high-risk category?)
Mitigation Strategies and Expert Panel Recommendations	• Embed the appropriate disclosures of the chatbot's function to patients and other users prior to use • Design chatbots to recognize their limitations, including prompts they cannot respond to (e.g., illegal activity, promotion of violence or harm) • Develop escalation pathways for chatbot users (e.g., identify when to contact a human clinician, or when a patient should go to an ED); chatbots must have a low threshold to escalate care when, for example, the patient provides information that might signal a clinical or mental health emergency • Calibrate and validate chatbot models for the local context to ensure safe and quality care for the population served • Human oversight ("human in the loop") of AI-driven chatbots is critical, including routine auditing of chatbot performance and clinical review of conversations to ensure that patients' needs are being met • Establish data-driven guidance on the use of chatbots, including clinical guardrails (e.g., chatbots recognize when they are practicing beyond their intended scope and connect patients to clinicians for further evaluation); policies and disclosures of data use, privacy, and protection; and escalation plans (e.g., when patient needs emergent care, how this is communicated and how quickly)

Appendix B: GenAI Use Cases

The IHI Lucian Leape Institute expert panel reviewed three use cases, selected for their broad representation of anticipated clinical uses of AI in the next several years: documentation support, clinical decision support, and patient-facing chatbots. Each use case promoted a discussion aimed at identifying specific risks as well as possible strategies to monitor (and hopefully mitigate) these risks. (Note: The use cases below have been edited for publication.)

GenAI Use Case #1: GenAI-Driven Documentation Support in Ambulatory Care

Instructions

The review of each case study is meant to foster discussion about the (safety) risks and benefits of generative AI (genAI) tools that will be implemented in health systems or used by patients in the near future.

This case study focuses on genAI tools that influence an individual ambulatory patient encounter — by summarizing a patient's EHR record, acting as a virtual scribe to create a note, and then sending that note (in various forms, depending on the audience) to key parties, including the patient.

Background

The year is 2025, and you're seeing patients in your busy primary care clinic. Your first patient is a 74-year-old woman with a history of heart failure, hypertension, type 2 diabetes, and COPD. She was recently hospitalized for a COPD exacerbation and pneumonia.

Your computer recognizes your face and, based on your schedule, opens up to the first patient's record. You say, "Please summarize the past record." A document, about half a page long, pops up on the screen highlighting the key elements of the patient's past history, including prior surgeries and hospitalizations. You then say, "Reconcile medications," and the AI pulls up an up-to-date medication list, reconciling prior prescriptions with the actual record of medications delivered from the patient's online pharmacy.

The patient enters your office and sits down. You say, "Standard clinical encounter template for returning patient, please." A light flashes on the voice-activated smart device sitting on the table and begins to record the conversation (starting with the patient giving her consent to do so). Following the 10 minutes of discussion and assessment, which includes questions and answers about any existing symptoms/conditions and new concerns, along with a brief physical examination, the clinician clicks on the button to produce a draft clinical note. The computer instantly produces a correctly formatted follow-up note, placing the patient's answers in the appropriate sections and omitting the initial small talk you had with the patient about her grandchildren and the weather.

After the visit, you quickly review and then approve the note. The visit focused on management of the patient's known problems, so you don't click on the "Suggest diagnoses" button. Instead you click the "Suggest actions" button. The computer produces a list of suggested tests ("Repeat chest x-ray in 3 weeks") and medications ("Refill medications") and you click to agree to these suggestions. A copy of the note is automatically forwarded to the specialists who follow the patient. In addition, per the patient's previously expressed preferences, a version of your note written at a sixth-grade reading level is forwarded to the patient's electronic portal.

Next Steps or Consideration for Discussion

Please discuss the potential safety benefits and risks of this new genAI tool. Once you've completed that, feel free to touch on other implications — such as the impact of such tools on the workforce and on clinician education. Finally, reserve some time to discuss strategies to mitigate any potential harms that you've identified.

GenAI Use Case #2: GenAI-Driven Decision Support

Instructions

The review of each use case is meant to foster discussion about the (safety) risks and benefits of generative AI (genAI) tools that will be implemented in health systems or used by patients in the near future.

This use case focuses on a genAI-based decision support tools that suggests diagnostic and management possibilities for clinicians. It takes the perspective of the health system's patient safety program, thereby broadening the discussion to go beyond the individual clinician-patient encounter to consider the implications of AI for monitoring clinical programs for foreseen and unanticipated risks.

This decision support use case includes a second scenario, in which a clinician asks the AI tool about a complex medical decision much like the clinician might ask a specialist colleague. In each scenario, we use the cases to promote a discussion aimed at identifying specific risks as well as possible strategies to monitor (and hopefully mitigate) these risks.

Background: Scenario 1

It's January 2026 and you work in a health care system that includes hospitals and ambulatory clinics delivering primary and speciality care. Six months earlier, your health system implemented a comprehensive genAI-driven tool for delivering point-of-care decision support to clinicians. The genAI tool works off the note a given clinician is entering, as well as past notes and test results for the patient. Once a clinician finishes the note, the AI tool offers "on demand" suggestions that include diagnoses to consider, tests to order, and potential treatments. Specifically, when reviewing the encounter note (whether in a clinic or hospital setting), the clinician has the option to click on a button to "Suggest diagnoses." There is also a button to "Suggest actions," which includes tests to order or treatments to pursue, as well as initiation of new treatments or modification of existing ones, like dose adjustments for currently prescribed medications.

For example, in a patient with heart failure who also has multiple sclerosis and chronic kidney disease, the AI tool suggests heart failure therapies informed by evidence-based guidelines, while also considering any particular risks or side effects owing to the patient's co-morbid conditions.

The AI tool vendor has configured the program so that it does not make suggestions unless the model deems these suggestions as supported by the diagnoses and patterns of care for at least 80 percent of similar patients. (The vendor's default threshold is 70 percent, but a working group of clinicians and executives from your health system asked the vendor to set the threshold to 80 percent.)

You are a member of the group tasked with producing a report on the performance of this genAI decision support tool 12 months post-implementation. This is the first meeting of the working group. You can use this hypothetical scenario as the context for your group discussion. Further details are as follows.

To help with the planned evaluation at one year, some specific strategies to identify and monitor for safety problems and benefits were put in place at the outset.

Brief surveys were sent via email to a 5 percent random sample of clinicians who used the AI tool at least twice in a given week and had not been surveyed within the last 12 weeks. These surveys were designed to minimize time burdens. They include only a few focused questions about any perceived benefits to patient care, errors or other problems noticed by the user, and a free text option to provide additional comments or feedback. In user testing, the time to reply was less than one minute and this seems roughly accurate based on the data at six months. Response rates were initially surprisingly high (around 60 percent), but declined to 30 percent after the first eight weeks.

The working group also suggested tracking a few measures aimed at capturing potentially unnecessary lab and radiologic testing, as well as tracking the percentage of patients receiving broad spectrum antibiotics and opiate analgesics.

Next Steps or Consideration for Discussion

At this first meeting of the evaluation working group, there is still time to modify these strategies so feel free to discuss if they are/were reasonable and/or how they might be improved, as well as other methods you might put in place to monitor for the benefits and risks (but especially risks) of this decision support tool. While our primary focus is patient safety, it's fine to touch on other potential problems.

Background: Scenario 2*

[*Note on AI response: The transcript below reflects the results obtained when querying ChatGPT-4 in October 2023. We ran it again a couple months later, and the result was more circumspect – it laid out the possible treatment options but did not make any statements about any particular treatment or combination of treatments likely representing the preferred approach. This demonstrates that these systems are improving rapidly, an important fact to take into account as we consider safety hazards and opportunities.]

A primary care physician will be seeing a 65-year-old man, a patient he has followed for many years, for the first visit after a recent biopsy showing that he has a new diagnosis of prostate cancer. The physician plans to refer the patient to specialists, likely a urologist and medical oncologist. But the physician also knows that the patient will understandably have many questions about the treatment options, and he does not want to make the patient wait days or weeks to hear what the specialists have to say.

To offer the patient a tentative outline of the specialists' likely recommendations, the physician types the following question into ChatGPT-4:

Clinician: What is the preferred strategy for a 65-year-old man with a prior TURP, moderate Parkinson's disease, who now has prostate cancer with a Gleason score of 8 and no evidence of metastatic disease?

AI Response: Generally speaking, high-risk prostate cancer (like those with a Gleason score of 8) often requires more aggressive treatment than low-risk cancers. Here are some considerations:

- **Radical prostatectomy:** Surgery can be an effective treatment for high-risk prostate cancer but given the patient's history of TURP and the potential for surgical complications and side effects, this may not be the first choice. Additionally, Parkinson's disease can complicate postoperative recovery and the rehabilitation process.

- **Radiation therapy with ADT:** A combination of external beam radiation therapy and androgen deprivation therapy (ADT) is often recommended for high-risk prostate cancer. This approach can provide good cancer control. The prior TURP does not preclude radiation, but specific radiation-related side effects would need to be weighed.

- **Androgen deprivation therapy alone:** While directed at controlling prostate cancer growth, using androgen deprivation therapy alone without radiation might be less effective at achieving long-term cancer control for high-risk prostate cancers. However, it could be considered if other treatments pose too many risks.

Given the context provided, **a combination of radiation therapy and hormone therapy might be the preferred approach** for many clinicians, considering its effectiveness for high-risk prostate cancer and the challenges posed by surgery in this patient. Radiation can provide local control of the disease, and the addition of hormone therapy can enhance this effect.

Clinician: Don't you think the hormones will exacerbate his weakness from his Parkinson's?

AI Response: You raise a valid point. Androgen deprivation therapy has a number of side effects; among them are fatigue, decreased muscle mass, and decreased bone density. These side effects can certainly impact individuals' overall strength and functional status, which compound the physical challenges faced by someone with Parkinson's disease.

Next Steps or Consideration for Discussion

We appreciate this second scenario is complex. We chose it because the first scenario involved an AI-driven tool for delivering more routine decision support — suggesting diagnostic

possibilities for the clinician to consider and next steps to take. This second scenario highlights the potential benefits and risks when clinicians call on genAI decision support tools in much the same way they might run a case by a specialist colleague (a so-called "hallway consult").

GenAI Use Case #3: Chatbots Answering Patients' Questions and Suggesting Diagnoses and Treatments

Instructions

The review of each use case is meant to foster discussion about the (safety) risks and benefits of generative AI (genAI) tools that will be implemented in health systems or used by patients in the near future.

This use case focuses on chatbots for patients. It includes two scenarios: 1) a patient needs advice from her primary care physician because she has developed diarrhea after starting antibiotics, and 2) a mother wants to know if she should take her 4-year-old child to the ED after swallowing a coin.

Background: Scenario 1

A patient has a regular primary care physician (PCP), who prescribed her an antibiotic during a recent visit for a skin infection on her leg. The patient now has diarrhea. She logs into her patient portal and tries to make an appointment with her PCP, but the next available appointment is in 10 days. She notices a button on the portal that says, "Send an email to your medical team." She clicks the button and types an email message.

Patient: I've been taking the antibiotic you prescribed for my skin infection for the past 5 days and my bowel movements are really loose. Should I be worried?

[She expects that she'll receive the response she has gotten for prior queries on the patient portal: "A member of your health care team will be in touch with you within 48 hours." She is surprised to immediately receive the following response:]

Chatbot Response: Thank you for your note. I'm sorry that you are having loose stools since starting on Septra (trimethoprim-sulfamethoxazole) 5 days ago. *[Note that the patient didn't name the antibiotic in her email.]* I'd like to ask you a few questions to help me respond in the most helpful way. Can you tell me how many bowel movements a day you've been having?

Patient: I've had 5 loose bowel movements in the past day.

Chatbot Response: Thank you. Do you have a fever?

Patient: No, I don't think so.

Chatbot Response: Are you having abdominal pain?

Patient: Not really, sometimes I feel a little bloated before I go to the bathroom, but not really pain.

Chatbot Response: Great, thanks very much. One last question about your diarrhea: have you noticed any blood?

Patient: No, it's brown, and sometimes a little yellow. But nothing red.

Chatbot Response: And how is the skin infection doing?

Patient: Much better, it's definitely less red than it was, and the swelling is down.

Chatbot Response: Great. Loose stools are a common side effect of antibiotics in general and Septra in particular. It looks like you should be done with your antibiotics in 2 days. As long as there is no fever and abdominal pain, I think it's fine for you to complete your course of antibiotics. Just stay hydrated since the biggest risk is that you'd become dehydrated from the diarrhea. Otherwise, please contact us again if you develop a fever, abdominal pain, or have increased diarrhea or notice any blood in your stool.

Patient: Great, thank you very much.

Background: Scenario 2

A young child has just swallowed a dime. It's 10:00 PM and the mother is loath to take her child to the ED, although that's what her pediatrician's answering service says to do. She opens ChatBot MD and types her question.

Mother: My 4-year-old son has swallowed a dime. What should I do?

Chatbot Response: I'm not a doctor, but I can offer some general guidance on what to do if your child swallows a foreign object like a dime. It's essential to consult with a health care professional for specific advice in your situation. However, here are some general steps you can consider:

- **Stay Calm:** Try to remain calm and reassure your child. Panicking can make the situation more stressful for both you and your child.
- **Assess the Situation:** Determine if your child is experiencing any immediate distress or difficulty breathing. If your child is choking, you should perform first aid for choking immediately.
- **Seek Medical Advice:** Contact your child's pediatrician or call your local poison control center for guidance. They can provide you with specific advice based on your child's age and the circumstances.
- **Monitor Symptoms:** Keep a close eye on your child for any signs of discomfort, pain, difficulty breathing, or other unusual symptoms. If you notice any concerning symptoms, seek immediate medical attention.
- **X-rays and Imaging:** Your health care provider may recommend an x-ray or other imaging studies to locate the dime and assess its position in your child's digestive system.
- **Follow Medical Advice:** Follow the recommendations provided by your health care provider. In some cases, the object may pass naturally through your child's digestive system, but in other cases, medical intervention may be necessary.

Remember that small objects like coins can sometimes pass through the digestive tract without causing harm, but it's crucial to consult with a health care professional to ensure the safety and

well-being of your child. Always prioritize your child's health and safety and seek immediate medical attention if you have any concerns.

Mother: Thank you for your advice. I contacted my child's pediatrics practice and the nurse practitioner on call advised I go to the nearest emergency department to have my son checked out (since their urgent care clinic has already closed for the night). I have a colicky baby at home, too, and have no childcare. Taking both the baby and my 4-year-old child to the ED tonight will be hard. My son seems fine. Is there a safe way I can just monitor him at home and only go to the ED if concerning signs develop?

Chatbot Response: I'm not a medical professional, but I can provide some general information that may help you make an informed decision. If the health care provider you spoke to recommended taking your child to an urgent care clinic or the nearest Emergency Department (ED), it's important to consider their advice seriously. They have the expertise to assess the situation and make recommendations based on your child's specific circumstances.

However, if your child appears to be comfortable and is not experiencing any immediate distress or symptoms such as choking, severe pain, or difficulty breathing, you can consider closely monitoring your child at home for any concerning signs. Here are some steps you can take if you decide to monitor your child at home:

- **Assess the Situation:** Determine if your child is experiencing any immediate distress or difficulty breathing. If your child is choking, you should perform first aid for choking immediately.
- **Seek Medical Advice:** Contact your child's pediatrician or call your local poison control center for guidance. They can provide you with specific advice based on your child's age and the circumstances.
- **Monitor Symptoms:** Keep a close eye on your child for any signs of discomfort, pain, difficulty breathing, or other unusual symptoms. If you notice any concerning symptoms, seek immediate medical attention.

[The chatbot offers some additional comments very similar to the ones in the initial response, along with recommendations to stay in touch with the person on-call for the practice, and, if there is any doubt, go to the ED. The chatbot then concludes with the following:]

It's essential to trust your instincts as a parent and prioritize your child's health. While it may be challenging to manage both your colicky baby and your 4-year-old, it's crucial to follow the medical advice provided by your health care provider to ensure the best possible care for your child who swallowed the dime. If you have concerns or questions about managing both children, you can discuss them with the health care professionals you consult with.

Next Steps or Consideration for Discussion

For this second scenario, we listed some specific risks to consider. We are not suggesting that these are the only risks or even the most important ones, but we thought it might help to have some risks to kick off the discussion for this use case.

- When fielding these sorts of calls at night, clinicians make a rough judgment about the patient's (or, in this case, the parent's) "reliability" – their level of understanding with respect to accurately conveying all the relevant details of the current situation and their likely ability to safely monitor for signs that necessitate going to a clinic or ED. This is not possible for the chatbot — at least not currently.

- What a parent may think is a coin, specifically in the case of a dime, may in fact be a button battery (also called a watch battery). If a button battery becomes impacted in the esophagus, it's life threatening due to the risk of chemical burns.

- Signs of respiratory distress can be subtle. We ran various versions of this query through the chatbot and all chatbot responses made mention of monitoring for immediate distress or difficulty breathing. But signs of respiratory distress in children can be subtle (e.g., refusing to lie flat, as opposed to more overt signs of difficulty breathing) and sick children can look fine but then quickly deteriorate.

- FDA approval will likely be required for any chatbot programs that do not involve clinicians approving (or modifying) responses before transmitting them to patients. It's possible that some vendors will market non-FDA-approved AI chatbots on the grounds that they advise their clients to always have a human in the loop — for example, an on-call clinician who receives a text alerting her to a draft message requiring urgent attention; she then reviews the content provided by the chatbot before signing off on the response sent to the patient. The safety risk to consider is the degree to which the clinician fully engages with the text to correct any oversights, including oversights they would not themselves make if they were actually speaking to the patient directly or seeing them in person (e.g., noting if the child is reluctant to lie flat or asking if it might have been a button battery and not a dime that the child swallowed). Keep in mind the risk of automation complacency in any system that depends on a clinician (or any human) overseeing and correcting the output of any AI system that is correct most of the time.

References

[1] Kohn LT, Corrigan JM, Donaldson MS (eds). (Committee on Quality of Health Care in America, Institute of Medicine). *To Err Is Human: Building a Safer Health System*. Washington, DC: National Academies Press; 2000:9728.

[2] Bates DW, Levine DM, Salmasian H, et al. The safety of inpatient health care. *N Engl J Med*. 2023;388(2):142-153.

[3] Grimm CA. *Adverse Events in Hospitals: A Quarter of Medicare Patients Experienced Harm in October 2018*. US Department of Health and Human Services, Office of Inspector General; 2022:1-117. https://oig.hhs.gov/oei/reports/OEl-06-18-00400.asp

[4] Gaffney A, Woolhandler S, Cai C, et al. Medical documentation burden among US office-based physicians in 2019: A national study. *JAMA Intern Med*. 2022;182(5):564.

[5] Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Ann Intern Med*. 2016;165(11):753.

[6] Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catal Innov Care Deliv*. 2024 Feb;5(3).

[7] Lohr S. AI may someday work medical miracles. For now, it helps do paperwork. *New York Times*. June 26, 2023.

[8] Adler-Milstein J, Redelmeier DA, Wachter RM. The limits of clinician vigilance as an AI safety bulwark. *JAMA*. 2024 Apr 9;331(14):1173-1174.

[9] Zenko M. How Not to Red Team. Council on Foreign Relations Blog. September 8, 2016. https://www.cfr.org/blog/how-not-red-team

[10] Fish JM. Understanding Airport Screening's 95% Failure Rate. Psychology Today Blog. July 5, 2016. https://www.psychologytoday.com/us/blog/looking-in-the-cultural-mirror/201607/understanding-airport-screenings-95-failure-rate

[11] Kwan JL, Lo L, Ferguson J, et al. Computerised clinical decision support systems and absolute improvements in care: Meta-analysis of controlled clinical trials. *BMJ*. 2020 Sep 17:370:m3216.

[12] Kwan JL, Lo L, Ferguson J, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. *BMJ*. 2020 Sep 17:370:m3216.

[13] Fazal MI, Patel ME, Tye J, Gupta Y. The past, present and future role of artificial intelligence in imaging. *European Journal of Radiology*. 2018;105:246-250.

[14] Downing NL, Bates DW, Longhurst CA. Physician burnout in the electronic health record era: Are we ignoring the real cause? *Ann Intern Med*. 2018;169(1):50.

[15] Wachter RM. *The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age*. McGraw-Hill Education; 2017.

[16] Ayers JW, Desai N, Smith DM. Regulate artificial intelligence in health care by prioritizing patient outcomes. *JAMA*. 2024;331(8):639.

[17] Karthikesalingam A, Natarajan V. AMIE: A Research AI System for Diagnostic Medical Reasoning and Conversations. Google Research Blog. January 12, 2024. https://blog.research.google/2024/01/amie-research-ai-system-for-diagnostic_12.html

[18] Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589.

[19] GPT-4 System Card. OpenAI; March 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf

[20] Goldberg CB, Adams L, Blumenthal D, et al. To do no harm — and the most good — with AI in health care. *Nat Med*. 2024 Mar;30(3):623-627.

[21] Health Care Artificial Intelligence Code of Conduct. National Academy of Medicine. https://nam.edu/programs/value-science-driven-health-care/health-care-artificial-intelligence-code-of-conduct/

[22] Adams L, Fontaine E, Lin S, Crowell T, Chung VCH, Gonzalez AA (eds). Artificial intelligence in health, health care and biomedical science: An AI code of conduct framework principles and commitments discussion draft. *NAM Perspectives*. Washington, DC: National Academy of Medicine; 2024. https://doi.org/10.31478/202403a

[23] *Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare (Version 1.0)*. Coalition for Health AI; April 2023. https://coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf

[24] Classen DC, Holmgren AJ, Co Z, et al. National trends in the safety performance of electronic health record systems from 2009 to 2018. *JAMA Netw Open*. 2020;3(5):e205547.

[25] The AI Collaborative. The Health Management Academy. https://hmacademy.com/ai-collaborative/

[26] Ratwani RM, Bates DW, Classen DC. Patient safety and artificial intelligence in clinical care. *JAMA Health Forum*. 2024 Feb 2;5(2):e235514.

[27] Mitchell I, Schuster A, Smith K, Pronovost P, Wu A. Patient safety incident reporting: A qualitative study of thoughts and perceptions of experts 15 years after '*To Err is Human*.' *BMJ Qual Saf*. 2016;25(2):92-99.

[28] Shojania KG. Incident reporting systems: What will it take to make them less frustrating and achieve anything useful? *Jt Comm J Qual Patient Saf*. 2021;47(12):755-758.

[29] Macrae C. The problem with incident reporting. *BMJ Qual Saf*. 2016;25(2):71-75.

[30] Gandhi TK, Schulson LB, Thomas AD. Bringing the equity lens to patient safety event reporting. *Jt Comm J Qual Patient Saf.* 2024;50(1):87-89.

[31] Classen D, Li M, Miller S, Ladner D. An electronic health record–based real-time analytics program for patient safety surveillance and improvement. *Health Affairs*. 2018;37(11):1805-1812.

[32] Casalino LP, Gans D, Weber R, et al. US physician practices spend moretThan $15.4 billion annually to report quality measures. *Health Affairs*. 2016;35(3):401-406.

[33] Saraswathula A, Merck SJ, Bai G, et al. The volume and cost of quality metric reporting. *JAMA*. 2023;329(21):1840.

[34] How AI Is Improving Diagnostics, Decision-Making and Care. American Hospital Association. May 9, 2023. https://www.aha.org/aha-center-health-innovation-market-scan/2023-05-09-how-ai-improving-diagnostics-decision-making-and-care

[35] Hunter B, Hindocha S, Lee RW. The role of artificial intelligence in early cancer diagnosis. *Cancers*. 2022;14(6):1524.

[36] Najjar R. Redefining radiology: A review of artificial intelligence integration in medical imaging. *Diagnostics*. 2023;13(17):2760.

[37] Bates DW, Levine D, Syrowatka A, et al. The potential of artificial intelligence to improve patient safety: A scoping review. *NPJ Digit Med*. 2021;4(1):54.

[38] Ferrara M, Bertozzi G, Di Fazio N, et al. Risk management and patient safety in the artificiallintelligence era: A systematic review. *Healthcare*. 2024;12(5):549.

[39] Ferrara M, Bertozzi G, Di Fazio N, et al. Risk management and patient safety in the artificial intelligence era: A systematic review. *Healthcare*. 2024;12(5):549.

[40] Fong A, Howe J, Adams K, Ratwani R. Using active learning to identify health information technology related patient safety events. *Appl Clin Inform*. 2017;26(01):35-46.

Additional Resources

Visit ihi.org/LLISafetyAI

IHI Innovation Report

Artificial Intelligence in Health Care: Implications for Patient and Workforce Safety

In Summer 2023, the IHI Lucian Leape Institute (LLI) commissioned the IHI Innovation Team to conduct a 90-day innovation project that aimed to identify the high-level applications of genAI along with their advantages and disadvantages, including unexpected consequences or new safety hazards for patients and the workforce. The resulting IHI Innovation Report, *Artificial Intelligence in Health Care: Implications for Patient and Workforce Safety*, serves as foundational content that the LLI expert panel reviewed and discussed, and informs the content of this LLI report.

Patient Safety and Artificial Intelligence: Considerations for Key Groups

As part of their work to explore the promise of genAI and its potential risks for patient safety, the IHI Lucian Leape Institute expert panel also reviewed considerations for key groups and provided specific recommendations and mitigation strategies for these audiences:

- Patients and Patient Advocates
- Clinicians
- Safety and Quality Professionals
- Health Care Systems
- GenAI Developers
- Researchers
- Regulators and Policymakers

Patient Safety and Artificial Intelligence
Considerations for Key Groups

Patients and Patient Advocates

As generative artificial intelligence (genAI) is introduced into health care, patient safety and well-being must remain the top priority. The IHI Lucian Leape Institute offers the following recommendations for patients and patient advocates:

- **Demand transparency and trust:** Patients have the right to be informed when genAI is used in their care. While this may become less necessary over time as genAI tools become a more routine part of patient care, at this nascent stage health care organizations must establish minimum standards for disclosure and consent to ensure that patients are informed about the use of genAI in their care. Health systems must also provide patients with a choice to opt out of genAI and prepare alternatives should a patient refuse AI-based services.

- **Educate and support patients:** Patients need to be educated about how genAI will be used to assist clinicians in their care. Clinicians and AI technology (such as chatbots) must create safe spaces for patients to ask questions and share concerns, which includes prompting patients (e.g., "What questions are still on your mind?" or "What are your concerns or worries?").

- **Enhance patient-centered care:** GenAI may relieve clinicians of time-consuming administrative tasks. While some of this time can be used to lower costs to patients and the health care system, a significant fraction of this time needs to be reallocated to improve the quality and patient-centeredness of care. Such efforts might include allocating more time to patient-clinician encounters, providing clinicians with additional time to review and edit documentation prior to and after an encounter, or providing clinicians with breaks to reduce their cognitive loads so they can more fully engage with patients.

IHI Lucian Leape Institute Expert Panel Report on Patient Safety and AI

In January 2024, the IHI Lucian Leape Institute convened an expert panel to further explore the promise of generative artificial intelligence (genAI) and its potential risks for patient safety.

The panel reviewed the literature on AI and patient safety and engaged in a robust discussion that focused on three likely use cases for genAI in health care: documentation support, clinical decision support, and patient-facing chatbots.

The panel also reviewed considerations for key groups and provided specific recommendations and mitigation strategies for these audiences.

Visit ihi.org/LLISafetyAI

- **Protect data privacy and security:** Health care systems and AI developers are responsible for ensuring that patient data is protected and used appropriately. Patients need to provide informed consent for their data to be used, including what data will be used, how it will be used, and how their data will be protected. Key groups also need to be proactive in identifying and addressing new data protection risks such as theft of audio recordings of patient-clinician conversations (e.g., through use of digital scribes).

- **Advocate for effective AI integration:** Patients and patient advocates can encourage the development of AI that enhances the patient-clinician relationship and support policies that maintain human connection and empathy in health care. While patients and advocates need to be included in genAI governance and advisory capacities, other key groups must remember that the responsibility and burden should not fall on patients and patient advocates to ensure that technology is safe for clinical use.

Patient Safety and Artificial Intelligence
Considerations for Key Groups

Institute for Healthcare Improvement

Clinicians

As use of generative artificial intelligence (genAI) becomes more widespread in health care, it is crucial for clinicians to actively learn about these new technologies and how to appropriately use them in care delivery. The IHI Lucian Leape Institute offers the following recommendations for clinicians:

- **Reinvest saved time into improving patient care:** AI-based tools can offer clinicians potential relief from time-consuming clerical work, potentially saving clinicians hours in their workday and contributing to improved working conditions, less clinician burnout and cognitive overload, improved patient experience, and better quality and safety. Yet, these benefits may be undermined if clinical workflows are not adjusted to align with new approaches to documentation or if all of the clinicians' newly available "free" time is reallocated to compensate for the cost of AI-based tools (e.g., see more patients) or save the system money. Clinicians need to advocate for any time made available by AI-generated efficiencies is reallocated, at least in part, to activities that support clinician well-being and provision of high-quality, safe care.

- **Advocate for continuous improvements:** Clinicians need to understand the limitations and challenges of genAI and advocate for or contribute to advances through research or improvement projects. Clinicians must employ their skills and expertise to better understand and ameliorate existing concerns such as the possibility of AI-generated bias; the lack of representative data sets for underserved populations; the absence of robust, rigorous testing and validation on accuracy of performance and outcomes; and the lack of transparency or explainability of AI-based recommendations or results.

IHI Lucian Leape Institute Expert Panel Report on Patient Safety and AI

In January 2024, the IHI Lucian Leape Institute convened an expert panel to further explore the promise of generative artificial intelligence (genAI) and its potential risks for patient safety.

The panel reviewed the literature on AI and patient safety and engaged in a robust discussion that focused on three likely use cases for genAI in health care: documentation support, clinical decision support, and patient-facing chatbots.

The panel also reviewed considerations for key groups and provided specific recommendations and mitigation strategies for these audiences.

Visit ihi.org/LLISafetyAI

- **Recognize that genAI tools may identify discrepancies that need to be reconciled:** While genAI tools may do an acceptable job in summarizing medical records, they may sometimes identify discrepancies such as in medications or code status. Instrument and implement AI tools to show clinicians (and, where appropriate, patients) these discrepancies and ask them to identify ground truth. Exercise care to ensure that this process does not create a tremendous amount of new work for clinicians and that, once reconciled, the record is modified to reflect the true state going forward.

- **Work to learn new skills and retain old skills:** Potential dependency on AI-based outputs and recommendations may lead to overreliance and clinical deskilling. Thus clinicians will need to continue training on basic competencies such as developing appropriate differential diagnoses and management plans, detection of patient deterioration, and even communicating effectively with patients, caregivers, and other clinicians. In addition, education on new genAI-related skills will need to be mastered, including basic genAI knowledge, AI and ethics, simulation practice with clinical genAI tools, and instruction on health care system practices and policies including those to be used in the event of system downtime or when patients do not consent to use of genAI tools in their care.

- **Consider AI tools as an aid, not a replacement:** Current genAI tools function best as support tools for clinicians, assisting with work tasks related to certain aspects clinical care. GenAI tools, at least at present, have not yet reached the level of maturity and accuracy needed to function independently of clinicians. Clinicians, in partnership with patients, must remain the final decision makers, utilizing their critical thinking skills and empathy to guide clinical care and health-related decisions. Health care systems must implement robust strategies to ensure that "clinician in the loop" systems that provide oversight for use of AI tools in clinical care lead to actual safety, and not just the appearance of safety.

Patient Safety and Artificial Intelligence
Considerations for Key Groups

Safety and Quality Professionals

By monitoring processes, identifying risks, and implementing best practices, patient safety and quality professionals prevent errors, reduce harm, and enhance patient outcomes. Tasked with this responsibility, these professionals can help mitigate risks potentially introduced by generative artificial intelligence (genAI) tools, while also employing this new technology to advance patient safety and quality of care. The IHI Lucian Leape Institute offers the following recommendations for safety and quality professionals:

- **Harness AI to advance safety and quality:** Patient safety and quality professionals often spend more than half their time gathering data for the purpose of reporting and monitoring. GenAI can enhance the detection and monitoring of patient safety issues and create efficiencies for mandatory reporting. AI tools may also make existing patient safety processes more efficient, such as collating incident reports, chart abstraction, and root cause analyses. It is important to ensure that a reasonable fraction of the time gained through these AI-generated efficiencies is used to re-task patient safety and quality personnel to develop effective strategies to mitigate recurring safety risks and hazards, instead of solely prioritizing reduced costs.

- **Rethink key paradigms in patient safety:** Before using AI to enhance the incident reporting process, it is worth rethinking the entire ascertainment model of patient safety. For example, historically health care organizations relied on voluntary incident reports to determine cases of healthcare-associated infections. Today, most use sophisticated chart and laboratory review methods rather than voluntary reporting, which is burdensome and neither sensitive nor specific. GenAI tools create opportunities for safety professionals to have "eyes

IHI Lucian Leape Institute Expert Panel Report on Patient Safety and AI

In January 2024, the IHI Lucian Leape Institute convened an expert panel to further explore the promise of generative artificial intelligence (genAI) and its potential risks for patient safety.

The panel reviewed the literature on AI and patient safety and engaged in a robust discussion that focused on three likely use cases for genAI in health care: documentation support, clinical decision support, and patient-facing chatbots.

The panel also reviewed considerations for key groups and provided specific recommendations and mitigation strategies for these audiences.

Visit ihi.org/LLISafetyAI

and ears" on the clinical workplace through nuanced and even real-time chart reviews. Ideally, patients and clinicians at risk for adverse events can be identified before actual harm occurs, creating opportunities to prevent rather than react to harm.

- **Collaborate to strengthen governance and oversight:** Human oversight and governance as a safety guardrail for AI-based tools may be a surprisingly weak protection, as humans are often poor at exercising vigilance. This might lead to perfunctory and ineffective human double checks of generally accurate AI outputs. Thus, patient safety and quality professionals need to collaborate with other key groups (e.g., patients and clinicians, information technology and informatics personnel, data scientists and analysts, leaders) to ensure that systems are in place to facilitate safe, high-quality care. This collaboration must ensure that human oversight of AI involves authentic partnerships and robust, ongoing efforts — not just nominal involvement, which may create the illusion of safety. In addition, engage patient safety and quality professionals in federal and local governance bodies tasked with overseeing AI use in health care to ensure that the design and implementation of genAI prioritizes safe and quality care, with realistic assessments of the true safety of systems that depend on humans acting as safety bulwarks in the face of periodically inaccurate AI outputs.

Patient Safety and Artificial Intelligence
Considerations for Key Groups

Institute for Healthcare Improvement

Health Care Systems

As health care systems embrace the use of generative artificial intelligence (genAI) tools in care delivery, they must remain steadfast in their mission to deliver safe care with quality outcomes. The IHI Lucian Leape Institute offers the following recommendations for health care systems:

- **Embrace AI deliberately:** While genAI promises to streamline operations and enhance care, health system leaders must navigate the associated risks. This includes ensuring that AI systems are developed, implemented, and used responsibly, ethically, and equitably; trustworthy and accurate; and cybersecurity risks have been mitigated. Health systems and their leaders must ensure that the use of genAI meets existing needs; create design and implementation plans that account for the possibility of flawed results and unanticipated consequences; and foster a culture that values patient safety, equitable care, and responsible use of genAI to enhance patient care. Health care systems must also resist the instinct to repurpose any AI-derived efficiencies into expectations of higher clinician throughput, instead reallocating some time efficiencies to reduce clinician burnout, improve the clinician-patient interaction, and meaningfully double-check AI results and recommendations.

- **Invest in AI education, training, and safeguards:** Health systems need to invest in educating their clinicians and safety and quality staff to build competencies for the effective use of genAI tools. This includes basic knowledge of AI, ethics and AI, and training and simulations on how to use system-approved AI-based tools. Also ensure that clinicians and staff maintain basic medical competencies and can function effectively with and without AI-based tools. Prioritize awareness of AI-related system policies and procedures,

IHI Lucian Leape Institute Expert Panel Report on Patient Safety and AI

In January 2024, the IHI Lucian Leape Institute convened an expert panel to further explore the promise of generative artificial intelligence (genAI) and its potential risks for patient safety.

The panel reviewed the literature on AI and patient safety and engaged in a robust discussion that focused on three likely use cases for genAI in health care: documentation support, clinical decision support, and patient-facing chatbots.

The panel also reviewed considerations for key groups and provided specific recommendations and mitigation strategies for these audiences.

Visit **ihi.org/LLISafetyAI**

including downtime plans and efforts to broadly enhance health literacy.[1] Since clinicians are highly likely to use genAI tools to interpret patient data (e.g., to suggest possible diagnoses or guide plans of care), health systems must ensure that they offer such tools inside institutional firewalls to decrease the probability of HIPAA violations.

- **Develop robust AI governance and promote interdisciplinary collaboration:** Establish governance, evaluation, and monitoring procedures to guide the use of genAI with clear policies on privacy, security (including cybersecurity), and data ownership and stewardship, as well as guidance for internal development of AI-based tools. The development and operationalization of governance must precede any AI clinical design or implementation efforts and help prioritize AI use cases, balancing the desire for rapid deployment with the need for caution. Governance bodies need to enlist a diverse group of interested parties while promoting learning within and across systems to maximize genAI's benefits. Governance also needs to develop and implement downtime procedures and AI audit and assessment processes.

- **Prioritize human-centered AI design:** Health care systems need to ensure that internally developed AI-based tools support and enhance the clinician-patient relationship, maintaining a focus on empathy and human connection in care while also improving efficiencies. Require external partners to demonstrate their use of human-centered AI design, which needs to be tested and validated before clinical implementation.

References

[1] Brach C, Keller D, Hernandez LM, et al. Ten Attributes of Health Literate Health Care Organizations. *NAM Perspectives*. Discussion Paper. Washington, DC: National Academy of Medicine; 2012. https://doi.org/10.31478/201206a

Patient Safety and Artificial Intelligence
Considerations for Key Groups

GenAI Developers

In the dynamic landscape of health care and technology, generative artificial intelligence (genAI) developers must balance innovation with human well-being and safety. The IHI Lucian Leape Institute offers the following recommendations for genAI developers:

- **Prioritize transparent and explainable AI design to build trust:** Develop genAI tools that are transparent in their operations and decisions, so patients and clinicians can understand, and build trust in, the outputs. Ideally, once implemented, AI tools provide users with an indication of how confident users should be with each answer or output and cite evidence-based sources, when appropriate. Conduct ongoing evaluations to ensure that AI tools are performing well and to implement improvements.

- **Establish two-way dialogue with users:** Collaborate closely with health care professionals, safety leaders, patients, and patient advocates to gather feedback and iteratively improve AI tools, ensuring they align with existing workflows and enhance patient care and safety. AI developers must be proactive with these efforts, continually iterating on improvements by developing effective feedback channels. In addition, implement mechanisms that give designers signals when AI results are flawed (e.g., frequent rejection by clinicians of a specific piece of AI decision support). Finally, developers need to build tools that increase the probability that "human in the loop" double-checks provide real safety, not the illusion of safety.

- **Commit to learning about and mitigating patient safety risks:** To create effective, ethical, and safe health care solutions, AI developers must ensure that they are knowledgeable about existing and potential patient safety risks. Developers also

IHI Lucian Leape Institute Expert Panel Report on Patient Safety and AI

In January 2024, the IHI Lucian Leape Institute convened an expert panel to further explore the promise of generative artificial intelligence (genAI) and its potential risks for patient safety.

The panel reviewed the literature on AI and patient safety and engaged in a robust discussion that focused on three likely use cases for genAI in health care: documentation support, clinical decision support, and patient-facing chatbots.

The panel also reviewed considerations for key groups and provided specific recommendations and mitigation strategies for these audiences.

Visit ihi.org/LLISafetyAI

need to contribute to ongoing monitoring of their AI-based tools, including continually testing AI tools to ensure safe and appropriate use and contributing to central monitoring systems to identify and mitigate risks across health care systems.

- **Focus on avoiding bias:** Because genAI results generally mirror past practices, they may produce biased results if these practices were biased. Also, if a genAI tool uses a dataset containing a biased sample (e.g., too few women or minoritized patients), the AI output will reflect these biases. Finally, because genAI results are likely to be persuasive, this increases the possibility that self-interested parties (e.g., corporations selling a relevant product) will try to insert themselves into the process to influence AI-generated recommendations. AI developers need to work to overcome and avoid these potential biases at all stages, including the creation of datasets, employing appropriate computational methods to correct for biases, and being transparent about and avoiding corporate influences on AI outputs.

- **Incorporate robust data protection:** Integrate advanced data security measures to protect sensitive patient data and address concerns about privacy and data ownership head-on. Developers must follow existing guidelines for cybersecurity across their product's life cycle and proactively identify potential risks with data, such as the emergence of new data types (e.g., audio recordings produced by digital scribes) and new threat actors and their tactics.[1]

- **Align with regulatory and ethical standards:** Stay informed about and compliant with health care regulations and guidelines to ensure AI tools meet safety, privacy, and ethical standards set by governing bodies. To further ensure that genAI is designed and implemented safely, developers must also consider other suggested or voluntary actions they can take, including the World Health Organization's *Regulatory Considerations on AI for Health*,[2] the National Academy of Medicine's Health Care Artificial Intelligence Code of Conduct,[3] and the Biden-Harris Administration's Executive Order 14110[4] and voluntary commitments[5].

References

[1] Farlow CS, Jump ML, Seeberger MS, Fitzgerald BJ. ANSI/AAMI SW96: Raising the Bar for Medical Device Security Risk Management. *Biomedical Instrumentation & Technology*. 2023;57(2):40-43.

[2] Regulatory Considerations on Artificial Intelligence for Health. World Health Organization; 2023. https://iris.who.int/handle/10665/373421

[3] Health Care Artificial Intelligence Code of Conduct. National Academy of Medicine. https://nam.edu/programs/value-science-driven-health-care/health-care-artificial-intelligence-code-of-conduct/

[4] Sendak M. Health AI Partnership: An Innovation and Learning Network for Health AI Software. Duke Institute for Health Innovation Blog. December 23, 2021. https://dihi.org/health-ai-partnership-an-innovation-and-learning-network-to-facilitate-the-safe-effective-and-responsible-diffusion-of-health-ai-software-applied-to-health-care-delivery-settings/

[5] *ANSI/AAMI SW96:2023; Standard for Medical Device Security—Security Risk Management for Device Manufacturers*. Association for the Advancement of Medical Instrumentation; 2022. https://array.aami.org/doi/10.2345/9781570208621.ch1

Patient Safety and Artificial Intelligence
Considerations for Key Groups

Institute for Healthcare Improvement

Researchers

The entry of generative artificial intelligence (genAI) into health care creates significant concerns regarding validity and effectiveness. Researchers have the opportunity to build a research base on genAI in health care as well as bridge the gap between research and practice. The IHI Lucian Leape Institute offers the following recommendations for researchers:

- **Build and ensure equitable functionality:** A concern raised about genAI is bias in the dataset, which impacts how AI-based tools perform. Research into mitigating the problem of biased datasets and inaccurate outputs of AI-based tools for underserved and racialized patients needs to be a central theme in genAI research efforts. Research in the equitable distribution of and access to quality AI-based tools also needs to be prioritized.

- **Harness validated evidence to build trust and confidence:** Inaccuracies, including hallucinations (situations in which genAI fabricates results), diminish the trustworthiness of AI tools. Researchers can help improve the trustworthiness of AI-based tools and systems by ensuring that data and outcomes, including recommendations on diagnosis or treatment, are accurate and based on the latest evidence-based data. A fruitful line of research is to develop ways of conveying levels of confidence for genAI outputs, such that users could ascertain at a glance (e.g., with color coding) how confident an AI-based tool is in specific statements and recommendations. In addition, researchers can help test and validate each tool's ability to handle conflicting information and the credibility of evidence, and how well AI-human dyads work in real-life practice settings.

IHI Lucian Leape Institute Expert Panel Report on Patient Safety and AI

In January 2024, the IHI Lucian Leape Institute convened an expert panel to further explore the promise of generative artificial intelligence (genAI) and its potential risks for patient safety.

The panel reviewed the literature on AI and patient safety and engaged in a robust discussion that focused on three likely use cases for genAI in health care: documentation support, clinical decision support, and patient-facing chatbots.

The panel also reviewed considerations for key groups and provided specific recommendations and mitigation strategies for these audiences.

Visit ihi.org/LLISafetyAI

- **Prioritize people:** Identifying ways to safely and effectively deliver AI-based tool decision support for patients and clinicians is another area that needs substantial study, using human factors expertise and human-centered design. For instance, would a clinician want an AI-driven alert to pop up automatically with a diagnostic or management recommendation, or would they want to ask for help before genAI provides a recommendation? Human factors issues related to patient preference, alert fatigue, and the impact of technology on the patient-clinician relationships also deserve further study.

- **Advise on the guardrails:** Another critical area of research is developing approaches for ensuring effective human oversight of AI-based tools. For instance, if an AI-based tool like a chatbot is generating responses to patient queries and the process is designed to ensure that clinicians review the AI output to identify inaccuracies and edit prior to responding to the patient, then strategies will be needed to ensure that this human review actually (and meaningfully) occurs. Researchers can also help identify strategies to avoid clinical overreliance or dependence on AI tools and resultant deskilling, as well as study potential AI biases and ways to overcome them.

Patient Safety and Artificial Intelligence
Considerations for Key Groups

Institute for Healthcare Improvement

Regulators and Policymakers

By fostering responsible generative artificial intelligence (genAI) adoption and use, regulators and policymakers can help patients, clinicians, and health care systems leverage the promise of genAI to enhance patient care and outcomes without compromising safety or quality. The IHI Lucian Leape Institute offers the following recommendations for regulators and policymakers:

- **Establish clear guidelines for ethical and trustworthy use of AI:** Develop comprehensive regulations for the development and deployment of genAI in health care, focusing on patient safety, data protection, and ethical use of technology. Regulators and policymakers need to collaborate with ongoing efforts in health care to develop guidelines such as the National Academy of Medicine's Health Care Artificial Intelligence Code of Conduct,[1] the guidelines and guardrails by the Coalition for Health AI,[2] and the Association for the Advancement of Medical Instrumentation's guidance on medical device cybersecurity (ANSI/AAMI SW96)[3].

- **Support transparency and accountability:** Mandate transparency in genAI systems to ensure that users understand how AI makes decisions, with mechanisms in place for accountability in case of errors or adverse outcomes. It is critical that regulators and policymakers, in a collaborative effort with AI developers and users, establish definitions of and thresholds for transparency. Because genAI results are likely to be persuasive, this increases the possibility that self-interested parties (e.g., corporations selling a relevant product) will try to insert themselves into the process to influence the AI recommendations. It is important that such efforts are anticipated and dealt with through appropriate regulations.

IHI Lucian Leape Institute Expert Panel Report on Patient Safety and AI

In January 2024, the IHI Lucian Leape Institute convened an expert panel to further explore the promise of generative artificial intelligence (genAI) and its potential risks for patient safety.

The panel reviewed the literature on AI and patient safety and engaged in a robust discussion that focused on three likely use cases for genAI in health care: documentation support, clinical decision support, and patient-facing chatbots.

The panel also reviewed considerations for key groups and provided specific recommendations and mitigation strategies for these audiences.

Visit **ihi.org/LLISafetyAI**

- **Promote AI literacy:** Advocate for and fund initiatives that enhance AI literacy among clinicians, health care staff, and patients, ensuring that they are informed about the benefits and limitations of genAI tools. These efforts can include facilitating training for health care professionals on genAI fundamentals, ethical considerations, and practical applications; public awareness campaigns; and incentives for learning such as accredited certification programs or open access continuing education credits on genAI in health care.

- **Incentivize AI development and research that prioritizes safety:** Encourage the creation of genAI tools that prioritize patient safety through incentives for AI developers and health care systems that meet high testing thresholds and safety and quality standards. Research should also be incentivized to ensure the creation of a robust, impartial evidence base.

- **Facilitate localized decision-making:** Recognize the importance of context and nuance in health care by allowing for local governance in the implementation of genAI, while providing a federal framework for overarching AI safety and ethical standards. This can help ease user anxiety and temper overexcitement while providing a standard framework that allows for safe design, implementation, and use of genAI tools across hospitals and health care systems. This effort can be further bolstered through required safety data reporting from health care systems and developers, overseen by appropriate regulatory and enforcement bodies.

References

[1] Health Care Artificial Intelligence Code of Conduct. National Academy of Medicine. https://nam.edu/programs/value-science-driven-health-care/health-care-artificial-intelligence-code-of-conduct/

[2] *Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare (Version 1.0)*. Coalition for Health AI; April 2023. https://coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf

[3] *ANSI/AAMI SW96:2023; Standard for Medical Device Security—Security Risk Management for Device Manufacturers*. Association for the Advancement of Medical Instrumentation; 2022. https://array.aami.org/doi/10.2345/9781570208621.ch1