American Journal of Healthcare Strategy
DOI 10.61449/ajhcs.2025.15 | E-ISSN 2995-6242
Published in Volume 1, Issue 3, on August 14, 2025

Artificial Intelligence in Healthcare: No Longer Optional But Neither Is Patient Safety

Ryan Sears, PharmD
Cooper University Hospital

Abstract

The use of artificial intelligence (AI) in healthcare comes with many potential benefits but there is potential for catastrophe if AI models are not properly designed and vigilantly monitored. This article cites several historical disasters involving healthcare in AI, including the disappointing performance of the Epic Sepsis Model and a care-management algorithm that unintentionally de-prioritized Black patients due to differences in healthcare spending from White patients. Health system leaders must form AI governance committees and appoint personnel to monitor performance of their AI models to prevent history from repeating itself.

Introduction

Though its presence is often unseen—sometimes even unknown—artificial intelligence (AI) weaves through nearly every layer of modern healthcare delivery, from the moment a triage nurse orders laboratory work to the instant a discharge prescription is printed. When these systems are properly validated and monitored, they can detect lung cancer on computed-tomography scans as accurately as expert radiologists in a fraction of the time, accelerating diagnosis and potentially improving survival[1]. Alternatively, the same computational speed that magnifies clinical benefit can also amplify harm, as illustrated by a series of high-profile failures in the last decade. Understanding why those missteps occurred, and how emerging regulatory and technical guardrails aim to prevent their recurrence, is now essential for any health-system leader or clinician seeking to use AI responsibly.

Foreshadowing: An Early Failure of Healthcare AI

In 2011, IBM's Watson computer system dazzled television viewers by defeating two human champions on *Jeopardy!*, a quiz competition game show. Eager to capitalize on the positive media

[1] Hammad M, ElAffendi M, El-Latif AAA, Ateya AA, Ali G, Plawiak P. Explainable AI for lung cancer detection via a custom CNN on CT images. Sci Rep. 2025;15(1):12707. Published 2025 Apr 13. doi:10.1038/s41598-025-97645-5

DOI 10.61449/ajhcs.2025.15 | E-ISSN 2995-6242

Published in Volume 1, Issue 3, on August 14, 2025

attention, the company quickly redirected Watson's natural-language technology toward the field of oncology, investing roughly $4 billion to create a system that could suggest personalized chemotherapy regimens. Internal company memoranda later revealed that Watson for Oncology had been trained on a narrow set of hypothetical cases and a small number of experts' opinions rather than population-level clinical data or clinical practice guidelines. As a result, it produced "unsafe and incorrect treatment recommendations"[2]. IBM was forced to dismantle the expensive program after participating hospitals withdrew, eventually selling off a large portion of its Watson Health business altogether[3]. The financial and reputational disaster underscores a recurrent theme: spectacular performance on synthetic tasks does not guarantee reliability at the bedside.

New AI Models, Same Old Problems

Subsequent generations of commercial tools, while more modern, have repeated similar errors in subtler ways. The proprietary Epic Sepsis Model, deployed in hundreds of United States hospitals, was marketed as a real-time early-warning system for patients experiencing sepsis. One audit of the model showed its algorithm missed two-thirds of patients who ultimately developed sepsis. Not only did it miss so many patients, it also repeatedly flagged patients without sepsis for review[4]. These alerts are distracting and may cause providers to take real alerts less seriously. Sepsis management is time-critical, where both false reassurance and alert fatigue can be deadly; thus, the model's inadequate calibration created a glaring patient safety risk.

Even when overall accuracy appears acceptable, seemingly benign design choices can entrench structural inequities. A widely used "high-risk care-management" algorithm evaluated by Obermeyer and colleagues prioritized patient care based on predicted health-care expenditures. Because Black patients historically accrue lower costs than comparably ill White patients, the model underestimated their disease burden by 26.3% at the threshold where patients were auto-identified for aid, thereby removing access to the follow-up care they desperately needed and would have otherwise been offered[5]. The algorithm developers assumed that cost was an neutral proxy for need; in reality, it encoded decades of healthcare disparities into a mathematical function.

---

[2] Hogan A. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. STAT News. Published July 25, 2018. Accessed June 25, 2025. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

[3] Taylor P. IBM sells off large parts of Watson Health business. PharmaPhorum. Published January 24, 2022. Accessed July 7, 2025. https://pharmaphorum.com/news/ibm-sells-off-large-parts-of-watson-health-business

[4] Wong A, Otles E, Donnelly JP, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients [published correction appears in JAMA Intern Med. 2021 Aug 1;181(8):1144. doi:10.1001/jamainternmed.2021.3907.]. JAMA Intern Med. 2021;181(8):1065-1070. doi:10.1001/jamainternmed.2021.2626

[5] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453. doi:10.1126/science.aax2342

American Journal of Healthcare Strategy
DOI 10.61449/ajhcs.2025.15 | E-ISSN 2995-6242
Published in Volume 1, Issue 3, on August 14, 2025

Following these results, AI safety advocates began focusing more heavily on the reduction of bias in both models and regulatory frameworks[6].

How are Regulators Responding?

Regulators have begun to respond to the growing incidence of undesirable AI events. In 2023, the U.S. National Institute of Standards and Technology released the voluntary AI Risk Management Framework (AI-RMF 1.0), which guides organizations through the identification, measurement, and mitigation of algorithmic hazards across the technology life-cycle[7]. Complementing this, the U.S. Food and Drug Administration, Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency jointly published ten Good Machine Learning Practice (GMLP) principles. These principles emphasize multidisciplinary design teams, representative training data, rigorous pre-deployment testing, and post-market performance monitoring; they also bear similarities to the controls applied to traditional medical devices[8]. Internationally, the World Health Organization's 2021 guidance places human rights and equity at the center of AI governance, calling for transparency, accountability, and corrective mechanisms for when systems fail[9].

A Call to Action for Health-System Leaders

Frameworks, however, cannot implement themselves. Health-system leaders must institutionalize oversight structures akin to infection-control committees. Institutions should assemble standing algorithm-governance boards, composed of clinicians, pharmacists, informaticians, data scientists, ethicists, and patient representatives. These boards can vet proposed models against predefined technical and ethical checklists, mandate external validation on local data, and set thresholds for acceptable performance drift. Model fact sheets should be created detailing data provenance, training-set demographics, known limitations, and intended clinical context in language accessible both to specialists and frontline users. Once a tool is live, automated dashboards should be utilized to track real-time sensitivity, specificity, and subgroup performance. When metrics deviate from

[6] Waithira J, Chweya R, Cyprian RM. Adversarial debiasing for bias mitigation in healthcare AI systems: a literature review. OALib Journal. 2025;12:1-13. doi:10.4236/oalib.1113340
[7] Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST Trustworthy and Responsible AI. National Institute of Standards and Technology; 2023. https://doi.org/10.6028/NIST.AI.100-1. Accessed June 26, 2025.
[8] U.S. Food and Drug Administration; Health Canada; United Kingdom Medicines and Healthcare products Regulatory Agency. Good Machine Learning Practice for Medical Device Development: Guiding Principles. Updated March 25, 2025. Accessed June 26, 2025. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles
[9] World Health Organization. Ethics and governance of artificial intelligence for health. Published June 28, 2021. Accessed June 26, 2025. https://www.who.int/publications/i/item/9789240029200

DOI 10.61449/ajhcs.2025.15 | E-ISSN 2995-6242

Published in Volume 1, Issue 3, on August 14, 2025

baselines established during testing, alerts should trigger pre-determined review and correction plans.

Vendor transparency is equally critical. While proprietary code does not necessarily need to be open-sourced, developers must provide sufficient technical documentation to allow independent replication of results and facilitate adverse-event investigations. Recent standards on decision-support safety by The Joint Commission (effective January 2024) already require accredited hospitals to demonstrate they have procedures in place to evaluate and monitor such tools throughout their life-cycle; failure to comply can jeopardize accreditation and reimbursement[10]. In parallel, the Algorithmic Accountability Act reintroduced in the U.S. Senate would mandate bias audits and impact assessments for high-risk health algorithms, signaling growing legislative interest[11].

How to Incorporate AI into Healthcare the Right Way

There are health systems rising to the challenge of creating responsible frameworks for AI model use. The benefits of doing so are not just limited to defense against regulation; when intentionally designed and monitored with vigilance, AI can improve patient outcomes. An example of a success story can be found at the University of California San Diego School of Medicine. UCSD researchers published an assessment of COMPOSER, their deep-learning model for the "early prediction of sepsis on patient outcomes." In a before-and-after study across two emergency departments, the COMPOSER model led to a 1.9% absolute drop in sepsis mortality in-hospital and a 5% absolute rise in sepsis bundle compliance[12].

The differences between the successful COMPOSER model and the failed Epic Sepsis Model? COMPOSER was built with safety and transparency front-and-center. Rather than flag all outlier patients as having sepsis, COMPOSER marked patients with data points outside of the training distribution as "indeterminate." This reduced the number of false alarms and helped maintain clinician trust. Each alert displayed the top variables contributing to the model's determination. These alerts were delivered through a nurse-facing Best Practice Advisory, averaging only 1.65 alerts per nurse per month. When nurses saw these alerts, they mostly selected "Will Notify MD

[10] The Joint Commission. National Patient Safety Goals®: Hospital Program. Effective January 2024. Published 2023. Accessed June 26, 2025. https://www.jointcommission.org/-/media/tjc/documents/standards/national-patient-safety-goals/2024/npsg_chapter_hap_jan2024.pdf

[11] Simonite T. Senators Protest a Health Algorithm Biased Against Black People. WIRED. Published December 3, 2019. Accessed June 26, 2025. https://www.wired.com/story/senators-protest-health-algorithm-biased-against-black-people/

[12] Boussina A, Shashikumar SP, Malhotra A, et al. Impact of a deep learning sepsis prediction model on quality of care and survival [published correction appears in NPJ Digit Med. 2024 Jun 12;7(1):153. doi: 10.1038/s41746-024-01149-x.]. NPJ Digit Med. 2024;7(1):14. Published 2024 Jan 23. doi:10.1038/s41746-023-00986-6

DOI 10.61449/ajhcs.2025.15 | E-ISSN 2995-6242

Published in Volume 1, Issue 3, on August 14, 2025

Immediately" as their response. Because of these design choices, and several others developed over 2-3 years prior to model deployment, the alerts were infrequent and seen as trustworthy. This led to more action being taken and an actual reduction of sepsis mortality as hoped[13].

Conclusion

The trajectory of AI in health care is unlikely to reverse as its analytical abilities expand with every passing week and each additional terabyte of clinical data. The pressing question is whether AI governance mechanisms will mature quickly enough to safeguard patients while still permitting innovation. By embedding rigorous validation protocols, continuous post-deployment surveillance, and transparent reporting into routine quality-improvement activities, institutions can convert AI from an unpredictable hazard into a dependable ally. The same vigilance that transformed hospital infection control from an afterthought into a cornerstone of patient safety must now be applied to algorithmic care. Only then will AI reliably amplify, rather than undermine, the expertise of the clinicians it is meant to serve.

[13] Boussina A, Shashikumar SP, Malhotra A, et al. Impact of a deep learning sepsis prediction model on quality of care and survival [published correction appears in NPJ Digit Med. 2024 Jun 12;7(1):153. doi: 10.1038/s41746-024-01149-x.]. NPJ Digit Med. 2024;7(1):14. Published 2024 Jan 23. doi:10.1038/s41746-023-00986-6