HealthManagement
Promoting Management and Leadership

Balancing AI Innovation with Patient Safety

Artificial intelligence has become deeply embedded in healthcare delivery, from triage and diagnostics to discharge. Its ability to process vast data rapidly offers opportunities to improve survival rates and efficiency, yet history shows that poorly designed or monitored systems can create severe risks. Cases of failed implementation illustrate that without robust governance, patient safety can be undermined. The challenge for leaders is to ensure that innovation progresses while safeguards are firmly in place.

Early Missteps in Healthcare AI

The promise of AI in healthcare gained early attention when IBM redirected its Watson system into oncology after its success on television. With a significant investment, the company sought to revolutionise cancer treatment recommendations. However, Watson for Oncology relied on a narrow set of hypothetical cases and limited expert opinions rather than broad population-level data. This led to unsafe and incorrect treatment suggestions, forcing hospitals to abandon the programme and IBM to dismantle much of the Watson Health business. The collapse demonstrated how technological spectacle does not equate to clinical reliability.

The risks did not end with Watson. The Epic Sepsis Model, deployed widely in American hospitals, was marketed as a real-time detection tool. Yet evaluations showed it missed most patients who developed sepsis while simultaneously over-flagging others, contributing to alert fatigue and dangerous complacency. Another example involved a care-management algorithm that unintentionally perpetuated structural inequities. By using healthcare costs as a proxy for patient need, it underestimated the disease burden of Black patients, reducing their access to vital care. These failures underscored that even modern AI can encode systemic biases or fail in calibration, endangering patient safety.

The Role of Regulation and Governance

In response to such failures, regulatory bodies have introduced frameworks to mitigate AI risks. In 2023, the U.S. National Institute of Standards and Technology issued the AI Risk Management Framework to help organisations assess and control algorithmic hazards across the life cycle of technologies. Complementing this, agencies in the United States, Canada and the United Kingdom jointly released principles of Good Machine Learning Practice. These emphasise representative training data, rigorous testing before deployment and continuous monitoring once systems are in use. The World Health Organization has added a global perspective, insisting on transparency, accountability and human rights at the core of AI governance.

Must Read: ***Patient Safety and High Reliability***

For these frameworks to succeed, health systems must create internal oversight structures similar to infection-control committees. Multidisciplinary algorithm-governance boards can evaluate proposed models against ethical and technical criteria, require external validation on local data and monitor ongoing performance through dashboards that track sensitivity, specificity and subgroup outcomes. Vendor transparency is another critical element. Developers must provide documentation sufficient to allow replication and facilitate investigation of adverse events. Standards from organisations such as The Joint Commission already require hospitals to evaluate and monitor decision-support systems throughout their life cycles, linking compliance to accreditation and reimbursement.

Pathways to Responsible AI in Healthcare

Despite early setbacks, some systems demonstrate how responsible AI use can deliver real improvements. The University of California San Diego School of Medicine reported success with its deep-learning model COMPOSER, designed to predict sepsis early. Unlike the Epic Sepsis Model, COMPOSER included transparency features such as marking indeterminate cases rather than forcing decisions and displaying the main

factors influencing alerts. With fewer but more meaningful alerts, clinicians trusted the system, acted on the guidance and achieved reduced sepsis mortality. The model's design choices, grounded in safety and transparency, highlight the benefits of careful development and prolonged preparation before deployment.

Such examples illustrate that integrating AI responsibly requires vigilance at every stage. Model fact sheets should clarify data sources, demographics, limitations and intended clinical use in accessible language. Automated monitoring must detect deviations from baseline performance, triggering predefined corrective actions. Leaders must prioritise diversity in design teams and ensure that equity considerations are embedded in algorithms. By doing so, AI can be used not only to comply with regulations but to improve quality of care and strengthen patient outcomes.

Artificial intelligence is set to play an increasingly central role in healthcare, but its risks are as significant as its promise. Past failures show that enthusiasm without oversight can lead to unsafe recommendations, missed diagnoses or entrenched inequities. Regulatory bodies and global organisations have responded with frameworks that place responsibility on both developers and health systems. Ultimately, success depends on vigilant governance, transparent processes and continuous monitoring. If patient safety is given the same priority in AI as in infection control, healthcare can move toward a future where algorithms enhance rather than compromise clinical expertise.

Source: **American Journal of Healthcare Strategy**

Image Credit: **iStock**

Published on : Tue, 9 Sep 2025