Journal of Translational Critical Care Medicine **2026** vol. 8(1) | e00021
DOI: 10.14218/JTCCM.2025.00021

Review Article

On the Purpose of Artificial Intelligence in Critical Care Medicine

Cristian Drudi[1*], Sarah Matta[2,3], Hyeonhoon Lee[4,5,6], Sharon C. O'Donoghue[7], Helen T. D'Couto[8], Amjad Hamza[9], Claribeth Arias Gutierrez[10,11], Rose Nakasi[12], Joseph Byers[13], Martin Tumukunde[14], Riccardo Barbieri[1] and Leo Anthony Celi[13,15,16]

[1]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy; [2]Université de Bretagne Occidentale, Brest, France; [3]INSERM, UMR 1101, Brest, France; [4]Department of Transdisciplinary Medicine, Seoul National University Hospital, Seoul, Korea; [5]Healthcare AI Research Institute, Seoul National University Hospital, Seoul, Korea; [6]Department of Medicine, Seoul National University College of Medicine, Seoul, Korea; [7]School of Nursing, Northeastern University, Boston, MA, USA; [8]School of Medicine, Georgetown University, Washington, DC, USA; [9]Faculty of Medicine, American University of Beirut, Beirut, Lebanon; [10]Sanitary Registration Group for Biological Medicines and Radiopharmaceuticals, INVIMA, Bogotá, Colombia; [11]AMCITECH-Biomedical Technology and Innovation in Critical Care Committee, Colombian Association of Critical Medicine and Intensive Care (AMCI), Bogotá, Colombia; [12]Makerere AI Health Lab, College of Computing and Information Sciences, Makerere University, Kampala, Uganda; [13]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA; [14]Faculty of Medicine, Mbarara University of Science and Technology, Mbarara, Uganda; [15]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA; [16]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA

Received: August 30, 2025 | **Revised:** January 07, 2026 | **Accepted:** January 28, 2026 | **Published online:** March 30, 2026

Abstract

The modern intensive care unit (ICU) inundates clinicians with large volumes of data, leading to cognitive overload and a gap between data availability and actionable insight. While artificial intelligence (AI) promises a solution, its clinical adoption is limited by systemic barriers, including algorithmic bias, a lack of trust, and validation failures. This paper argues that a design philosophy that envisions AI as an autonomous decision-maker, rather than an integrated collaborative tool, has hindered its clinical adoption. We propose an alternative: a collaborative framework designed to augment the intensivist's expertise by offloading specific cognitive burdens. This framework redefines AI's purpose as managing data-intensive tasks, illustrated through four collaborative example roles: a synthesizer to create coherent clinical narratives, a sentinel for proactive deterioration surveillance, a simulator to forecast patient responses to interventions, and a stratifier to identify meaningful subphenotypes within complex syndromes. By delegating these computational tasks, this collaborative model frees clinicians to focus on complex synthesis, nuanced judgment, and compassionate communication. Realizing this vision requires a deliberate translational pathway focused on robust data infrastructure, human-centered design, and rigorous validation through prospective clinical trials. Ultimately, the successful integration of AI in critical care depends not on replacing clinicians but on empowering them, creating a more functional ICU in which technology supports the delivery of safer, more precise, and more humane care.

Introduction

The modern intensive care unit (ICU) represents one of the most data-dense and technologically complex environments in modern medicine. It is a domain of unprecedented technological sophistication, where continuous multi-organ support and high-frequency monitoring have dramatically improved survival rates over the past three decades.[1] Yet, this very progress has precipitated a crisis of its own making. The contemporary intensivist is inundated by a high-velocity, high-dimensional data stream, a relentless flow of information that overwhelms the finite capabilities of the human mind.[2] This data-rich environment paradoxically fosters an information-poor reality, leading to cognitive overload, diagnostic inertia, and clinician burnout. This phenomenon, in which an abundance of data fails to translate into actionable insights, represents a missed opportunity and a fundamental challenge to the

Keywords: Critical care; Artificial intelligence; Human-AI collaboration; Decision-making; Clinical decision support; Cognitive overload.

***Correspondence to:** Cristian Drudi, Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan 20133, Italy. ORCID: https://orcid.org/0009-0000-1733-8823. E-mail: cristian.drudi@polimi.it

How to cite this article: Drudi C, Matta S, Lee H, O'Donoghue SC, D'Couto HT, Hamza A, *et al*. On the Purpose of Artificial Intelligence in Critical Care Medicine. *J Transl Crit Care Med* 2026;8(1):e00021. doi: 10.14218/JTCCM.2025.00021.

delivery of safe and effective care.[3]

The history of health information technology provides a crucial cautionary tale regarding the sociotechnical complexities of implementation. The widespread adoption of electronic health records (EHRs), while offering undeniable benefits in data accessibility and standardization, often came at a significant cost.[4] Studies have extensively documented the unintended negative consequences, including disrupted clinical workflows, a high prevalence of alert fatigue from poorly tuned decision support systems, and a significant increase in documentation burden.[5,6] This time spent on documentation outside of direct work hours has been directly linked to the erosion of professional satisfaction, which is a major contributor to the "quadruple aim" challenge of improving clinician well-being.[7] This example teaches a vital lesson: Technology naively deployed, without deep consideration for human factors and workflow integration, can create as many problems as it solves.

Into this complex environment, artificial intelligence (AI) has arrived with the promise of a second, more profound transformation. AI research in medicine has seen major breakthroughs in recent years, with AI systems demonstrating expert-level performance in various domains, including dermatology, ophthalmology, and oncology.[8] Many of these applications have successfully obtained regulatory approval, such as Conformité Européenne (CE) marking or Food and Drug Administration (FDA) clearance in the United States.[9,10] The fundamental promise is transformative: to enhance diagnostic precision, streamline operational efficiencies, optimize resource allocation, and significantly alleviate the burden of long patient waiting times.[11] However, a deep and troubling gap persists between the booming research landscape and the reality of clinical practice. Despite rapid technological progression and regulatory clearance, the integration of AI into routine clinical care remains limited and fraught with challenges.[12]

This gap between AI's potential and its real-world application stems from a complex interplay of systemic barriers. Foremost among them are issues of algorithmic integrity; models trained on unrepresentative datasets risk perpetuating and even exacerbating health disparities, a concern substantiated by reviews showing significant demographic reporting gaps in FDA-approved devices.[13,14] This is compounded by persistent validation deficiencies, where models that perform well in controlled settings fail to generalize to diverse, real-world clinical environments—a problem of distributional shift that undermines model safety and reliability.[15,16] External validation studies have revealed potential safety concerns when algorithms are deployed in unseen domains, underscoring the urgent need for prospective, interventional trials to assess their effectiveness in clinical practice, even post-FDA approval.[17] Beyond these technical hurdles lies a profound deficit of trust among clinicians and patients, who are skeptical of opaque "black box" systems and anxious about the erosion of humanistic care.[18–20] This mistrust is exacerbated by unresolved questions of legal liability, lagging regulatory frameworks, and the absence of clear reimbursement models to incentivize adoption.[21–23]

Intensivists are expected to perform high-dimensional, long-range counterfactual reasoning, for instance, predicting how a single intervention like increasing the vasopressor dose might interact with multiple comorbidities (*e.g.*, underlying diastolic dysfunction) and therapies (*e.g.*, continuous renal replacement therapy) to affect a patient's trajectory not just over hours, but over days.[24] The inherent difficulty of this task, coupled with known cognitive biases,[25] highlights the precise need for computational support that can model complex downstream effects and extend clinical foresight. The question is not if computational support is needed, but how it should be designed and for what ultimate purpose.

This paper argues that the chasm between promise and practice exists, in large part, because the dominant narrative framing AI as an autonomous replacement for the clinician is fundamentally flawed. This perspective ignores the irreducible uncertainty of critical illness and the essential role of human virtues like empathy, compassion, and ethical judgment.

We posit that the true and highest purpose of AI in critical care is not to replace the irreplaceable but to function as a collaborator. This framework recasts AI as a powerful tool designed to augment the intensivist's expertise by offloading the immense cognitive burden of data management and pattern recognition. In doing so, it frees the clinician to focus on the higher-order tasks of clinical assessment, complex synthesis, nuanced decision-making, and compassionate communication. To achieve this vision, this paper will first deconstruct the systemic barriers hindering AI adoption. We will then critically examine the telos of AI in critical care before proposing a concrete framework for the realization of this symbiotic collaboration between clinicians and AI. Finally, we will outline the necessary translational pathway "from code to bedside" required to build a future in which technology fosters a more precise, proactive, and humane ICU.

Deconstructing the gap: Systemic barriers to clinical translation

The translation of AI from controlled research environments to the chaotic, high-stakes reality of the ICU is not a simple matter of technological scaling; it is a complex sociotechnical challenge fraught with systemic barriers.[11] While AI models frequently demonstrate expert-level performance on curated datasets, their real-world clinical adoption remains profoundly limited. As illustrated in Figure 1, the gap between promise and practice can be deconstructed into three interdependent domains of failure: fundamental issues of algorithmic integrity and validation, a deep-seated deficit of trust rooted in opacity and human factors, and significant socioeconomic and regulatory hurdles that stifle implementation.

Algorithmic integrity

The performance of AI models can be deceptively fragile. This epistemological fragility is a primary contributor to the research-practice gap; models may fail when transitioned from controlled, retrospective datasets to the dynamic and heterogeneous environment of real-world clinical practice. This failure to translate performance into the real world is a predictable outcome of two deeply interconnected issues: pervasive algorithmic bias inherent in the training data and a systemic validation gap that masks the true performance of models on unseen populations.[26]

Pervasive algorithmic bias

Algorithmic bias refers to systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one group of users over others.[13] In medical AI, this represents a critical threat to health equity, with the potential to encode, perpetuate, and even amplify existing societal and health disparities. This bias can manifest in several insidious forms throughout the data lifecycle.[27]

Sampling and representation bias

The foundation of most machine learning models is the training data, and if this foundation is skewed, the resulting model will also be skewed. Models are frequently trained on data from a small number of large, urban, academic medical centers, which systematically

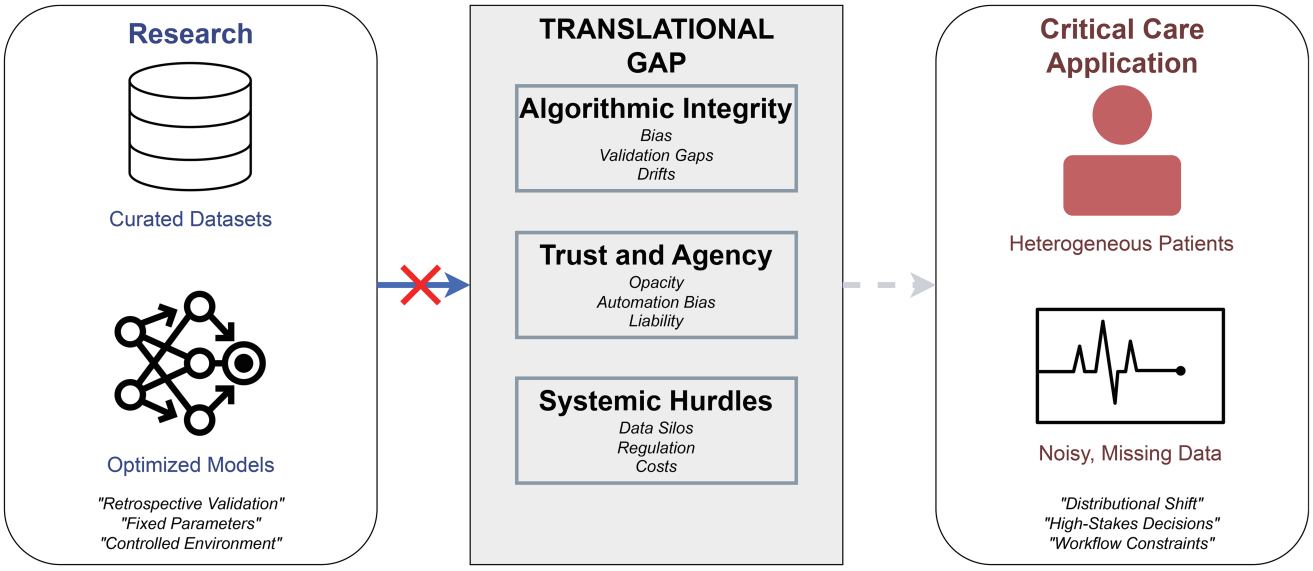

Fig. 1. The translational gap in AI for critical care. A conceptual diagram illustrating the barriers between research and critical care applications. The figure highlights three primary hurdles: algorithmic integrity, trust deficits, and systemic barriers. AI, artificial intelligence.

underrepresent rural populations, specific racial and ethnic minorities, and patients from lower socioeconomic strata.[28] A canonical example of this was demonstrated in a study of a commercial algorithm used to manage the health of populations.[28] The algorithm used healthcare cost as a proxy for health need, but because Black patients at a given level of illness tend to incur lower healthcare costs than White patients (due to factors like access and historical mistrust), the algorithm falsely concluded they were healthier, leading to significantly less care being allocated to them.[28]

In the ICU, an acute respiratory distress syndrome (ARDS) prediction model trained exclusively at a tertiary referral center with a high volume of complex medical ARDS may perform poorly in a community hospital ICU that sees a higher proportion of patients with postoperative or trauma-related respiratory failure, as the underlying patient characteristics and etiologies are fundamentally different. An AI model trained on such data does not learn useful information about critically ill patients; it may also learn the systemic biases of the healthcare system that generated the data.

Measurement and feature bias

Bias can also be embedded in the very measurements used as features. The data points are not objective reflections of biology but are filtered through measurement devices and clinical practices. A clear example is the known racial bias in pulse oximetry, which is less accurate in patients with darker skin pigmentation and can lead to the underdetection of occult hypoxemia.[29] An AI model trained on this data has no way of knowing the measurement is flawed; it treats the biased SpO_2 value as ground truth, thereby learning a dangerous and incorrect association that could systematically endanger Black patients by delaying intubation or undertitrating oxygen. Similarly, variations in imaging equipment, acquisition protocols, or even laboratory assay manufacturers can introduce systematic, nonpathological variations in data that a model might erroneously learn as a signal for disease.[30]

Label and annotation bias

The "ground truth" labels used to train and evaluate models are often far from true. Diagnostic labels like "sepsis" or "pneumonia" are the product of human interpretation, which is subject to variability and error. In a prior study, it was found that a model for pneumonia detection had learned a spurious correlation: It associated the presence of a portable chest X-ray machine (a confounder) with a higher likelihood of pneumonia, not because of the lung pathology itself, but because sicker patients who are more likely to have pneumonia are also more likely to receive a portable X-ray at the bedside.[30] The model learned a clever shortcut to minimize its loss function, not clinical medicine. This is further complicated by significant interrater variability among expert clinicians; an AI trained on the annotations of one pathologist may perform poorly when judged against the standards of another, making its "truth" contingent and arbitrary.[31]

The validation gap and distributional shift

A critical contributing factor to this discrepancy is the limitation of standard validation practices, which often yield inflated performance estimates that fail to generalize to real-world clinical environments.

Internal *vs.* external validation

Most studies report performance based on internal validation, where a model is tested on a held-out portion of the same dataset on which it was trained. While a necessary step to check for overfitting, this provides very little information about how the model will perform on data from a different hospital, a different patient demographic, or even the same hospital a year later.[15] The true test of a model's robustness is rigorous external validation on completely independent datasets. Such studies frequently reveal a significant degradation in performance, a phenomenon that has been well documented across medical specialties.[16,32]

The challenge of distributional shift

This performance degradation is a direct result of distributional shift, the statistical mismatch between the training data distribution and the deployment data distribution.[15] This shift can occur in

several ways. Covariate shift happens when the distribution of patient characteristics (*e.g.*, age or comorbidities) changes. Concept drift occurs when the underlying relationship between features and outcomes changes over time, for instance, due to new clinical guidelines or treatments. For example, an acute kidney injury (AKI) prediction model developed before the widespread adoption of balanced crystalloids might overestimate risk in a hospital that has since shifted away from saline-based resuscitation.[33] A static AI model deployed in such a dynamic environment will inevitably become outdated and potentially unsafe.

Choice of appropriate validation metrics

Selecting validation metrics that reflect clinically meaningful performance is essential for assessing the real-world utility of AI models. Commonly reported metrics, such as the area under the receiver operating characteristic curve, can overestimate usefulness in imbalanced datasets,[34] which are prevalent in critical care settings (*e.g.*, predicting rare events like pulmonary embolism). In such contexts, metrics like the area under the precision-recall curve (AUPRC), calibration measures (*e.g.*, Brier score or calibration slope), and decision-analytic approaches (*e.g.*, net benefit *via* decision curve analysis) provide more informative evaluations.[35] Moreover, reporting multiple complementary metrics helps capture different aspects of performance, thereby enabling a more comprehensive and transparent appraisal of model readiness for deployment.

The clinical implications of this gap are profound. A multicenter, head-to-head study of seven different FDA-cleared AI systems for diabetic retinopathy screening found wide variability in performance and identified potential safety concerns when these systems were applied in real-world clinical settings.[17] This demonstrates that regulatory approval, while an important milestone, is not a final guarantee of clinical robustness or safety. The combination of inherent data bias and validation practices that fail to account for distributional shift creates a perfect storm that stands as one of the most formidable barriers to the safe and equitable translation of AI into critical care. This necessitates a paradigm shift toward continuous model monitoring, prospective trial designs, and a commitment to algorithmic auditing as standard practice.

A lack of trust: Model opacity, accountability, and sociotechnical friction

Even a technically robust and externally validated algorithm will fail if clinicians do not trust it and patients do not accept it. The deficit of trust is arguably the most important nontechnical barrier to AI adoption, representing a complex web of sociotechnical friction. This friction arises from the inherent opacity of many AI models, the resulting ambiguity in accountability, and the complex human factors that govern the acceptance of new technology. At the heart of the trust deficit lies the problem of epistemic opacity, or the "black box" nature of many high-performance AI models.[19] While deep neural networks can achieve remarkable performance metrics, their internal decision-making processes often defy human-readable explanation. This opacity is fundamentally at odds with the epistemological foundations of medicine, which demand not just a correct answer but a physiologically plausible and defensible rationale. A clinician cannot responsibly act on a high-stakes recommendation, such as initiating vasopressors or de-escalating life support, without understanding its clinical basis.

The limitations of explainable AI (XAI)

The field of XAI has emerged to address this challenge, offering methods such as gradient-weighted class activation mapping (Grad-CAM),[36] Shapley additive explanations (SHAP),[37] and local interpretable model-agnostic explanations (LIME)[38] to provide post hoc feature attributions.[39] However, these methods have significant limitations in the clinical context. They can show what features the model weighed most heavily, but they cannot reveal the model's underlying causal logic or guarantee that the reasoning is medically sound.[19] For instance, an XAI technique might highlight a patient's heart rate as important for a sepsis prediction, but it cannot explain why or confirm that the model is not simply exploiting a spurious correlation (*e.g.*, associating tachycardia with sepsis while failing to recognize it as a compensatory response to hypovolemia from a different cause). This has led prominent researchers to argue that for high-stakes decisions, inherently interpretable models should be prioritized over attempting to explain opaque black-box models.[40]

Prioritizing interpretability does not necessarily mean a compromise in performance. While deep neural networks undeniably outperform traditional methods in high-dimensional perception tasks (*e.g.*, image analysis or raw waveform processing), recent systematic reviews suggest that for prediction tasks based on structured tabular data, complex black-box models rarely yield a clinically meaningful performance advantage over well-tuned interpretable models.[40,41] This phenomenon, often described as the Rashomon set, implies that for many clinical problems, a multitude of models exist with approximately equal accuracy, allowing us to select the one that is most intelligible to the human clinician. Furthermore, in the high-stakes context of the ICU, a black-box model that achieves marginally higher statistical accuracy by leveraging confounders (*e.g.*, treatment artifacts) actually possesses lower clinical utility and safety than a transparent model that aligns with physiological reasoning.

Erosion of professional autonomy

When a clinician is asked to act on an opaque recommendation, it undermines their professional autonomy and judgment. It reduces the clinician from a learned professional engaged in critical reasoning to a mere executor of algorithmic outputs. This dynamic erodes epistemic trust—the trust that the AI system's knowledge and reasoning processes are valid and reliable—which is a prerequisite for meaningful clinical collaboration.

The opacity of AI creates a dangerous accountability vacuum. If an AI-driven error leads to patient harm, the lines of responsibility become blurred. Is the intensivist who accepted the recommendation liable? The hospital that implemented the system? The company that developed the algorithm? This legal and ethical ambiguity is a major deterrent to adoption.[21] This ambiguity can create what has been termed "moral crumple zones", where the human operator, the clinician, is positioned to absorb the blame for a systemic failure, even if the root cause was a flawed algorithm.[42] This challenge is compounded by the predictable cognitive biases that emerge during human-algorithm interaction.

Automation bias

Automation bias is the tendency for humans to over-rely on automated systems, leading to errors of commission when they uncritically accept incorrect algorithmic recommendations.[43] For instance, a fatigued clinician might accept an AI's suggestion to withhold a fluid bolus, overlooking new signs like profound skin mottling that would otherwise prompt resuscitation.[44] Such seemingly authoritative AI recommendations can also induce confirmation bias, where clinicians selectively seek supporting information

while neglecting contradictory evidence. Over time, this dynamic normalizes reduced critical appraisal, increasing vulnerability to systematic errors in patient management, particularly in rapidly evolving clinical scenarios.[45]

Algorithm aversion

Algorithm aversion is the tendency to reject algorithmic advice, particularly after seeing it make an error, even if the algorithm is statistically superior to human judgment over time.[46] This can lead to errors of omission, where a clinician dismisses a valid early warning from AI, thereby losing the potential benefit of the technology. If an AI sepsis alert is triggered by a transient artifact in the heart rate data, the entire clinical team may subsequently ignore the next 10 valid alerts, viewing the system as unreliable. In team-based critical care, public recognition of an AI error can amplify skepticism across the group, fostering a culture of avoidance that persists beyond the initial incident.[47] Furthermore, in the high-pressure ICU environment, where every second is critical, clinicians may also avoid engaging with AI tools that require substantial additional data entry, navigation, or workflow changes, perceiving these as distractions from direct patient care.

Skill atrophy (deskilling)

A longer-term concern is the potential for overreliance on AI to lead to the erosion of fundamental clinical skills among trainees and junior clinicians. If diagnostic reasoning or electrocardiogram (ECG) interpretation is consistently offloaded to an algorithm, opportunities for deliberate practice and the development of deep, intuitive expertise may be lost. For instance, if an AI perfectly identifies patient-ventilator asynchrony from waveform data, future intensivists may lose the ability to recognize double-triggering or flow starvation at the bedside by observing the patient and the ventilator screen. This poses an insidious threat to the future of the medical profession.[48]

The patient perspective

Surveys and qualitative studies consistently show that while patients are optimistic about AI's potential to improve diagnostics, they harbor significant anxieties about its role in their care.[20,49] The collaborative bond between patient and clinician is a powerful determinant of health outcomes,[50] and an AI intermediary may weaken this critical relationship, replacing nuanced human interaction with a cold, impersonal process. Ultimately, healthcare is a profoundly human endeavor. The trust deficit is thus deeply rooted in anxieties about the potential for AI to disrupt the therapeutic relationship and the core identity of the medical professional.[51]

The clinician perspective

For clinicians, systematic reviews of their attitudes toward AI reveal a similar duality of optimism and apprehension.[52] While many recognize the potential for AI to improve efficiency and diagnostic accuracy, significant barriers to acceptance remain. Documented anxieties include concerns over the potential devaluation of their diagnostic expertise, shifts in professional roles and responsibilities, accuracy of outcomes, and unresolved legal and ethical liabilities. These factors contribute to a climate of professional apprehension that significantly hinders the enthusiastic adoption and successful integration of AI tools into clinical workflows.

Socioeconomic and systemic hurdles

Beyond the challenges of algorithmic integrity and user trust, the translation of AI into clinical practice is profoundly obstructed by formidable systemic barriers. These hurdles are foundational obstacles related to the very infrastructure, regulatory frameworks, and economic models that govern modern healthcare. They represent the pragmatic, real-world friction that can grind even the most promising technological innovations to a halt. Effective AI is predicated on the availability of vast quantities of high-quality, standardized, and accessible data. In most healthcare systems, this prerequisite is far from being met. The challenge lies in overcoming decades of accumulated technical debt from legacy systems and achieving true semantic interoperability.

Data silos and lack of standardization

Healthcare data are notoriously fragmented, residing in disparate, siloed systems (*e.g.*, EHRs, laboratory information systems, and picture archiving and communication system [PACS] for imaging) that were not designed to communicate with one another. Much of this infrastructure still relies on older standards like HL7v2, which lack the flexibility and semantic richness required for modern data science.[53] While newer standards such as Fast Healthcare Interoperability Resources (FHIR) offer a path forward by providing a standardized API for health data exchange, their adoption is resource-intensive and far from universal.[54] In critical care settings, continuous physiologic monitoring devices such as bedside patient monitors, ventilators, and infusion pumps often produce high-frequency waveform and telemetry data in vendor-specific proprietary formats.[55] These data streams are rarely integrated into the main hospital EHR in real time and lack widely adopted semantic standards equivalent to HL7 or FHIR for structured clinical data, making cross-device interoperability and downstream AI applications particularly challenging.[56]

The "unglamorous" work of data engineering

A significant hidden cost of AI implementation is the immense effort required for data engineering. Raw clinical data are inherently messy, characterized by missing values, inconsistent terminology, and measurement errors. The process of creating analysis-ready datasets through ETL (extract, transform, load) pipelines, data cleaning, harmonization, and annotation is a labor-intensive and highly specialized task that can consume up to 80% of the total time and resources in a data science project.[57] This foundational work is often underestimated in institutional planning, leading to project delays and budget overruns. The rapid, iterative pace of AI development is fundamentally asynchronous with the deliberate, slower pace of regulatory oversight, creating a landscape of uncertainty for developers and healthcare institutions alike.

The challenge of regulating adaptive AI

Regulatory bodies like the US FDA have established frameworks for software as a medical device (SaMD), but these were initially designed for static, locked algorithms.[22] Regulating adaptive AI—models that continuously learn and change their performance based on new data—poses a much greater challenge. The FDA has proposed a predetermined change control plan to address this, allowing manufacturers to prespecify planned modifications, but this remains a complex and evolving area.[58] The development of good machine learning practice principles aims to provide guidance, but clear, enforceable standards are still emerging.

Global regulatory fragmentation

The regulatory landscape is also fragmented globally. In Europe, the General Data Protection Regulation (GDPR) imposes strict constraints on the use of personal health data, creating significant

hurdles for data access and model training. More recently, the EU AI Act has introduced a risk-stratified approach, classifying medical AI devices as "high-risk" systems subject to stringent requirements for data quality, transparency, human oversight, and post-market surveillance. Navigating these different and sometimes conflicting international regulations adds significant complexity and cost for AI developers aiming for a global market.[59,60]

Perhaps the most pragmatic and powerful barrier is economic. The path from a validated algorithm to a sustainably implemented clinical tool is littered with financial obstacles, creating an economic "valley of death" that many promising technologies fail to cross.[61]

The reimbursement dilemma

An important challenge is the lack of clear and consistent reimbursement pathways. Payers, including government bodies like the Centers for Medicare & Medicaid Services (CMS) and private insurers, are often hesitant to establish new coverage policies for AI-powered procedures without extensive evidence of clinical utility and cost-effectiveness.[23] This creates a classic chicken-and-egg dilemma: Payers require robust evidence from large-scale trials for reimbursement, but generating this evidence often requires the very investment that is untenable without a clear reimbursement pathway. For an ICU-specific AI tool that optimizes sedation to reduce delirium and ventilator days, there is no specific current procedural terminology (CPT) code to bill for its use, rendering it financially unsustainable in fee-for-service models, despite its clear clinical value. One possible solution would be modifier codes for existing CPT codes to indicate the use of AI in an existing billing code.

Misaligned incentives and value proposition

The value proposition of an AI tool is often not aligned across stakeholders. For example, an AI that accurately predicts and helps prevent hospital readmissions provides immense value to an insurer by reducing payouts. However, for a hospital operating under a fee-for-service model, preventing readmission represents a loss of potential revenue. This misalignment of incentives can disincentivize a hospital from investing in and adopting the technology, even if it improves patient outcomes. While the shift toward value-based care models aims to correct this, such models are not yet universal, and the upfront costs of AI implementation remain a significant barrier for resource-constrained health systems.

Reflecting on the telos: The purpose of AI in critical care

The discourse surrounding AI in medicine has been overwhelmingly dominated by a technological imperative, a focus on what AI can do, measured by metrics of predictive accuracy, speed, and computational power. This focus, while important, has largely overshadowed a more fundamental and urgent inquiry: What should AI do to bridge the chasm between promise and practice? We must pivot from a purely technical conversation to a normative one, critically examining the telos, or ultimate purpose, of these powerful tools within the ethical framework of medicine. The successful integration of AI is not a matter of optimizing algorithms alone; it is a matter of aligning them with the core professional and humanistic goals of healing, alleviating suffering, and respecting the dignity of the patient.[62]

The risks of misaligned objectives

At its core, a machine learning model is an optimization engine. It is designed to maximize or minimize a predefined objective function, a mathematical expression of a desired goal.[63] The peril lies in the fact that the goals of a healthcare system are often complex, multifactorial, and not easily quantified. In the increasingly commercialized landscape of modern medicine, there is immense pressure to define these goals in economic terms; for instance, many organizations view AI primarily as a means to optimize revenue cycles and control costs.[64] When the objective function of an AI system is aligned with financial metrics rather than with the patient's best interest, a dangerous misalignment occurs.

This is not a theoretical concern. Consider an AI model deployed in an ICU with the objective function of minimizing hospital length of stay, a common metric of operational efficiency. Such a model might learn to identify patients for early discharge. While this could be appropriate in many cases, it could also learn to flag patients for discharge prematurely, subtly encouraging decisions that increase the risk of post-discharge complications or readmission, simply because that outcome falls outside its optimization window.[65] Worse, it might learn that patients who transition to comfort care have the shortest length of stay, and could begin to subtly flag patients with certain characteristics as having a high probability of "short stay", potentially biasing a goals-of-care discussion before it even begins. The algorithm, in its amoral pursuit of the defined objective, would be functioning perfectly, yet it would be failing the patient. A normatively aligned approach would define the objective function in terms of patient-centered clinical outcomes.

Instead of minimizing length of stay, a more appropriate objective function might be to maximize ventilator-free days, minimize the incidence of delirium, or reduce the long-term risk of post-ICU cognitive impairment.[66] Carefully selected focused clinical outcomes are essential to ensuring AI models minimize biased outcomes and do not overstep the holistic clinical judgment of intensivists. This requires a conscious and deliberate choice to prioritize the physician's fiduciary duty to the patient over institutional financial incentives. Without this explicit normative alignment at the design stage, AI risks becoming a powerful tool for optimizing the business of medicine at the expense of the practice of medicine.

The fallacy of the technological fix: AI and system-level pathologies

A second critical error in the current discourse is the framing of AI as a panacea for deeply entrenched systemic problems—a phenomenon known as technosolutionism.[67] The belief that a sophisticated algorithm can solve problems that are fundamentally human, organizational, or structural is a fallacy. Layering AI onto a dysfunctional system will, at best, fail to produce meaningful change and, at worst, amplify the underlying pathology.[68]

The recent history of the EHR serves as a powerful and cautionary analog. The EHR was promoted as a solution to problems of fragmentation and medical error, yet its implementation in the absence of corresponding workflow redesign and attention to human factors created a new epidemic of clinician burnout and documentation burden.[7] It did not fix the underlying communication challenges; in many cases, it exacerbated them by replacing face-to-face conversation with a fragmented and cumbersome digital interface.

Similarly, consider an AI-powered early warning system for sepsis. The algorithm may perfectly identify patients at high risk hours before clinical signs become apparent. However, if the hospital suffers from chronic nursing shortages, a dysfunctional pharmacy workflow that delays antibiotic delivery, or a culture of poor interprofessional communication, the alert is clinically useless. The system correctly identifies the problem but cannot trigger an

Table 1. Comparison of design philosophies: current standard AI development and the proposed collaborative framework

Dimension	Current standard (autonomous focus)	Proposed model (collaborative focus)
Core philosophy	Replacement: AI functions as an oracle intended to automate decision-making	Augmentation: AI functions as a cognitive prosthetic intended to extend human foresight
Primary task	Prediction: Binary classification of outcomes (*e.g.*, sepsis/no sepsis)	Synthesis: Pattern recognition, data abstraction, and counterfactual forecasting
Interaction	Unidirectional: The system pushes alerts to the clinician (often causing alarm fatigue)	Bidirectional: The clinician queries the system; the system provides rationale
Uncertainty	Opaque: Outputs raw probability scores often without context	Transparent: Outputs confidence intervals and feature attributions
Success metric	Statistical accuracy: AUROC/sensitivity on retrospective datasets	Clinical utility: Reduction in cognitive load and improved patient outcomes

AI, artificial intelligence; AUROC, area under the receiver operating characteristic curve.

effective response. The failure is not algorithmic; it is systemic. Before implementing AI, healthcare organizations must first address foundational challenges in data quality, organizational readiness, and clinical workflows.[69] To do otherwise is to invest in a sophisticated solution to the wrong problem, mistaking a symptom for the disease.

Offloading cognitive burden to amplify expertise

The final and most important normative turn involves redefining the relationship between AI and the clinician, moving away from a narrative of replacement and toward a pragmatic framework of cognitive division of labor.[61] This framework is built on the recognition that humans and AI possess distinct and complementary cognitive strengths. As a computational system, AI excels at tasks characterized by high volume, high velocity, and the detection of complex, high-dimensional patterns within structured data. The human clinician, in contrast, excels at tasks requiring integrative synthesis, causal reasoning, the navigation of profound uncertainty, and affective intelligence.[70] The goal should not be to automate the clinician but to strategically liberate them from tasks that are cognitively burdensome. This fundamental shift in design philosophy, from replacement to augmentation, is detailed in Table 1.

We call this strategic separation cognitive unbundling. In the current standard of care, the clinician is burdened with a monolithic set of tasks ranging from data surveillance to complex ethical judgment. Our framework explicitly separates these functions. It delegates high-velocity, high-dimensional data processing tasks, such as continuous waveform monitoring and information retrieval, to computational agents. This unbundling is distinct from generic "augmentation", which often simply adds new alerts to an already overwhelmed workflow. Instead, it aims to subtract the computational load, liberating the clinician's finite cognitive resources for the tasks that remain irreducible to algorithms: semantic reasoning, causal synthesis, and compassionate communication.[71]

The practice of modern critical care forces clinicians to spend an inordinate amount of time and finite cognitive resources on functions that are computationally intensive and ill-suited to the human mind. This includes administrative and computational duties that consume a disproportionate share of clinical time,[72] such as the rote generation of daily progress notes from structured data, the meticulous but formulaic calculation of clinical severity scores such as sequential organ failure assessment (SOFA) or APACHE, the reconciliation of medication lists upon admission or transfer, and triaging noncritical laboratory results and system alerts to find the one crucial signal hidden within the noise. These functions are major contributors to the high intrinsic cognitive load of critical care, which has been shown to deplete finite mental resources like working memory and executive function, thereby increasing the risk of error and burnout.[73,74]

This strategic offloading is a mechanism for cognitive restoration and the amplification of unique human expertise. By delegating the burdensome tasks of data surveillance, documentation, and routine calculation to AI, the clinician's cognitive capacity is freed to be deployed where it is most needed and most effective: domains that require higher-order reasoning. This includes the complex synthesis of disparate and often conflicting information, integrating a subtle physical exam finding, an ambiguous imaging report, a critical lab value, and the patient's own narrative into a coherent diagnostic hypothesis. This is a form of abductive and causal reasoning that seeks to understand the "why" behind the data, a function distinct from the correlational pattern recognition performed by AI.[70] Furthermore, the clinician's role is defined by the ability to navigate profound uncertainty, making high-stakes judgments with incomplete information and weighing competing values in situations where there is no single "correct" answer.

Perhaps most critically, this cognitive liberation allows clinicians to rededicate themselves to the essential affective dimensions of care. This includes engaging in nuanced and empathetic goals-of-care discussions, translating complex medical information into understandable terms for distressed families, and providing compassionate support that forms the bedrock of the therapeutic alliance, a factor with a demonstrable impact on patient outcomes.[50] This strategic division of labor reframes AI as a powerful cognitive tool. It augments the clinician's capacity by managing the data, which allows them to dedicate their irreplaceable expertise to the uniquely human domains of wisdom, judgment, and compassion. This model does not diminish the role of the clinician; it elevates it, enabling them to practice medicine at the peak of their capabilities.

A framework for the critical care AI collaborator

Moving beyond critique, we now envision how AI can be practically and ethically integrated into the ICU workflow. Building on the collaborative model outlined in Section 3.3, this framework strategically delegates specific computational tasks to AI agents, thereby liberating the clinical team to focus on the irreplaceable human domains of synthesis, judgment, and empathy in critical care.

This principle can be illustrated through four distinct roles for a critical care AI collaborator, as shown in Figure 2. The four roles

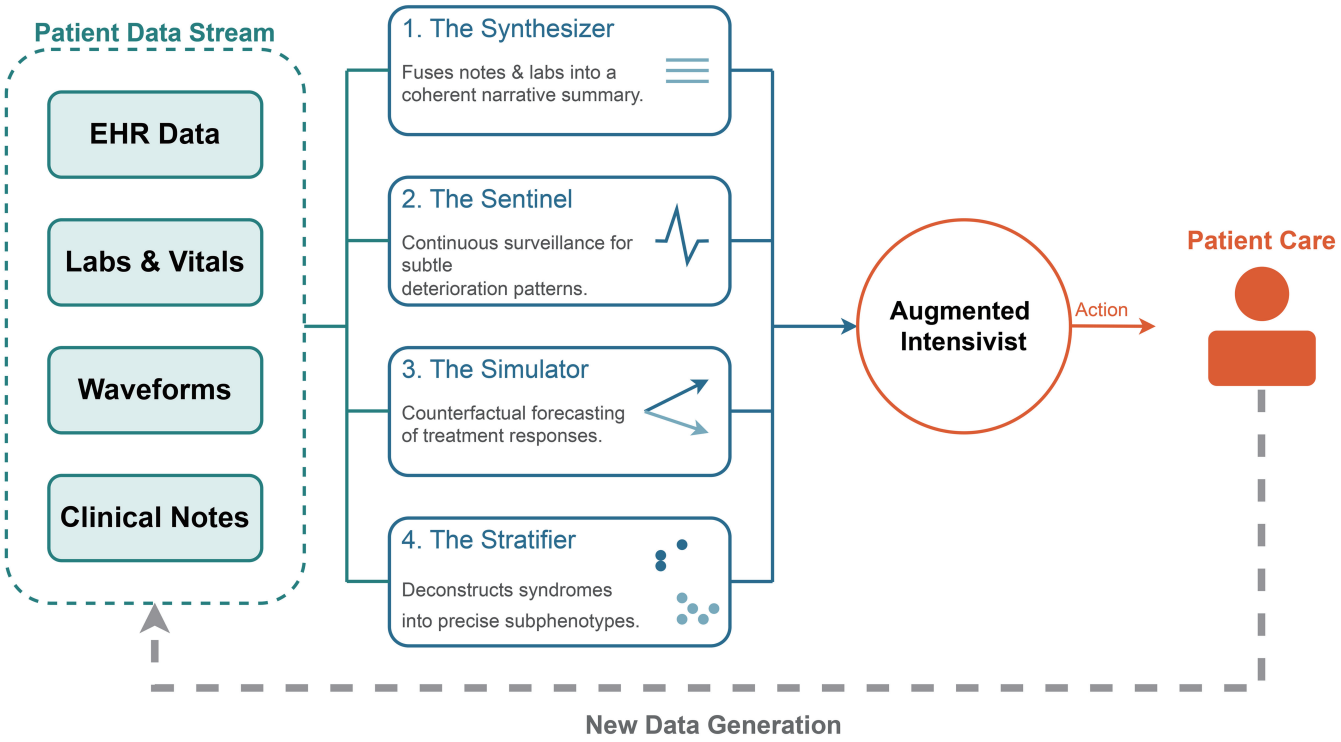

Fig. 2. The collaborative AI framework. A schematic overview of the four proposed AI roles. High-frequency physiological data and unstructured records are processed by specialized agents: the synthesizer (narrative generation), sentinel (surveillance), simulator (forecasting), and stratifier (phenotyping). These outputs converge to augment the intensivist's decision-making, rather than replacing it. AI, artificial intelligence; EHR, electronic health record.

described—the synthesizer, sentinel, simulator, and stratifier—are intended as non-exhaustive examples. They exemplify a design philosophy centered on augmenting specific cognitive processes unique to the ICU and serve as concrete illustrations of how AI can be purposefully engineered to address the challenges of cognitive overload and physiological data chaos, collectively forming a system that enhances situational awareness, proactivity, and personalization at the bedside.

The synthesizer: from data chaos to coherent patient narrative

The first possible role of the AI collaborator is that of the synthesizer, directly confronting the problem of data fragmentation and the cognitive burden of information retrieval that define the start of every ICU shift.[7] An intensivist's initial task is to develop situational awareness, a coherent mental model of the patient's current state and trajectory. This is currently a manual, time-consuming process of navigating separate data sources.[14] The synthesizer automates this by abstracting temporal data and fusing information from multiple sources.

Leveraging natural language processing (NLP), the synthesizer can parse unstructured text from radiology reports, consultant notes, and nursing documentation, extracting key concepts like "new right ventricular strain" from an echocardiogram report or "agitated, pulling at lines" from a nursing note.[75] Recent empirical work validates the feasibility of this role; for instance, adapted large language models have been shown to generate clinical summaries that are rated by physician evaluators as superior to those written by human experts in terms of completeness and correctness.[76] It then integrates these concepts with structured, high-frequency data streams unique to the ICU: second-by-second arterial line waveforms, breath-by-breath ventilator graphics, and minute-by-minute lab results. The output is a concise, dynamically updated clinical narrative. For example, instead of forcing a clinician to manually trend a lactate, review nursing notes, check vasopressor doses, and analyze ventilator mechanics, the synthesizer could generate a summary, as follows.

Over the past 6 h, despite stable norepinephrine requirements (0.1 mcg/kg/min), the patient has developed a new metabolic acidosis (lactate rising from 2.1 to 4.5 mmol/L) and oliguria. This is concurrent with a new documentation of "mottled extremities" in the nursing notes and a 25% increase in ventilator driving pressure.

This synthesized narrative does not provide a diagnosis; it presents salient evidence from multiple disparate domains (hemodynamics, metabolism, perfusion, and respiratory mechanics) in a clinically relevant context. By transforming the data into a coherent story, the synthesizer accelerates the intensivist's ability to grasp the clinical picture, allowing them to immediately ask the higher-order question, "Why is this happening" and formulate a differential diagnosis for this new shock state.

The sentinel: proactive surveillance beyond simple alarms

Building on the synthesized data stream, the sentinel addresses the profound failure of current alarm systems and the resulting epidemic of alarm fatigue.[6] Traditional ICU alarms are based on simple, static, univariate thresholds (*e.g.*, heart rate > 120 bpm) that generate a high volume of false positives and fail to detect subtle, complex patterns of deterioration.[77] The sentinel functions as a sophisticated early warning system, employing multivariate time-series analysis to provide proactive, personalized surveillance.

Instead of relying on population-level thresholds, AI can learn

each patient's unique physiological baseline and model their individual trajectory. It continuously analyzes dozens of variables simultaneously, integrating subtle shifts in heart rate variability, arterial waveform morphology (*e.g.*, dicrotic notch position and pulse pressure variation), and respiratory rate patterns to find multidimensional signatures of impending decompensation. It could, for instance, detect the early, subtle signs of patient-ventilator asynchrony by analyzing pressure and flow waveforms, flagging a patient for a sedation adjustment or ventilator mode change long before they develop overt respiratory distress or barotrauma. For syndromes like sepsis or AKI, these models can identify at-risk patients hours before they meet conventional diagnostic criteria.[78] Similarly, models have been developed to predict neurological outcomes in comatose patients in the ICU by analyzing complex patterns in brain connectivity data from electroencephalogram (EEG) recordings, offering a prognostic window that extends beyond traditional clinical signs.[79] However, the transition from research to bedside is fraught with challenges; external validation studies of widely implemented proprietary sepsis models have revealed significant performance degradation when they are deployed across diverse hospital systems.[80]

This underscores that a true sentinel cannot simply be a black-box alarm; it must be explainable, providing the interpretable "why" and a clear rationale behind the risk to allow for clinician validation. For instance: Alert—high probability (85%) of hemodynamic decompensation within 2 h. Key drivers—30% decrease in heart rate variability over the last hour, progressive decrease in pulse pressure variation despite no fluid administration, and a 15% increase in norepinephrine requirement to maintain MAP goal.

This probabilistic and interpretable warning serves as a cognitive prompt, not a command. It directs the intensivist's limited attention to the patient at the highest risk and provides a starting point for investigation. The clinician's role is to validate the alert at the bedside, integrate it with their physical examination and holistic assessment, and ultimately decide on the appropriate course of action. The sentinel transforms ICU monitoring from a reactive, threshold-based process into a proactive, pattern-based partnership.

The simulator: In silico trials for personalized intervention

The most forward-looking role is that of the simulator, which allows intensivists to move beyond population-based protocols and toward truly personalized therapeutic decisions. Many core ICU interventions, such as titrating mechanical ventilation or vasopressors, are performed with significant uncertainty regarding an individual patient's response. The simulator addresses this by enabling *in silico* counterfactual reasoning, allowing clinicians to conduct virtual "what-if" scenarios before intervening on the actual patient.[81]

Using patient-specific "digital twins", hybrid models integrating mechanistic physiological principles with the patient's own data, the clinician could query the simulator as follows for a mechanical ventilation use case.

Forecast the probabilistic distribution of this patient's PaO_2/FiO_2 ratio, driving pressure, and cardiac index over the next 2 h if positive end-expiratory pressure (PEEP) is increased from 10 to 14 cmH_2O.

The simulator would return a personalized forecast, complete with confidence intervals, quantifying the likely trade-off between improved oxygenation and potential hemodynamic compromise.

For the hemodynamic management use case, using techniques like reinforcement learning, an AI agent could learn a patient's unique vasopressor sensitivity. Such models have already been proposed to identify optimal personalized strategies for administering vasopressors and intravenous fluids in sepsis, moving beyond static protocols to policies learned directly from patient data.[82–84] The clinician could ask: "What is the optimal MAP target for this specific patient to maximize lactate clearance while minimizing the total norepinephrine dose over the next 4 h".

AI does not choose the optimal setting; the clinician remains responsible for the final decision. It provides a personalized risk-benefit analysis that augments the clinician's judgment. The intensivist integrates this forecast with their knowledge of the patient's specific pathology (*e.g.*, focal *vs.* diffuse ARDS, underlying right ventricular dysfunction) and makes the final responsible decision. This transforms therapeutic decision-making from an act of recall based on population averages to a data-driven exploration of personalized futures.

The stratifier: Deconstructing syndromes for precision critical care

Finally, the stratifier addresses one of the greatest challenges in critical care: the profound heterogeneity of clinical syndromes. We currently treat complex conditions such as sepsis and ARDS with broad, protocol-based therapies, despite overwhelming evidence that they are not single diseases but umbrella terms for multiple, distinct underlying pathophysiologies.[85] The stratifier employs unsupervised machine learning to perform computational phenotyping, identifying clinically meaningful patient subgroups that are invisible to the human eye.

Using techniques such as clustering on high-dimensional data (clinical variables, biomarkers, and genomics), the stratifier can deconstruct a heterogeneous syndrome into more homogeneous subphenotypes. Previous research has already demonstrated the power of this approach, identifying distinct sepsis phenotypes (*e.g.*, α, β, γ, δ) with different host-response patterns and mortality rates,[86] as well as ARDS subphenotypes (*e.g.*, hyperinflammatory *vs.* hypoinflammatory) that respond differently to therapies like PEEP or simvastatin.[87]

The role of the stratifier is to provide the clinician with a new, deeper layer of diagnostic information. By identifying a patient as belonging to the "hyperinflammatory" ARDS subphenotype, for example, the AI collaborator provides a powerful rationale for considering enrollment in a clinical trial of an anti-inflammatory agent or for choosing a specific ventilation strategy. It does not dictate therapy but provides the crucial stratification needed to move critical care from a one-size-fits-all paradigm toward the long-sought goal of precision medicine.

These four roles are not an exhaustive catalog but illustrative examples of this collaborative framework. They represent a way of thinking about AI development that prioritizes augmenting specific, high-burden cognitive tasks. The unifying principle is the targeted application of AI to offload computational burdens, thereby amplifying the diagnostic, therapeutic, and humanistic capabilities of the intensivist and the entire ICU team.

A strategic implementation framework: From code to bedside

The vision of an AI collaborator in critical care, while conceptually compelling, faces a challenging translational reality. The gap between a validated algorithm and a sustainably implemented clinical tool is magnified by resource constraints, fragmented data infrastructure, and misaligned stakeholder incentives.[88] To bridge this gap, we propose a strategic framework for implementation. This section outlines a phased, resource-conscious pathway that

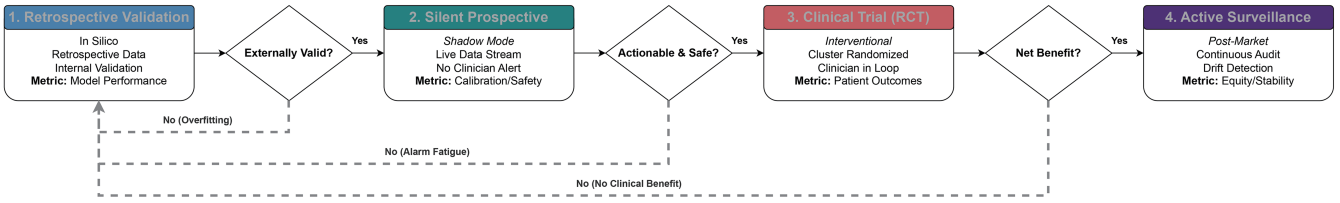

Fig. 3. The translational pathway from code to bedside. A staged implementation framework. The process moves from retrospective analysis (Stage 1) to "silent" shadow-mode validation (Stage 2) to ensure safety before any patient interaction. Clinical utility is established via randomized controlled trials (Stage 3), followed by continuous post-market algorithmic auditing (Stage 4) to detect drift. Feedback loops (dashed lines) indicate where models must be sent back for retraining if they fail a safety gate. RCT, randomized clinical trial.

allows institutions to begin with high-impact interventions while building toward a sustainable ecosystem (Fig. 3). We draw upon principles of implementation science to propose concrete strategies for overcoming the coordination problems inherent in modern healthcare.

The foundational imperative: A robust and equitable data infrastructure

The performance and equity of any clinical AI system are fundamentally constrained by the quality and accessibility of the underlying data.[89] Building a trustworthy AI collaborator is therefore predicated on a foundational institutional commitment to creating a robust data infrastructure. This is not merely a technical prerequisite but a strategic imperative for any modern healthcare organization.

Adherence to FAIR principles

The guiding philosophy for this infrastructure must be the FAIR data principles, ensuring that data are findable, accessible, interoperable, and reusable.[90] This requires moving beyond siloed, proprietary data systems toward a more open, standardized architecture. The adoption of modern interoperability standards like HL7 FHIR is critical, as this provides a standardized API for exchanging health data. For the ICU, this is necessary for integrating not just EHR and lab data but also the high-frequency, high-value data streams from bedside devices—ventilator waveforms, infusion pump rates, and continuous cardiac output monitors—which are often trapped in proprietary, inaccessible formats.[54]

Overcoming technical debt

Most health systems are burdened by significant technical debt. The "unglamorous" but essential work of data engineering, including the development of robust ETL pipelines, data cleaning, and harmonization of parameters from different generations of ventilators or monitors, represents a massive upfront investment. Furthermore, establishing clear data provenance is essential for model reproducibility and auditing, ensuring that every data point used for training or inference can be traced back to its source. Without this meticulous groundwork, any AI system built on top will inherit the flaws of the data, rendering it unreliable and potentially unsafe.

Human-centered design and cognitive integration

The history of health IT is littered with examples of technologically sound systems that failed due to poor design and a lack of integration into clinical workflows. The clinician burnout epidemic, exacerbated by cumbersome EHRs, serves as a stark reminder that any new tool must be designed with a deep understanding of cognitive ergonomics, the science of fitting a system to the user's cognitive capabilities and limitations.[91]

Co-design and participatory development

To avoid repeating past mistakes, AI tools must be developed through a process of co-design, involving intensivists, nurses, respiratory therapists, and other end users from the earliest stages of development. This participatory approach ensures that the tool is designed to solve real clinical problems and that its outputs are intuitive, nondisruptive, and actionable within the high-pressure context of the ICU. The goal must be to reduce, not increase, the clinician's cognitive load.

Designing for actionable intelligence

The presentation of AI-generated insights is as important as their accuracy. Raw probability scores or complex feature attribution maps are of little use at the bedside; the user interface must translate algorithmic outputs into actionable intelligence. For example, an AI alert should present not just a risk but also its key contributing factors in clear, clinical language and a concise narrative, rather than a data table. This human-computer interaction focus is critical for effectively translating AI's computational power into improved human decision-making.

Furthermore, successful coordination requires preserving the clinician's sense of agency. Subjective resistance often arises when AI is perceived as an opaque authority rather than a supportive tool. To mitigate this, systems must be designed for bidirectional interaction. Rather than passively receiving alerts, clinicians should have the ability to provide real-time feedback, flagging false positives or correcting contextual errors (*e.g.*, "patient is on palliative care"). This "human-in-the-loop" feedback not only facilitates local model calibration but psychologically reinforces the clinician's status as the ultimate decision-maker, transforming the interaction from compliance to collaboration.[92]

Finally, reducing obstacles to coordination requires strict adherence to cognitive ergonomics. The AI's output must align with the clinician's existing mental models of disease. An alert that contradicts a clinician's intuition without explanation creates cognitive dissonance and rejection. Therefore, the interface must present the "why" alongside the "what" grouping data into physiologically coherent clusters (*e.g.*, preload, afterload, contractility) rather than arbitrary lists, ensuring the AI output speaks the language of the clinicians.

A hierarchy of validation: Building trust through rigorous evidence

Trust is the currency of medicine, and for an AI tool, it must be earned through a transparent and rigorous hierarchy of clinical validation. Moving beyond simplistic, retrospective metrics of accuracy is necessary to prove that an AI collaborator is not just sta-

tistically sound but clinically effective and safe.[93]

From retrospective analysis to prospective evaluation

The validation pathway should be a staged process. It begins with retrospective validation on large, diverse datasets to establish initial performance. The next step is prospective, "silent" validation, in which the model runs in the background on live ICU data, allowing researchers to assess its real-world performance without influencing clinical care. This allows for crucial safety checks, such as determining the false alert rate of a sepsis model in real time or measuring how often an extubation failure predictor would have been correct. This provides a critical safety check and a more realistic measure of its accuracy under current clinical conditions.[16]

The imperative of randomized controlled trials

The ultimate arbiter of clinical utility is the randomized controlled trial. For system-level interventions like AI decision support, cluster-randomized trials, where ICUs are randomized, are often the most appropriate design.[94] These trials must be designed to measure what truly matters: patient-centered clinical outcomes (*e.g.*, mortality, ventilator-free days, delirium-free days, and length of stay), process-of-care metrics (*e.g.*, time to antibiotics for sepsis), and clinician-centered outcomes (*e.g.*, measures of burnout, diagnostic confidence, and time spent on documentation). Only through such rigorous prospective evidence can we confidently conclude that an AI tool provides a net benefit to patients and the health system.

Ethical and regulatory guardrails: Ensuring accountable governance

The deployment of powerful AI systems into the high-stakes environment of the ICU necessitates a robust governance structure to ensure they are used safely, equitably, and responsibly. This requires a proactive approach to ethics and regulation that extends throughout the AI lifecycle.

Algorithmic auditing and post-market surveillance

AI models are not static entities; their performance can degrade over time due to concept drift.[93] Therefore, a one-time validation is insufficient. Regulatory institutions must establish a process for continuous algorithmic auditing and post-market surveillance to monitor model performance, detect drift, and identify any emergent biases or failures on a national scale.[8] The COVID-19 pandemic provided a dramatic example of concept drift, where ARDS prediction models trained on pre-2020 data performed poorly on a novel disease with a different pathophysiology.[95] This requires close collaboration between data scientists, clinicians, institutional review boards, and federal-level regulation, which must develop the necessary expertise to oversee AI-based research and implementation. National-level standards are crucial for minimizing disparities in implementation and outcomes. Larger urban health systems are likely to have far more resources for oversight than rural systems. As such, the burden and responsibility for oversight must fall to federal regulation and monitoring.

Navigating the regulatory landscape

Adherence to evolving regulatory frameworks is mandatory. This includes complying with data privacy regulations like GDPR and Health Insurance Portability and Accountability Act (HIPAA laid out by bodies like the FDA).[22] As these frameworks evolve to address adaptive AI, continuous engagement with regulatory science will be necessary for any institution seeking to deploy these technologies responsibly.

Democratizing oversight: Infrastructure for equity and sustainability

Effective governance must extend beyond safety to address the economic and coordination challenges threatening equitable AI implementation. Regulatory frameworks should actively prevent a "digital divide" in critical care, ensuring that advanced AI systems are accessible beyond well-resourced academic centers. We advocate for regulatory incentives, such as expedited FDA review for AI tools demonstrating equitable performance across diverse populations, and public reporting requirements to ensure algorithmic transparency and reward equity prioritization. Furthermore, national agencies should coordinate to establish a centralized AI infrastructure, including algorithmic audit systems and open-access validation datasets. This approach distributes oversight costs across stakeholders, transforming regulation from a barrier to adoption into an enabler of safe, equitable, and sustainable implementation.

A tiered implementation model: From single institutions to national systems

A "big bang" approach to AI adoption, attempting to overhaul infrastructure and deploy complex autonomous agents simultaneously, is unlikely to succeed due to the immense resource investment required. Instead, we propose a three-tier implementation model that provides pragmatic entry points for institutions at varying levels of digital maturity.

Tier 1: Institutional-level implementation—building trust with existing data

The most immediate opportunities lie in leveraging data that are already digitized but underutilized. At this tier, institutions focus on "readily achievable targets"—tools that require minimal integration overhead and deliver immediate benefits. For example, a synthesizer application using NLP to summarize nursing notes and lab trends requires read-only access to the EHR and does not demand real-time streaming telemetry. Such tools can be deployed with minimal local investment, yet they serve a critical strategic function: they demonstrate immediate return on investment by reducing documentation time, thereby building the clinical trust and institutional political capital necessary for more ambitious projects.

Tier 2: Regional consortia and federated learning

To overcome the limitations of single-center datasets and the high cost of validation, institutions must transition from isolated silos to collaborative networks. The primary barrier here is data privacy and the logistical challenge of sharing patient records. The solution lies in federated learning (FL). FL allows AI models to be trained across multiple institutions without patient data leaving the local firewall; only the model characteristics (gradients/weights) are shared.[96] This level envisions regional consortia where academic medical centers and community hospitals collaborate. A community hospital, which may lack the resources to build a sentinel model from scratch, contributes to the federation and, in return, gains access to a model trained on a diverse, generalized population. This approach democratizes access to high-quality AI while solving the "small data" problem of individual ICUs.

Tier 3: National policy and infrastructure

The ultimate realization of the AI collaborator requires systemic support that individual hospital systems cannot generate. This tier

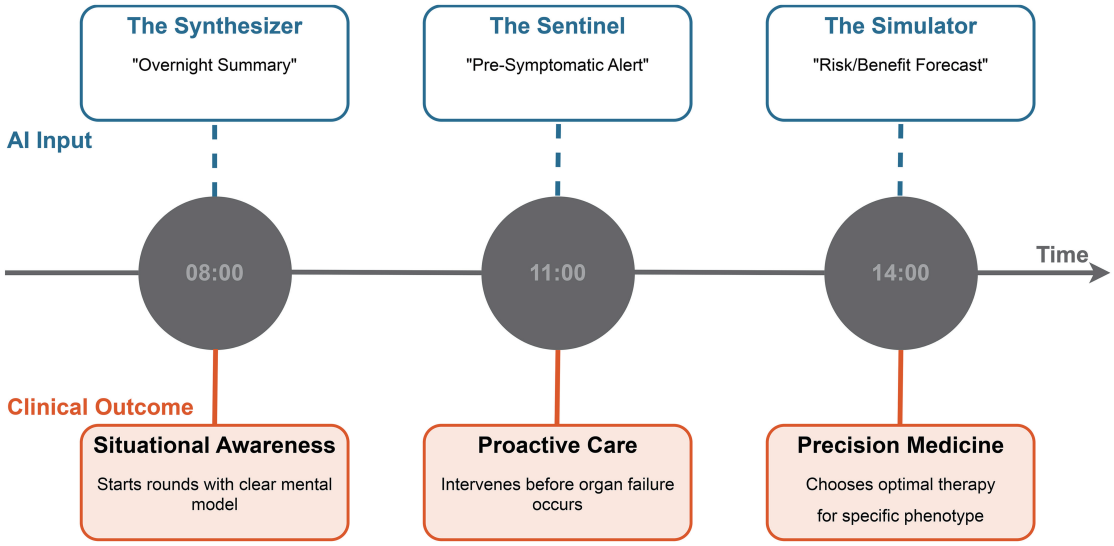

Fig. 4. The augmented ICU workflow. A timeline illustrating the integration of AI tools into daily clinical routine. At 08:00, the synthesizer reduces cognitive load during rounds. At 11:00, the sentinel enables proactive intervention before overt decompensation. At 14:00, the simulator supports precision decision-making for complex interventions. AI, artificial intelligence; ICU, intensive care unit.

focuses on the "public goods" of AI infrastructure. This includes the creation of national validation datasets (analogous to the MIMIC-IV database but on a national scale) and centralized algorithmic auditing bodies.[97] Similarly, as the FDA's Sentinel Initiative monitors drug safety using distributed data,[98] a national AI surveillance infrastructure could monitor algorithmic drift and safety signals across the health system, distributing the cost of oversight that is currently too heavy for any single institution to bear.

Overcoming coordination barriers

Even with a phased technical model, implementation will stall without addressing the fundamental socioeconomic friction: the difficulty of coordinating interests across fragmented systems and the lack of sustainable business models. The gap between academic validation and commercial deployment is often unbridgeable without external capital and deliberate institutional design.

The learning health system consortium

We must abandon the model of competitive isolation in favor of the learning health system.[99] Successful precedents such as the Observational Health Data Sciences and Informatics (OHDSI) collaborative and PCORnet demonstrate that large-scale interoperability is possible when governance is shared.[100] We propose the formation of Critical Care AI Consortia that share not just data standards (such as Observational Medical Outcomes Partnership [OMOP] or FHIR) but implementation blueprints. By open-sourcing the "connective tissue" code, the software that pipes data from a ventilator to a model and back to the clinician, consortia can reduce the redundant engineering costs that currently affect every new AI project.

Public-private partnerships to bridge the translational gap

The gap between academic validation and commercial deployment is often unbridgeable without external capital. Public-private partnerships offer a mechanism to de-risk this transition. Government initiatives, such as the National Institutes of Health (NIH)'s Bridge2AI program in the US,[101] can support the high-risk phase of multi-site validation. In return, private industry partners (EHR vendors or device manufacturers) can handle the "last mile" of integration and user interface design. This symbiosis ensures that tools are clinically validated (academia), scalable (industry), and equitable (government oversight).

Realigning financial incentives

Finally, we must address the economic reality: In many fee-for-service systems, efficiency is penalized. An AI tool that prevents readmissions or reduces length of stay may paradoxically reduce hospital revenue. Sustainable adoption requires aligning AI with value-based care payment models. Policymakers and payers should consider specific reimbursement mechanisms for AI-augmented care. In the US context, this might look like CPT modifier codes for AI-assisted decision-making; in the UK (or in other Beveridge systems), this could involve enhanced tariffs for ICUs meeting digital maturity standards. When the financial viability of hospitals depends on patient outcomes rather than service volume, the AI sentinel and simulator can transform them from cost centers into essential assets for financial survival.

The vision: Augmented intensivists and the future of the ICU

The prospect of integrating AI into the ICU often brings to mind a sterile, automated environment where clinicians are reduced to data entry clerks for an algorithmic oracle. This fear, while understandable, is not a foregone conclusion. It is a potential future born from a failure of purpose and design. The potential framework we have proposed charts a more practical course. The goal is not a futuristic, automated ICU but a more functional one, an environment where technology is purposefully designed to minimize cognitive load, allowing the entire clinical team to practice at the top of their license.

To envision this future, as illustrated in Figure 4, one needs only imagine a better, more efficient workday for the ICU team. The day no longer begins with a 30 min struggle against the EHR, piecing together a patient's story from a dozen different screens. Instead, the augmented intensivist starts rounds by reviewing a one-page

Table 2. Clinical vignettes: standard care *vs*. AI-augmented workflows

AI role	Clinical scenario	Current standard of care	AI-augmented workflow
The synthesizer	Overnight admission of a septic patient	Fragmented: Clinician toggles between nursing notes, vitals trends, and lab tabs. Risk of missing the link between rising lactate and subtle agitation	Coherent: AI generates a narrative: "Lactate rose 2.1→4.5 mmol/L concurrent with new agitation and 20% increase in vasopressors". Clinician starts rounds with immediate situational awareness
The sentinel	Post-op cardiac surgery (occult bleeding)	Reactive: Alarms trigger only when MAP < 65 mmHg. By then, the patient is already hypotensive and requires reactive fluid resuscitation	Proactive: AI detects "multivariate signature" of volume loss (decreasing pulse pressure variation + rising HR) 60 min before hypotension. Alert: "85% risk of hypovolemia". Team intervenes early
The simulator	ARDS: Titrating PEEP	Trial and error: Clinician increases PEEP based on generic protocol. Must wait 1–2 h to see if oxygenation improves or if hemodynamic collapse occurs	Counterfactual forecasting: Clinician queries AI, "Forecast driving pressure if PEEP increased to 14". AI predicts, "Driving pressure will likely increase to 18 cmH_2O (high risk)". Clinician chooses prone positioning instead
The stratifier	Sepsis heterogeneity	One-size-fits-all: All septic patients receive the same 30 cc/kg fluid bolus and broad-spectrum antibiotics, regardless of underlying biology	Precision phenotyping: AI classifies patient into "hyperinflammatory" *vs.* "immunosuppressed" subphenotype. Clinician tailors therapy (*e.g.*, steroids for the former, avoiding them for the latter)

AI, artificial intelligence; ARDS, acute respiratory distress syndrome; HR, heart rate; MAP, mean arterial pressure; PEEP, positive end-expiratory pressure.

summary for each patient, automatically generated overnight by the synthesizer. This summary flags the key events, trends the net fluid balance against vasopressor requirements, and synthesizes the overnight nursing concerns. This process takes five minutes, not thirty, allowing the team to walk into the patient's room with a clear head and a solid grasp of the clinical situation.

Later that morning, an alert from the sentinel flags a patient at high risk of developing AKI. The alert does not simply sound an alarm; it provides a clear, concise reason: "Risk elevated due to 24 h of piperacillin-tazobactam exposure combined with a subtle but persistent drop in mean arterial pressure". This prompt leads the team to proactively review the patient's medications and fluid status, potentially averting a serious complication hours before the creatinine would have started to rise.

Around midday, the team discusses weaning a patient from mechanical ventilation. Instead of relying solely on the rapid shallow breathing index, they consult an AI prediction model (the simulator) that has integrated dozens of variables: the patient's cough strength (analyzed *via* a bedside microphone), recent delirium scores, and the minute-to-minute variability of their respiratory rate. The model provides a nuanced prediction: "85% probability of successful extubation, but a 40% probability of post-extubation dysphagia based on age and duration of intubation". This prompts an early swallow evaluation, turning a simple wean/fail decision into a more holistic plan for recovery.

When faced with a fragile ARDS patient, the team does not have to rely solely on generalized protocols. Before increasing the PEEP, the intensivist asks the simulator, "What is the likely impact of this change on driving pressure and right ventricular strain over the next hour" AI provides a risk-benefit forecast based on that specific patient's data, helping the team make a more informed, personalized decision. AI does not make the choice; it provides personalized information to support the team's clinical judgment.

The true impact of this vision relates to what this cognitive support allows the human clinician to do better. By automating tedious, repetitive, and mentally draining tasks, such as data searching, the manual calculations, and the triage of low-value alerts, the AI collaborator frees up the intensivist's most valuable resource: their mental energy. This allows them to reinvest their focus in the work that truly requires human expertise.

AI can present the data, but it takes a human mind to synthesize them into a coherent diagnosis. This is the work of connecting the subtle physical exam findings with the ambiguous lab result and the family's description of the patient's symptoms to determine the "why" behind the numbers.

The ICU is defined by uncertainty. Deciding how to proceed when the data are incomplete or some treatment options have significant risks, or when to recommend withdrawing life-sustaining care are decisions that require judgment, experience, and ethical reasoning—not just probabilistic calculation. Furthermore, with less time spent on data entry, senior intensivists have more time for teaching at the bedside, mentoring trainees through difficult cases, and leading multidisciplinary rounds that are focused on collaborative problem-solving.

Most importantly, by reducing the time they spend at a computer, this model gives clinicians the time to be present with their patients. It allows them to sit, make eye contact, and listen without the distraction of a screen. It creates the space to have difficult but necessary conversations with families, explaining complex situations with clarity and compassion. These are the foundational acts of care that build trust and alleviate suffering, and the main reason most clinicians enter the profession.[50] To illustrate the practical impact of this transition, we present four concrete clinical vignettes in Table 2, contrasting the current standard of care with the proposed AI-augmented workflow.

Ultimately, the success of AI in critical care will be measured by its practical impact on the daily work of clinicians and the experience of patients. The goal is to use sophisticated technology to make the practice of medicine simpler, safer, and more focused on the patient. The vision is for an ICU where the most powerful tool is not the computer but the mind of a clinician who is supported, not burdened, by technology, and is fully empowered to practice medicine with skill, wisdom, and compassion.

Conclusions

AI has arrived in critical care, but thus far, it has promised more than it has delivered. There is a significant gap between the im-

pressive results seen in research studies and the limited, often frustrating reality of using these tools in ICUs. This paper has argued that the problem lies in a flawed goal: the idea that AI should be developed to replace the human clinician. This approach has produced tools that are often biased, that fail to work reliably in different hospitals, and whose black-box nature makes it impossible to trust with high-stakes clinical decisions. Too often, these systems seem designed to serve a balance sheet rather than a patient.

We have proposed an example of a more practical and powerful path forward: treating AI as a smart assistant, not an artificial doctor. The goal is to build a genuine partnership between the clinician and the machine based on a clear division of labor. This means designing AI to handle the exhausting, repetitive, and data-heavy tasks that contribute to clinician burnout. By letting AI manage the data, summarize records, flag subtle changes, and run routine calculations, we free up clinicians to do the work that only humans can do: complex problem-solving, making tough judgment calls under pressure, and, most importantly, providing compassionate care to patients and their families.

Getting this right requires serious and sustained commitment. We must first fix our underlying data systems so that the information feeding these tools is reliable and fair. We must build these tools with clinicians and nurses from day one, not hand them down from a technology department. We must demand that these systems be tested in rigorous, real-world clinical trials that prove they are safe and effective and actually help patients. In addition, we need clear rules and oversight to ensure they are used responsibly.

The choice before us is not about whether to use AI but about how we use it. We can continue to chase the fantasy of a fully automated ICU, or we can start building practical tools that support expert clinicians. The ultimate goal of AI in critical care should not be to create a more complicated system but a more functional one. Its success will be measured by its ability to help skilled clinicians make better decisions and have more time to connect with the patients in their care.

Acknowledgments

During the preparation of this work, the authors used the LLM "gemini-3-pro-preview" to improve the readability and language of the manuscript and of the review letter. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

LAC is funded by the National Institutes of Health through DS-I Africa U54 TW012043-01 and Bridge2AI OT2OD032701, the National Science Foundation through ITEST #2148451, and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant number: RS-2024-00403047).

Conflict of interest

The authors have no conflicts of interest to declare.

Author contributions

Conceptualization (LAC), project administration (CD), original draft preparation (CD, SM, SCO, AH, CAG, JB, RN, MT), writing and editing (CD, SM, HL, HTD, RB), supervision (LAC, RB). All authors reviewed and approved the final version of the manuscript.

References

[1] Zimmerman JE, Kramer AA, Knaus WA. Changes in hospital mortality for United States intensive care unit admissions from 1988 to 2012. Crit Care 2013;17(2):R81. doi:10.1186/cc12695, PMID:23622086.
[2] Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine Learning and Decision Support in Critical Care. Proc IEEE Inst Electr Electron Eng 2016;104(2):444–466. doi:10.1109/JPROC.2015.2501978, PMID:27765959.
[3] Asgari E, Kaur J, Nuredini G, Balloch J, Taylor AM, Sebire N, *et al*. Impact of Electronic Health Record Use on Cognitive Load and Burnout Among Clinicians: Narrative Review. JMIR Med Inform 2024;12:e55499. doi:10.2196/55499, PMID:38607672.
[4] Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011;4:47–55. doi:10.2147/RMHP.S12985, PMID:22312227.
[5] Lindsay MR, Lytle K. Implementing Best Practices to Redesign Workflow and Optimize Nursing Documentation in the Electronic Health Record. Appl Clin Inform 2022;13(3):711–719. doi:10.1055/a-1868-6431, PMID:35668677.
[6] Cvach M. Monitor alarm fatigue: an integrative review. Biomed Instrum Technol 2012;46(4):268–277. doi:10.2345/0899-8205-46.4.268, PMID:22839984.
[7] Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, *et al*. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. Ann Intern Med 2016;165(11):753–760. doi:10.7326/M16-0961, PMID:27595430.
[8] Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31–38. doi:10.1038/s41591-021-01614-0, PMID:35058619.
[9] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–118. doi:10.1038/nature21056, PMID:28117445.
[10] Yan S, Yu Z, Primiero C, Vico-Alonso C, Wang Z, Yang L, *et al*. A multimodal vision foundation model for clinical dermatology. Nat Med 2025;31(8):2691–2702. doi:10.1038/s41591-025-03747-y, PMID:40481209.
[11] Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A. A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare. Cureus 2023;15(10):e46454. doi:10.7759/cureus.46454, PMID:37927664.
[12] Esmaeilzadeh P. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: A perspective for healthcare organizations. Artif Intell Med 2024;151:102861. doi:10.1016/j.artmed.2024.102861, PMID:38555850.
[13] Cross JL, Choma MA, Onofrey JA. Bias in medical AI: Implications for clinical decision-making. PLOS Digit Health 2024;3(11):e0000651. doi:10.1371/journal.pdig.0000651, PMID:39509461.
[14] Muralidharan V, Adewale BA, Huang CJ, Nta MT, Ademiju PO, Pathmarajah P, *et al*. A scoping review of reporting gaps in FDA-approved AI medical devices. NPJ Digit Med 2024;7(1):273. doi:10.1038/s41746-024-01270-x, PMID:39362934.
[15] Matta S, Lamard M, Zhang P, Le Guilcher A, Borderie L, Cochener B, *et al*. A systematic review of generalization research in medical image classification. Comput Biol Med 2024;183:109256. doi:10.1016/j.compbiomed.2024.109256, PMID:39427426.
[16] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(1):195. doi:10.1186/s12916-019-1426-2, PMID:31665002.
[17] Lee AY, Yanagihara RT, Lee CS, Blazes M, Jung HC, Chee YE, *et al*. Multicenter, Head-to-Head, Real-World Validation Study of Seven Automated Artificial Intelligence Diabetic Retinopathy Screening Systems. Diabetes Care 2021;44(5):1168–1175. doi:10.2337/dc20-1877, PMID:33402366.

[18] Sagona M, Dai T, Macis M, Darden M. Trust in AI-assisted health systems and AI's trust in humans. NPJ Health Syst 2025;2(1):10. doi:10.1038/s44401-025-00016-5.
[19] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021;3(11):e745–e750. doi:10.1016/S2589-7500(21)00208-9, PMID:34711379.
[20] Robertson C, Woods A, Bergstrand K, Findley J, Balser C, Slepian MJ. Diverse patients' attitudes towards Artificial Intelligence (AI) in diagnosis. PLOS Digit Health 2023;2(5):e0000237. doi:10.1371/journal.pdig.0000237, PMID:37205713.
[21] Price WN 2nd, Cohen IG. Privacy in the age of medical big data. Nat Med 2019;25(1):37–43. doi:10.1038/s41591-018-0272-7, PMID:30617331.
[22] Bouderhem R. Shaping the future of AI in healthcare through ethics and governance. Humanit Soc Sci Commun 2024;11:416. doi:10.1057/s41599-024-02894-w.
[23] Wu G, Segovis CS, Nicola LP, Chen MM. Current Reimbursement Landscape of Artificial Intelligence. J Am Coll Radiol 2023;20(10):957–961. doi:10.1016/j.jacr.2023.07.018, PMID:37604328.
[24] Lighthall GK, Vazquez-Guillamet C. Understanding Decision Making in Critical Care. Clin Med Res 2015;13(3-4):156–168. doi:10.3121/cmr.2015.1289, PMID:26387708.
[25] Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak 2016;16(1):138. doi:10.1186/s12911-016-0377-1, PMID:27809908.
[26] Naser MZ. A Guide to Machine Learning Epistemic Ignorance, Hidden Paradoxes, and Other Tensions. Wires Data Min Knowl Discov 2025;15:e70038. doi:10.1002/widm.70038.
[27] Dankwa-Mullan I. Health Equity and Ethical Considerations in Using Artificial Intelligence in Public Health and Medicine. Prev Chronic Dis 2024;21:E64. doi:10.5888/pcd21.240245, PMID:39173183.
[28] 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, PMID:31649194.
[29] Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial Bias in Pulse Oximetry Measurement. N Engl J Med 2020;383(25):2477–2478. doi:10.1056/NEJMc2029240, PMID:33326721.
[30] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15(11):e1002683. doi:10.1371/journal.pmed.1002683, PMID:30399157.
[31] Sylolypavan A, Sleeman D, Wu H, Sim M. The impact of inconsistent human annotations on AI driven clinical decision making. NPJ Digit Med 2023;6(1):26. doi:10.1038/s41746-023-00773-3, PMID:36810915.
[32] Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, *et al*. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020;368:l6927. doi:10.1136/bmj.l6927, PMID:32198138.
[33] Funke BE, Jackson KE, Self WH, Collins SP, Saunders CT, Wang L, *et al*. Effect of balanced crystalloids versus saline on urinary biomarkers of acute kidney injury in critically ill adults. BMC Nephrol 2021;22(1):54. doi:10.1186/s12882-021-02236-x, PMID:33546622.
[34] Liu Y, Li Y, Xie D. Implications of imbalanced datasets for empirical ROC-AUC estimation in binary classification tasks. J Stat Comput Simul 2024;94(1):183–203. doi:10.1080/00949655.2023.2238235.
[35] Vickers AJ, Holland F. Decision curve analysis to evaluate the clinical benefit of prediction models. Spine J 2021;21(10):1643–1648. doi:10.1016/j.spinee.2021.02.024, PMID:33676020.
[36] Suara S, Jha A, Sinha P, Sekh AA. Is Grad-CAM explainable in medical images? In: Kaur H, Jakhetiya V, Goyal P, Khanna P, Raman B, Kumar S (eds). Computer Vision and Image Processing: 8th International Conference, CVIP 2023, Jammu, India, November 3-5, 2023. Cham: Springer; 2024:124–135. doi:10.1007/978-3-031-58181-6_11.
[37] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds). Advances in Neural Information Processing Systems 30 (NIPS 2017). Red Hook (NY): Curran Associates, Inc; 2017:4765–4774.
[38] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the predictions of any classifier. In: DeNero J, Finlayson M, Reddy S (eds). Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. San Diego (CA): Association for Computational Linguistics; 2016:97–101. doi:10.18653/v1/N16-3020.
[39] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 2018;6:52138–52160. doi:10.1109/ACCESS.2018.2870052.
[40] Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell 2019;1(5):206–215. doi:10.1038/s42256-019-0048-x, PMID:35603010.
[41] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12–22. doi:10.1016/j.jclinepi.2019.02.004, PMID:30763612.
[42] Elish MC. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. Engag Sci Technol Soc 2019;5:40–60. doi:10.17351/ests2019.260.
[43] Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012;19(1):121–127. doi:10.1136/amiajnl-2011-000089, PMID:21685142.
[44] Ait-Oufella H, Bourcier S, Alves M, Galbois A, Baudel JL, Margetis D, *et al*. Alteration of skin perfusion in mottling area during septic shock. Ann Intensive Care 2013;3(1):31. doi:10.1186/2110-5820-3-31, PMID:24040941.
[45] Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: A Bowtie analysis. J Saf Sci Resil 2024;5(4):460–469. doi:10.1016/j.jnlssr.2024.06.001.
[46] Cadario R, Longoni C, Morewedge CK. Understanding, explaining, and utilizing medical artificial intelligence. Nat Hum Behav 2021;5(12):1636–1642. doi:10.1038/s41562-021-01146-0, PMID:34183800.
[47] Bienefeld N, Keller E, Grote G. Human-AI Teaming in Critical Care: A Comparative Analysis of Data Scientists' and Clinicians' Perspectives on AI Augmentation and Automation. J Med Internet Res 2024;26:e50130. doi:10.2196/50130, PMID:39038285.
[48] Niroda K, Drudi C, Byers J, Johnson J, Cozzi G, Celi LA, *et al*. Artificial Intelligence in Cardiology: Insights From a Multidisciplinary Perspective. J Soc Cardiovasc Angiogr Interv 2025;4(3Part B):102612. doi:10.1016/j.jscai.2025.102612, PMID:40230667.
[49] Esmaeilzadeh P. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. BMC Med Inform Decis Mak 2020;20(1):170. doi:10.1186/s12911-020-01191-1, PMID:32698869.
[50] Kelley JM, Kraft-Todd G, Schapira L, Kossowsky J, Riess H. The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and meta-analysis of randomized controlled trials. PLoS One 2014;9(4):e94207. doi:10.1371/journal.pone.0094207, PMID:24718585.
[51] Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. BMC Med Inform Decis Mak 2023;23(1):73. doi:10.1186/s12911-023-02162-y, PMID:37081503.
[52] Hassan N, Slight R, Bimpong K, Bates DW, Weiand D, Vellinga A, *et al*. Systematic review to understand users perspectives on AI-enabled decision aids to inform shared decision making. NPJ Digit Med 2024;7(1):332. doi:10.1038/s41746-024-01326-y, PMID:39572838.
[53] Agrawal R, Prabakaran S. Big data in digital healthcare: lessons learnt and recommendations for general practice. Heredity (Edinb) 2020;124(4):525–534. doi:10.1038/s41437-020-0303-2, PMID:32139886.
[54] Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016;23(5):899–908. doi:10.1093/jamia/ocv189, PMID:26911829.

[55] Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, *et al*. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. Crit Care Med 2011;39(5):952–960. doi:10.1097/CCM.0b013e31820a92c6, PMID:21283005.
[56] De Georgia MA, Kaffashi F, Jacono FJ, Loparo KA. Information technology in critical care: review of monitoring and data acquisition systems for patient care and research. ScientificWorldJournal 2015;2015:727694. doi:10.1155/2015/727694, PMID:25734185.
[57] Wickham H. Tidy data. J Stat Soft 2014;59(10):1–23. doi:10.18637/jss.v059.i10.
[58] Babendererde N, Daum M, Reinke A, Samala RK, Zamzmi G. FDA's PCCP: Opportunities and Gaps. In: Zamzmi G, Reinke A, Samala RK (eds). Bridging Regulatory Science and Medical Imaging Evaluation. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2025:98–114. doi:10.1007/978-3-032-05663-4_6.
[59] Li H, Yu L, He W. The impact of GDPR on global technology development. J Glob Inf Technol Manag 2019;22(1):1–6. doi:10.1080/1097198X.2019.1569186.
[60] van Kolfschooten H, van Oirschot J. The EU Artificial Intelligence Act (2024): Implications for healthcare. Health Policy 2024;149:105152. doi:10.1016/j.healthpol.2024.105152, PMID:39244818.
[61] Ghenimi N, Govender R, Moodley K. The paradox of artificial intelligence (AI) and narrative-based medicine: challenges and potential for enhanced patient care. AI & Soc 2026;41(1):251–257. doi:10.1007/s00146-025-02418-3.
[62] Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implement Sci 2024;19(1):27. doi:10.1186/s13012-024-01357-9, PMID:38491544.
[63] Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge (MA): MIT Press; 2016.
[64] Healthcare Financial Management Association. Most healthcare organizations are adopting AI in the revenue cycle: HFMA poll. 2025. Available from: https://www.hfma.org/technology/most-healthcare-organizations-are-adopting-ai-in-the-revenue-cycle-hfma-poll/. Accessed July 30, 2025.
[65] Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. N Engl J Med 2017;376(26):2507–2509. doi:10.1056/NEJMp1702071, PMID:28657867.
[66] Needham DM, Davidson J, Cohen H, Hopkins RO, Weinert C, Wunsch H, *et al*. Improving long-term outcomes after discharge from intensive care unit: report from a stakeholders' conference. Crit Care Med 2012;40(2):502–509. doi:10.1097/CCM.0b013e318232da75, PMID:21946660.
[67] Sætra HS. Technology and Sustainable Development: The Promise and Pitfalls of Techno-Solutionism. 1st ed. Abingdon: Routledge; 2023. doi:10.1201/9781003325086.
[68] Stark P. Addressing Healthcare's Root Problems Over AI Band-Aids. SSRN; 2025. Accessed December 14, 2025. doi:10.2139/ssrn.576144.
[69] Alami H, Lehoux P, Denis JL, Motulsky A, Petitgand C, Savoldelli M, *et al*. Organizational readiness for artificial intelligence in health care: insights for decision-making and practice. J Health Organ Manag 2021;35(1):106–114. doi:10.1108/JHOM-03-2020-0074, PMID:33258359.
[70] Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med 2019;380(14):1347–1358. doi:10.1056/NEJMra1814259, PMID:30943338.
[71] Kerasidou A. Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. Bull World Health Organ 2020;98(4):245–250. doi:10.2471/BLT.19.237198, PMID:32284647.
[72] Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. J R Soc Med 2019;112(1):22–28. doi:10.1177/0141076818815510, PMID:30507284.
[73] Shanafelt TD, Boone S, Tan L, Dyrbye LN, Sotile W, Satele D, *et al*. Burnout and satisfaction with work-life balance among US physicians relative to the general US population. Arch Intern Med 2012;172(18):1377–1385. doi:10.1001/archinternmed.2012.3199, PMID:22911330.
[74] Baddeley A. Working memory: theories, models, and controversies. Annu Rev Psychol 2012;63:1–29. doi:10.1146/annurev-psych-120710-100422, PMID:21961947.
[75] Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, *et al*. Clinical information extraction applications: A literature review. J Biomed Inform 2018;77:34–49. doi:10.1016/j.jbi.2017.11.011, PMID:29162496.
[76] Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, *et al*. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024;30(4):1134–1142. doi:10.1038/s41591-024-02855-5, PMID:38413730.
[77] Siebig S, Kuhls S, Imhoff M, Gather U, Schölmerich J, Wrede CE. Intensive care unit alarms—how many do we need? Crit Care Med 2010;38(2):451–456. doi:10.1097/CCM.0b013e3181cb0888, PMID:20016379.
[78] Koyner JL, Carey KA, Edelson DP, Churpek MM. The Development of a Machine Learning Inpatient Acute Kidney Injury Prediction Model. Crit Care Med 2018;46(7):1070–1077. doi:10.1097/CCM.0000000000003123, PMID:29596073.
[79] Sampaio IW, Leccardi M, Drudi C, Liu J, Righetti F, Bianchi AM, *et al*. Predicting comatose patient's outcome using brain functional connectivity with a random forest model. Comput Cardiol 2023;50:1–4. doi:10.22489/CinC.2023.372.
[80] Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, *et al*. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med 2021;181(8):1065–1070. doi:10.1001/jamainternmed.2021.2626, PMID:34152373.
[81] Laubenbacher R, Niarakis A, Helikar T, An G, Shapiro B, Malik-Sheriff RS, *et al*. Building digital twins of the human immune system: toward a roadmap. NPJ Digit Med 2022;5(1):64. doi:10.1038/s41746-022-00610-z, PMID:35595830.
[82] Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018;24(11):1716–1720. doi:10.1038/s41591-018-0213-5, PMID:30349085.
[83] Drudi C, Mollura M, Lehman LH, Barbieri R. A Reinforcement Learning Model for Optimal Treatment Strategies in Intensive Care: Assessment of the Role of Cardiorespiratory Features. IEEE Open J Eng Med Biol 2024;5:806–815. doi:10.1109/OJEMB.2024.3367236, PMID:39559781.
[84] Mollura M, Drudi C, Lehman LW, Barbieri R. Optimal fluid and vasopressor interventions in septic ICU patients through reinforcement learning model. Comput Cardiol 2022;49:1–4. doi:10.22489/CinC.2022.189.
[85] Cavaillon JM, Chrétien F. From septicemia to sepsis 3.0-from Ignaz Semmelweis to Louis Pasteur. Genes Immun 2019;20(5):371–382. doi:10.1038/s41435-019-0063-2, PMID:30903106.
[86] Komorowski M, Green A, Tatham KC, Seymour C, Antcliffe D. Sepsis biomarkers and diagnostic tools with a focus on machine learning. EBioMedicine 2022;86:104394. doi:10.1016/j.ebiom.2022.104394, PMID:36470834.
[87] Maddali MV, Churpek M, Pham T, Rezoagli E, Zhuo H, Zhao W, *et al*. Validation and utility of ARDS subphenotypes identified by machine-learning models using clinical data: an observational, multicohort, retrospective analysis. Lancet Respir Med 2022;10(4):367–377. doi:10.1016/S2213-2600(21)00461-6, PMID:35026177.
[88] Hoffman J, Wenke R, Angus RL, Shinners L, Richards B, Hattingh L. Overcoming barriers and enabling artificial intelligence adoption in allied health clinical practice: A qualitative study. Digit Health 2025;11:20552076241311144. doi:10.1177/20552076241311144, PMID:39906878.
[89] de Andrade JBC, de Medeiros Cavalcante MAO, Lopes TLM, Treigher JMS, Balsells MD, Vasconcelos JL, *et al*. Discovery of data quality issues in electronic health records: profound consequences for critical care medicine applications - a systematized review. Crit Care 2026;30(1):19. doi:10.1186/s13054-025-05677-0, PMID:41508097.
[90] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, *et al*. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018. doi:10.1038/sdata.2016.18, PMID:26978244.
[91] Amaral AC, Taneva S, Morita PP, Lane-Fall M. Who Is Responsible for Mitigating Work Fatigue of Critical Care Clinicians-the Individual or the System? Ann Am Thorac Soc 2016;13(9):1453–1455. doi:10.1513/AnnalsATS.201606-517ED, PMID:27627474.

[92] Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. "Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. Proc ACM Hum-Comput Interact 2019;3:1–24. doi:10.1145/3359206.
[93] Mathur P, Burns ML. Artificial Intelligence in Critical Care. Int Anesthesiol Clin 2019;57(2):89–102. doi:10.1097/AIA.0000000000000221, PMID:30864993.
[94] Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ 2015;350:h391. doi:10.1136/bmj.h391, PMID:25662947.
[95] Kore A, Abbasi Bavil E, Subasri V, Abdalla M, Fine B, Dolatabadi E, *et al*. Empirical data drift detection experiments on real-world medical imaging data. Nat Commun 2024;15(1):1887. doi:10.1038/s41467-024-46142-w, PMID:38424096.
[96] Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, *et al*. The future of digital health with federated learning. NPJ Digit Med 2020;3:119. doi:10.1038/s41746-020-00323-1, PMID:33015372.
[97] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, *et al*. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023;10(1):1. doi:10.1038/s41597-022-01899-x, PMID:36596836.
[98] Platt R, Brown JS, Robb M, McClellan M, Ball R, Nguyen MD, *et al*. The FDA Sentinel Initiative - An Evolving National Resource. N Engl J Med 2018;379(22):2091–2093. doi:10.1056/NEJMp1809643, PMID:30485777.
[99] Greene SM, Reid RJ, Larson EB. Implementing the learning health system: from concept to action. Ann Intern Med 2012;157(3):207–210. doi:10.7326/0003-4819-157-3-201208070-00012, PMID:22868839.
[100] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, *et al*. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574–578. doi:10.3233/978-1-61499-564-7-574, PMID:26262116.
[101] Suran M. New NIH Program for Artificial Intelligence in Research. JAMA 2022;328(16):1580. doi:10.1001/jama.2022.15483, PMID:36282268.