A Clinician's Guide to Understanding Bias in Critical Clinical Prediction Models

João Matos, MSc[a,b,c], Jack Gallifant, MBBS, MSc[c,d], Anand Chowdhury, MD, MMCi[e], Nicoleta Economou-Zavlanos, PhD[f], Marie-Laure Charpignon, MS[g], Judy Gichoya, MD[h], Leo Anthony Celi, MD, MS, MPH[c,i,j], Lama Nazer, PharmD[k], Heather King, PhD[l,m,n], An-Kwok Ian Wong, MD, PhD[e,o,*]

KEYWORDS

• Bias • Prediction models • Artificial intelligence • AI • Machine learning

KEY POINTS

- This narrative review focuses on the role of clinical prediction models in supporting informed clinical decision-making in critical care, emphasizing their 2 forms: traditional scores and artificial intelligence-based models.
- Clinicians should evaluate these prediction models for their validity in ways similar to how ICU clinicians assess validity of pulse pressure variation.
- The assessment of pulse pressure variation is one of the many tasks critical care practitioners perform daily.
- Clinical prediction models play a crucial role in handling complex data to support clinicians to make more informed and timely decisions.

[a] University of Porto (FEUP), Porto, Portugal; [b] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Porto, Portugal; [c] Laboratory for Computational Physiology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA; [d] Department of Critical Care, Guy's and St Thomas' NHS Trust, London, UK; [e] Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, Duke University, Durham, NC, USA; [f] Duke Health, AI Health, Durham, NC, USA; [g] Institute for Data Systems and Society, Massachusetts Institute of Technology, Cambridge, MA, USA; [h] Department of Radiology, Emory University, Atlanta, GA, USA; [i] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; [j] Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA; [k] Department of Pharmacy, King Hussein Cancer Center, Amman, Jordan; [l] Durham VA Health Care System, Health Services Research and Development, Center of Innovation to Accelerate Discovery and Practice Transformation (ADAPT), Durham, NC, USA; [m] Department of Population Health Sciences, Duke University, Durham, NC, USA; [n] Division of General Internal Medicine, Duke University, Duke University School of Medicine, Durham, NC, USA; [o] Department of Biostatistics and Bioinformatics, Duke University, Division of Translational Biomedical Informatics, Durham, NC, USA
* Corresponding author. Duke University, 2 Genome Court, Box 103000 Durham, NC 27710.
E-mail address: med@aiwong.com

Crit Care Clin 40 (2024) 827–857
https://doi.org/10.1016/j.ccc.2024.05.011 **criticalcare.theclinics.com**

INTRODUCTION

With the rapid deployment of medical sensors, devices, and software systems in hospitals, the practice of critical care medicine has evolved and now relies extensively on the use of scoring tools and models to better monitor or predict clinical endpoints. For example, the assessment of volume status is critical to determine whether a hypotensive patient requires either more fluid or the initiation of vasopressors. To address this need, models such as pulse pressure variation (PPV) were created to predict volume responsiveness.[1,2] In clinical medicine, we are taught that PPV is predictive of fluid responsiveness but that certain conditions must be met to ensure its validity.[1,2] For example, initial studies required 2 criteria: positive pressure ventilation of 8 to 12 cc/kg and a regularly regular heart rate (ie, not in atrial fibrillation). Outside of these conditions, the accuracy of PPV is debatable; therefore, using this surrogate metric to guide clinical decision-making may not result in the intended effects.[1,2]

The assessment of PPV is one of the many tasks critical care practitioners perform daily. While this tool and other scoring systems (eg, Sequential Organ Failure Assessment [SOFA], Pneumonia Patient Outcomes Research Team [PORT] cohort study, and pneumonia severity index [PSI]) were studied in great detail during clinical training and subsequently put into practice, a large number of the prediction models embedded into the workflow involve recent advances in artificial intelligence and machine learning (AI/ML) approaches (eg, epic sepsis model [ESM][3] and Glucommander[4]). Further, although clinical teams are familiar with physiology-based measures such as PPV, they may not be equally versed in model-based tools. Given the speed at which new model architectures emerge in computer science, continued education is needed not only to familiarize with the content and methods underlying clinical prediction models but also to become aware of the associated risk of bias and inaccuracy.[3,5]

Given the unprecedented complexity, and potential for widespread impact, many governments are taking steps to regulate AI that may impact the daily lives of their citizens. The EU AI Act, approved by the European Parliament in June 2023, establishes obligations for providers and users depending on the level of risk from AI.[6] Just recently, in October 2023, the Biden–Harris Administration issued an executive order on safe, secure, and trustworthy AI,[7] reflecting the global concern over the potential implications of unregulated AI advancement.

As the field advances, it is critical for clinicians to understand the applicability and limitations of the many prediction models used in the intensive care unit (ICU), especially those based on AI/ML. In this narrative review, we take the perspective of critical care clinicians evaluating the practical aspects of a clinical prediction model that is available for use in the ICU. Through case studies and a listing of existing educational materials, our objective is to raise awareness and encourage the clinical end-user to be more inquisitive when apprehending a new prediction model.

CHALLENGES INHERENT TO THE INTENSIVE CARE UNIT

Clinical prediction tools, encompassing various scoring systems and models, play an indispensable role in the field of critical care.[8] In the ICU, clinicians face challenges akin to analyzing "big data," due to their quantity, sampling frequency, multimodality, and varying resolution and quality.[9] Health care practitioners in the critical care setting must incorporate information from multiple data sources, ranging from patient interviews to physical examinations, laboratory results, imaging, consultant reports, physiologic sensors, and scientific evidence. The complexities arising from this wide array of data are further compounded by patient heterogeneity, ranging from clinical

features such as comorbidities and surgical histories to vital signs[10] and, to a lesser extent in critical care, social determinants of health.[11]

Furthermore, observational studies of physician decision-making over time and in cognitively demanding clinical settings have suggested that repeated engagement in cognitively intense thinking can lead to a degradation in the quality of decisions.[12] In the demanding milieu of critical care, clinicians are burdened with multifaceted goals. Balancing patient safety, optimizing postillness outcomes, employing resources judiciously, and tailoring care through personalized medicine necessitate processing this vast amount of information. The criticality of each decision is heightened by the inherent ambiguity and challenge of establishing causal connections between treatment and outcomes.[13]

To navigate the challenges posed by big data and mitigate the impact of cognitive limitations, clinicians turn to clinical prediction tools. These tools assist in identifying the salient aspects of the data that are most critical for a particular decision.[14] Similar to applying a filter to a database search, clinical prediction models seek to clarify which data elements contain the most information pertinent to a particular problem and isolate these essential data into easy-to-interpret scales, such as categories or percentage risks. This summarization process involves strategically discarding nonessential information, ensuring that what remains is of utmost importance for the intended decision.

CLINICAL PREDICTION MODELS IN THE INTENSIVE CARE UNIT

Clinical prediction models have the potential to enhance the quality of care delivery and contribute to improved patient outcomes within the dynamic and demanding environment of critical care.[15]

Traditionally, these models are score-based, meaning they consist of a set of operations that consider various clinical variables, ultimately yielding a numerical score. By assessing the scale and distribution of data, thresholds can be established to facilitate informed decision-making. These models can aid in critical decisions involving risks versus benefits of specific treatments (eg, MELD,[16] congestive heart failure, hypertension, age $\geq$75 [doubled], diabetes, stroke [doubled], vascular disease, age 65 to 74 and sex category [female] [$CHADS_2$-VASC], and Hypertension, Abnormal Renal/Liver Function, Stroke, Bleeding History or Predisposition, Labile INR, Elderly, Drugs/Alcohol Concomitantly [HAS-BLED][17]), detection of early or atypical disease presentations (eg, the laboratory risk indicator for necrotizing fasciitis [LRINEC] score for necrotizing soft tissue infection,[18] or Hscore for hemophagocytic lymphohistiocytosis[19]), test selection and interpretation (eg, Wells score for PE prediction), prognosis assessment (eg, Acute Physiology and Chronic Health Evaluation [APACHE] IV[20] and Oxford Acute Severity of Illness Score (OASIS)[21] for mortality risk assessment), and risk adjustment for benchmarking and comparison (eg, Medicare Severity Diagnosis Related Group [MS-DRG],[22] Charlson comorbidity,[23] and Elixhauser scores[24]). At times, these models have been used to inform decisions related to resource allocation, even when not originally designed for such purposes (eg, SOFA for extracorporeal cardiopulmonary resuscitation [eCPR][25] and extracorporeal membrane oxygenation [ECMO][26]).

As AI/ML technologies become the foundation of these prediction models, some of the more recent developments have shifted toward leveraging advanced computational algorithms to handle complex, high-dimensional data and to extract intricate patterns that might not be discernible through conventional statistical methods.[8] AI-based models, distinguished from their traditional score-based counterparts, can tackle a wider range of tasks, adapt to evolving clinical environments, and refine their predictive accuracy over time. Examples of predictive models of this nature include

monitoring, early diagnosis (eg, sepsis), treatment decision support systems (eg, onset of mechanical ventilation), and outcome and prognosis assessment (eg, in-hospital mortality).[27]

These AI-based models can incorporate sophisticated modeling approaches such as deep learning architectures[28] or reinforcement learning techniques,[29] enabling a more dynamic and adaptive approach to data analysis and decision-making within critical care settings. AI-based models can also integrate multimodal input,[30] real-time data streams, and offer personalized predictions, thereby contributing to a more precise and tailored approach to patient care, treatment optimization, and resource allocation.[31]

In this section, we explore both "traditional score-based" and "AI-based" prediction models, which reveal distinct approaches and capabilities (**Table 1**). Traditional score-based models are typically limited in their scope, addressing specific problems such as mortality prediction, illness severity, and early warning scores (EWS).[37] These models rely on prespecified patient characteristics and are relatively simple, often summarized as a sequence of operations and easily computed using tools like MDCalc.[32] On the other hand, AI-based models have a much broader range of tasks, including monitoring, diagnosis, treatment, and outcome prediction.[27] However, their complexity lies in their often opaque, "black-box" structures, which require significant computational resources for both training and inference.[33]

Although traditional models generally do not consider fairness during development,[38] AI-based models are starting to incorporate fairness metrics from their initial design, even though there is still progress to be made in this area.[34] In terms of longevity and generalization, traditional score-based models are commonly used across various geographic and clinical settings, often remaining relevant for decades, even when not designed with that objective.[23,25,26] In contrast, AI-based models are more customized and designed to be adaptable, often tailored to specific populations, hospitals, or units and intended for iterative improvement.[35] Finally, while traditional models heavily emphasize clinical expertise, the development of AI-based models necessitates collaboration between individuals with clinical expertise and those possessing data science skills.[36]

Traditional Scores as Clinical Prediction Models

Clinical prediction models in the form of traditional scores have long served as vital tools for informing clinical decision-making. These scores often serve the purpose of either providing valuable information or aiding in specific decision-making processes, such as treatment assignment and resource allocation within a defined time frame. However, there are instances where these models, initially developed for specific purposes, may be repurposed or applied in different clinical contexts. An example of this is the utilization of the SOFA score for patient triage, highlighting the versatility of these tools beyond their original intended scope. In this section, we review some common score-based prediction models in the ICU, highlighting some of their limitations and factors to consider when using them in clinical practice.

Ashana and colleagues' investigation of the SOFA score's predictive capabilities for in-hospital mortality risks unearths marked racial disparities.[39] Their findings show that the risk of mortality is frequently underestimated among White patients but overestimated among Black patients. In scenarios where crisis standards of care (CSCs) apportion resources based on predicted mortality risks, this bias could inadvertently skew resource allocation. This is particularly evident among patients with projected mortality rates under 30%—the demographic arguably benefiting most from intensive care.[39] This bias persisted after adjustment for age, sex, and comorbidities, hinting at

Table 1
Comparison of "traditional" with "artificial intelligence-based" clinical prediction models

	"Traditional Score-based" Prediction Models	"AI-based" Prediction Models
Range of addressed problems	Limited range: eg, mortality prediction, illness severity, or EWS	Wide range of tasks, encompassing monitoring, diagnosis, treatment, and outcomes
Underlying patient characteristics	Often prespecified	May not always be specified
Complexity and interpretability	Simple, usually summarized as a sequence of operations and easily computed with tools like MDCalc[32]	Often resemble "black-box" models,[33] with complex structures and higher computational times both for training and inference
Fairness	Generally not considered during development; mostly evaluated post hoc and after the scores have been deployed	Although fairness metrics have a long way to go,[34] it is beginning to be considered from the initial design
Longevity and generalization	Used across diverse geographic and clinical settings. Use often grows stale and lasts for decades	More customized, "disposable," often limited to certain populations, hospitals, or units. Designed to be reiterated[35]
Talent and teams	Emphasis on clinical expertise	Collaboration between clinical expertise and data science skills for development, and implementation scientists for deployment[36]

Abbreviation: AI, artificial intelligence.

systemic factors like structural racism influencing mortality differentials. Ashana and colleagues' study demonstrates the pernicious consequences of this miscalibration on resource allocation, highlighting the pitfalls of using a model outside of its intended scope, and arguing for a reassessment of the SOFA score's place within CSCs.[40]

Model predictions can be skewed by information unrelated to the patient's clinical condition, such as the rate at which data are sampled. A case in point is the utilization of the APACHE II and simplified acute physiology score (SAPS) II severity scores in intensive care contexts.[37] Suistomaa and colleagues' exploration at a university hospital's ICU involved varying the sampling rates of laboratory and hemodynamic data and observing the consequential effects on severity scores.[37] Three distinct scoring paradigms were assessed: traditional scores (manual hemodynamic data paired with sporadic laboratory values), clinical information management system (CIMS) scores (2 minute median hemodynamic data with laboratory values based on clinical needs), and high rate scores (2 minute median hemodynamic data with 2 hourly laboratory assessments). The results revealed that increasing the sampling rate for hemodynamic monitoring and laboratory testing amplified the APACHE II and SAPS II scores considerably, leading to heightened predicted probabilities of hospital deaths. Notably, these increased scores did not correspond to heightened mortality rates, suggesting that predictive overestimations can distort clinical judgments. Additionally, the APACHE II score, despite its widespread use, harbors intrinsic limitations: operational complexity (its intricate nature poses operational challenges to routine use), predictive limitations (not a reliable prognostic tool, especially within the first 24 hours postadmission), and generalizability issues (initial validation was tailored for ICU-admitted patients, thereby reducing its efficacy for patients transferred from other wards or institutions). Suistomaa and colleagues' findings, juxtaposed with the inherent limitations of APACHE II, underscore the necessity for methodological rigor when interpreting the predictions from this model in the context of delivering care to individual patients.

EWS, like the United Kingdom's National Early Warning Score 2 (NEWS2), have become indispensable for identifying early decompensation in a complicated clinical milieu. The instrumentality of oxygen saturation by pulse oximetry (SpO_2), a core component, accentuates its utility in assessing respiratory functions. Nonetheless, recent studies spotlight biases in pulse oximetry, especially pertinent during the coronavirus disease 2019 (COVID-19) pandemic.[41] A retrospective analysis of 7126 patients with COVID-19 revealed a concerning racial bias in oxyhemoglobin measurement by pulse oximeters, with the device disproportionately overestimating arterial oxygen saturation for Asian, Black, and Hispanic patients vis-à-vis White patients.[41] These miscalibrations led to a substantial number of Black and Hispanic patients being overlooked for COVID-19-specific treatments. An exhaustive cross-sectional study further substantiated these discrepancies and highlighted the "hidden hypoxemia" phenomenon, which portends dire clinical ramifications.[42] These findings underscore the necessity for continuous re-evaluation of scores like NEWS2, with special emphasis on rectifying inherent biases to ensure clinical equity. In summation, it is paramount for clinicians to continuously scrutinize and understand the intricacies, correct application, and potential biases of traditional clinical prediction models. Doing so ensures that these tools maintain their efficacy and reliability, ultimately safeguarding the quality and equity of patient care.

Artificial Intelligence-based Clinical Prediction Models

In recent years, the landscape of clinical risk prediction models in critical care has witnessed a significant shift toward AI-based solutions. This precedent—including

successes and failures in design and/or implementation—can inform the training and deployment of new models, especially in situations where traditional scores might come short and alternatives are urgently needed.[8,43] Unlike traditional scores that can be easily computed,[32] AI-based models often necessitate integration within the health care system infrastructure, making them less readily available for scrutiny or interpretation by individual clinicians. This trend toward AI-based models also raises concerns about the purchase of newly developed and commercialized medical software and devices by health care institutions. Specifically, the selection of commercial products for use in the ICU may not be by end-user clinicians, although they should be part of the decision-making process.[44] Assessing the risks associated with these models becomes a critical consideration in this context, especially given the increased complexity of the algorithms, as well as the varying expertise of the teams developing them, which may differ significantly from traditional medical expertise.[36]

In this section, we explore the realm of clinical prediction models, emphasizing the role of AI in shaping their evolution. **Table 2** outlines the main categories for the tasks where AI/ML is being leveraged, similar to the taxonomy proposed by Hong and colleagues.[27] The 4 categories outlined can be grouped into 2 broader classifications. The first category is related to a "current assessment," involving (1) *real-time monitoring*, which evaluates the progression of patients' physiologic variables (22), or the settings of an ongoing treatment like mechanical ventilation.[46,48] The reviewed studies often utilized simpler modeling approaches. On the other hand, the second broad category focuses on predicting the "future" state of the patient and encompasses (2) *early diagnosis*, (3) *treatment decision support systems*, and (4) *outcome assessment*. These have recently garnered significant attention in research, leveraging state-of-the-art technologies in the realm of AI/ML.

Examples of prediction models for early diagnosis include acute kidney injury,[49] sepsis,[85] and respiratory disease,[56,57] all of which are associated with increased mortality in the ICU, and abnormal blood glucose levels.[86] As for treatment decision support systems, our review describes prediction models related to therapies that have a decisive impact on the management and outcomes of critically ill patients, including mechanical ventilation,[60–62,87] antibiotics dosing,[63] intravenous fluids and vasopressor administration,[64,65] heparin dosing,[66] morphine dosing,[67] and insulin dosing.[68,70] Regarding outcome prediction, prevalent tasks identified in the literature include predicting ICU and in-hospital mortality (20), ICU length of stay,[78–80] ICU readmission,[82,83] and long-term survival and quality of life.[84]

These prediction models can potentially improve patient outcomes, the salience of information, and, thus, the quality of decisions taken by the clinical teams and enhance bed management, aiding in resource allocation. Yet, the current reality remains that the algorithms prominently featured in research literature are largely impractical for direct implementation at the forefront of clinical practice.[44,88,89] Implementation may often be significantly harder than development on retrospective data; data management, model development, and clinical workflow implementation are 3 common hurdles that must all be passed.[88] In the following section, we explore the current limitations, challenges, and suggestions for clinicians to mitigate the risk of bias associated with such prediction tools in the ICU.[45,47,50–55,58,59,69,71–77,81,89]

RISK OF BIAS: RECOMMENDATIONS FOR A CLINICIAN USING AN ARTIFICIAL INTELLIGENCE TOOL IN THE INTENSIVE CARE UNIT

In this section, we apply the taxonomy proposed by Nazer and colleagues[90] on the bias in the AI/ML development pipeline, with a focus on an example application of

Table 2
Nonexhaustive description of categories of intensive care unit artificial intelligence-based clinical prediction models in literature

Category	Specific Task	Examples in Research Literature
1. Real-time monitoring	Physiologic indicators	Zhang and Szolovits[45] proposed patient-specific, bedside, real-time alarm algorithms based on neural network learning for adaptive monitoring in the ICU
	Mechanical ventilation settings	Kwok et al.[46] used a linear regression model and a nonlinear adaptive neuro-fuzzy inference system to estimate Fio_2; Rehm et al.[47] and Gholami et al.[48] created an ML classifier to detect patient-ventilator asynchrony
2. Early diagnosis	Acute kidney injury (AKI)	Sun et al.[49] proposed the use of clinical notes and deep learning for an early detection of AKI onset; Sanchez-Pinto and Khemani[50] delved into AKI prediction among critically ill children, using multivariable logistic regression
	Sepsis and infection	Desautels et al.[51] presented "InSight," a gradient-boost ML model to predict sepsis using a minimal set of EHR variables. Calvert et al.[52] studied the same model among an alcohol use disorder patient population. Mao et al.[53] from the same company, validated the same model across different centers in the United States. Ghosh et al.[54] explored coupled hidden Markov models to predict septic shock in the ICU. Bedoya et al.[55] developed a multioutput Gaussian process and recurrent neural network to predict sepsis upon emergency department admission. Wong et al.[3] attempted to externally validate the ESM, a proprietary early warning system for sepsis that has shown poor discrimination and calibration in predicting sepsis
	Respiratory disease	Le et al.[56] proposed gradient-boosted tree models for early prediction of acute respiratory distress syndrome (ARDS) in the ICU. Sauthier et al.[57] used random forest models to predict prolonged acute hypoxemic respiratory failure in influenza-infected critically ill children
	Abnormal glucose	Tang et al.[58] used deep neural networks to predict blood glucose concentrations after short-acting insulin injections
3. Treatment decision support system	Mechanical ventilation timing, duration, weaning, reinitiation	Miu et al.[59] created a multivariable logistic regression model to predict the need for reintubation in the ICU Ghazal et al.[60] trained bagged complex trees to predict SpO2 value after a ventilator setting change. Yu et al.[61] 2020 proposed a supervised-actor-critic reinforcement learning modeling approach to aid in the decision-making problems of ventilation and sedative dosing in the ICU. Sayed et al.[62] used gradient-boosted tree models to predict invasive mechanical ventilation duration after ARDS onset

	Antibiotics dosing	Janssen et al.[63] proposed a framework for informed precision dosing, requiring accurate pharmacokinetic or ML
	Intravenous (IV) fluid and vasopressor administration	Komorowski et al.[64] developed a reinforcement learning agent to achieve optimal administration of IV fluids and vasopressors. Srinivasan and Doshi-Velez[65] developed a novel interpretable batch variant of Adversarial Inverse Reinforcement Learning algorithm to optimize vasopressor and IV fluid administration in the ICU
	Heparin dosing	Nemati et al.[66] developed a deep reinforcement learning model to learn an optimal heparin dosing policy in the ICU
	Morphine dosing	Lopez-Martinez et al.[67] proposed a decision-making framework for opioid dosing based on reinforcement learning
	Insulin dosing	DeJournett et al.[68] proposed an AI-based closed-loop glucose controller for an ICU setting using an adaptive modeling approach proposed by Van Herpe et al.[69] Nguyen et al.[70] proposed an ensemble model to predict patients requiring more than 6 units of total daily insulin dose
4. Outcome Assessment	In-hospital and ICU mortality/survival	Hsieh et al.[71] created a Fuzzy Hyper-Rectangular Composite Neural Network to predict the survival of ICU patients in a Taiwanese center. Johnson and Mark[72] developed a gradient-boosting model to predict mortality among ICU patients in MIMIC-III.[73] Monteiro et al.[74] proposed the use of a linear support-vector machine model coupled with a multivariate feature selection process to predict ICU mortality using the 3 datasets of the PhysioNet/Computing in Cardiology Challenge.[75] Iwase et al.[76] created random forest models to predict ICU mortality and length of stay in a Japanese center. Choi et al.[77] trained, among others, light gradient-boosted machine models to predict ICU mortality in 2 university hospitals in South Korea
	ICU length of stay (LoS)	Abd-Elrazek et al.[78] employed fuzzy logic to predict LoS in the ICU using general admission features Alghatani et al.[79] created a binary model to predict whether the ICU stay is short or long, using MIMIC-III Hempel et al.[80] found random forest models to attain the highest performance for ICU LoS prediction using MIMIC-IV[81]
	ICU readmission	Rojas et al.[82] proposed a gradient-boosted machine model to predict ICU readmission using MIMIC-III. Lin et al.[83] used recurrent neural networks with long short-term memory to predict unplanned readmission using MIMIC-III
	Long-term survival and quality of life	Oeyen et al.[84] developed a prediction model for quality of life 1 y after ICU discharge based upon data available at the first ICU day using Lasso regression

Abbreviations: AKI, acute kidney injury; ARDS, acute respiratory distress syndrome; ESM, epic sepsis model; HMM, hidden Markov models; ICU, intensive care unit; LoS, length of stay; MIMIC-III, medical information mart for intensive care-III.

Taxonomy based on Hong and colleagues.[27]

this framework in the ICU. We emphasize how to effectively utilize these tools while maintaining a critical stance that addresses the potential risks of bias. *The premise is that a new AI tool has just been deployed in an ICU*. Prior to deployment, it has presumably been carefully analyzed by the hospital administration, who reviewed the framing of the problem and model, ensured that the modeling overarching approach is well-suited to solve the problem at hand; the team of developers was well suited for the task; and the methodology was sound and well executed. It is imperative for the clinicians to be well-versed in their institution's governance process, as it plays a pivotal role in querying developers and vendors. Following this initial step, we analyze the different steps of the ML development pipeline in the context of critical care, as defined by Nazer and colleagues.[90] For each step, we provide specific ICU examples; recommendations for bias risk assessment; and highlight how these can be put into practice in the context of the well-known overhaul and withdrawal of the epic sepsis prediction model.[3,91]

Guidelines for the ethical and equitable development, implementation, utilization, and governance of AI/ML models in the health care sector as a whole have attracted considerable attention in scholarly studies in recent times. Wiens and colleagues[92] presented guidelines on the translation of ML-based interventions into health care. Faes and colleagues[93] focused on promoting clinicians' critical appraisal studies of clinical applications of ML. Van de Sande and colleagues[94] summarized current guidelines, challenges, regulatory documents, and good practices that are needed to develop and safely implement AI in medicine. Nazer and colleagues[90] highlighted sources of bias within the process of developing AI algorithms in health care. Hassan and colleagues[95] provided a road map to develop predictive models that can be used in clinical practice.

With the increasing recognition of the importance of prioritizing fairness and uncovering biases among AI/ML developers,[96] reviews similar to the one performed by Nazer and colleagues[90] have been extensively performed across medical specialties. Arbet and colleagues[97] outlined common misconceptions about ML studies using electronic health record (EHR) data. Similarly, Sauer and colleagues[98] elaborated on potential pitfalls to be avoided when dealing with leveraging EHR data. Roberts and colleagues[99] conducted a systematic review that showed that all examined models intended to detect COVID-19 presented methodological flaws that hampered their utility. Delgado and colleagues[100] reviewed biases of AI algorithms developed for contact tracing and medical triage for COVID-19. Drukker and colleagues[101] delved into the different sources of bias in medical imaging-based ML methods. Gichoya and colleagues[102] reviewed pitfalls framed in the larger AI lifecycle for radiology applications. Nakayama and colleagues[103] listed the biases that can lurk in the AI lifecycle in ophthalmology. This abundance of studies suggests that the field is cognizant of the need for more structure; however, consensus is still needed for a set of common operating principles.[104–114]

Table 3 outlines the main sources of bias for an ML-based critical clinical prediction model. In the context of critical care, we explored: *data sources*, which include limitations related to selection bias,[104,111] unequally performing medical devices (80), and label bias; *data preprocessing*, where missingness handling (82, 83) and outlier removal can drive harmful spurious correlations; *model development*, which encompasses understanding the input features and their potential to leak information that compromises the utility of the model in real clinical practice (84); and *model validation* and *implementation*, which are associated with the performance of the algorithms, external validation, and postdeployment monitoring.[3,35,114]

Table 3
A checklist for intensive care unit clinicians to evaluate artificial intelligence-based clinical prediction models

ML Step	Risk of Bias/Challenge	What to do about it as a Clinician in the ICU	Case Study: ESM[3,91]
Data sources	Selection bias A mismatch between the training set and the real-world target; it can occur due to data and clinician drifts,[104] population shift,[105] and others	• Understand the population and datathon model was trained on • Compare the ICU typical composition with the cohorts behind the algorithms • Does the task at hand require any exclusion that I should be aware of?	Underlying patient characteristics are not reported (typically done in "Table 1"[106,107]) *"This model was developed and validated by Epic Systems Corporation based on data from 405,000 patient encounters across 3 health systems from 2013 to 2015."* We are unable to assess potential selection bias, which would require looking at the inclusion/exclusion criteria of the cohort and represented demographics
	Biased medical devices Pulse oximeters, ECG, EEG, temporal thermometers, and sphygmometers are ubiquitously used in the ICU but have been shown to yield inaccuracies among certain subpopulations[108]	• Are these device limitations taken into account in the model? • Is the model's incorporation of these limitations enough to produce biased results? • Which groups of patients should I be especially worried about?	There is limited information on the input data due to the proprietary nature of ESM *"Data elements included vital signs, medication orders, lab values, comorbidities, and demographic information."* The inclusion of vital signs could raise concerns as, for example, pulse oximetry readings are likely to be biased against Black and Hispanic patients[42]
	Label bias The label (or ground-truth) may be missing or inaccurate, thus leading to assumptions or limitations.	• Is the ground-truth a good ground-truth? • Is there any problem with the way we document the label that the model uses? • Is there anything I can do to increase the accuracy of the way possible labels are reported in the future?	*"(…) sepsis was defined as any encounter associated with an International Classification of Diseases (ICD-9) code indicating diagnosis of sepsis. Time of sepsis onset was defined as 6 h prior to clinical intervention (…)"* The label seems to be derived from billing information and based on ICD-9, which

(*continued on next page*)

Table 3
(*continued*)

ML Step	Risk of Bias/Challenge	What to do about it as a Clinician in the ICU	Case Study: ESM[3,91]
			may mean the model was not trained on patients with sepsis documented differently. This could raise questions about mismatches with actual practice
Data preprocessing	Handling missingness Some variables may not be missing at random, driving spurious correlations[109,110]; eg, the measurement of arterial blood gas in the ICU seems to be less likely among certain subpopulations[42] Outlier removal In the ICU, extreme values often take place (eg, blood pressure, glycemia)	• How does the model handle missingness? • Does it mirror my practice? • Does it embed any biases? • Is it consistent across subgroups? • Should I change anything in the way I report or not readings/measurements? • Does the outlier handling remove data points that I should be aware of? • Can I trust the model for such edge cases? • Can I recognize such edge cases and know when to fully ignore the model?	Data preprocessing is not mentioned by the vendors. *"This limited information is of concern because proprietary models are difficult to assess owing to their opaque nature and have been shown to decline in performance over time."* We recommend not to accept a model with such opaqueness and requesting these details from the development team
Model development	Diagnostic suspicion bias[111] An uneven diagnostic procedure in the target population, where some of the variables used to train the model (eg, timing and results of a test order) already convey information about the outcome, which constitutes a subtle yet common example of data leakage and limiting performance in patients where staff are not already suspicious Included variables The input data must contain relevant predictors that avoid leakages and go in line with clinical causal intuition	• Does the model require any variables that depend on my suspicion? • Is there any variable that reflects my bias and can be further confirmed or reinforced by the model's output? • Is there any reinforcement loop that I can or should avoid? • Do the included features follow a causal rationale that makes sense according to my clinical intuition? • Understand that variable importance tools (eg, SHAP[112]) are not causal	The only information on the modeling is: *"The ESM is a penalized logistic regression model (…)"* As the input features are not disclosed, assessing the soundness of this step is very challenging, posing a significant concern for this model. The study conducted by Wong et al.[3] suggested that one input variable to the model was antibiotic orders by a provider—a classic example of diagnostic suspicion bias since the suspicion of the clinical team would be necessary for the model to work. Further, we are unable to assess whether the possible relations between the features and the output have a causal rationale behind them

Model validation	Model performance The model will never work for 100% of the patients. External validation can help assess the utility of a model,[3] but it does not represent a definitive answer[35,113]	• What was the reported performance of the model? • Does the evaluation look sound? • Was it equally effective across groups? • To what extent can I trust this model?	In the final ESM, *"(…) AUC ranged between 0.76–0.83."*, which seems to be fairly calibrated. However, the external validation performed by Wong et al.[3] showed a significant deterioration, with poor discrimination in predicting the onset of sepsis—AUC of 0.63—and a large burden of alert fatigue. Critically appraising these metrics is fundamental to gaining trust
Implementation	Postdeployment monitoring Models are prone to drifts of different kinds over time[114]	• Is the model being updated? • Is it adaptable to possible changes? • Can I trust it in the long term? • Are there plans for real world monitoring?	This model bypassed peer review and regulatory oversight. The postdeployment monitoring seems to have been poorly conducted, considering the drop in performance verified by Wong et al.[3]

Abbreviations: AUC, area under the curve; ECG, electrocardiogram; EEG, electroencephalogram; ESM, epic sepsis model; ICD, International Classification Of Diseases; ICU, intensive care unit.

Taxonomy of ML steps based on Nazer and colleagues.[90]

As a case study, we delved into the ESM, which has been withdrawn following a publication from Wong and colleagues.[3] It presents various risk factors, biases, and, more importantly, uncertainties that raise significant structural concerns for its use in the ICU setting. First, the lack of transparency on the sources for the model's training data hampers our ability to assess the introduction of bias.[115] Additionally, the model's reliance on medical devices that were later shown to introduce racial bias, such as pulse oximeters, poses a risk since such medical devices may yield biased readings for certain patient groups as described earlier.[108] The use of billing data for label generation could also raise questions about the accuracy of the ground-truth, and the lack of information on missingness handling and outlier removal complicates the model's reliability.[103] The model's dependence on diagnostic suspicion poses a serious limitation to its real-world applicability and may perpetuate biases inherent in clinical decision-making.[3] The absence of information on model variables and causal rationale further challenges its clinical applicability. Furthermore, the discrepancy between vendor-reported and externally validated performance metrics, along with the apparent lack of postdeployment monitoring and transparency, diminishes trust in the model's long-term reliability.[3]

Despite all these concerns, the ESM was implemented in hundreds of US hospitals, bypassing peer review and regulatory oversight.[3] In light of the concerns and potential risks associated with the ESM, our recommendations for a clinician utilizing a similar model emphasize 3 core principles: curiosity, questioning, and skepticism. It is imperative for clinicians to actively engage with the model's documentation, approaching the model's reported training processes and underlying architecture with a critical and inquisitive mindset. This involves probing the model's data sources, underlying assumptions, and algorithms, as well as seeking transparency and detailed information from the developers regarding data preprocessing, feature selection, model development, and validation. Clinicians should continuously question the model's accuracy, especially in the context of their specific ICU patient population and remain vigilant for any potential biases or limitations.

HOW TO LEARN ABOUT CLINICAL PREDICTION MODELS AND SOURCES OF BIAS?

AI in health care is a field that is developing and expanding rapidly, and therefore, clinicians should be constantly updated on the most recent advances in this field, as well as understand the various sources of bias and potential strategies to mitigate them. This has traditionally been taught in the framework of scientific articles.[116] Historically, continued education and professional development used to be limited to individuals and institutions that can cover the cost of training and educational resources; however, over the past several years, the increased availability of open-access resources and virtual conferences/webinars has facilitated upskilling in data science for health care for practitioners and institutions within various resource settings. **Table 4** outlines major resources that clinicians may utilize to advance their skills and knowledge in the field of AI and the potential sources of bias. These include dedicated textbooks and journals as well as more modern resources such as datathons and workshops, which allow interactions around real-time, hands-on case studies. The launch of the first SCCM datathon in August 2023 illustrates the importance of the venue in fostering collegial information sharing and learning about the development, implementation, and evaluation of clinical prediction models in critical care.

Textbooks have traditionally been considered as the primary source of knowledge; however, as with all textbooks, the information quickly becomes outdated.[120] Though

Table 4
Resources available for clinicians to learn more about artificial intelligence in critical care and potential biases[a]

Source	Examples
Books	Chayakrit Krittanawong. Artificial intelligence in clinical practice. 1st Edition, 2023[117] MIT Critical Data. Secondary Analysis of Electronic Health Records.[118] Asselbergs FW. Clinical Applications of Artificial Intelligence in Real-World Data. 1st Edition, 2023
Journals[b]	Journal of American Medical Informatics Association Lancet Digital Health PLOS Digital Health BMC Digital Health BMJ Health & Care Informatics
Preprint servers	arXiv medRxiv
News Web sites	Stat News Guardian Technology MIT Technology Review Stanford HAI News ProPublica Technology Wired Science
Social media (Linkedin, X (Twitter), others)	SCCM ESICM YouTube—for interest-oriented learning. Keywords relevant to these topics include "AI Bias," "ML Fairness," "Health Equity"
Societies/professional groups	SCCM Discovery Data Science ESICM Data Science Section BrainX WiDS
Others	Datathons—eg, MIT Critical Datathon 2023, focused on Pulse Oximetry Bias; and SCCM Discovery Datathon 2023, which included Patient Safety and Health Equity tracks Conferences Coursera—eg, Emma Pierson's "Practical Steps for Building Fair AI Algorithms" course[119] Udemy

Abbreviations: ESICM, European Society of Intensive Care Medicine; SCCM, Society of Critical Care Medicine.

[a] This is a nonexhaustive list of common sources.

[b] All critical care journals have been publishing articles related to AI and ML in critical care.

most textbooks may be expensive to purchase, there are a few AI textbooks that are freely available, such as *Secondary Analysis of Electronic Health Records*.[118] Journals are another major resource that clinicians constantly rely on to stay up to date on recent science. In general, most critical care journals publish in the field of AI, but there are also journals that specialize in digital health and AI, such as the Journal of the American Medical Informatics Association, Lancet Digital Health, PLoS Digital Health, and BMC Digital Health. However, not all of these journals are open access. Preprint servers, such as arXiv or medRix, serve as valuable platforms for researchers to rapidly disseminate their findings to the scientific community before formal peer review, fostering early communication and collaboration. However, while enabling swift

knowledge sharing, preprint servers may present challenges in terms of quality control and the potential spread of unvalidated or misleading information.[121]

News Web sites also play an interesting role in scientific dissemination by simplifying complex technical articles and making them accessible to diverse audiences. News Web sites offer curated information sources, facilitating access to complex technical and conceptual material for those who may find it challenging to navigate on their own; as an example in the scope of technology and critical care, technical computer science articles tailored to the comprehension of a medical audience can often be found within these resources. Additionally, news Web sites serve as a means for the early distribution of preprints and diverse viewpoints, contributing to the rapid flow of information within the scientific community. Although institutions usually have partnerships with these Web sites, hefty subscription fees may pose an obstacle to accessibility. Examples include *Stat News*,[122] *MIT Technology Review*,[123] *Stanford HAI News*,[124] *Guardian Technology*,[125] and *ProPublica Technology*.[126] Social media is becoming a major educational source for health care practitioners as it provides an update on what has been recently published in science, as well as creates a platform for discussing various aspects. The diversity of members on social media in terms of their backgrounds and settings creates an enriching platform to understand limitations and bias within various fields, including AI. Most critical care societies and journals post on social media platforms, mainly LinkedIn and X/Twitter, and to a lesser extent on Facebook and Instagram. However, though there is significant value to learning through social media, one should keep in mind that the content does not undergo any form of peer review and, therefore, should be carefully assessed for its validity.

Societies are also an important venue for various educational programs. There are data science groups within societies that conduct various educational activities during their annual conferences as well as webinars and other educational sessions. For example, the Society of Critical Care Medicine (SCCM) has a Data Science Campaign through their Discovery Research Section,[127] and the European Society of Intensive Care Medicine (ESICM) has a Data Science section.[128] However, activities through such societies and sections are limited to those who have membership. There are also other societies that are more specialized in AI and big data, such as the BrainX Community[129] and Women in Data Science (WiDS),[130] both of which offer free membership. In addition, they both provide various educational programs, many of which require no registration fees.

Datathons are a helpful way to build capacity and collaborations.[131–133] Conceived in 2016, it places clinical staff and data scientists/informaticists in direct contact. As opposed to a clinician who blindly relies on a black box/"magical" thinking, codevelopment and working together breaks down silos and provides firsthand experience with the process of model development. In-person datathons create fellow student and researcher teams so that data scientists and clinicians can combine their skills when addressing a problem. This unique opportunity to bring clinicians and data scientists together allows for the creation of an interface layer.[131] Not all academic centers have both in sufficient concentrations or naturally encounter each other. Working together on a project through a datathon can be a time-efficient manner to increase effective learning.[134] Datathon organizers include institutions (eg, Massachusetts Institute of Technology) and clinical societies (eg, ESICM and SCCM).[135] Datathons have been conducted in-person and virtually and have been received positively.[117,119,136] They have also generated diverse groups of research teams that continue to work together after the datathons. However, such datathons are restricted to a small group of participants, given that most of them require financial support for the travel of the

participants to the site of the event or the travel of instructors. While some virtual datathons have been organized, their impact remains to be assessed and compared with that of in-person events.

DISCUSSION

Modeling Limitations

Despite rapid progress in ML for health in the last decade, estimating the causal effects of interventions taking place in ICU settings remains challenging. Indeed, ICU patients often present multiple comorbidities upon admission, and their status may further complicate during their stay, resulting in a large number of time-varying confounders of any treatment-outcome relationships of interest. Moreover, ICU clinicians often prescribe several treatments concurrently (eg, antibiotics, anticoagulants, and antiarrhythmics).[137] The complexity of ICU pharmacotherapy thus makes isolating the effect of a single drug difficult and preventing harmful drug–drug interactions complicated.[138–140]

Beyond the presence of measurement errors emanating from numerous medical devices used in ICU settings (eg, pulse oximeters and sphygmomanometers),[141,142] observational studies conducted in critical care are also more prone to immortal time bias[143] than in other fields of medicine due to a high mortality rate in the first 24 hours following ICU admission. Indeed, a recent retrospective cohort study performed in Alberta, Canada, found that patients who die within 1 day comprise one-third of ICU deaths.[144] Therefore, if we were interested in evaluating the effect of an intervention only made available after the first day on ICU length-of-stay, patients who survived to their first ICU day would have a period of unexposed immortal time before receiving the intervention, an easily missed sampling bias. In 2009, Shintani and colleagues[145] had already warned about the prevalent but misleading use of standard Cox regression models in ICU survival analyses, showcasing the extent of bias when using time-fixed covariates to analyze the effect of a time-varying exposure on ICU length of stay. In a newly published perspective,[146] Vail and colleagues have again called for increased attention to immortal time bias in critical care, illustrating their argumentation with flawed observational studies of exposure to hydrocortisone, ascorbic acid, and thiamine therapy among patients with sepsis and septic shock—all published between 2017 and now. The authors took a step forward by providing a checklist for clinicians to more easily evaluate the characteristics of study design and analysis that may result in immortal time bias or detect a lack of sufficient reporting to rule out its absence. Two simple recommendations emanate from the studies of Shintani and colleagues and Vail and colleagues: first, carefully checking the methods section of any clinical article to ensure that time-varying analytical techniques were used appropriately, and second, ensuring that follow-up begins after the intervention eligibility period ends and at a time that is aligned across all patients.

For practitioners who are also greatly involved in research, the detailed specification of a "target trial" is advisable when retrospectively analyzing routinely collected patient data, that is, following the same procedure as when writing the detailed protocol of a randomized controlled trial. Practices such as listing inclusion/exclusion criteria, describing the static or dynamic treatment strategies under investigation, defining the follow-up period, and eliciting the causal estimands of interest all contribute to improving the transparency of statistical inferences. For instance, by considering more realistic treatment eligibility criteria and strategies, Wanis and colleagues[147] have shown that ICU patients captured in the medical information mart for intensive care (MIMIC)-IV database who were intubated earlier versus later during their stay

had similar 30 day mortality rates. Their findings contrast with prior studies, which often used infeasible treatment strategies, and highlight the sensitivity of treatment effect estimates to critical but often neglected study design decisions.

The challenges faced in the ICU, including those related to the complexity of establishing causality within a context of multiple concurrent treatments, biases from medical devices, and the presence of immortal time bias, have the potential to generate misleading and harmful spurious correlations. Spurious correlations are noncausal relationships between the input and the outcome, which may shift in deployment.[105] These spurious correlations are particularly concerning, especially when they arise from systemic social discrimination, as seen in the case of bias in critical care medical devices.[108] Allowing the embedding of these errors, biases, and limitations in subsequent AI models could perpetuate and exacerbate existing disparities.[148,149] Therefore, modeling efforts in the realm of critical care must be approached with caution,[150] and a careful inspection of such sources of bias must be conducted a priori.[151]

Geographic Variability

Another challenge to the body of knowledge of clinical prediction models in the ICU is the significant variability observed across different ICU units, hospitals, and geographic locations. Numerous factors contribute to this variation, encompassing aspects such as differential patient illness severity, clinical outcomes, hospital type (eg, academic, community), size, number of beds, occupancy, staffing coverage, weekend coverage, demographics of the served population, reasons for ICU admission, or types of ICU units within the hospital.[152] For instance, comparing the health care systems in the United Kingdom and the United States reveals substantial dissimilarities.[153,154] The United States has 7 times as many ICU beds per capita as the United Kingdom.[153] In the United Kingdom, hospital stays before ICU admission are longer and the severity of illnesses is heightened.[153] This diversity in health care settings presents significant challenges in developing clinical prediction models that aim to effectively function across different contexts. An illustrative example is the NEWS2 in the United Kingdom, which, despite probably not being equally performant for all the different settings and populations,[41] is used nationwide as a guideline. As a result, it is imperative for clinicians to evaluate the architecture of their clinical prediction models critically. The multitude of reasons why a model may not be effective in a new setting underscores the need for a nuanced understanding of the local dynamics. Therefore, any model must be conscious of these challenges, be grounded in its local context, and aim to accommodate the intricacies of geographic variability from design. Similarly, the methodologies and recommendations suggested herein may not universally apply to all settings.

Challenges Related to Explainability, Generalizability, and External Validation

Explainability methodologies are argued to build trust among health care professionals, offering transparency in AI/ML decision-making, and potentially reducing bias.[155] In fact, recent Food and Drug Administration guidance recommends incorporating explanations into clinical decision support software so that clinicians are informed about the foundations of recommendations.[156] Nevertheless, the added value of such explanations remains debatable. In fact, a recent randomized clinical survey conducted by Jabbour and colleagues[157] showed that AI model explanations did not aid clinicians in identifying systematically biased models. In the absence of suitable explainability techniques, it is argued that the emphasis should be on careful internal and external validation of clinical models.[158]

Generalizability, that is, the ability of AI/ML models to extrapolate their knowledge to unobserved data, has also attracted considerable attention from researchers. Futoma and colleagues[113] highlighted that generalizability is not a binary concept but a multifaceted one, involving not only temporal considerations like prospective application within the original center but also external validation across new centers and timeframes. However, neglecting such limitations in generalizability could lead to missed opportunities for leveraging AI/ML models in situations with potential clinical utility. Instead, narrow, "overfit," local models that work under certain circumstances and for certain subpopulations may actually be acceptable and yield value in real-world ICU practice.[35]

This has been confirmed in subsequent studies such as Youssef and colleagues[35] that argue that external validation of a clinical prediction model does not necessarily imply that it is useful in real-world settings. To ensure the practical usefulness of AI-based clinical models, we recommend complementing offline internal and external validations by the implementation of prospective impact studies; these can subsequently be used to timely determine the need to retrain the model locally. Wide adoption of the proposed recurring and local validation framework should allow for addressing distribution shifts in treatment, outcome, or both. In addition, if a change in health insurance contracts affects the mix of patients coming to the ICU, if critical care protocols are updated (eg, following the modification to sepsis recommendation guidelines), or if a hospital deploys a new EHR system, a timely update to the model using local patient cohorts would incorporate these new operational inputs.

Hence, as clinicians critically appraises a clinical prediction model, despite extensive external validation, it is crucial to approach any prediction model with caution and skepticism, as success in various centers and environments does not guarantee optimal performance within a specific ICU setting.

Postdeployment Detection and Mitigation of Disparities

Detecting and mitigating disparities after model deployment involves a multistep process, from data collection to data analysis, model correction development, dashboard creation, and near real-time monitoring of the revised models once implemented. While existing models may take at most a week to get updated and released on Hugging Face (Brooklyn, New York, NY), methods to evaluate the extent of their biases have not been standardized, and there is no platform where investigators can similarly post the results of model investigations or stress tests.

Disparity Dashboards

Despite the limited offer, a few initiatives have recently emerged. For example, Yi and colleagues[159] have described the steps needed to design and develop a digital equity dashboard for the emergency department of UC San Francisco hospitals. The use of disparity dashboards in clinical care delivery is growing. To sustain such efforts, Gallifant and colleagues[160] have recommended the setup of incentive systems to accelerate health data collection and reporting and of rewards that acknowledge successful mitigation of health disparities. Nonetheless, certain biases are more subtle and may remain undetected.[161] For instance, cognitive biases may affect the way clinicians handle conversations regarding end-of-life care with a patient's family.

Frameworks and Guidance Initiatives

High-level frameworks, recommendations, and guidance initiatives are also being designed to address these issues. Specifically, initiatives like STANDINGTogether[162] have the objective of ensuring the comprehensive representation of diverse

populations in health datasets for the development of AI systems, which could solve part of the problem. The primary focus lies in offering guidance on the collection and reporting of crucial demographic details, including but not limited to gender, race, ethnicity, and others. Emphasizing transparency, the recommendations advocate for clear disclosure of any limitations within the dataset. This transparency facilitates informed decision-making by developers when selecting datasets for their AI models or tools. Furthermore, the STANDINGTogether guidelines provide insights into identifying potential harm to specific groups when employing medical AI systems, thereby contributing to the responsible and ethical use of such technologies. Other initiatives, such as the Coalition for Healthcare AI, are addressing this by convening experts from health care systems experts from multiple institutions representing health care systems, academia, government, and industry to identify problems and propose solutions to enable trustworthy AI in health care.[163] By developing a framework for an assurance standard and releasing a blueprint as a first step to building consensus on the execution, they attempt to develop an executable path toward assurance laboratories for continued assessment and monitoring of deployed and implemented systems.

The Potential of Artificial Intelligence/Machine Learning Models to Help Level the Playing Field

Chen and colleagues[164] have also argued that AI can help address health disparities, including by identifying and mitigating well-documented societal bias. A foundational study by Obermeyer and colleagues[165] estimated the calibration bias of an algorithm used to predict the health needs of insured patients 1 year ahead and showed significant differences based on race. Practically, a Black patient with the same algorithmic risk score as a White patient would on average have worse outcomes than their counterpart a year later. This retrospective analysis suggests that the insurer's model was underestimating the health needs of Black patients. Because they also had access to yearly health care costs per patient, the authors were able to identify the source of this bias, namely the use of individual-level health costs as a misleading proxy for health needs.

When the mechanisms underlying existing disparities can be interrogated and the sources of bias can be identified even partially, the development of correction models is facilitated. For example, underrepresentation of women and minority groups in clinical trials for cardiovascular diseases is known to affect the fairness of risk prediction models for atherosclerotic cardiovascular disease; yet, explicit adjustment in new models can alleviate the repercussions of a lack of inclusion in past trials.[166] Using the Southern Community Cohort Study, Zink and colleagues[167] identified differences in data quality as another source of bias in colorectal cancer risk prediction models.

Detecting or addressing disparities or biased practices sometimes involves stratifying or adjusting for race, ethnicity, and other social determinants of health. However, the decision of using race and ethnicity as input variables in risk prediction models remains highly contentious[168] and should be made on a case-by-case basis, with desired health outcome targets and fairness metrics clearly stated. Indeed, while the push[169,170] to remove such variables from risk scoring systems is legitimate, simply omitting race and ethnicity could yield worse prediction accuracy for racially minority groups,[171,172] as recently demonstrated by Khor and colleagues[173] in the context of a risk prediction model for colorectal cancer recurrence. Similarly, Zink and colleagues showed that implementing race-based corrections into colorectal cancer risk prediction models can counterbalance differences in data collection (eg, missingness, quality) by race.

In Pursuit of Fair, Performant, Sustainable, and Transparent Models

What do we ultimately seek from critical clinical prediction models? The answer, we posit, is 4 fold: fairness, performance, sustainability, and transparency. Fairness is essential to prevent the perpetuation and exacerbation of harmful societal biases within our models. Performance is crucial for ensuring the reliability and accuracy of these prediction models. Sustainability is pivotal to enable the necessary continual, automated updates of the models. Finally, transparency will democratize the ability to examine the underlying cohorts, methodologies, and architectures, which will ultimately foster fairness, performance, and sustainability.

While implementation poses challenges, it is imperative to enhance the hospital's capacity to accommodate the demands of these ever-evolving, lifelong learning AI/ML models. This requires not only building and improving data infrastructures within our hospitals but also providing comprehensive training to clinicians, specifically intensivists in the context of this review, to enable them to critically analyze the risk of bias, and effectively utilize the new generation of ICU tools.

SUMMARY

Clinical prediction models play a crucial role in handling complex data to support clinicians to make more informed and timely decisions. These models come in 2 forms: traditional scores and AI-based, each addressing a different range of tasks, with varying levels of complexity, interpretability and generalizability; these differences are typically inherent to the differences of expertise between the development teams of each type of model. Bias is not limited to either traditional or AI-based models, as both types have been found to potentially perpetuate harmful societal biases. Mitigating bias in AI models requires collaboration among diverse teams well-versed in understanding the underlying datasets and AI methodologies, as well as the critical appraisal of these tools by both hospital leadership and clinicians, particularly in the ICU. As bias can emerge at every stage of the AI lifecycle, from data sources to model deployment, we outline 6 steps, accompanied by examples that serve as a scaffold to design strategies to manage the risk of bias. In a more holistic view, bias mitigation will require ensuring the sustainability of clinical data pipelines within hospitals, prioritizing transparency and fairness in model development, and providing a more interdisciplinary training to clinicians. For clinicians interested in expanding their understanding of bias, resources like books, journals, social media platforms, professional societies, and events like datathons can be valuable sources of information.

CLINICS CARE POINTS

- Critical clinical prediction models enable clinicians to distill complex data into actionable insights, facilitating well-informed and timely decisions. These algorithms can be "traditional score-based" and "AI-based," yielding different properties regarding their range of addressed problems, underlying patient characteristics, generalization capabilities, development teams, fairness assessment, and complexity and interpretability.
- Biases can be present in traditional clinical prediction models (eg, SOFA, NEWS), as well as in AI-based models (eg, ESM), and both models have demonstrated the risk to perpetuate harmful societal biases.
- Effectively mitigating bias and reducing potential harm in AI-based models necessitates the collaboration of diverse teams possessing expertise in understanding both underlying

datasets and AI methodologies. Equally important is the critical evaluation of these tools in the ICU by hospital leadership and clinicians. Biases, spanning the entire AI lifecycle, originate from data sources, preprocessing, model development, evaluation, and deployment stages.

- Clinicians seeking to expand their knowledge on bias can explore resources such as books, journals, social media platforms, societies and professional groups, and other decentralized events like datathons.
- Enhancing transparency and fairness in the development of predictive models, ensuring sustainability in hospitals' clinical data pipelines, and providing comprehensive training to clinicians are fundamental steps to identify and mitigate biases in critical clinical prediction models.

DISCLOSURE

A.I. Wong holds equity and management roles in Ataia Medical. A.I. Wong is supported by the Duke CTSI by the National Center for Advancing Translational Sciences, United States (NCATS) of the National Institutes of Health, United States under UL1TR002553 and REACH Equity under the National Institute on Minority Health and Health Disparities, United States (NIMHD) of the National Institutes of Health under U54MD012530. All other authors have no conflicts to disclose.

REFERENCES

1. Myatra SN, Prabu NR, Divatia JV, et al. The changes in pulse pressure variation or stroke volume variation after a "tidal volume challenge". Reliably predict fluid responsiveness during low tidal volume ventilation 2017. https://doi.org/10.1097/CCM.0000000000002183.
2. De Backer D, Heenen S, Piagnerelli M, et al. Pulse pressure variations to predict fluid responsiveness: influence of tidal volume. Intensive Care Med 2005;31(4):517–23.
3. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021;181(8):1065–70.
4. Davidson PC, Steed RD, Bode BW. Glucommander: a computer-directed intravenous insulin system shown to be safe, simple, and effective in 120,618 h of operation. Diabetes Care 2005;28(10):2418–23.
5. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464). https://doi.org/10.1126/science.aax2342.
6. EU AI Act: first regulation on artificial intelligence. Available at: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. [Accessed 1 November 2023].
7. The White House. President biden issues executive order on safe, secure, and trustworthy artificial intelligence. Available at: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/. [Accessed 1 November 2023].
8. Johnson AEW, Ghassemi MM, Nemati S, et al. Machine learning and decision support in critical care. Proc IEEE Inst Electr Electron Eng 2016;104(2):444–66.
9. Celi LA, Mark RG, Stone DJ, et al. "Big data" in the intensive care unit. Closing the data loop. Am J Respir Crit Care Med 2013;187(11):1157–60.

10. Balogh EP, Miller BT, Ball JR, et al. The diagnostic process. (US): National Academies Press; 2015.
11. Katz A, Chateau D, Enns JE, et al. Association of the social determinants of health with quality of primary care. Ann Fam Med 2018;16(3):217–24.
12. Zheng B, Kwok E, Taljaard M, et al. Decision fatigue in the Emergency Department: how does emergency physician decision making change over an eight-hour shift? Am J Emerg Med 2020;38(12):2506–10.
13. Han PKJ, Klein WMP, Arora NK. Varieties of uncertainty in health care: a conceptual taxonomy. Med Decis Making 2011;31(6):828–38.
14. Delétang G, Ruoss A, Duquenne P-A, et al. Language modeling is compression. arXiv [csLG] 2023.
15. Meissen H, Gong MN, Wong A-KI, et al. The future of critical care: optimizing technologies and a learning healthcare system to potentiate a more humanistic approach to critical care. Crit Care Explor 2022;4(3):e0659.
16. Kamath PS, Wiesner RH, Malinchoc M, et al. A model to predict survival in patients with end-stage liver disease. Hepatology 2001;33(2):464–70.
17. Pisters R, Lane DA, Nieuwlaat R, et al. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. Chest 2010;138(5):1093–100.
18. Tarricone A, Mata KDL, Gee A, et al. A systematic review and meta-analysis of the effectiveness of LRINEC score for predicting upper and lower extremity necrotizing fasciitis. J Foot Ankle Surg 2022;61(2):384–9.
19. Knaak C, Nyvlt P, Schuster FS, et al. Hemophagocytic lymphohistiocytosis in critically ill patients: diagnostic reliability of HLH-2004 criteria and HScore. Crit Care 2020;24(1):244.
20. Zimmerman JE, Kramer AA, McNair DS, et al. Acute Physiology and Chronic Health Evaluation (Apache) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 2006;34(5):1297–310.
21. Johnson AEW, Kramer AA, Clifford GD. A new severity of illness scale using a subset of Acute Physiology and Chronic Health Evaluation data elements shows comparable predictive accuracy. Crit Care Med 2013;41(7):1711–8.
22. Goldfield N. The evolution of diagnosis-related groups (DRGs): from its beginnings in case-mix and resource use theory, to its implementation for payment and now for its current utilization for quality within and outside the hospital. Qual Manag Health Care 2010;19(1):3–16.
23. Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373–83.
24. Elixhauser A, Steiner C, Harris DR, et al. Comorbidity measures for use with administrative data. Med Care 1998;36(1):8–27.
25. Staudacher D, Supady A, Schroth F, et al. Performance of SOFA, SAVE, and SAPS2 score in venoarterial extracorporeal membrane oxygenation (VA-ECMO) for cardiogenic shock and extracorporeal cardiopulmonary resuscitation (eCPR). Resuscitation 2018;130:e5–6.
26. Hick JL, Rubinson L, O'Laughlin DT, et al. Clinical review: allocating ventilators during large-scale disasters–problems, planning, and process. Crit Care 2007;11(3):217.
27. Hong N, Liu C, Gao J, et al. State of the art of machine learning-enabled clinical decision support in intensive care units: literature review. JMIR Med Inform 2022;10(3):e28781.
28. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.

29. Sutton RS, Barto AG. Reinforcement learning. MIT press. Available at: https://mitpress.mit.edu/9780262193986/reinforcement-learning/. [Accessed 21 October 2023].
30. Acosta JN, Falcone GJ, Rajpurkar P, et al. Multimodal biomedical AI. Nat Med 2022;28(9):1773–84.
31. Johnson KB, Wei W-Q, Weeraratne D, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021;14(1):86–93.
32. Elovic A, Pourmand A. MDCalc medical calculator app review. J Digit Imaging 2019;32(5):682–4.
33. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1(5):206–15.
34. Mbakwe AB, Lourentzou I, Celi LA, et al. Fairness metrics for health AI: we have a long way to go. EBioMedicine 2023;90. https://doi.org/10.1016/j.ebiom.2023.104525.
35. Youssef A, Pencina M, Thakur A, et al. External validation of AI models in health should be replaced with recurring local validation. Nat Med 2023;1–2.
36. Quinn TP, Senadeera M, Jacobs S, et al. Trust and medical AI: the challenges we face and the expertise needed to overcome them. J Am Med Inf Assoc 2021;28(4):890–4.
37. Suistomaa M, Kari A, Ruokonen E, et al. Sampling rate causes bias in Apache II and SAPS II scores. Intensive Care Med 2000;26(12):1773–8.
38. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ Digit Med 2020;3:99.
39. Ashana DC, Anesi GL, Liu VX, et al. Equitably allocating resources during crises: racial differences in mortality prediction models. Am J Respir Crit Care Med 2021;204(2):178–86.
40. Miller WD, Han X, Peek ME, et al. Accuracy of the sequential organ failure assessment score for in-hospital mortality by race and relevance to Crisis standards of care. JAMA Netw Open 2021;4(6):e2113891.
41. Fawzy A, Wu TD, Wang K, et al. Racial and ethnic discrepancy in pulse oximetry and delayed identification of treatment eligibility among patients with COVID-19. JAMA Intern Med 2022;182(7):730–8.
42. Wong A-KI, Charpignon M, Kim H, et al. Analysis of discrepancies between pulse oximetry and arterial oxygen saturation measurements by race and ethnicity and association with organ dysfunction and mortality. JAMA Netw Open 2021;4(11):e2131674.
43. Eini-Porat B, Amir O, Eytan D, et al. Tell me something interesting: clinical utility of machine learning prediction models in the ICU. J Biomed Inform 2022;132:104107.
44. Kellogg KC, Sendak M, Balu S. AI on the front lines. MIT Sloan Manag Rev 2022;63(4):44–50. Cambridge.
45. Zhang Y, Szolovits P. Patient-specific learning in real time for adaptive monitoring in critical care. J Biomed Inform 2008;41(3):452–60.
46. Kwok HF, Linkens DA, Mahfouf M, et al. Adaptive ventilator FiO2 advisor: use of non-invasive estimations of shunt. Artif Intell Med 2004;32(3):157–69.
47. Gholami B, Phan TS, Haddad WM, et al. Replicating human expertise of mechanical ventilation waveform analysis in detecting patient-ventilator cycling asynchrony using machine learning. Comput Biol Med 2018;97:137–44.

48. Rehm GB, Han J, Kuhn BT, et al. Creation of a robust and generalizable machine learning classifier for patient ventilator asynchrony. Methods Inf Med 2018;57(4):208–19.
49. Sun M, Baron J, Dighe A, et al. Early prediction of acute kidney injury in critical care setting using clinical notes and structured multivariate physiological measurements. Stud Health Technol Inform 2019;264:368–72.
50. Sanchez-Pinto LN, Khemani RG. Development of a prediction model of early acute kidney injury in critically ill children using electronic health record data. Pediatr Crit Care Med 2016;17(6):508–15.
51. Desautels T, Calvert J, Hoffman J, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Med Inform 2016;4(3):e28.
52. Calvert J, Desautels T, Chettipally U, et al. High-performance detection and early prediction of septic shock for alcohol-use disorder patients. Ann Med Surg (Lond) 2016;8:50–5.
53. Mao Q, Jay M, Hoffman JL, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open 2018;8(1):e017833.
54. Ghosh S, Li J, Cao L, et al. Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns. J Biomed Inform 2017;66:19–31.
55. Bedoya AD, Futoma J, Clement ME, et al. Machine learning for early detection of sepsis: an internal and temporal validation study. JAMIA Open 2020;3(2):252–60.
56. Le S, Pellegrini E, Green-Saxena A, et al. Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS). J Crit Care 2020;60:96–102.
57. Sauthier MS, Jouvet PA, Newhams MM, et al. Machine learning predicts prolonged acute hypoxemic respiratory failure in pediatric severe influenza. Crit Care Explor 2020;2(8):e0175.
58. Tang B, Yuan Y, Yang J, et al. Predicting blood glucose concentration after short-acting insulin injection using discontinuous injection records. Sensors 2022;22(21). https://doi.org/10.3390/s22218454.
59. Frandes M, Timar B, Lungeanu D. A risk based neural network approach for predictive modeling of blood glucose dynamics. Stud Health Technol Inform 2016;228:577–81.
60. Ghazal S, Sauthier M, Brossier D, et al. Using machine learning models to predict oxygen saturation following ventilator support adjustment in critically ill children: a single center pilot study. PLoS One 2019;14(2):e0198921.
61. Yu C, Ren G, Dong Y. Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC Med Inform Decis Mak 2020;20(Suppl 3):124.
62. Sayed M, Riaño D, Villar J. Predicting duration of mechanical ventilation in acute respiratory distress syndrome using supervised machine learning. J Clin Med Res 2021;10(17). https://doi.org/10.3390/jcm10173824.
63. Janssen A, De Waele JJ, Elbers PWG. Towards adequate and automated antibiotic dosing. Intensive Care Med 2023;49(7):853–6.
64. Komorowski M, Celi LA, Badawi O, et al. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018;24(11):1716–20.

65. Srinivasan S, Doshi-Velez F. Interpretable batch IRL to extract clinician goals in ICU hypotension management. AMIA Jt Summits Transl Sci Proc 2020;2020:636–45.
66. Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. Conf Proc IEEE Eng Med Biol Soc 2016;2016:2978–81.
67. Lopez-Martinez D, Eschenfeldt P, Ostvar S, et al. Deep reinforcement learning for optimal critical care pain management with morphine using dueling double-deep Q networks. Conf Proc IEEE Eng Med Biol Soc 2019;2019:3960–3.
68. DeJournett L, DeJournett J. In silico testing of an artificial-intelligence-based artificial pancreas designed for use in the intensive care unit setting. J Diabetes Sci Technol 2016;10(6):1360–71.
69. Van Herpe T, Espinoza M, Haverbeke N, et al. Glycemia prediction in critically ill patients using an adaptive modeling approach. J Diabetes Sci Technol 2007;1(3):348–56.
70. Nguyen M, Jankovic I, Kalesinskas L, et al. Machine learning for initial insulin estimation in hospitalized patients. J Am Med Inf Assoc 2021;28(10):2212–9.
71. Hsieh Y-Z, Su M-C, Wang C-H, et al. Prediction of survival of ICU patients using computational intelligence. Comput Biol Med 2014;47:13–9.
72. Johnson AEW, Mark RG. Real-time mortality prediction in the intensive care unit. AMIA Annu Symp Proc 2017;2017:994–1003.
73. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035.
74. Monteiro F, Meloni F, Baranauskas JA, et al. Prediction of mortality in Intensive Care Units: a multivariate feature selection. J Biomed Inform 2020;107:103456.
75. Silva I, Moody G, Scott DJ, et al. Predicting in-hospital mortality of ICU patients: the PhysioNet/computing in cardiology challenge 2012. Comput Cardiol 2012;39:245–8.
76. Iwase S, Nakada T-A, Shimada T, et al. Prediction algorithm for ICU mortality and length of stay using machine learning. Sci Rep 2022;12(1):1–9.
77. Choi MH, Kim D, Choi EJ, et al. Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records. Sci Rep 2022;12(1):1–11.
78. Abd-Elrazek MA, Eltahawi AA, Abd Elaziz MH, et al. Predicting length of stay in hospitals intensive care unit using general admission features. Ain Shams Eng J 2021;12(4):3691–702.
79. Alghatani K, Ammar N, Rezgui A, et al. Predicting intensive care unit length of stay and mortality using patient vital signs: machine learning model development and validation. JMIR Med Inform 2021;9(5):e21347.
80. Hempel L, Sadeghi S, Kirsten T. Prediction of intensive care unit length of stay in the MIMIC-IV dataset. NATO Adv Sci Inst Ser E Appl Sci 2023;13(12):6930.
81. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023;10(1):1.
82. Rojas JC, Carey KA, Edelson DP, et al. Predicting intensive care unit readmission with machine learning using electronic health record data. Ann Am Thorac Soc 2018;15(7):846–53.
83. Lin Y-W, Zhou Y, Faghri F, et al. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PLoS One 2019;14(7):e0218942.
84. Oeyen S, Vermeulen K, Benoit D, et al. Development of a prediction model for long-term quality of life in critically ill patients. J Crit Care 2018;43:133–8.

85. Moor M, Rieck B, Horn M, et al. Early prediction of sepsis in the ICU using machine learning: a systematic review. Front Med 2021;8:607952.
86. Zale A, Mathioudakis N. Machine learning models for inpatient glucose prediction. Curr Diab Rep 2022;22(8):353–64.
87. Miu T, Joffe AM, Yanez ND, et al. Predictors of reintubation in critically ill patients. Respir Care 2014;59(2):178–85.
88. van de Sande D, van Genderen ME, Huiskens J, et al. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med 2021;47(7):750–60.
89. Johnson A.E.W., Pollard T.J., Mark R.G., Reproducibility in critical care: a mortality prediction case study, Machine learning for healthcare conference, 18–19 Aug 2017;68:361–376.
90. Nazer LH, Zatarah R, Waldrip S, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digit Health 2023;2(6):e0000278.
91. Habib AR, Lin AL, Grant RW. The epic sepsis model falls short—the importance of external validation. JAMA Intern Med 2021;181(8):1040–1.
92. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25(9):1337–40.
93. Faes L, Liu X, Wagner SK, et al. A clinician's guide to artificial intelligence: how to critically appraise machine learning studies. Transl Vis Sci Technol 2020;9(2):7.
94. van de Sande D, Van Genderen ME, Smit JM, et al. Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. BMJ Health Care Inform 2022;29(1). https://doi.org/10.1136/bmjhci-2021-100495.
95. Hassan N, Slight R, Morgan G, et al. Road map for clinicians to develop and evaluate AI predictive models to inform clinical decision-making. BMJ Health Care Inform 2023;30(1):e100784.
96. Ferryman K, Mackintosh M, Ghassemi M. Considering biased data as informative artifacts in AI-assisted health care. N Engl J Med 2023;389(9):833–8.
97. Arbet J, Brokamp C, Meinzen-Derr J, et al. Lessons and tips for designing a machine learning study using EHR data. J Clin Transl Sci 2020;5(1):e21.
98. Sauer CM, Chen L-C, Hyland SL, et al. Leveraging electronic health records for data science: common pitfalls and how to avoid them. The Lancet Digital Health 2022;4(12):e893–8.
99. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021;3(3):199–217.
100. Delgado J, de Manuel A, Parra I, et al. Bias in algorithms of AI systems developed for COVID-19: a scoping review. J bioeth Inq 2022;407–19. https://doi.org/10.1007/s11673-022-10200-z.
101. Drukker K, Chen W, Gichoya J, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. J Med Imaging (Bellingham) 2023;10(6):061104.
102. Gichoya JW, Thomas K, Celi LA, et al. AI pitfalls and what not to do: mitigating bias in AI. Br J Radiol 2023;96(1150):20230023.
103. Nakayama LF, Matos J, Quion J, et al. Unmasking biases and navigating pitfalls in the ophthalmic artificial intelligence lifecycle: a review. arXiv [csCY] 2023.
104. Hegedus EJ, Moody J. Clinimetrics corner: the many faces of selection bias. J Man Manip Ther 2010;18(2):69–73.

105. Yang Y, Zhang H, Katabi D, et al. Change is hard: a closer look at subpopulation shift. arXiv 2023.
106. Yoshida K., Bohn J. Tableone: create “table 1” to describe baseline characteristics. R Package Version n.d.
107. Pollard TJ, Johnson AEW, Raffa JD, et al. tableone: an open source Python package for producing summary statistics for research papers. JAMIA Open 2018;1(1):26–31.
108. Charpignon M-L, Byers J, Cabral S, et al. Critical bias in critical care devices. Crit Care Clin 2023;39(4):795–813.
109. Nijman S, Leeuwenberg AM, Beekers I, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. J Clin Epidemiol 2022;142:218–29.
110. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. BMJ 2016;352:i1981.
111. Delgado-Rodríguez M, Llorca J. Bias. J Epidemiol Community 2004;58(8):635–41.
112. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. arXiv 2017.
113. Futoma J, Simons M, Panch T, et al. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020;2(9):e489–92.
114. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021;385(3):283.
115. Daneshjou R, Smith MP, Sun MD, et al. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. JAMA Dermatol 2021;157(11):1362–9.
116. Young JM, Solomon MJ. How to critically appraise an article. Nat Clin Pract Gastroenterol Hepatol 2009;6(2):82–91.
117. Krittanawong C, editor. *Artificial Intelligence in Clinical Practice: How AI Technologies Impact Medical Research and Clinics*. Amsterdam, Netherlands: Elsevier; 2023.
118. MIT Critical Data. *Secondary analysis of electronic health records*. Berlin, Germany: Springer Nature; 2016. p. 427.
119. Practical steps for building fair AI algorithms. Coursera. Available at: https://www.coursera.org/learn/algorithmic-fairness. [Accessed 6 January 2024].
120. Greene P. Bill gates says the textbook is dying. Is He right? Forbes Magazine 2019.
121. Nabavi Nouri S, Cohen YA, Madhavan MV, et al. Preprint manuscripts and servers in the era of coronavirus disease 2019. J Eval Clin Pract 2021;27(1):16–21.
122. Facher L, Garde D, Silverman E, et al. Stat. Stat. Available at: https://www.statnews.com/. [Accessed 3 November 2023].
123. Magazine series, MIT Technology Review.
124. News. Stanford Institute for human-centered artificial intelligence. Available at: https://hai.stanford.edu/news. [Accessed 3 November 2023].
125. Magazine series, Guardian Technology.
126. Technology. ProPublica. Available at: https://www.propublica.org/topics/technology. [Accessed 3 November 2023].
127. SCCM. Society of critical care medicine (SCCM). Available at: https://sccm.org/Research/Discovery-Research-Network/datascience. [Accessed 3 November 2023].

128. Data science. ESICM. Available at: https://www.esicm.org/groups/data-science/. [Accessed 3 November 2023].
129. Home. BrainX community. Available at: https://brainxai.org/. [Accessed 3 November 2023].
130. WiDS Worldwide. WiDS worldwide. Available at: https://www.widsworldwide.org/. [Accessed 3 November 2023].
131. Sobel J, Almog R, Celi L, et al. How to organise a datathon for bridging between data science and healthcare? Insights from the Technion-Rambam machine learning in healthcare datathon event. BMJ Health Care Inform 2023;30(1). https://doi.org/10.1136/bmjhci-2023-100736.
132. Aboab J, Celi LA, Charlton P, et al. A "datathon" model to support cross-disciplinary collaboration. Sci Transl Med 2016;8(333):333ps8.
133. Luo EM, Newman S, Amat M, et al. MIT COVID-19 Datathon: data without boundaries. BMJ Innov 2021;7(1):231–4.
134. Piza FM, Celi LA, Deliberato RO, et al. Assessing team effectiveness and affective learning in a datathon. Int J Med Inf 2018;112:40–4.
135. Datathon. Society of critical care medicine (SCCM). Available at: https://sccm.org/Research/Discovery-Research-Network/datascience/Datathon. [Accessed 3 November 2023].
136. Lyndon MP, Pathanasethpong A, Henning MA, et al. Measuring the learning outcomes of datathons. BMJ Innovations 2022;8(2). https://doi.org/10.1136/bmjinnov-2021-000747.
137. Zhou S, Skaar DJ, Jacobson PA, et al. Pharmacogenomics of medications commonly used in the intensive care unit. Front Pharmacol 2018;9:1436.
138. Bakker T, Abu-Hanna A, Dongelmans DA, et al. Clinically relevant potential drug-drug interactions in intensive care patients: a large retrospective observational multicenter study. J Crit Care 2021;62:124–30.
139. Moore P, Burkhart K. Adverse drug reactions in the intensive care unit. Critical Care Toxicol 2017;693–739.
140. Wang H, Shi H, Wang N, et al. Prevalence of potential drug - drug interactions in the cardiothoracic intensive care unit patients in a Chinese tertiary care teaching hospital. BMC Pharmacol Toxicol 2022;23(1):39.
141. Charpignon M-L, Carrel A, Jiang Y, et al. Going beyond the means: exploring the role of bias from digital determinants of health in technologies. PLOS Digit Health 2023;2(10):e0000244.
142. Liu J, Li Y, Li J, et al. Sources of automatic office blood pressure measurement error: a systematic review. Physiol Meas 2022;43(9). https://doi.org/10.1088/1361-6579/ac890e.
143. Yadav K, Lewis RJ. Immortal time bias in observational studies. JAMA 2021;325(7):686–7.
144. Andersen SK, Montgomery CL, Bagshaw SM. Early mortality in critical illness - a descriptive analysis of patients who died within 24 hours of ICU admission. J Crit Care 2020;60:279–84.
145. Shintani AK, Girard TD, Eden SK, et al. Immortal time bias in critical care research: application of time-varying Cox regression for observational cohort studies. Crit Care Med 2009;37(11):2939–45.
146. Vail EA, Gershengorn HB, Wunsch H, et al. Attention to immortal time bias in critical care research. Am J Respir Crit Care Med 2021;203(10):1222–9.
147. Wanis KN, Madenci AL, Hao S, et al. Emulating target trials comparing early and delayed intubation strategies. Chest 2023;164(4):885–91.

148. Angwin J, Larson J, Kirchner L, et al. Machine bias. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. [Accessed 22 October 2023].
149. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. J Glob Health 2019;9(2):010318.
150. Iqbal U, Celi LA, Hsu Y-HE, et al. Healthcare artificial intelligence: the road to hell is paved with good intentions. BMJ Health Care Inform 2022;29(1). https://doi.org/10.1136/bmjhci-2022-100650.
151. Teotia K, Jia Y, Woite NL, et al. Variation in monitoring: glucose measurement in the ICU as a case study to preempt spurious correlations. bioRxiv 2023. https://doi.org/10.1101/2023.10.12.23296568.
152. Critical Care Statistics. Society of critical care medicine (SCCM). Available at: https://www.sccm.org/Communications/Critical-Care-Statistics. [Accessed 24 October 2023].
153. Wunsch H, Angus DC, Harrison DA, et al. Comparison of medical admissions to intensive care units in the United States and United Kingdom. Am J Respir Crit Care Med 2011;183(12):1666–73.
154. Angus DC, Shorr AF, White A, et al. Critical care delivery in the United States: distribution of services and compliance with Leapfrog recommendations. Crit Care Med 2006;1016–24.
155. Amann J, Blasimme A, Vayena E, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20(1):310.
156. Center for Devices. Radiological Health. Clinical decision support software - guidance. U.S. Food and Drug Administration. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software. [Accessed 6 January 2024].
157. Jabbour S, Fouhey D, Shepard S, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. JAMA 2023;330(23):2275–84.
158. 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):e745–50.
159. Yi S, Burke C, Reilly A, et al. Designing and developing a digital equity dashboard for the emergency department. J Am Coll Emerg Physicians Open 2023;4(4):e12997.
160. Gallifant J, Kistler EA, Nakayama LF, et al. Disparity dashboards: an evaluation of the literature and framework for health equity improvement. Lancet Digit Health 2023;5(11):e831–9.
161. Harleen Kaur Johal CD. Challenging cognitive biases in the intensive care unit. BMJ | Journal of Medical Ethics 2020. Available at: https://blogs.bmj.com/medical-ethics/2020/07/14/challenging-cognitive-biases-in-the-intensive-care-unit/. [Accessed 6 January 2024].
162. Ganapathi S, Palmer J, Alderman JE, et al. Tackling bias in AI health datasets through the STANDING Together initiative. Nat Med 2022;28(11):2232–3.
163. Chai. Available at: https://www.coalitionforhealthai.org/. [Accessed 7 January 2024].
164. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. Nat Med 2020;26(1):16–7.

165. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;447–53. https://doi.org/10.1126/science.aax2342.
166. Pfohl S, Marafino B, Coulet A, et al. Creating fair models of atherosclerotic cardiovascular disease risk, . Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. New York, NY, USA: Association for Computing Machinery; 2019. p. 271–8.
167. Zink A, Obermeyer Z, Pierson E. Race corrections in clinical algorithms can help correct for racial disparities in data quality. bioRxiv 2023. https://doi.org/10.1101/2023.03.31.23287926.
168. Manski CF, Mullahy J, Venkataramani AS. Using measures of race to make clinical predictions: decision making, patient health, and fairness. Proc Natl Acad Sci U S A 2023;120(35):e2303370120.
169. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020;383(9):874–82.
170. Diao JA, Inker LA, Levey AS, et al. In search of a better equation - performance and equity in estimates of kidney function. N Engl J Med 2021;384(5):396–9.
171. Stevens ER, Caverly T, Butler JM, et al. Considerations for using predictive models that include race as an input variable: the case study of lung cancer screening. J Biomed Inform 2023;147:104525.
172. Hammond G, Johnston K, Huang K, et al. Social determinants of health improve predictive accuracy of clinical risk models for cardiovascular hospitalization, annual cost, and death. Circ Cardiovasc Qual Outcomes 2020;13(6):e006752.
173. Khor S, Haupt EC, Hahn EE, et al. Racial and ethnic bias in risk prediction models for colorectal cancer recurrence when race and ethnicity are omitted as predictors. JAMA Netw Open 2023;6(6):e2318495.