REVIEW

Open Access

Artificial Intelligence in Critical Care Medicine

Joo Heung Yoon[1*], Michael R. Pinsky[2] and Gilles Clermont[2]

Abstract

This article is one of ten reviews selected from the Annual Update in Intensive Care and Emergency Medicine 2022. Other selected articles can be found online at https://www.biomedcentral.com/collections/annualupdate2022. Further information about the Annual Update in Intensive Care and Emergency Medicine is available from https://link.springer.com/bookseries/8901.

Introduction

The clinical outcome of critically ill patients has improved significantly and to an unprecedented level as standards of care have improved [1]. However, conventional critical care practice still has limitations in understanding the complexity of acuity, handling extreme individual heterogeneity, anticipating deterioration, and providing early treatment strategies before decompensation. Critical care medicine has seen the arrival of advanced monitoring systems and various non-invasive and invasive treatment strategies to provide timely intervention for critically-ill patients. Whether the mergence of such systems represents the next step in improving bedside care is an existing, yet unproven possibility.

The simplified concept of artificial intelligence (AI) is to allow computers to find patterns in a complex environment of multidomain and multidimensional data, with the prerequisite that such patterns would not be recognized otherwise. Previously, applying the concept in real life required a tremendous amount of computing time and resources. This could only be done in limited fields, including physics or astronomy. However, with recent exponential growth in computing power and portability, the power of AI became available to many fields, including critical care medicine where data are vast, abundant, and complex [2]. More and more clinical investigations are being performed using AI-driven models to leverage the data in the intensive care unit (ICU), but our understanding of the power and utility of AI in critical care medicine is still quite rudimentary. In addition, there are many obstacles and pitfalls for AI to overcome before becoming a core component of our daily clinical practice.

In this chapter, we seek to introduce the roles of AI with the potential to change the landscape of our conventional practice patterns in the ICU, describe its current strengths and pitfalls, and consider future promise for critical care medicine.

Applications of AI in Critical Care
Disease Identification

Oftentimes, finding the root cause of clinical deterioration from the exhaustive list of differential diagnoses is challenging, because of the insidious characteristic of early disease progression or the presence of co-existing conditions masking the main problem (Fig. 1). More than anything, the underlying context should be deciphered correctly, often a challenging task. For example, pulmonary infiltrates cannot simply be assumed to represent excessive alveolar fluid. They could indicate pulmonary edema from a cardiac cause, pleural effusion, parapneumonic fluid from inflammation or infection, or in some cases collections of blood as a result of trauma.

*Correspondence: yoonjh@upmc.edu
[1] Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA
Full list of author information is available at the end of the article
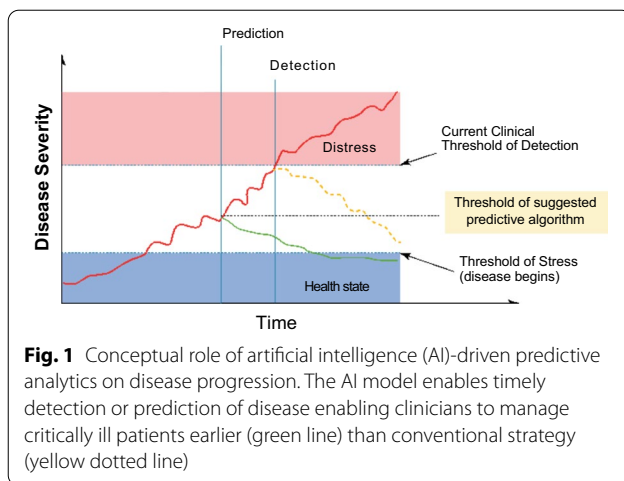
Fig. 1 Conceptual role of artificial intelligence (AI)-driven predictive analytics on disease progression. The AI model enables timely detection or prediction of disease enabling clinicians to manage critically ill patients earlier (green line) than conventional strategy (yellow dotted line)

Without clinical context and further testing, adequate and timely management could be delayed. AI could assist in such cases by obtaining a more precise diagnosis, given advanced text and image processing capability. The presence of congestive heart failure (CHF) could be differentiated from other causes of lung disease using a machine learning model [3], and amounts of pulmonary edema secondary to the CHF could be quantified with semi-supervised machine learning using a variational autoencoder [4]. During the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic, imaging data from patients admitted to hospital were processed to detect coronavirus disease 2019 (COVID-19) using an AI model [5]. With recent efforts in image segmentation and quantification of lesions by convolutional neural networks, a type of algorithm particularly apt at interpreting images, the presence of traumatic brain injury (TBI) on head computed tomography (CT) could be evaluated with higher accuracy than manual reading [6]. Similarly, traumatic hemoperitoneum was quantitatively visualized and measured using a multiscale residual neural network [7].

Disease Evolution Prediction

Disease detection and prediction of disease evolution is one of the holy grails for critically ill patients. Given that the disease process is a continuum, instability of clinical condition can take various paths, even prior to ICU admission [8]. In a series of step-down unit patients who experienced cardiorespiratory instability (defined as hypotension, tachycardia, respiratory distress, or a desaturation event using numerical thresholds), a dynamic model using a random forest classification showed that a personalized risk trajectory predicted deterioration 90 min ahead of the crisis [9].

In the ICU, rapid clinical deterioration is common, and the result can be irreversible and even lead to mortality if detected late. Thus, efforts are being made to predict such hemodynamic decompensation. Tachycardia, one of the most commonly observed deviations from normality prior to shock, was predicted 75 min prior to development using a normalized dynamic risk score trajectory with a random forest model [10]. Hypotension, a manifestation of shock, was also predicted in the operating room [11]. The utility of a machine learning model in reducing intraop-erative hypotension was further confirmed in a randomized controlled trial in patients having intermediate and high-risk surgery, with hypotension occurring in 1.2% of patients managed with an AI-driven intervention, versus 21.5% using conventional methods [12]. Hypotension events have also been predicted in the ICU where vital sign granularities are lower and datasets contain more noise. Using electronic health record (EHR) as well as physiologic numeric vital sign data, clinically relevant hypotension events were predicted with a random forest model, achieving a sensitivity of 92.7%, with the average area under the curve (AUC) of 0.93 at 15 min before the actual event [13].

Hypoxia and respiratory distress have also been major targets for prediction, the roles of which have expanded during the recent coronavirus pandemic. In the first few months of the pandemic, AI-driven models were used to predict the progression of COVID-19, using imaging, biological, and clinical variables [14]. Cardiac arrest has also been predicted using an electronic Cardiac Arrest Risk Triage (eCART) score from EHR data, showing non-inferior scores compared to conventional early warning scoring systems [15]. Sepsis has also been predicted, with an AUC of 0.85 using Weibull-Cox proportional hazards model on high-resolution vital sign time series data and clinical data [16]. Other clinical outcomes may be predicted using AI models, including mortality after TBI [17] or mortality of COVID-19 patients with different risk profiles [18].
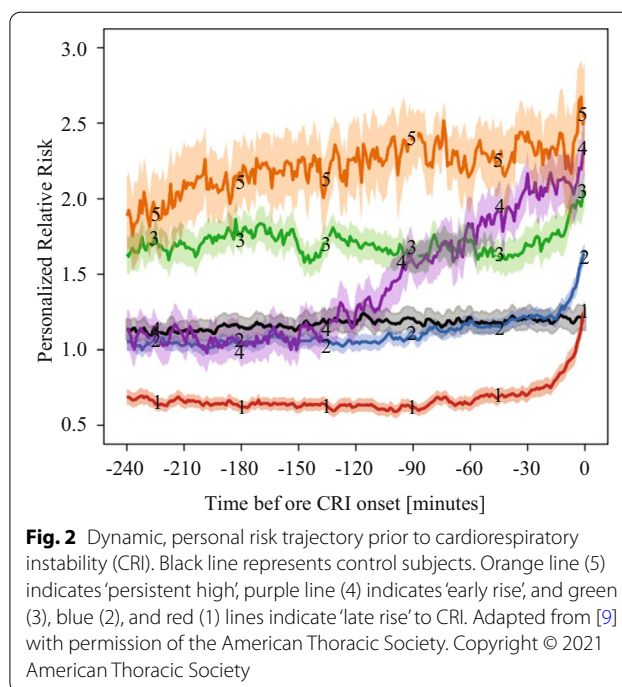
Disease Phenotyping

Critical illness is complex and its manifestations can rarely be reduced to typical presentations. Rather, critical illness manifests in a lot of different ways (inherent heterogeneity), and carries significant risks for organ dysfunction that can subsequently complicate the underlying disease process or recovery processes. Such syndromes should not be treated blindly without careful consideration of underlying etiologies or clinical conditions for a given individual. Moreover, the complex critical states change over time, such that clinicians cannot rely on assessments from even a few hours earlier. Yet, evidence-based guidelines should be followed whenever these

Yoon *et al. Critical Care*   (2022) 26:75

Page 3 of 9

exist. With its strong capability of pattern recognition from complex data, AI could delineate distinctive phenotypes or endotypes that could reflect influences from the critical state and hence open up avenues to personalize management, integrated into existing guidelines.

Sepsis, one of the most common ICU conditions, is a highly heterogenous syndrome, and has been a favorite target of AI algorithms. Recently, using different clinical trial cohorts, sepsis was clustered into four phenotypes (α, 3, γ, and 6) by consensus K-means clustering, a type of unsupervised machine learning model. The phenotypes had distinctive demographic characteristics, different biochemical presentations, correlated to host-response patterns, and were eventually associated with different clinical outcomes [19]. Such phenotypes are of a descriptive nature, are useful in describing case-mix, and could represent targets for predictive enrichment of clinical trials. However, they are not at this juncture based in mechanism and thus are not therapeutically actionable. Nevertheless, further explorations using richer data might allow a greater degree of actionability.

In the acute respiratory distress syndrome (ARDS), latent class analysis (LCA) revealed two subtypes (hypo- and hyper-inflammatory subtypes) linked with different clinical characteristics, treatment responses, and clinical outcomes [20]. A parsimonious model was developed and achieved similar performance to the initial LCA using a smaller set of classifier variables (interleukin [IL]-6, -8), protein C, soluble tumor necrosis factor (TNF) receptor 1, bicarbonate, and vasopressors). This result was validated in a secondary analysis of three different randomized clinical trials [21]. This machine learning-driven ARDS phenotyping has expanded our knowledge in assessing and treating complex disease, and become one of the criteria for predictive enrichment of future clinical trials.

Dynamic phenotyping for prediction of clinical deterioration can be performed on time series data. Using analysis of 1/20 Hz granular physiologic vital sign data, several unique phenotypes, including persistently high, early onset, and late onset deterioration, were identified prior to overt cardiorespiratory deterioration, using K-means clustering (Fig. 2) [9]. Time-series of images can be clustered for dynamic phenotyping, as performed using transesophageal echocardiographic monitoring images in patients with septic shock: using a hierarchical clustering method, septic shock was clustered into three cardiac deterioration patterns and two responses to interventions, which were linked with clinical outcome with different day 7 and ICU mortality rates [22].

Fig. 2 Dynamic, personal risk trajectory prior to cardiorespiratory instability (CRI). Black line represents control subjects. Orange line (5) indicates 'persistent high', purple line (4) indicates 'early rise', and green (3), blue (2), and red (1) lines indicate 'late rise' to CRI. Adapted from [9] with permission of the American Thoracic Society. Copyright © 2021 American Thoracic Society

Guiding Clinical Decisions

For a complex problem, one-size-fits-all solutions do not work well. Over the last decade, research has failed to improve the outcome of septic shock with different treatment guidelines [23, 24]. The extreme heterogeneity of septic shock, various underlying conditions, and different host-responses could be at least partially addressed by AI to provide individualized solutions using reinforcement learning. The algorithm in reinforcement learning is designed to detect numerous variables in a given state to build an action model, which then learns from the reward or penalty from the results of the action. Applying this to the sepsis population, reinforcement learning could provide optimal sequential decision-making solutions for sepsis treatment, showcasing the potential impact of AI to generate personalized solutions [25]. In patients receiving mechanical ventilation, time series data with 44 features were extracted and reinforcement learning (Markov Decision Process) resulted in better results compared to physicians' standard clinical care, with target outcomes of 90-day and ICU mortality [26]. These examples demonstrate the role AI may have in guiding important decision-making for critically ill patients. The notion of AI's therapeutic utility could be more pronounced and provocative in different clinical environments, such as critical case scenarios in remote areas where clinicians are not available and patient transfer is not possible, or resource-limited settings where treatment options are limited. Because the optimality

Yoon *et al. Critical Care* (2022) 26:75

Page 4 of 9

of such treatment recommendations is computed from retrospective and observational datasets, it is imperative that recommendation sequences or policies originating from such AI systems be fully analyzed and then tested prospectively before clinical implementation.

Implementation

An important consideration in successful deployment of AI at the bedside is system usability and trustworthiness. Deployment of such systems should involve all stakeholders, including clinicians and patients (end users), researchers (producers), and hospital administrators (logistics and management). In a specific research project, an implementation strategy entails creating models with adequate amounts of information (not a 'flood' of information with alarms), delivered with understandable (interpretable) logics, and placed on a visually appealing vehicle or dashboard—a graphic user interface. These systems, when deployed as alerting tools, must be accurate enough and parsimonious enough to prevent alarm fatigue, which leads to delays in detecting, and intervening for, developing crises [27]. In recent work on prediction of hypotension in the ICU, researchers found that AI-generated alerts could be reduced tenfold while maintaining sensitivity, when they used a stacked random forest model, or a model checking on another model before generating alerts [13].

Understanding AI-derived predictions and recommendations is arguably an important component of AI acceptance at the bedside. Although complex models can be thought of as 'black-boxes', an enormous effort is underway to enhance model interpretability and explainability. For example, in a recent report on hypoxia detection, researchers adopted concepts from game theory to differentially weight predictive physiological readouts during surgery, as an attempt to interpret the clinical drivers of hypoxic alarming from an AI system [28]. Creating the graphic user interface is necessary not only for the AI output to be delivered to the bedside, but also to improve hospital workflow and alleviate nursing burden. As shown in recent work, deep learning could be used to analyze fiducial points from the face, postures, and action of patients, and from environmental stimuli to discriminate delirious and non- delirious ICU patients [29]. Future ICU design should embrace the functionalities of AI solutions to enable clinicians to react earlier to any potential deterioration, and researchers to build models that perform better using more comprehensive data, and presented in such a way that it will be readily available, highly accurate, and trusted by bedside clinicians.

Pitfalls of AI in Critical Care

As much as the power of the AI model changes the current landscape of data analysis and plays an important role in assisting early diagnosis and management, there are many road blocks that should not be overlooked when introducing AI models for critically ill patients.

Explainability and Interpretability

Many AI models have complex layers of nodes that enable the characteristics of input data to be more meaningful in revealing hidden patterns. While the model may produce seemingly accurate output through that process, oftentimes the rationale of the computation cannot be provided to the end users. In the clinical environment, this can create strong resistance to accepting AI models into daily practice, as clinicians fear that performing unnecessary interventions or changing a treatment strategy without supporting scientific evidence could easily violate the first rule of patient care, *primum non nocere*. In critical care medicine, such a move could be directly and rapidly associated with mortality. On the other hand, many novel treatments did not have enough evidence when first introduced in the history of medicine, and 'black-box' models do not need to be completely deciphered, advocating the use of inherently interpretable AI [30]. Another recent approach argues that providing detailed methodologies for the model validation, robustness of analysis, history of successful/unsuccessful implementations, and expert knowledge could alleviate epistemological and methodological concerns and gain reliability and trust [31].

Multiple efforts have been introduced to overcome the complexity of deep learning models. Explaining feature contribution on a dynamic time series dataset became possible by leveraging game theory into measuring feature importance, when predicting near-term hypoxic events during surgery [32]. In that report, contributing features explained by SHapley Additive exPlanation (SHAP) showed consistency with the literature and prior knowledge from anesthesiologists for upcoming hypoxia risks. Moreover, anesthesiologists were able to make better clinical decisions to prevent intraoperative hypoxia when assisted by the explainable AI model.

Lack of Robustness

The readiness of AI for the real-life clinical environment is limited by the lack of adequate clinical experiments and trials, with a disappointingly low rate of reproducibility and prospective analyses. In a recent review of 172 AI-driven solutions created from routinely collected chart data, the clinical readiness level for AI was low. In that study, the maturity of the AI was classified into nine stages corresponding to real world application [33].

Yoon *et al. Critical Care*   (2022) 26:75

Page 5 of 9

Strikingly, around 93% of all analyzed articles remained below stage 4, with no external validation process, and only 2% of published studies had performed prospective validation. Thus, current AI models in critical care medicine have largely been generated using retrospective data, without external validation or prospective evaluation.

Reproducibility of AI solutions is not guaranteed and no clear protocols exist to examine this thoroughly. As mentioned above, AI solutions already have limitations in terms of data openness and almost inexplicable algorithmic complexity, so the lack of reproducibility on top of these factors could significantly impact the fidelity of the AI model. A recent study attempted to reproduce 38 experiments for 28mortality prediction projects using the Medical Information Mart for Intensive Care (MIMIC-III) database, and reported large sample size differences in about a half the experiments [34]. This problem highlights the importance of accurate labeling, understanding the clinical context to create the study population, as well as precise reporting methods including data pre-processing and featurization.

Adherence to reporting standards and risks of bias is also sub-optimal, as a study that analyzed 81 non-randomized and 10 randomized trials using deep learning showed only 6 of 81 non-randomized studies had been tested in a real-world clinical setting and 72% of studies showed high risks of bias [35]. Hence, considering the scientific rigor of conventional randomized controlled trials needed to prove scientific hypothesis, the maturity and robustness of AI-driven models would be even less convincing for everyday practice.

More complex and sophisticated AI models, like reinforcement learning are also not free from challenges, as such intricate models require a lot of computational resources and are difficult to test on patients in order to train or test the models in a clinical environment. Inverse reinforcement learning, which infers information about rewards, could be a new model-agnostic reinforcement learning approach to constructing decision-making trajectories, because this approach alleviates the stress of manually designing a reward function [36]. With those algorithmic advances, decision-assisting engines can be more robust and reliable when input data varies, which may be a great asset to critical care data science where work is conducted with enormous quantity and extreme heterogeneity of data.

Ethical Concerns

Use of AI in critical care is still a new field to most researchers and clinicians. We will not really appreciate what ethical issues we will encounter until AI becomes more widely used and apparent in the development pipeline and bedside applications. However, given the nature of AI characteristics and current AI-driven solutions, a few aspects can be discussed to look around the corner into likely ethical dilemmas of AI models in critical care. The first issue is in data privacy and sharing. Innovation in data science allows us to collect and manipulate data to find hidden patterns, during which course collateral data leakage could pose threats, especially in its pre-processing and in external validation steps towards generalization. It is very hard to remove individual data points from the dataset once they are already being used by the AI model. De-identification and parallel/distributive computing could provide some solutions to data management, and novel models, including federated learning, might minimize data leakage and potentially speed up the multicenter validation process.

A second issue in ethics is safety of the AI model at the bedside. To semi-quantitatively describe the safety of the model, the analogous maturity metric used by self-driving cars was used for clinical adaptability of AI-driven solutions, with 6 levels [37]: 0 (no automation) to 2 (partial automation) represent situations where the human driver monitors the environment; 3 (conditional automation) to 5 (full automation) represent situations where the system is monitoring the environment rather than any human involvement. According to this scale, if used in real life, most of the AI-driven solutions developed would fall into categories 1 or 2. This concept signifies that the safety and accountability of the AI model cannot be blindly guaranteed, and decision making by clinicians remains an integral part of patient care. Also, the autonomy of individual patients has never been more important, including generating informed consent or expressing desire to be treated in life-threatening situations—here the AI recommendations might not be aligned with those of the patient. Recognition of such ethical issues and preparing for potential solutions to overcome limitations of AI, as well as understanding more about patient perspectives could allow researchers and clinicians to develop more practical and ethical AI solutions.

Future Tasks of AI in Critical Care

Data De-identification/Standardization/Sharing Strategy

Like any other clinical research, AI solutions need validation from many different angles. External validation, which uses input data from other environments, is one of the most common ways to generalize a model. Although external validation and prospective study designs certainly require collaborative data pools and concerted efforts, creating such a healthy ecosystem for AI research in critical care demands considerable groundwork.

De-identification of the healthcare data is probably the first step to ascertain data privacy and usability. The Society of Critical Care Medicine (SCCM)/European Society

Yoon *et al. Critical Care*   (2022) 26:75

Page 6 of 9

of Intensive Care Medicine (ESICM) Joint Data Science Task Force team has published the process to create a large-scale database from different source databases, including the following steps: (1) using an anonymization threshold, personal data are separated from anonymous data; (2) iterative, risk-based process to de-identify personal data; (3) external review process to ensure privacy and legal considerations to abide within the European General Data Protection Regulation (GDPR) [38]. Such a de-identification process would ascertain safe data transfer and could further facilitate high-quality AI model training.

Other important groundwork for the multi-center collaboration is data standardization. Individual hospital systems have developed numerous different data labeling strategies in different EHR layers. Even within the same hospital system, small discrepancies, including the number of decimals, commonly used abbreviations, and data order within the chart, could be stumbling blocks for systematic data standardization. In addition, data with higher granularity, including physiologic waveform data, are even harder to standardize, as there are no distinctive labels to express the values in a structured way. To address this, international researchers have developed a standardized format to facilitate efficient exchange of clinical and physiologic data [39]. In this Hierarchical Data Format-version 5 (HDF5)-based critical care data exchange format, multiparameter data could be stored, compressed, and streamed real time. This type of data exchange format would allow integration of other types of large-scale datasets as well, including imaging or genomics.

While one cannot completely remove the data privacy and governance concerns, rapid collaboration can be facilitated when those are less of an issue. An example is federated learning, where models can be designed to be dispatched to local centers for training, instead of data from participating centers collected to one central location for model training. While the data are not directly exposed to the outside environment, the model could still be trained by outside datasets with comparable efficacy and performance [40]. Federated learning could be even more useful when the data distribution among different centers is imbalanced or skewed, demonstrating the real-world collaboration environment [41]. A comprehensive federated learning project was performed during the COVID-19 pandemic. Across the globe, 20 academic centers collaborated to predict clinical outcomes from COVID-19 by constructing federated learning within a strong cloud computing system [42]. During the study phase, researchers developed an AI model to predict the future oxygen requirements of patients with symptomatic COVID-19 using chest X-ray data, which was then

dispatched to participating hospitals. The trained model was calibrated with shared partial-model weights, then the averaged global model was generated, while privacy was preserved in each hospital system. In that way, the AI model achieved an average AUC > 0.92 for predicting 24- to 72-h outcomes. In addition, about a 16% improvement in average AUC, with a 38% increase in generalizability was observed when the model was tested with federated learning compared to the prediction model applied to individual centers. This report exemplifies the potential power of a federated learning-based collaborative approach, albeit the source data (chest X-ray and other clinical data) are relatively easy to standardize for the federated learning system to work on.

Novel AI Models and Trial Designs

Labeling target events for AI models is a daunting, labor-intensive task, and requires a lot of resources. To make the task more efficient, novel AI models, such as weakly supervised learning, have been introduced. Weakly supervised learning can build desired labels with only partial participation of domain experts, and may potentially preserve resource use. One example was provided by performing weakly supervised classification tasks using medical ontologies and expert-driven rules on patients visiting the emergency department with COVID-19 related symptoms [43]. When ontology-based weak supervision was coupled with pretrained language models, the engineering cost of creating classifiers was reduced more than for simple weakly supervised learning, showing an improved performance compared to a majority vote classifier. The results showed that this AI model could make unstructured chart data available for machine learning input, in a short period of time, without an expert labeling process in the midst of a pandemic. Future clinical trials could also be designed with AI models, especially to maximize benefits and minimize risks to participants, as well as to make the best use of limited resources. One example of such an innovative design is the REMAP-CAP (Randomized Embedded Multifactorial Additive Platform for Community-Acquired Pneumonia), which adopted a Bayesian inference model. In detail, this multicenter clinical trial allows randomization with robust causal inference, creates multiple intervention arms across multiple patient subgroups, provides response-adaptive randomization with preferential assignment, and provides a novel platform with perpetual enrollment beyond the evaluation of the initial treatments [44]. The platform, initially developed to identify optimal treatment for community-acquired pneumonia, continued to enroll throughout the COVID-19 pandemic, and has contributed to improved survival among critically ill COVID-19 patients [45, 46].

Yoon *et al. Critical Care*   (2022) 26:75

Page 7 of 9

Real-Time Application

To establish a valuable AI system in the real-life setting, the model should be able to deliver important information in a timely manner. In critically ill patients, the feedback time should be extremely short, sometimes less than a few minutes. Prediction made too early would have enough time to formulate the model, but have less predictive power, and prediction made very close to target events would have higher performance, with no time to curate the input data and run the model for output.

To be used in the real-life environment, a real time AI model should be equipped with a very fast data pre-processing platform, and able to parsimoniously featurize to update the model with new input data simultaneously. The output should also be delivered to the bedside rapidly. In that strict sense of real time, almost no clinical studies have accomplished real-time prediction. A few publications claim real-time prediction, but most of them used retrospective data, and failed to demonstrate continuous real-time data pre-processing without time delay. Using a gradient boosting tree model, one study showed dynamic 'real-time' risks of the onset of sepsis from a large retrospective dataset. The duration of ICU stay was divided into three periods (0–9 h, 10–49 h, and more than 50 h of ICU stay), partitioned to reflect different sepsis onset events, and resulted in different utility scores in each phase [47]. While this provides valuable information on different performances of the AI model for different durations of ICU stay, use of continuous real-time pre-processing without a time delay was not demonstrated. Another study using a recurrent neural network on postoperative physiologic vital sign data produced a high positive predictive value of 0.90 with sensitivity of 0.85 in predicting mortality, and was superior to the conventional metric to predict mortality and other complications [48]. The study also showed that the predictive difference between the AI solution and conventional methods was evident from the beginning of the ICU stay. However, prediction from the earliest part of the ICU stay also does not qualify as true real-time prediction. Although this is a challenging task for current technology, application of the real- time AI model to the critical care environment could yield significant benefit in downstream diagnostic or therapeutic options without time delay.

Quality Control After Model Deployment

Once the AI model achieves high performance and is deemed to be useful in a real-life clinical setting, implementation strategies as well as quality assessment efforts should follow. Anticipating such changes in the clinical/administrative landscape, the National Academy of Medicine of the United States has published a white paper on AI use in healthcare, in which the authors urge the development of guidelines and legal terms for safer, more efficacious, and personalized medicine [49]. In particular, for the maturity of AI solutions and their integration with healthcare, the authors suggested addressing implicit and explicit bias, contextualizing a dialogue of transparency and trust, developing and deploying appropriate training and educational tools, and avoiding over-regulation or over-legislation of AI solutions.

Conclusion

Rapid development and realization of AI models secondary to an unprecedented increase in computing power has invigorated research in many fields, including medicine. Many research topics in critical care medicine have employed the concept of AI to recognize hidden disease patterns among the extremely heterogeneous and noise-prone clinical datasets. AI models provide useful solutions in disease detection, phenotyping, and prediction that might alter the course of critical diseases. They may also lead to optimal, individualized treatment strategies when multiple treatment options exist. However, at the current stage, development and implementation of AI solutions face many challenges. First, data generalization is difficult without proper groundwork, including de-identification and standardization. Second, AI models are not robust, with sub-optimal adherence to reporting standards, a high risk of bias, lack of reproducibility, and without proper external validation with open data and transparent model architecture. Third, with the nature of the obscurity and probabilistic approach, AI models could lead to unforeseen ethical dilemmas. For the successful implementation of AI into clinical practice in the future, collaborative research efforts with plans for data standardization and sharing, advanced model development to ascertain data security, real-time application, and quality control are required.

Yoon *et al. Critical Care*   (2022) 26:75

Page 8 of 9

Declarations

Author details
[1]Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [2]Department of Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA, USA.

References

1. Zimmerman JE, Kramer AA, Knaus WA. Changes in hospital mortality for United States intensive care unit admissions from 1988 to 2012. Crit Care. 2013;17:R81.
2. Yoon JH, Pinsky MR. Predicting adverse hemodynamic events in critically ill patients. Curr Opin Crit Care. 2018;24:196–203.
3. Seah JCY, Tang JSN, Kitchen A, Gaillard F, Dixon AF. Chest radiographs in congestive heart failure: visualizing neural network learning. Radiology. 2019;290:514–22.
4. Horng S, Liao R, Wang X, Dalal S, Golland P, Berkowitz SJ. Deep learning to quantify pulmonary edema in chest radiographs. Radiol Artif Intell. 2021;3:e190228.
5. Li L, Qin L, Xu Z, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. Radiology. 2020;296:E65-e71.
6. Monteiro M, Newcombe VFJ, Mathieu F, et al. Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: an algorithm development and multicentre validation study. Lancet Digit Health. 2020;2:e314–22.
7. Dreizin D, Zhou Y, Fu S, et al. A multiscale deep learning method for quantitative visualization of traumatic hemoperitoneum at CT: assessment of feasibility and comparison with subjective categorical estimation. Radiol Artif Intell. 2020;2:e190220.
8. Vincent JL. The continuum of critical care. Crit Care. 2019;23(Suppl 1):122.
9. Chen L, Ogundele O, Clermont G, Hravnak M, Pinsky MR, Dubrawski AW. Dynamic and personalized risk forecast in step-down units. Implications for monitoring paradigms. Ann Am Thorac Soc. 2017;14:384–91.
10. Yoon JH, Mu L, Chen L, et al. Predicting tachycardia as a surrogate for instability in the intensive care unit. J Clin Monit Comput. 2019;33:973–85.
11. Wijnberge M, Geerts BF, Hol L, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoper-ative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. JAMA. 2020;323:1052–60.
12. Joosten A, Rinehart J, Van der Linden P, et al. Computer-assisted individualized hemodynamic management reduces intraoperative hypotension in intermediate- and high-risk surgery: a randomized controlled trial. Anesthesiology. 2021;135:258–72.
13. Yoon JH, Jeanselme V, Dubrawski A, Hravnak M, Pinsky MR, Clermont G. Prediction of hypotension events with physiologic vital sign signatures in the intensive care unit. Crit Care. 2020;24:661.
14. Lassau N, Ammari S, Chouzenoux E, et al. Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. Nat Commun. 2021;12:634.
15. Bartkowiak B, Snyder AM, Benjamin A, et al. Validating the electronic cardiac arrest risk triage (eCART) score for risk stratification of surgical inpatients in the postoperative setting: retrospective cohort study. Ann Surg. 2019;269:1059–63.
16. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med. 2018;46:547–53.
17. Raj R, Luostarinen T, Pursiainen E, et al. Machine learning-based dynamic mortality prediction after traumatic brain injury. Sci Rep. 2019;9:17672.
18. Banoei MM, Dinparastisaleh R, Zadeh AV, Mirsaeidi M. Machine-learning-based COVID-19 mortality prediction model and identification of patients at low and high risk of dying. Crit Care. 2021;25:328.
19. Seymour CW, Kennedy JN, Wang S, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA. 2019;321:2003–17.
20. Calfee CS, Delucchi KL, Sinha P, et al. Acute respiratory distress syndrome subphenotypes and differential response to simvastatin: secondary analysis of a randomised controlled trial. Lancet Respir Med. 2018;6:691–8.
21. Sinha P, Delucchi KL, McAuley DF, O'Kane CM, Matthay MA, Calfee CS. Development and validation of parsimonious algorithms to classify acute respiratory distress syndrome phenotypes: a secondary analysis of randomised controlled trials. Lancet Respir Med. 2020;8:247–57.
22. Geri G, Vignon P, Aubry A, et al. Cardiovascular clusters in septic shock combining clinical and echocardiographic parameters: a post hoc analysis. Intensive Care Med. 2019;45:657–67.
23. Rivers E, Nguyen B, Havstad S, et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. N Engl J Med. 2001;345:1368–77.
24. Rowan KM, Angus DC, Bailey M, et al. Early, goal-directed therapy for septic shock—a patient-level meta-analysis. N Engl J Med. 2017;376:2223–34.
25. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24:1716–20.
26. Peine A, Hallawa A, Bickenbach J, et al. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. NPJ Digit Med. 2021;4:32.
27. Hravnak M, Pellathy T, Chen L, et al. A call to alarms: current state and future directions in the battle against alarm fatigue. J Electrocardiol. 2018;51:S44–8.
28. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digit Health. 2020;2:e179–91.
29. Davoudi A, Malhotra KR, Shickel B, et al. Intelligent ICU for autonomous patient monitoring using pervasive sensing and deep learning. Sci Rep. 2019;9:8020.
30. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206–15.
31. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics. 2021;47:329–35.
32. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2:749–60.
33. Fleuren LM, Thoral P, Shillan D, Ercole A, Elbers PWG. Machine learning in intensive care medicine: ready for take-off? Intensive Care Med. 2020;46:1486–8.
34. Johnson AEW, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. PMLR. 2017;68:361–76.
35. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689.
36. Fu J, Luo K, Levine S. Learning robust rewards with adversarial inverse reinforcement learning. 2017. arXiv:1710.11248.
37. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44–56.
38. Thoral PJ, Peppink JM, Driessen RH, et al. Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: the Amsterdam university medical centers database (AmsterdamUMCdb) example. Crit Care Med. 2021;49:e563–77.
39. Laird P, Wertz A, Welter G, et al. The critical care data exchange format: a proposed flexible data standard for combining clinical and high-frequency physiologic data in critical care. Physiol Meas. 2021;42:065002.

Yoon *et al. Critical Care*   (2022) 26:75

Page 9 of 9

40. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3:119.
41. Lee GH, Shin SY. Federated learning on clinical benchmark data: performance assessment. J Med Internet Res. 2020;22:e20891.
42. Dayan I, Roth HR, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med. 2021;27:1735–43.
43. Fries JA, Steinberg E, Khattar S, et al. Ontology-driven weak supervision for clinical entity classification in electronic health records. Nat Commun. 2021;12:2017.
44. Angus DC, Berry S, Lewis RJ, et al. The REMAP-CAP (randomized embedded multifactorial adaptive platform for community-acquired pneumonia) study. Rationale and design. Ann Am Thorac Soc. 2020;17:879–91.
45. Angus DC, Derde L, Al-Beidh F, et al. Effect of hydrocortisone on mortality and organ support in patients with severe COVID-19: the REMAP-CAP COVID-19 corticosteroid domain randomized clinical trial. JAMA. 2020;324:1317–29.
46. Gordon AC, Mouncey PR, Al-Beidh F, et al. Interleukin-6 receptor antagonists in critically ill patients with Covid-19. N Engl J Med. 2021;384:1491–502.
47. Li X, Xu X, Xie F, et al. A time-phased machine learning model for real-time prediction of sepsis in critical care. Crit Care Med. 2020;48:e884–8.
48. Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med. 2018;6:905–14.
49. Matheny ME, Whicher D, Thadaney IS. Artificial intelligence in health care: a report from the National Academy of Medicine. JAMA. 2020;323:509–10.

Publisher's Note