Critical Care Explorations
June 2025 • Volume 7 • Number 6 • e1268

EDITORIAL

OPEN

Predictions With a Purpose: Elevating Standards for Clinical Modeling Research

Patrick G. Lyons, MD, MSc

KEYWORDS: artificial intelligence; community-acquired pneumonia; hospital outcomes; medical informatics; predictive modeling

Artificial intelligence (AI) predictive models have been proposed—and in many cases are already being adopted—as potential solutions to numerous challenges in the ICU (1). Sepsis care and high-risk medication management, for example, have benefited from thoughtfully implemented AI-based clinical decision support systems (2, 3). For every successful implementation, however, scores of predictive tools never advance beyond retrospective validation (4). This leaky pipeline often results from poorly defined use cases (e.g., predictions that do not inform clinical action), inadequate reporting, and methodological shortcomings (5).

It is essential, therefore, to reconcile the promise of AI predictive modeling with an ongoing commitment to rigor, reproducibility, and novelty in research. This imperative is particularly relevant for the Society of Critical Care Medicine's family of journals (6). Indeed, *Critical Care Explorations* has devoted an entire category of articles specifically to predictive modeling research, including AI approaches.

A notable contribution to this category of research appears in this compendium of *Critical Care Explorations*. In a large multicenter cohort, Chen et al (7) demonstrate that AI models can accurately predict specific critical care interventions in hospitalized patients with community-acquired pneumonia (CAP). The authors used clinical data from CAPTIVATE—a 5-year cohort of almost 4500 hospitalized patients with pneumonia from 16 Canadian hospitals (8)—to train and evaluate classifier models for invasive mechanical ventilation (IMV), vasopressor use, and renal replacement therapy (RRT) on hospital day 1 among patients not receiving these interventions on day 0. Consistent with findings from other critical illness scenarios (9), tree-based models showed particularly strong discrimination when predicting receipt of these specific interventions.

Several aspects of the study by Chen et al (7) are worth highlighting. First, most prediction models in critical care have been broadly aimed, either at all comers or at patient cohorts defined by heterogeneous syndromes (e.g., sepsis, acute respiratory distress syndrome) rather than specific diagnoses. Here, the authors adopt a more targeted approach, evaluating whether models tailored to a single clinical diagnosis (albeit one with substantial inherent host- and pathogen-level heterogeneity) could add value. Pneumonia is a leading cause of hospitalization and poor health outcomes, some of which may be modifiable through early interventions (10). Recognizing that appropriate triage might optimize early pneumonia care, Chen et al (7) hypothesize that targeted predictions of subsequent-day organ support could benefit

DOI: 10.1097/CCE.0000000000001268

CAP patients not requiring these therapies on the day of hospitalization. If this strategy ultimately lives up to its promise, it could enable personalized care pathways, avert failure-to-rescue events, and promote more efficient resource use.

A second strength lies in the study's thoughtful cohort curation. The authors evaluated model performance across several well-defined groups: two temporally distinct cohorts of patients with severe acute respiratory syndrome coronavirus 2 pneumonia and a separate cohort with pneumonia from other pathogens. High discrimination across these prospectively collected cohorts—which differ by both time period and pneumonia etiology—strengthens the case for their models' temporal robustness and generalizability beyond a single pathogen or case mix. The use of multiple validation cohorts increases confidence that the models are capturing true clinical signal rather than reflecting idiosyncrasies of a particular dataset.

Third, the authors strengthen the rigor and reproducibility of their work through several commendable practices, laying groundwork for more advanced applications. They share code to support transparency and enable external validation—key steps for building trust. Methodologically, they treat in-hospital death as a competing risk, conservatively assuming that patients who died before receiving IMV, vasopressors, or RRT would have gone on to receive these interventions; this sound choice reduces survivorship bias. The authors also report misclassification-associated model confidence in the supplemental materials. Disaggregating false positives and false negatives is important because the clinical consequences of these errors are rarely symmetric; depending on the implementation context, clinicians might prioritize sensitivity over specificity (or vice versa). Quantifying uncertainty in this way mirrors real-world decision-making and might be an effective way to decrease false-positive alerts (11).

The authors' approach generates interesting hypotheses but also faces several challenges. First, predicting discretionary interventions like intubation is inherently more complex than predicting unambiguous events like mortality. Clinicians differ in their thresholds for many interventions, making it difficult to know whether outcome labels reflect appropriate patient management or subjective decisions. Models adept at anticipating clinician actions may thus risk encoding bias or erroneous clinical decisions. Closely related is the broader question of whether we should predict interventions at all; ideally, predictive models would identify which patients will benefit from intervention, rather than those who will receive it. Reliance on discretionary clinical decisions as model outcomes also threatens generalizability; heterogeneous practice patterns may contribute to model performance degradation in new settings (12). As our methodological toolkit evolves, new approaches can help address these challenges. For instance, anchoring outcome labels on objective endpoints (or composites thereof) can increase consistency (13), while using causal inference techniques appropriately can separate idiosyncratic clinical decisions from underlying patient risk (14).

Second, modeling IMV, vasopressor initiation, and RRT as independent outcomes overlooks their inherent interconnectedness. These interventions often arise from a shared trajectory of physiologic deterioration, making it unsurprising that, for example, respiratory features were among the strongest predictors of vasopressor use in the study by Chen et al (7). Modeling these kinds of outcomes separately can lead to inefficiencies, miscalibration, and even contradictory predictions (15). Advances in AI now support multitask models that learn shared representations to predict several related outcomes simultaneously. By capturing overlapping physiologic signals and clinical decision pathways, multitask models might improve accuracy, calibration, and prediction coherence across interventions.

Chen et al (7) strike several right chords for an exploratory journal, advancing thoughtful hypotheses and engaging with many principles that characterize current best practices in predictive modeling. These core priorities—clarity of the use case, appropriate data selection, and methodologically sound model development—remain central to producing rigorous and clinically meaningful predictive modeling research.

First and foremost, a predictive tool must address a well-articulated clinical use case (how the model's output could meaningfully inform care). Strong use cases share three features: 1) the model would predict outcomes that matter to patients and clinicians; 2) the outcomes are plausibly modifiable through available interventions; and 3) there is a mechanism by which accurate predictions could influence decision-making or behavior. Most modeling efforts fall into one of two categories—prognostic models, which estimate the likelihood of an outcome within a given timeframe, or predictive models, which estimate the probability of response to an intervention. Another useful distinction is whether the model is intended to guide decisions for individual patients or to characterize patterns at the population level. Regardless of category, a clear rationale for modeling is foundational.

Second, the data underlying the model must meet several fundamental requirements. Outcome labels should be accurate, consistent, and have face validity for representing the clinical concept being predicted. Predictors should be available within the right prediction horizon: before observing the outcome and within a clinically sensible time frame (e.g., several hours before deterioration is evident). Finally, the data must be sufficient for demonstrating some basic level of validity (16). At minimum, this requirement indicates the need for a separate validation cohort. Even stronger are strategies that align dataset selection explicitly with the objectives of the analysis. As demonstrated by Chen et al (7), purposeful dataset selection can strengthen inferences regarding temporal model stability, geographic generalizability, and applicability to related clinical contexts. Although data sharing challenges can hinder acquisition of validation data, federated learning and other privacy-preserving methods may help overcome these barriers (17).

Finally, the scientific approach should adhere to modern standards for methodological rigor and transparency (6). Key best practices include establishing an empirical rationale for the number of candidate predictor variables (18), choosing performance measures appropriate for the clinical use case (19), and consistently following established reporting guidelines such as Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (20).

As *Critical Care Explorations* continues to provide a platform for early-stage and hypothesis-generating work, future contributors should do the same: articulate a compelling clinical use case, choose data that support both validity and generalizability, and align modeling methods with established best practices. While implementation studies remain the gold standard, articles that thoughtfully bridge innovation and discipline—as this one does—play a critical role in advancing the science of predictive modeling in critical care.

Division of Pulmonary, Allergy, and Critical Care Medicine, Oregon Health & Science University, Portland, OR.

Dr. Lyons is supported by K08CA270383.

REFERENCES

1. Pinsky MR, Bedoya A, Bihorac A, et al: Use of artificial intelligence in critical care: Opportunities and obstacles. *Crit Care* 2024; 28:113
2. Boussina A, Shashikumar SP, Malhotra A, et al: Impact of a deep learning sepsis prediction model on quality of care and survival. *NPJ Digit Med* 2024; 7:14
3. Bakker T, Klopotowska JE, Dongelmans DA, et al; SIMPLIFY study group: The effect of computerised decision support alerts tailored to intensive care on the administration of high-risk drug combinations, and their monitoring: A cluster randomised stepped-wedge trial. *Lancet* 2024; 403:439–449
4. Wynants L, Van Calster B, Collins GS, et al: Prediction models for diagnosis and prognosis of Covid-19: Systematic review and critical appraisal. *BMJ* 2020; 369:m1328
5. van Royen FS, Moons KGM, Geersing G-J, et al: Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. *Eur Respir J* 2022; 60:2200250
6. Leisman DE, Harhay MO, Lederer DJ, et al: Development and reporting of prediction models: Guidance for authors from editors of respiratory, sleep, and critical care journals. *Crit Care Med* 2020; 48:623–633
7. Chen G, Lee T, Tsang JL, et al; CAPtivate Investigators: Machine learning accurately predicts the need for critical care support in patients admitted to the hospital for community-aquired pneumonia. *Crit Care Explor* 2025; 7:e1262
8. Tsang JLY, Rego K, Binnie A, et al; For CAPTIVATE Investigators: Community versus academic hospital community-acquired pneumonia patients: A nested cohort study. *Pneumonia* 2024; 16:31
9. Churpek MM, Carey KA, Snyder A, et al: Multicenter development and prospective validation of eCARTv5: A gradient-boosted machine-learning early warning score. *Crit Care Explor* 2025; 7:e1232
10. Wang Y, Eldridge N, Metersky ML, et al: Analysis of hospital-level readmission rates and variation in adverse events among patients with pneumonia in the United States. *JAMA Netw Open* 2022; 5:e2214586
11. Shashikumar SP, Wardi G, Malhotra A, et al: Artificial intelligence sepsis prediction algorithm learns to say "I don't know." *NPJ Digit Med* 2021; 4:134
12. Lyons PG, Hofford MR, Yu SC, et al: Factors associated with variability in the performance of a proprietary sepsis prediction model across 9 networked hospitals in the US. *JAMA Intern Med* 2023; 183:611–612
13. Verghis R, Blackwood B, McDowell C, et al: Heterogeneity of surrogate outcome measures used in critical care studies: A systematic review. *Clin Trials* 2023; 20:307–318
14. Walker V, Sanderson E, Levin MG, et al: Reading and conducting instrumental variable studies: Guide, glossary, and checklist. *BMJ* 2024; 387:e078093

15. Roy S, Mincu D, Loreaux E, et al: Multitask prediction of organ dysfunction in the intensive care unit using sequential subnetwork routing. *J Am Med Inform Assoc* 2021; 28:1936–1946
16. Debray TPA, Vergouwe Y, Koffijberg H, et al: A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015; 68:279–289
17. Rojas JC, Lyons PG, Chhikara K, et al: A common longitudinal intensive care unit data format (CLIF) for critical illness research. *Intensive Care Med* 2025; 51:556–569
18. Riley RD, Ensor J, Snell KIE, et al: Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020; 368:m441
19. Steyerberg EW, Vickers AJ, Cook NR, et al: Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* 2010; 21:128–138
20. Collins GS, Reitsma JB, Altman DG, et al: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD statement. *Ann Intern Med* 2015; 162:55–63