INVITED EDITORIAL

Machine Learning for Predicting Intubations in Heart Failure Patients: the Challenge of the Right Approach

Sai Nikhila Ghanta[1] · Nitesh Gautam[1] · Jawahar L. Mehta[2] · Subhi J. Al'Aref[2]

Heart failure (HF) remains a major contributor to national and international morbidity and mortality, affecting nearly 6.2 million people in the USA and incurring around $43 billion in healthcare costs in 2020 alone [1, 2]. The progressive and multifaceted nature of HF is part of the reason behind the high morbidity and mortality associated with HF. While the 1-year HF mortality among Medicare beneficiaries declined slightly from 1998 to 2008 but remained high at 29.6%, the progressive nature of HF results in frequent decompensations and hospitalizations [3, 4]. For instance, HF exacerbations amount to nearly 1 million hospitalizations every year in individuals older than 65 years of age [5]. Acute decompensated HF is characterized by the presence of respiratory symptoms, given the underlying derangements in cardiac structure, function, and hemodynamics. For example, worsening dyspnea is present in approximately 90% of decompensated HF exacerbations, but less than 50% present with respiratory failure secondary to hypoxemia, hypercapnia, or both [6, 7]. As such, acute HF syndromes with elevated filling pressures (such as the pulmonary capillary wedge pressure) and/or systemic circulatory failure (as a result of a diminished cardiac output) can lead to respiratory failure requiring positive pressure ventilatory support, with the goal of intervening early enough to prevent the need for mechanical ventilation. Recent decades have witnessed a steady increase in the usage of non-invasive ventilation (NIV) and mechanical ventilation in patients with acute decompensated HF [8]. Several randomized control trials and meta-analyses have documented the superiority of NIV in reducing mortality, endotracheal intubation rates, and improvement in respiratory status over conventional oxygen therapy [9, 10]. Accordingly, early application of NIV in the pre-hospital setting has dramatically increased with a tendency to reduce the need for mechanical ventilation [11, 12].

Nevertheless, not all individuals presenting with acute decompensated HF are candidates for NIV. For example, patients presenting with altered sensorium cannot protect their airways and maintain spontaneous breathing, requiring the initiation of invasive mechanical ventilation. While mechanical ventilation can be lifesaving in decompensated HF, prolonged mechanical ventilation (PMV) is associated with increased morbidity and mortality and a significant burden on healthcare system [13, 14]. Some of the adverse events associated with mechanical ventilation include the development of ventilator-associated pneumonia, ventilator-induced lung injury, oxygen toxicity, hypotensive effects of sedative agents, as well as unfavorable right ventricular afterload in the context of positive pressure ventilation [15, 16]. In addition, PMV has been associated with adverse outcomes in HF patients, despite the inconsistent definition of PMV across the literature [17, 18]. As such, accurate and validated tools are needed for predicting patients at high risk for PMV, which would help allocate resources and guide management, especially in HF patients.

To this end, Li and colleagues developed and externally validated a novel prediction model using machine learning (ML) to identify patients at a high risk of PMV. The study titled: "Machine learning-based model for predicting prolonged mechanical ventilation in patients with congestive heart failure" utilized previously collected data on 4533 mechanically ventilated HF patients included in MIMIC-IV (Medical Information Mart for Intensive Care IV) database.

✉ Subhi J. Al'Aref
   sjalaref@uams.edu

   Sai Nikhila Ghanta
   sainikhila.g@gmail.com

   Nitesh Gautam
   Ngautam@uams.edu

   Jawahar L. Mehta
   mehtajl@uams.edu

1   Department of Internal Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, USA

2   Department of Medicine, Division of Cardiology, University of Arkansas for Medical Sciences, 4301 W. Markham St, Little Rock, AR, USA

Patients included in the study had been admitted to intensive care units (ICUs) at a single tertiary care center in the USA between 2008 and 2019 and had received mechanical ventilation in the first 24 h of the ICU admission. The external validation cohort included patients in the eICU Collaborative Research Database from 2014 to 2015, a multi-center ICU database from 208 US hospitals. They employed 12 ML algorithms and used LASSO regression for feature selection. Importantly, they included patients receiving mechanical ventilation in the first 24 h of the ICU admission and sought to predict mechanical ventilation ≥ 4 days (defined as PMV). They found that the CatBoost algorithm performed best, with an area-under-the-curve (AUC) of 0.766 on the training set and 0.733 on the external validation set. Important features associated with PMV included the presence of pneumonia, sepsis, and the use of inotropes (Odds ratio [OR] of 2.18, 1.75, and 1.49 for the presence of pneumonia, sepsis, and the use of inotropes, respectively, p-value < 0.001 for all). The authors also showed that the ML model accurately predicted in-hospital mortality, with an AUC of 0.844.

The strength of the study by Li et al. stems from several factors. For instance, the authors employed LASSO regularization prior to model training for feature selection, thereby mitigating the chances of overfitting. Moreover, the authors assessed the robustness of multiple ML models, of which the CatBoost model [19] emerged as the best predictor model. Furthermore, the CatBoost model had a superior performance in predicting the need for PMV than the traditionally established risk score models such as sequential organ failure assessment (SOFA) score [20], simplified acute physiology score (SAPS-II) [21], logistic organ dysfunction system (LODS) score [22] (AUC of 0.817, 0.697, 0.581, and 0.707 for CatBoost, SOFA, SAPS-II, and LODS score respectively). In addition, a decision tree nomogram was utilized to risk stratify patients. The authors concluded that patients in the high-risk category were at a ninefold higher risk for PMV than the low-risk patients.

Despite the strengths of the investigation, it is important to mention some of the major limitations of the study. Firstly, while the CatBoost model was shown to predict PMV with an accuracy of 69.4%, there is a lack of a discussion of a clinical context. For instance, while the ML-based model predicted PMV better than already published benchmark models, how useful is an accuracy of 69.4% in predicting prolonged mechanical ventilation? If the model is inaccurate 30.6% of the time in predicting mechanical ventilation ≥ 4 days, is that a useful benchmark in such a critical juncture in a patient's clinical trajectory? Furthermore, prediction accuracy estimated in the study can be helpful at population/group level, but the investigators failed to highlight diagnostic parameters that are useful at the individual patient level (such as positive predictive value (PPV) and negative predictive value (NPV)). Furthermore,

the study cohort was poorly defined, as there is no data on the baseline New York Heart Association (NYHA) functional classification of included patients, nor is there data on their clinical trajectory prior to study inclusion (i.e., whether the study cohort includes patients with chronic stable HF with an index decompensation or whether the study cohort includes patients with frequent decompensations and hospitalizations). Second, the investigation did not differentiate between factors precipitating the need for mechanical ventilation. Patients with impending respiratory failure due to superimposed infective etiologies (e.g., pneumonia and sepsis) progress toward mechanical ventilation due to a complex interplay of factors (e.g., the timing of antibiotic therapy, antimicrobial coverage, and the extent of infection). These factors need to be controlled for, and ideally, for a model to influence clinical decision-making for a clinician, stricter inclusion criteria comprising isolated decompensated HF etiologies are necessary, thereby minimizing the influence of other acute infectious pathologies. Third, NIV coupled with appropriate pharmacotherapy can prevent the need for mechanical ventilation in a significant proportion of HF exacerbation patients. However, the present investigation did not select for a specific indication for mechanical ventilation in the first 24 h, or whether these were patients who failed an initial strategy of NIV. Furthermore, pertinent HF features such as left ventricular ejection fraction (LVEF) [23], NYHA class [24], previous HF hospitalizations, device therapies, and the need for inotropic therapy were not included in the research methodology. Importantly, the decision to initiate mechanical ventilation, apart from clinical judgment, relies on multiple time-dependent variables such as the Glasgow Coma Score (GCS), blood gases/chemistries, and vital signs. While the study has included these variables as a snapshot (i.e., at the time of admission), the full predictive power of such data lies in the longitudinal trajectory of such variables over time. Future ML-based models that are specifically designed to incorporate time-dependent variables, such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTMs), will likely produce more useful and robust predictive modeling, especially in the context of complex clinical conditions such as HF [25]. Additionally, in the setting of a dynamic clinical environment, the data itself generated might change over time (data drift), and so can the relationship between episodic characteristics and target outcomes (concept drift) due to multiple reasons [26]. For example, prediction accuracies of ML models developed prior to coronavirus-19 (COVID-19) pandemic can be inaccurate if employed in the post-COVID-19 era. This means that proactive monitoring capabilities need to be in place to detect data drift and ensure that such models generalize appropriately to temporarily changing scenarios and outcomes. Duckworth et al. demonstrated how Shapley adaptive explanations (SHAP) can be used as a complementary

metric to detect data drift for emergency department admissions during the COVID-19 pandemic [26]. Finally, while the literature is divided on the accepted timeframe for PMV, it would have been clinically more useful to use a time stamp that influences clinical decisions, rather than use a cutoff of 4 days. For example, the decisions on placement of a feeding tube, tracheostomy, and/or initiation of goals of care discussions would be better suited as a timeframe for such a prediction model.

Although various protocols have been in place to predict intubations and successful extubations [27], in day-to-day clinical practice such a process primarily remains a clinical judgment that is affected by many time-varying factors. In addition, such a process and decision-making varies from facility to facility and is influenced by local clinical expertise/experience. There is no global consensus on a single protocol to identify patients at risk for PMV, or even mechanical ventilation alone. As a result, several studies have sought to harness the power of ML modeling to develop improved prediction modeling. However, ML has its own intrinsic limitations, and the usefulness of the produced model is often as good as the data it is fed, and the undertaken approach. For example, if a ML model is too conservative with its prediction of the need for PMV, it can lead to extubation failures and pre-mature tracheostomies. Furthermore, many of the ML models are constructed using retrospectively collected data, or even clinical trial data that might not inherently represent real-world patients and practices, and the resultant dilemma is whether it is ethically right to decide on intubation/PMV solely based on the result of an ML algorithm. For instance, the widely implemented proprietary sepsis prediction model, based on a big data approach, was recently shown to have poor sensitivity when externally examined [28]. Future studies evaluating the synergy between ML models' predictions and physicians' decisions to intubate are needed to verify the validity of ML modeling in clinical practice. As an example, Sax et al. identified barriers and opportunities while implementing a ML-based risk stratification tool for acute HF admissions in the emergency department: their ML model was paired with clinical decision support resulting in broader acceptance and adoption [29]. In addition, ML algorithms are prone to the inherent complexity and experts have described various artificial intelligence-enabled decision support and reporting guidelines to improve the transparency, reproducibility, and validity of prediction models [30–33]. Loftus et al. described the following checklist for ideal prediction models: explainable, dynamic, precise, autonomous, fair, and reproducible [34]. Furthermore, the U.S. food and drug administration (FDA) developed ten good machine learning practice (GMLP) guidelines to provide helpful framework for investigators and identify areas where collaborative bodies can work to advance GMLP.

Nevertheless, the study by Li and colleagues can guide future prospective studies and quality improvement ventures. The study methodologies can also be useful in designing models and risk scores that can predict the need for mechanical ventilation in other acute distress conditions such as (ARDS) or COVID-19 pneumonia, as well as chronic obstructive pulmonary disease (COPD) exacerbations.

Declarations

References

1. Virani SS, Alonso A, Benjamin EJ, et al. Heart disease and stroke statistics-2020 update: a report from the American Heart Association. Circulation. 2020;141:e139–596.
2. Benjamin EJ, Muntner P, Alonso A, et al. Heart disease and stroke statistics-2019 update: a report From the American Heart Association. Circulation. 2019;139:e56–528.
3. Tsao CW, Aday AW, Almarzooq ZI, et al. Heart disease and stroke statistics-2022 update: a report from the American Heart Association. Circulation. 2022;145:e153–639.
4. Heidenreich PA, Albert NM, Allen LA, et al. Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association. Circ Heart Fail. 2013;6:606–19.
5. Roger VL. Epidemiology of heart failure. Circ Res. 2013;113:646–59.
6. Mebazaa A, Pang PS, Tavares M, et al. The impact of early standard therapy on dyspnoea in patients with acute heart failure: the URGENT-dyspnoea study. Eur Heart J. 2010;31:832–41.
7. Siniorakis E, Arvanitakis S, Tsitsimpikou C, et al. Acute heart failure in the emergency department: respiratory rate as a risk predictor. In Vivo. 2018;32:921–5.
8. Tavazzi G. Mechanical ventilation in cardiogenic shock. Curr Opin Crit Care. 2021;27:447–53.
9. Gray A, Goodacre S, Newby DE, et al. Noninvasive ventilation in acute cardiogenic pulmonary edema. N Engl J Med. 2008;359:142–51.

10. Weng CL, Zhao YT, Liu QH, et al. Meta-analysis: noninvasive ventilation in acute cardiogenic pulmonary edema. Ann Intern Med. 2010;152:590–600.
11. Demoule A, Chevret S, Carlucci A, et al. Changing use of noninvasive ventilation in critically ill patients: trends over 15 years in francophone countries. Intensive Care Med. 2016;42:82–92.
12. Ducros L, Logeart D, Vicaut E, et al. CPAP for acute cardiogenic pulmonary oedema from out-of-hospital to cardiac intensive care unit: a randomised multicentre study. Intensive Care Med. 2011;37:1501–9.
13. Kuhn BT, Bradley LA, Dempsey TM, Puro AC, Adams JY. Management of mechanical ventilation in decompensated heart failure. J Cardiovasc Dev Dis. 2016;3(4):33. https://doi.org/10.3390/jcdd3040033.
14. Beduneau G, Pham T, Schortgen F, et al. Epidemiology of weaning outcome according to a new definition. The WIND Study. Am J Respir Crit Care Med. 2017;195:772–83.
15. Schmitt JM, Vieillard-Baron A, Augarde R, Prin S, Page B, Jardin F. Positive end-expiratory pressure titration in acute respiratory distress syndrome patients: impact on right ventricular outflow impedance evaluated by pulmonary artery Doppler flow velocity measurements. Crit Care Med. 2001;29:1154–8.
16. Slutsky AS, Ranieri VM. Ventilator-induced lung injury. N Engl J Med. 2013;369:2126–36.
17. Chatila WM, Criner GJ. Complications of long-term mechanical ventilation. Respir Care Clin N Am. 2002;8:631–47.
18. Lai CC, Shieh JM, Chiang SR, et al. The outcomes and prognostic factors of patients requiring prolonged mechanical ventilation. Sci Rep. 2016;6:28034.
19. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data. 2020;7:94.
20. Chapalain X, Vermeersch V, Egreteau PY, et al. Association between fluid overload and SOFA score kinetics in septic shock patients: a retrospective multicenter study. J Intensive Care. 2019;7:42.
21. Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA. 1993;270:2957–63.
22. Le Gall JR, Klar J, Lemeshow S, et al. The logistic organ dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. JAMA. 1996;276:802–10.
23. Mele D, Nardozza M, Ferrari R. Left ventricular ejection fraction and heart failure: an indissoluble marriage? Eur J Heart Fail. 2018;20:427–30.
24. Caraballo C, Desai NR, Mulder H, et al. Clinical Implications of the New York Heart Association Classification. J Am Heart Assoc. 2019;8:e014240.
25. Rush B, Celi LA, Stone DJ. Applying machine learning to continuously monitored physiological data. J Clin Monit Comput. 2019;33:887–93.
26. Duckworth C, Chmiel FP, Burns DK, et al. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. Sci Rep. 2021;11:23017.
27. Baptistella AR, Sarmento FJ, da Silva KR, et al. Predictive factors of weaning from mechanical ventilation and extubation outcome: a systematic review. J Crit Care. 2018;48:56–62.
28. 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;181:1065–70.
29. Sax DR, Sturmer LR, Mark DG, Rana JS, Reed ME. Barriers and opportunities regarding implementation of a machine learning-based acute heart failure risk stratification tool in the emergency department. Diagnostics (Basel). 2022;12(10):2463. https://doi.org/10.3390/diagnostics12102463.
30. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health. 2020;2:e549–60.
31. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26:1364–74.
32. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26:1320–4.
33. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11:e048008.
34. Loftus TJ, Tighe PJ, Ozrazgat-Baslanti T et al. Ideal algorithms in healthcare: explainable, dynamic, precise, autonomous, fair, and reproducible. PLOS Digit Health 2022;1(1):e0000006. https://doi.org/10.1371/journal.pdig.0000006.