`Predicting Failure of Noninvasive Respiratory Support Using Deep Recurrent Learning

Patrick Essay, PhD[1], Amin Nayebi[1], Julia M. Fisher, PhD[2], Jarrod M. Mosier, MD[3,4], Vignesh Subbian, PhD[1,5, 6]

[1] Department of Systems and Industrial Engineering, College of Engineering, The University of Arizona, Tucson, AZ

[2] Statistics Consulting Laboratory, BIO5 Institute, The University of Arizona, Tucson, AZ

[3] Department of Emergency Medicine, The University of Arizona College of Medicine, Tucson, AZ

[4] Division of Pulmonary, Allergy, Critical Care, and Sleep, Department of Medicine, The University of Arizona College of Medicine, Tucson, AZ

[5] Department of Biomedical Engineering, College of Engineering, The University of Arizona, Tucson, AZ

[6] BIO5 Institute, The University of Arizona, Tucson, AZ

Corresponding Author:
Jarrod M. Mosier, MD FCCM
Department of Emergency Medicine
1501 N. Campbell Ave., AHSL 4171D
PO Box 245057
Tucson, AZ 85724-5057
Phone: 520-626-2038
jmosier@aemrc.arizona.edu
ORCHID ID: 0000-0002-5371-0845

Author Contributions: PE, JMM, and VS conceived the study idea and conducted literature search. PE, VS, and JMM developed the phenotyping algorithm. PE and JMF collected and preprocessed the data. PE, AN, and VS developed the prediction models. JM and PE drafted the initial manuscript, and all authors participated in manuscript revisions.

Source of Support: This work was supported by an Emergency Medicine Foundation grant sponsored by Fisher & Paykel, and in part by the National Science Foundation under grant #1838745 and the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number 5T32HL007955. Neither funding agency or sponsor was involved in the design or conduct of the study or interpretation and presentation of the results.

Competing interests: JMM, VS, and JMF received grant support for this work by the Emergency Medicine Foundation. Authors have no conflicts of interest to declare.

Running head: Predicting failure of noninvasive respiratory support.

Word count: 3163

Abstract word count: 236

Data Availability: The datasets generated and/or analyzed during the current study are not publicly available to protect individual participant privacy but are available from the corresponding author on reasonable request and with appropriate institutional review board approval and data use agreement.

Acknowledgements: The authors want to acknowledge the support of the ARCS Foundation Phoenix Chapter as well as Don Saner and Mario Arteaga at Banner Health for their support of this work through the clinical data warehouse.

Keywords: machine learning, deep neural network, intensive care unit, respiratory failure, mechanical ventilation

Abstract

Background

Noninvasive respiratory support is increasingly used to support patients with acute respiratory failure. However, noninvasive support failure may worsen outcomes compared to primary support with invasive mechanical ventilation. Therefore, there is a need to identify patients noninvasive respiratory support failure so that treatment can be reassessed or altered. The objective of this study was to develop and evaluate three recurrent neural network models to predict noninvasive respiratory support failure.

Methods

This is a cross-sectional observational study to evaluate the ability of deep recurrent neural network models (long short-term memory, gated recurrent unit, and gated recurrent unit with trainable decay) to predict failure of noninvasive respiratory support. Data were extracted from electronic health records from all adult ($\geq$ 18 years) patient records requiring any type of oxygen therapy or mechanical ventilation between November 1, 2013, and September 30, 2020 across >20 hospitals in a single healthcare network.

Results

Time series data from electronic health records were available for 22,075 patients. The highest accuracy and area under the receiver operating characteristic curve were for the long short-term memory model (94.04% and 0.9636, respectively). Accurate predictions were made 12 hours after ICU admission and performance remained high well in advance of noninvasive respiratory support failure.

Conclusion

Recurrent neural network models using routinely collected time-series data can accurately predict noninvasive respiratory support failure well before intubation. This lead time may provide an opportunity to intervene to optimize patient outcomes.

Introduction

Acute respiratory failure is a common reason for admission to the intensive care unit, is costly for the health care system, and often exposes patients to high morbidity and mortality.[1-4] Patients with acute respiratory failure are commonly treated with a noninvasive respiratory support (NIRS) strategy to support the work of breathing, improve gas exchange, and avoid the undesirable consequences of invasive mechanical ventilation.[5]

Overall, NIRS strategies (noninvasive positive pressure ventilation or high flow nasal oxygen) reduce the need for intubation and consequently lower mortality, but the benefits are skewed heavily by helmet noninvasive positive pressure ventilation and comparisons show no difference between noninvasive positive pressure ventilation by facemask and high flow nasal oxygen.[6-9] While NIRS may benefit patient outcomes, failure of these therapies carries a cost of increased mortality.[10-12] Failure of noninvasive positive pressure ventilation in patients with acute hypoxemic respiratory failure is often associated with prolonged ICU stays and increased mortality.[12-15] Thus, predicting who will likely fail a NIRS strategy has important implications for clinical care and potentially patient outcomes.

Previous studies have evaluated factors associated with or predictive of NIRS failure with variable success and many limitations.[10, 15-25] Physiology studies have provided some evidence as to the mechanism of failure and injury associated with failure,[23, 26] but there are no routinely available clinical data that can be used as substitutes for measurements such as transpulmonary pressure and respiratory effort. The objective of this study is to evaluate the predictive value of a deep learning model

(recurrent neural network) to predict NIRS failure using clinically available time series data from electronic health records.

Materials and Methods

Study Setting

Clinical data were obtained from Banner Health Network clinical data warehouse. Banner Health Network represents >25 hospitals across six states in the western United States. All adult patients (≥18 years of age) requiring any form of oxygen therapy or mechanical ventilation were extracted between November 1, 2013, and September 30, 2020. Data consist of deidentified structured data generated from the Cerner electronic health record (Cerner Corporation, North Kansas City, MO, USA). This study was approved by both the University of Arizona Institutional Review Board and Banner Health Institutional Review Board and granted waiver from informed consent. All methods were carried out in accordance with relevant guidelines and regulations, and aligned with the TRIPOD statement for predictive modeling [27] and recommended reporting guidelines for machine learning algorithms.[28]

Study Participants

We generated seven cohorts of patients using a previously developed phenotyping algorithm,[29] which uses the sequence of therapy received to generate the following cohorts: 1. invasive mechanical ventilation only, 2. noninvasive positive pressure ventilation only, 3. high flow nasal oxygen only, 4. noninvasive positive pressure ventilation requiring subsequent invasive mechanical ventilation, 5. high flow

nasal oxygen requiring subsequent invasive mechanical ventilation, 6. invasive mechanical ventilation extubated to noninvasive positive pressure ventilation, and 7. invasive mechanical ventilation extubated to high flow nasal oxygen. All patients in Cohorts 2 and 3 (NIRS success) and 4 and 5 (NIRS failure) were included in this analysis. Patients that alternated between noninvasive positive pressure ventilation and high flow nasal oxygen during a single ICU admission were excluded. Readmissions to the ICU were also excluded as readmission may alter treatment trajectory or decision-making by a clinician regarding if and when to intubate as compared to a first ICU admission patient.

Predictors and Data Pre-processing

Failure of NIRS therapy requiring invasive mechanical ventilation at any time during an ICU stay was the prediction outcome of interest (i.e., patients that failed and patients that did not fail NIRS therapy). We trained and tested all models using three different outcomes, namely, noninvasive positive pressure ventilation failure, high flow nasal oxygen failure, and NIRS failure. The latter is the combination of noninvasive positive pressure ventilation and high flow nasal oxygen failures. Data were extracted from the entirety of each ICU visit, but only variables measured prior to NIRS failure were used for model derivation. Model inputs consisted of the twelve most common laboratory measurements and vital signs: chloride, creatinine, albumin, respiratory rate, heart rate, pulse oximetry oxygen saturation (SpO_2), fraction of inspired oxygen (FiO_2), oxygen saturation, and two measurements each (point-of-care and laboratory measurement) of partial pressure of carbon dioxide and partial pressure of arterial

oxygen from an arterial blood gas. Inputs were both regularly and irregularly sampled. While the frequency of input data for each feature varied across patients, most patients had at least one data point for each (online resource Table E1), and patients were not excluded due to missing data. All data preprocessing and prediction modeling was performed in Python (v.2.7.14; Python Software Foundation) using the Pandas (v.0.23.4),[30] Seaborn (v.0.9.0),[31] and sci-kit learn package (v.0.19) [32] libraries.

Prediction Modeling

Recurrent neural networks (RNN) are complex predictive models that are particularly well-suited for predicting outcomes over long time periods.[33, 34] We trained three recurrent neural network variations for multivariate time series predictions: 1) long short-term memory (LSTM),[35] 2) gated recurrent unit (GRU),[36] and 3) gated recurrent unit with trainable decay (GRU-D).[37] Recurrent neural network architectures and hyperparameters were held constant between the models for comparison, although long short-term memory and gated recurrent unit models used a global average pooling layer and gated recurrent unit with trainable decay did not (online resource Figure E1). The pooling layer for long short-term memory and gated recurrent unit reduces dimensionality by averaging across model inputs. By taking the average of two activation function weights within the networks, average pooling also helps avoid overfitting. The Adam optimizer and binary cross entropy loss function were used with accuracy as the target training metric which is common for binary classification. The hidden layers in the network structures used linear activation functions and the output layer used a sigmoid activation function.

We used a variable observation window (i.e., input feature extraction time window) and time at-risk (i.e., temporal distance between the end of the observation window and the time of failure) to evaluate clinical validity (Figure 1). For patients that did not fail NIRS, we extracted time series data from an observation window of 1 to 72 hours from the initiation of NIRS. NIRS initiation was used as a reference point for data extraction and to avoid complications that may have led to the patient dying even if they were not intubated. For NIRS failure patients, we extracted time series data from an observation window of 1 to 72 hours starting at the point of failure and moving back in time away from that point. We also varied the time at-risk from 1 to 72 hours effectively moving the observation window away from the time of failure. Data recorded during the time at-risk were not used as model inputs. Rather time at-risk was used to determine how far in advance accurate failure predictions can be made. This approach ensures consistent time-at-risk period for all patients that failed NIRS (i.e., consistent time between prediction and failure) and allows for side-by-side comparison between prediction models with predictions being made at the end of the observation window. Observation windows for NIRS failure patients at the beginning of NIRS therapy would result in variable time between prediction and failure. This could bias results toward models better suited for longer or shorter-term predictions. Thus, time-at-risk in our approach is constant across all comparisons.

We determined the highest performance for each model using the full observation window (72 hours) and a failure cohort time-at-risk period of one hour. We then shortened the observation window in decrements to a minimum of one hour and tested all models at each time decrement. We tested time-at-risk incrementally to the

maximum of 72 hours prior to failure using a fixed observation window size. In all experiments, the observation window for patients that did not fail NIRS (or NIRS success) mirrored the observation window of the NIRS failure patients in terms of window size but remained at the beginning of NIRS therapy for training and testing.

Model Training and Evaluation

For comparison, we used logistic regression and random forest models with discrete input data. Model inputs for logistic regression and random forest included demographics (i.e., age, gender) and Acute Physiology and Chronic Health Evaluation (APACHE) data.[38, 39] We used ten estimators for random forest and fully expanded trees to maximum depth (i.e., expanded until leaves contained less than two samples). The highest area under the receiver operating characteristic curve and the precision-recall curve were calculated and reported for all models. The area under the curve is reported for the three RNN models for variable observation window with time-at-risk held constant and for constant observation window with variable time-at-risk to illustrate performance over time.

Model training and testing was performed with 66% of the total population using 0.33 validation split. The remaining 33% of the total population was used as an additional hold-out test set. The train test split was stratified such that the proportion of failure patients in each set was constant. We performed an exhaustive grid search to determine batch size and number of epochs. Number of epochs was verified by graphically comparing training and testing accuracy and loss to avoid overfitting. Batch size was then held to 150 patients and 20 epochs were used for model training in all trials. Long short-term memory testing results were then used to further evaluate patient

outcomes after model prediction. Patient outcomes of mortality and length of stay were calculated for the test set output resulting in four groups of patients: 1) patients predicted to fail NIRS and failed, 2) patients predicted to fail NIRS and did not fail, 3) patients predicted not to fail NIRS and failed, and 4) patients predicted not to fail NIRS and did not fail.

Results

Descriptive Statistics

A total of 22,075 patients fit inclusion criteria for NIRS and NIRS failure groups (Table 1). The failure rate of noninvasive positive pressure ventilation was 26% and the failure rate of high flow nasal oxygen was 50%, respectively, for an overall failure rate of 42%. Generally, patient characteristics across both NIRS modalities were similar and not expected to negatively impact prediction performance.

Prediction Results and Model Comparisons

Long short-term memory and gated recurrent unit models had the highest area under the curve and best precision and recall for all observation window and time-at-risk trials (Figure 2). The long short-term memory model outperformed gated recurrent unit and gated recurrent unit with trainable decay in terms of prediction accuracy and area under the curve using data from a 72-hour observation window and a 1-hour time-at-risk window (Table 2). We compared results across all models for noninvasive positive pressure ventilation and high flow nasal oxygen separately and for the combined NIRS

group. The gated recurrent unit with trainable decay model was comparable to the two baseline models, logistic regression, and random forest.

Timeline Analysis

An observation window of 12-18 hours yields nearly the same results as using 72 hours of available data prior to NIRS failure (Figure 3). Performance begins to diminish with an observation window <12 hours. Using a fixed observation window size of 12 hours and moving the observation window away from the time of NIRS failure by increasing the time-at-risk period showed an initial drop in area under the curve but sustained performance beyond a 9-hour time-at-risk (Figure 3). This temporal relationship was seen consistently throughout all trials suggesting that accurate predictions can be made well in advance of failure using a trailing 12 hours of time series input variables. Using less than 12 hours of data still returns reasonable results but performance decreases across all three models.

Patient Outcomes

Mortality and ICU length-of-stay (Online Resource, Table E4) were analyzed for patients in the NIRS success and NIRS failure test set using the long short-term memory model with a 12-hour observation window and 1-hour time-at-risk period. Patients that failed NIRS therapy had a mortality rate around 30% in both predicted outcomes. Patients that were predicted not to fail but failed NIRS had a slightly higher mortality rate. On the other hand, for patients that did not fail NIRS, the predicted success patients had a lower mortality than the predicted failure patients.

ICU length-of-stay did not show the same relationship. The patients that failed NIRS but were predicted not to fail had a lower LOS than the patients that were successfully treated with NIRS. The reverse relationship was seen for patients predicted to fail NIRS where patients that failed had a longer LOS than patients that did not fail NIRS.

Discussion

Our results show that time series based deep learning model (long short-term memory) outperforms baseline models for predicting failure of NIRS in patients with acute respiratory failure, and the model performance remains high over relatively long observation windows. Other NIRS failure models have been tested near the time of failure [17] but not extended to test lengthy observation and time-at-risk windows for earlier prediction. We use twelve commonly available measurements that allows for use of a pooling layer in recurrent neural network design and compensates for missing variables. Other NIRS failure models use a smaller number of input features but are unable to make predictions if even one variable is missing. Lastly, our approach does not require that a patient receive NIRS or any other treatment prior to making predictions. The time series inputs are independent of treatment path and thus could predict decompensation for any patient if tested in a prospective study design or implemented in real-time. These results demonstrate that early prediction of failure can potentially impact patient outcome. However, there is not a consistent duration of time between prediction of failure and observed failure.

One important consideration for all prediction models in this application is that they are not directly predicting physiological decompensation but rather the clinical

determination of failure and the need for invasive mechanical ventilation. This decision carries subjectivity and variability between clinicians and institutions, and the subjectivity will likely evolve with new knowledge and experience. These models are durable, however, and may be retrained as practice changes, allowing the model to evolve based on how input data relates to the decision to intubate a patient. Predicting failure using widely available clinical data is challenging but has practical applications for differentiating patients likely to fail NIRS at the cost of increased mortality.

Several studies have identified factors associated with failure or predictive of failure for both noninvasive positive pressure ventilation and high flow nasal oxygen (see online resource Tables E2 and E3).[10, 11, 15-22, 24, 25] The ROX index, which is the ratio of oxygen saturation/FiO2 and the respiratory rate, is a recently derived and validated prediction tool, albeit not under protocolized conditions, for determining if a patient is likely to succeed or fail high flow nasal oxygen.[21] A value >4.88 at 2, 6, or 12 hours has good predictive value for not requiring intubation. Values <2.85 at 2 hours, <3.47 at 6 hours, and < 3.85 at 12 hours were predictors of failure. Unfortunately, the ROX index is only developed for one specific type of high flow nasal cannula system and is largely flow dependent with increases in ROX index when going from 30 to 60 liters per minute of flow indicating higher severity of lung disease.[19] In one retrospective study, high flow nasal oxygen was more likely to fail in patients with a significant increase in respiratory rate or decrease in ROX index within 3 days in patients with COVID-19.[40] However, lower oxygen saturation at admission were only significantly associated with failure after adjustment, and failure was associated with a 30% increase in mortality.

For our application, we used timeline adjustments to evaluate clinical validity of our approach and improve robustness. The variable observation window answered how much temporal data is required to make an accurate prediction; and how early accurate predictions can be made. This has potential to be employed clinically. A prediction of failure can be made early to allow an opportunity to alter strategies and potentially improve outcomes. Though, our current approach does not predict how long until failure.

Observation windows and time at-risk from one to 72 hours were selected for several reasons. Making predictions inside of one hour from the time of failure does not allow sufficient time for clinicians to alter treatment path to minimize potential impact to patient-centered outcomes (e.g., mortality). In other words, too short of a time window predicts what clinicians likely already know. Predicting outside of 72 hours allows for too much variability in the trajectory of a critically ill patient and is well beyond typical decision-making timelines. Clinicians will likely wait and evaluate whether a patient improves or worsens even if a failure prediction is made 72 hours in advance. In addition to practical implications of observation and at-risk window sizes of 1-72 hours, computational factors must be considered. For example, a newly admitted patient will require a prediction to be made soon after admission rather than waiting for enough data to be recorded and extending the amount of time series data being used increases computational load required to make a prediction.

There are several important limitations to these results. Missing time series data varies across datasets and among individual patients. Presumably, increased measurements would not adversely affect the training and testing aspects of model

development. Reproducibility and future implementation, however, could be affected if input data is insufficient for the model to accurately make predictions. This is reflected in our timeline adjustments with observation windows <12 hours. In addition, timeline adjustments resulted in fewer patients used for training and testing due to variations in ICU lengths-of-stay. For example, if a patient was in the ICU for 18 hours and then died (before or after NIRS failure), that patient was dropped from timeline experiments using >18 hours of time series data. Our total population, however, is large enough that missing patients did not change overall characteristics of training and testing sets nor of the NIRS success and failure groups. Presumably, this had minimal impact on testing outcome as the timeline adjustments progressed. Correlating prediction results to the clinical outcomes (e.g. mortality) in the same dataset has limited interpretability. In-depth interpretability efforts are a part of our ongoing work as these results require both external and prospective validation before they can be used clinically as a decision support tool on an electronic health record platform.

In conclusion, recurrent neural networks are promising for predicting NIRS failure in patients with acute respiratory failure. Long short-term memory and gated recurrent unit outperformed gated recurrent unit with trainable decay and baseline comparison models in predicting NIRS failure soon after ICU admission. Prediction performance remained high until using observation window sizes of twelve hours or fewer near time of failure. Prediction performance minimally decreased as observation window was moved away from the time of failure suggesting that the combination of deep learning model inputs captures sufficient information to predict NIRS failure regardless of temporal proximity to the time of failure. Predictions can potentially be early enough for

patient level impact and outperforms previously developed predictions for NIRS therapies.

References

1. Cartin-Ceba R, Kojicic M, Li G, Kor DJ, Poulose J, Herasevich V, et al. Epidemiology of critical care syndromes, organ failures, and life-support interventions in a suburban US community. Chest 2011;140(6):1447-1455.
2. Frat JP, Thille AW, Mercat A, Girault C, Ragot S, Perbet S, et al. High-flow oxygen through nasal cannula in acute hypoxemic respiratory failure. N Engl J Med 2015;372(23):2185-2196.
3. Stefan MS, Shieh MS, Pekow PS, Rothberg MB, Steingrub JS, Lagu T, et al. Epidemiology and outcomes of acute respiratory failure in the United States, 2001 to 2009: a national survey. Journal of hospital medicine : an official publication of the Society of Hospital Medicine 2013;8(2):76-82.
4. Storms AD, Chen J, Jackson LA, Nordin JD, Naleway AL, Glanz JM, et al. Rates and risk factors associated with hospitalization for pneumonia with ICU admission among adults. BMC pulmonary medicine 2017;17(1):208.
5. Piraino T. Noninvasive Respiratory Support. Respiratory care 2021;66(7):1128-1135.
6. Ferreyro BL, Angriman F, Munshi L, Del Sorbo L, Ferguson ND, Rochwerg B, et al. Association of Noninvasive Oxygenation Strategies With All-Cause Mortality in Adults With Acute Hypoxemic Respiratory Failure: A Systematic Review and Meta-analysis. JAMA : the journal of the American Medical Association 2020;324(1):57-67.
7. Ni YN, Luo J, Yu H, Liu D, Ni Z, Cheng J, et al. Can High-flow Nasal Cannula Reduce the Rate of Endotracheal Intubation in Adult Patients With Acute Respiratory Failure Compared With Conventional Oxygen Therapy and Noninvasive Positive Pressure Ventilation?: A Systematic Review and Meta-analysis. Chest 2017;151(4):764-775.
8. Shen Y, Zhang W. High-flow nasal cannula versus noninvasive positive pressure ventilation in acute respiratory failure: interaction between PaO2/FiO2 and tidal volume. Crit Care 2017;21(1):285.
9. Zhao H, Wang H, Sun F, Lyu S, An Y. High-flow nasal cannula oxygen therapy is superior to conventional oxygen therapy but not to noninvasive mechanical ventilation on intubation rate: a systematic review and meta-analysis. Crit Care 2017;21(1):184.
10. Antonelli M, Conti G, Moro ML, Esquinas A, Gonzalez-Diaz G, Confalonieri M, et al. Predictors of failure of noninvasive positive pressure ventilation in patients with acute hypoxemic respiratory failure: a multi-center study. Intensive Care Med 2001;27(11):1718-1728.
11. Carrillo A, Gonzalez-Diaz G, Ferrer M, Martinez-Quintana ME, Lopez-Martinez A, Llamas N, et al. Non-invasive ventilation in community-acquired pneumonia and severe acute respiratory failure. Intensive Care Med 2012;38(3):458-466.
12. Demoule A, Girou E, Richard JC, Taille S, Brochard L. Benefits and risks of success or failure of noninvasive ventilation. Intensive Care Med 2006;32(11):1756-1765.
13. Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, et al. Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries. JAMA : the journal of the American Medical Association 2016;315(8):788-800.

14. Bellani G, Laffey JG, Pham T, Madotto F, Fan E, Brochard L, et al. Noninvasive Ventilation of Patients with Acute Respiratory Distress Syndrome. Insights from the LUNG SAFE Study. Am J Respir Crit Care Med 2017;195(1):67-77.
15. Thille AW, Contou D, Fragnoli C, Cordoba-Izquierdo A, Boissier F, Brun-Buisson C. Non-invasive ventilation for acute hypoxemic respiratory failure: intubation rate and risk factors. Crit Care 2013;17(6):R269.
16. Carteaux G, Millan-Guilarte T, De Prost N, Razazi K, Abid S, Thille AW, et al. Failure of Noninvasive Ventilation for De Novo Acute Hypoxemic Respiratory Failure: Role of Tidal Volume. Crit Care Med 2016;44(2):282-290.
17. Duan J, Han X, Bai L, Zhou L, Huang S. Assessment of heart rate, acidosis, consciousness, oxygenation, and respiratory rate to predict noninvasive ventilation failure in hypoxemic patients. Intensive Care Med 2017;43(2):192-199.
18. Frat JP, Ragot S, Coudroy R, Constantin JM, Girault C, Prat G, et al. Predictors of Intubation in Patients With Acute Hypoxemic Respiratory Failure Treated With a Noninvasive Oxygenation Strategy. Crit Care Med 2018;46(2):208-215.
19. Mauri T, Carlesso E, Spinelli E, Turrini C, Corte FD, Russo R, et al. Increasing support by nasal high flow acutely modifies the ROX index in hypoxemic patients: A physiologic study. Journal of critical care 2019;53:183-185.
20. Panadero C, Abad-Fernandez A, Rio-Ramirez MT, Acosta Gutierrez CM, Calderon-Alcala M, Lopez-Riolobos C, et al. High-flow nasal cannula for Acute Respiratory Distress Syndrome (ARDS) due to COVID-19. Multidiscip Respir Med 2020;15(1):693.
21. Roca O, Caralt B, Messika J, Samper M, Sztrymf B, Hernandez G, et al. An Index Combining Respiratory Rate and Oxygenation to Predict Outcome of Nasal High-Flow Therapy. Am J Respir Crit Care Med 2019;199(11):1368-1376.
22. Sztrymf B, Messika J, Bertrand F, Hurel D, Leon R, Dreyfuss D, et al. Beneficial effects of humidified high flow nasal oxygen in critical care patients: a prospective pilot study. Intensive Care Medicine 2011;37(11):1780-1786.
23. Tonelli R, Fantini R, Tabbi L, Castaniere I, Pisani L, Pellegrino MR, et al. Early Inspiratory Effort Assessment by Esophageal Manometry Predicts Noninvasive Ventilation Outcome in De Novo Respiratory Failure. A Pilot Study. Am J Respir Crit Care Med 2020;202(4):558-567.
24. Varipapa RJ, DiGiacomo E, Jamieson DB, Desale S, Sonti R. Fluid Balance Predicts Need for Intubation in Patients with Respiratory Failure Initiated on High Flow Nasal Cannula. Respiratory care 2020.
25. Zucman N, Mullaert J, Roux D, Roca O, Ricard JD, Contributors. Prediction of outcome of nasal high flow use during COVID-19-related acute hypoxemic respiratory failure. Intensive Care Med 2020;46(10):1924-1926.
26. Grieco DL, Menga LS, Raggi V, Bongiovanni F, Anzellotti GM, Tanzarella ES, et al. Physiological Comparison of High-Flow Nasal Cannula and Helmet Noninvasive Ventilation in Acute Hypoxemic Respiratory Failure. Am J Respir Crit Care Med 2020;201(3):303-312.
27. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162(1):55-63.
28. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for Reporting Machine Learning Analyses in Clinical Research. Circulation Cardiovascular quality and outcomes 2020;13(10):e006556.
29. Essay P, Mosier J, Subbian V. Rule-Based Cohort Definitions for Acute Respiratory Failure: Electronic Phenotyping Algorithm

30. van der Walt S, Millman J. Python in Science. In: van der Walt S, Millman J editors|. Secondary Title|. Vol. Volume|. Place Published|: Publisher|, Year|:Pages|.
31. Waskom M, Botvinnik O, Hobson P, Cole JB, Halchenko Y, Hoyer S, et al. seaborn: v0.5.0 (November 2014)
32. Pedregosa F, Vincent M, Thirion B, Grisel O, Blondel M, Prettenhofer P, et al. Scikit-learn: Machine Learning in Python
33. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc 2017;24(2):361-370.
34. Maragatham G, Devi S. LSTM Model for Prediction of Heart Failure in Big Data
35. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735-1780.
36. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation
37. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Scientific reports 2018;8(1):6085.
38. Balkan B, Essay P, Subbian V. Evaluating ICU Clinical Severity Scoring Systems and Machine Learning Applications: APACHE IV / IVa Case Study
39. Goldhill DR, Sumner A. APACHE II, data accuracy and outcome prediction. Anaesthesia 1998;53(10):937-943.
40. Xia J, Zhang Y, Ni L, Chen L, Zhou C, Gao C, et al. High-Flow Nasal Oxygen in Coronavirus Disease 2019 Patients With Acute Hypoxemic Respiratory Failure: A Multicenter, Retrospective Cohort Study. Crit Care Med 2020;48(11):e1079-e1086.

Quick Look

Current Knowledge

Patients that fail NIRS have disproportionately worse outcomes than patients successfully treated with NIRS. Early prediction of failure for patients with acute respiratory failure on noninvasive respiratory support can potentially optimize the balance between improved outcomes with NIRS success and disproportionately worse outcomes with NIRS failure.

What This Paper Contributes to Our Knowledge

This study shows that neural networks using clinically available data can be used to predict NIRS failure. Multiple RNN models showed varying levels of prediction accuracy with sufficient lead time to potentially impact patient outcomes.

Figure 1. Timeline of NIRS failure patients and those patients not requiring intubation post NIRS therapy. Observation window and time at-risk varied from 1 to 72 hours (shown in gray).

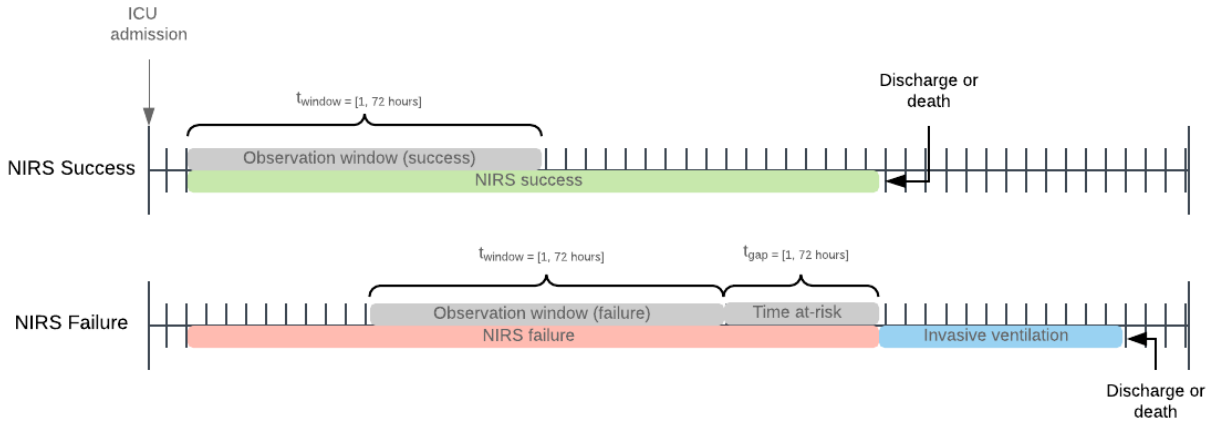

Figure 2. ROC and Precision-Recall curves for all models using 72-hour observation window and 1-hour time-at-risk for combined NIRS group.

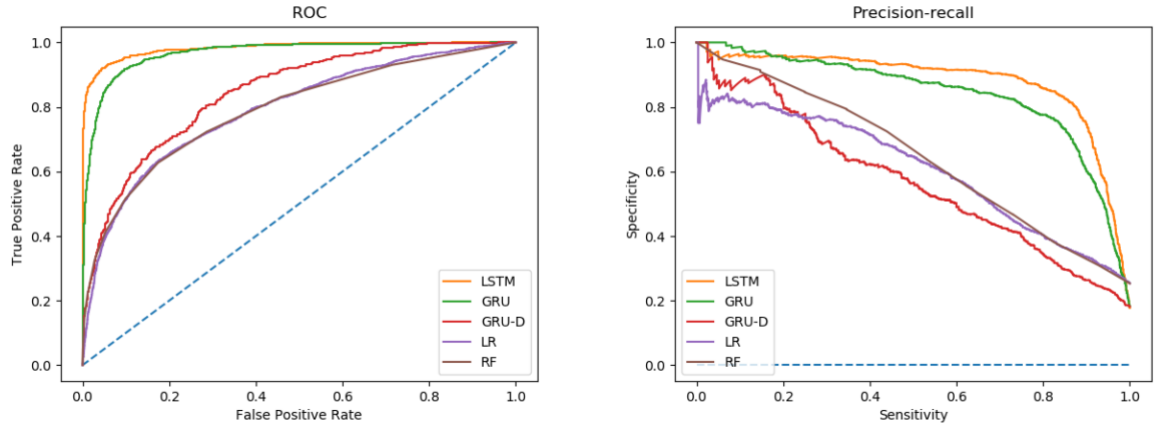

Figure 3. AUC change over variable observation window from 1 to 72 hours and time-at-risk window of 1 hour (left); AUC with constant 12-hour observation window size and variable time-at-risk window from 1 to 72 hours (i.e., observation window moving away from the time of failure) (right).

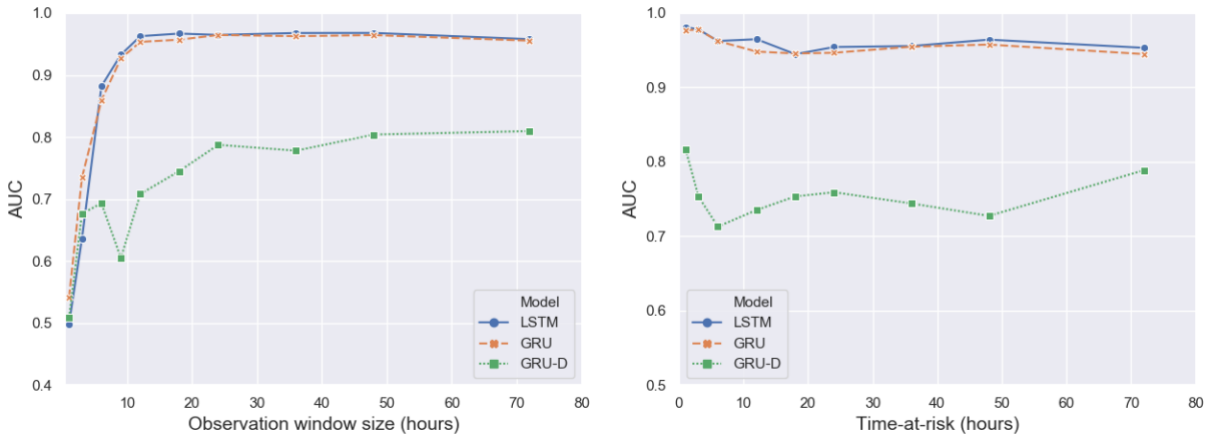

Table 1. Patient characteristics within each noninvasive ventilation group.

Parameters	Total	NIPPV Success	NIPPV Failure	HFNO Success	HFNO Failure
Patients, n (%)	22075	14,168	5087	1401	1419
Male, %	54.43	53.28	56.38	54.82	58.56
Age, median (IQR)	69 (58-78)	70 (59-79)	66 (56-75)	71 (59-81)	64 (52-74)
Race, %					
African American	5.53	5.87	4.92	5.80	3.97
Asian	0.94	0.86	1.15	0.86	1.13
Hispanic	0.06	0.07	0.04	0.00	0.14
Native American	2.55	0.13	3.08	2.01	5.32
Other/unknown	3.43	3.13	3.66	3.58	5.46
White	87.49	87.93	87.15	87.75	83.97
Ethnic group, %					
Hispanic or Latino	15.53	14.76	15.98	15.05	22.02
Not Hispanic or Latino	84.25	85.09	83.82	84.73	76.92
Unable or unwilling to answer	0.22	0.15	0.20	0.22	1.07
APACHE IVa score, median (IQR)	55 (41-69)	50 (39-63)	66 (51-84)	56 (43-70)	81 (59-105)
Respiratory rate, bpm, med. (IQR)	30 (14-36)	29 (13-35)	32 (17-38)	31 (24-36)	31 (16-37)
SpO2/FiO2 (first available)	194 (134-245)	196 (143-248)	184 (108-240)	147 (99-191)	150 (100-200)
Total therapy duration, days, med. (IQR)	1.62 (0-5)	0.74 (0-2)	3.01 (1-8)	0.17 (0-1)	1.26 (0-4)
Duration of IMV, days, med. (IQR)	1.87 (1-5)	-	2.05 (1-5)	-	1.23 (0-4)
Time to IMV from NIV start, hours, med. (IQR)	-	-	5.1 (2-23)	-	3.8 (2-22)
ICU length of stay, days, med. (IQR)	6.93 (4-12)	6.20 (4-10)	11.79 (6-20)	6.12 (3-11)	4.27 (2-11)
Mortality, %	17.83	8.34	24.89	34.63	71.98

APACHE – acute physiology and chronic health evaluation. ICU – intensive care unit. IQR – interquartile range.
Categorical variables are reported as proportion and continuous variables are reported as medians with interquartile range
NIPPV= noninvasive positive pressure ventilation
HFNO= high flow nasal oxygen

Table 2. Accuracy and ROC comparison between all models (Recurrent Neural Network and baseline) for three train and test cohorts (NIRS total, noninvasive positive pressure ventilation, and high flow nasal oxygen) using 72-hour observation window and 1-hour time-at-risk.

Cohort:	**NIRS**		**NIPPV**		**HFNO**	
Model	Accuracy, %	AUC	Accuracy, %	AUC	Accuracy, %	AUC
LSTM	94.04	0.9636	94.37	0.9666	82.12	0.8833
GRU	92.80	0.9538	93.66	0.9582	76.22	0.8668
GRU-D	83.37	0.7901	83.61	0.8149	77.08	0.6318
LR	84.56	0.7950	83.98	0.7894	74.54	0.7617
RF	84.56	0.7962	84.77	0.7868	77.16	0.7904

NIRS= Noninvasive respiratory support
NIPPV= Noninvasive positive pressure ventilation
HFNO= High flow nasal oxygen
LSTM= Long short-term memory
GRU= Gated recurrent unit
GRU-D= Gated recurrent unit with trainable decay
LR= Logistic regression
RF= Random forest