

---

## Research and Applications

# Respiratory support status from EHR data for adult population: classification, heuristics, and usage in predictive modeling

Sean C. Yu <sup>1,2</sup>, Mackenzie R. Hofford <sup>2,3</sup>, Albert M. Lai <sup>2</sup>, Marin H. Kollef<sup>4</sup>, Philip R.O. Payne <sup>2</sup>, and Andrew P. Michelson<sup>2,4</sup>

<sup>1</sup>Department of Biomedical Engineering, School of Engineering, Washington University School in St. Louis, St. Louis, Missouri, USA, <sup>2</sup>Institute for Informatics, Department of Medicine, Washington University School of Medicine in St. Louis, St. Louis, Missouri, USA, <sup>3</sup>Division of General Medicine, Department of Medicine, Washington University School of Medicine in St. Louis, St. Louis, Missouri, USA, and <sup>4</sup>Division of Pulmonary and Critical Care, Department of Medicine, Washington University School of Medicine in St. Louis, St. Louis, Missouri, USA

Corresponding Author: Sean C. Yu, MS, Washington University School of Medicine in St. Louis, 4444 Forest Park Avenue, Suite 6318, St. Louis, MI 63108, USA; [Sean.Yu@wustl.edu](mailto:Sean.Yu@wustl.edu)

Received 5 October 2021; Revised 9 December 2021; Editorial Decision 7 January 2022; Accepted 11 January 2022

### ABSTRACT

**Objective:** Respiratory support status is critical in understanding patient status, but electronic health record data are often scattered, incomplete, and contradictory. Further, there has been limited work on standardizing representations for respiratory support. The objective of this work was to (1) propose a practical terminology system for respiratory support methods; (2) develop (meta-)heuristics for constructing respiratory support episodes; and (3) evaluate the utility of respiratory support information for mortality prediction.

**Materials and Methods:** All analyses were performed using electronic health record data of COVID-19-tested, emergency department-admit, adult patients at a large, Midwestern healthcare system between March 1, 2020 and April 1, 2021. Logistic regression and XGBoost models were trained with and without respiratory support information, and performance metrics were compared. Importance of respiratory-support-based features was explored using absolute coefficient values for logistic regression and SHapley Additive exPlanations values for the XGBoost model.

**Results:** The proposed terminology system for respiratory support methods is as follows: Low-Flow Oxygen Therapy (LFOT), High-Flow Oxygen Therapy (HFOT), Non-Invasive Mechanical Ventilation (NIMV), Invasive Mechanical Ventilation (IMV), and ExtraCorporeal Membrane Oxygenation (ECMO). The addition of respiratory support information significantly improved mortality prediction (logistic regression area under receiver operating characteristic curve, median [IQR] from 0.855 [0.852–0.855] to 0.881 [0.876–0.884]; area under precision recall curve from 0.262 [0.245–0.268] to 0.319 [0.313–0.325], both  $P < 0.01$ ). The proposed generalizable, interpretable, and episodic representation had commensurate performance compared to alternate representations despite loss of granularity. Respiratory support features were among the most important in both models.

**Conclusion:** Respiratory support information is critical in understanding patient status and can facilitate downstream analyses.

**Key words:** respiratory support, oxygen support, supplemental oxygen, electronic health records, predictive analytics, machine learning

---

## INTRODUCTION

Managing respiratory status and providing appropriate respiratory support to prevent or mitigate hypoxemia is a critical aspect of clinical management, especially for patients suffering from respiratory conditions such as coronavirus disease 2019 (COVID-19).<sup>1</sup> Leveraging respiratory support information not only provides a more complete clinical picture, but it can also be used in downstream analyses such as sub-phenotyping or predictive modeling as a source of features or to identify endpoints.<sup>2,3</sup> However, generating generalizable conclusions from respiratory support information is difficult due to the heterogeneity of respiratory support methods and settings, the lack of standardized representations, and the poor quality of data regarding respiratory support.<sup>4</sup>

Standardization of patient data representation has accelerated knowledge discovery, replication of results, and translation into practice. While many parts of healthcare information have been standardized—such as the International Classification of Diseases (ICD) for diagnoses, LOINC for measurements and observations, or the all-encompassing UMLS—much of the patient information in EHR data, such as clinical cultures and respiratory support methods, remains unmapped.<sup>5,6</sup> Standardization of respiratory support is made challenging due to the inherent diversity of methods, especially when considering methods designed for subpopulations, variations, and modifiable settings. Worse yet, the lack of a widely adopted standardization schema has resulted in the usage of highly heterogeneous terms in literature and practice, even for identical concepts—for example, heated and humidified high-flow nasal cannula has been described as high-flow nasal cannula, high humidity nasal cannula, high-flow nasal oxygen therapy, or referred to by brand names such as Airvo<sup>TM</sup> or Optiflow<sup>TM</sup>.<sup>7–9</sup> Prior efforts to standardize respiratory support terms resulted in high granularity, which while necessary for coverage, can be excessive for downstream tasks. For example, the “Respiratory Therapy” concept in SNOMED-CT contains 14 children, of which “Oxygen Therapy” itself contains 14 children.<sup>10</sup> Thus, there is a need for a pragmatic and parsimonious standardization of respiratory support terms.

Beyond the lack of standardization, extraction of respiratory support status from electronic health records (EHRs) data is made challenging due to scattered, incomplete, and often contradictory documentation which necessitates the usage of auxiliary documentation and heuristics to determine respiratory support status.<sup>11–13</sup> Prior heuristic development work, however, are dataset-specific and focus on endotracheal intubation thus fail to address other strata of respiratory support.<sup>11,12</sup>

Therefore, the objective of the study was to (1) propose a preliminary, parsimonious, and pragmatic terminology system for respiratory support stratified by severity of hypoxemia; (2) develop (meta-)heuristics for the construction of respiratory support episodes from raw and heterogeneous EHR data; and (3) evaluate the terminology system and heuristics by measuring its impact on 30-day mortality prediction through the assessment of feature importance and feature ablation studies.

## MATERIALS AND METHODS

### Study design, data sources, and population

All patients  $\geq 18$  years of age admitted to all hospitals within a large Midwestern healthcare system serving the metropolitan St. Louis, mid-Missouri, and southern Illinois regions between March 1, 2020

and April 1, 2021 were eligible for inclusion. Patients were included if they had a COVID-19 PCR or antigen test, positive or negative, within 14 days prior to or 7 days after hospital admission. Patients were excluded if they had a hospital length of stay (LOS)  $< 24$  h to allow for a sufficient observation window for the predictive modeling study. Patients without associated demographics, comorbidities, or location data were excluded. Only patients with at least 5 heart rate and 5 SpO<sub>2</sub> measurements during the first 24 h of hospital arrival were included. EHR data, including demographics, vital signs, lab results, flowsheet entries, and so on, were extracted for all included subjects. This project was approved with a waiver of informed consent by the Washington University in St. Louis Institutional Review Board.

### Classification system of respiratory support methods

To facilitate generalizability and reproducibility of studies leveraging respiratory support information, we propose the following terminologies, in increasing severity: Low-Flow Oxygen Therapy (LFOT), High-Flow Oxygen Therapy (HFOT), Non-Invasive Mechanical Ventilation (NIMV), Invasive Mechanical Ventilation (IMV), and ExtraCorporeal Membrane Oxygenation (ECMO).<sup>11,14</sup>

### Meta-heuristics for identification of respiratory support episodes

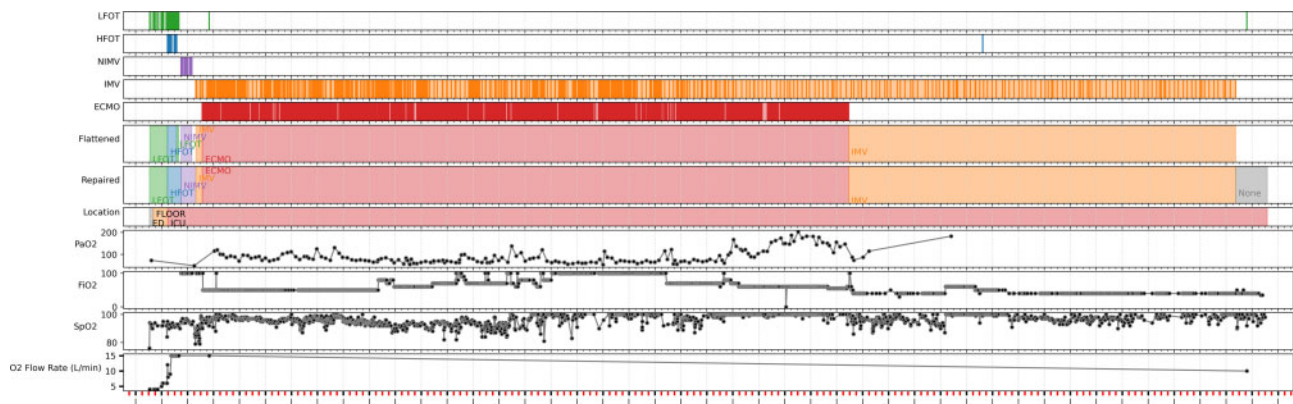
The authors of MIMIC-III, in addition to the data, also published code for data processing and analysis to expedite and encourage collaborative research.<sup>11</sup> Among the SQL scripts in their GitHub repository is one for calculating mechanical ventilation duration. Essentially, their logic was to chain together proximal pieces of documentation that are indicative of mechanical ventilation to form episodes with start and end times. Their heuristic has since been used successfully by researchers using the MIMIC-III dataset.<sup>13</sup> To generalize the heuristic for other forms of respiratory support beyond mechanical ventilation, and for other datasets beyond MIMIC-III, we developed a generalized version of the MIMIC-III heuristic—a “meta-heuristic”—to guide the development of heuristics for the assembly of respiratory support episodes as follows:

1. Define 2 parameters:
  - a. MIN\_DURATION for the minimum episode duration
  - b. EXTENSION\_TOLERANCE for the maximum allowable time gap between documentation for the formation of episodes
2. Identify timestamped documentation that are indicative of the presence of respiratory support
3. Link consecutive documentation occurring within EXTENSION\_TOLERANCE into episodes
4. Discard any episodes with duration less than or equal to MIN\_DURATION

Next, as we conceived of the respiratory support methods as being mutually exclusive, respiratory support episodes are “flattened” into a single timeline such that at any given time, a patient is on either no respiratory support or a single respiratory support method, by giving higher severity methods priority. Finally, the respiratory support trajectories are “repaired” such that gaps between episodes with a duration less than EXTENSION\_TOLERANCE are filled by extending the preceding episode. Also, gaps at the beginning and end of the patient stay (between encounter start time and first respiratory support episode start time, and between last respiratory sup-

**Table 1.** Respiratory support terminologies and heuristics

Term	Institution-specific evidence
Low-Flow Oxygen Therapy (LFOT)	Oxygen delivery method documentation: nasal cannula, non-rebreather mask, simple mask, venturi mask, aerosol mask, face tent.
High-Flow Oxygen Therapy (HFOT)	Oxygen flow rate $\leq 15$ L/min Oxygen delivery method documentation: high-flow nasal cannula, high humidity nasal cannula, optiflow
Non-Invasive Mechanical Ventilation (NIMV)	Oxygen flow rate $> 15$ L/min
Invasive Mechanical Ventilation (IMV)	Documentation of “NPPV Status” as “In Use” in the “Adult NPPV/NIV” flowsheet
ExtraCorporeal Membrane Oxygenation (ECMO)	Documentation of “Vent Status” as “In Use” in the “Ventilator Documentation” flowsheet Documentation of “Pump Flow (L/min)” or “ECMO Pump Speed (RPM)” in the ECMO or VAD flowsheets



**Figure 1.** Respiratory support trajectory example. This plot demonstrates the application of the respiratory support trajectory heuristics on a full, single patient encounter using data elements from Table 1, MIN\_DURATION of 6 h, and EXTENSION\_TOLERANCE of 24 h. The x-axis indicates time with each black tick indicating 24 h and red tick indicating 6 h. The top 5 subplots each pertain to a single respiratory support method where vertical lines indicate the times at which pieces of documentation serving as evidence for respiratory support were documented. The individual sub-trajectories are all merged into a single timeline as shown in the “flattened” subplot, after which it is repaired as according to the heuristic. The subplots below the “repaired” subplot provide context for the patient, showing patient location and measurements pertaining respiratory status. Abbreviations: ECMO: extracorporeal membrane oxygenation; ED: emergency department; HFOT: high-flow oxygen therapy; ICU: intensive care unit; IMV: invasive mechanical ventilation; LFOT: low-flow oxygen therapy; NIMV: non-invasive mechanical ventilation.

port episode end time and encounter end time), if less than MIN\_DURATION, are filled by extending the first or last episode, respectively.

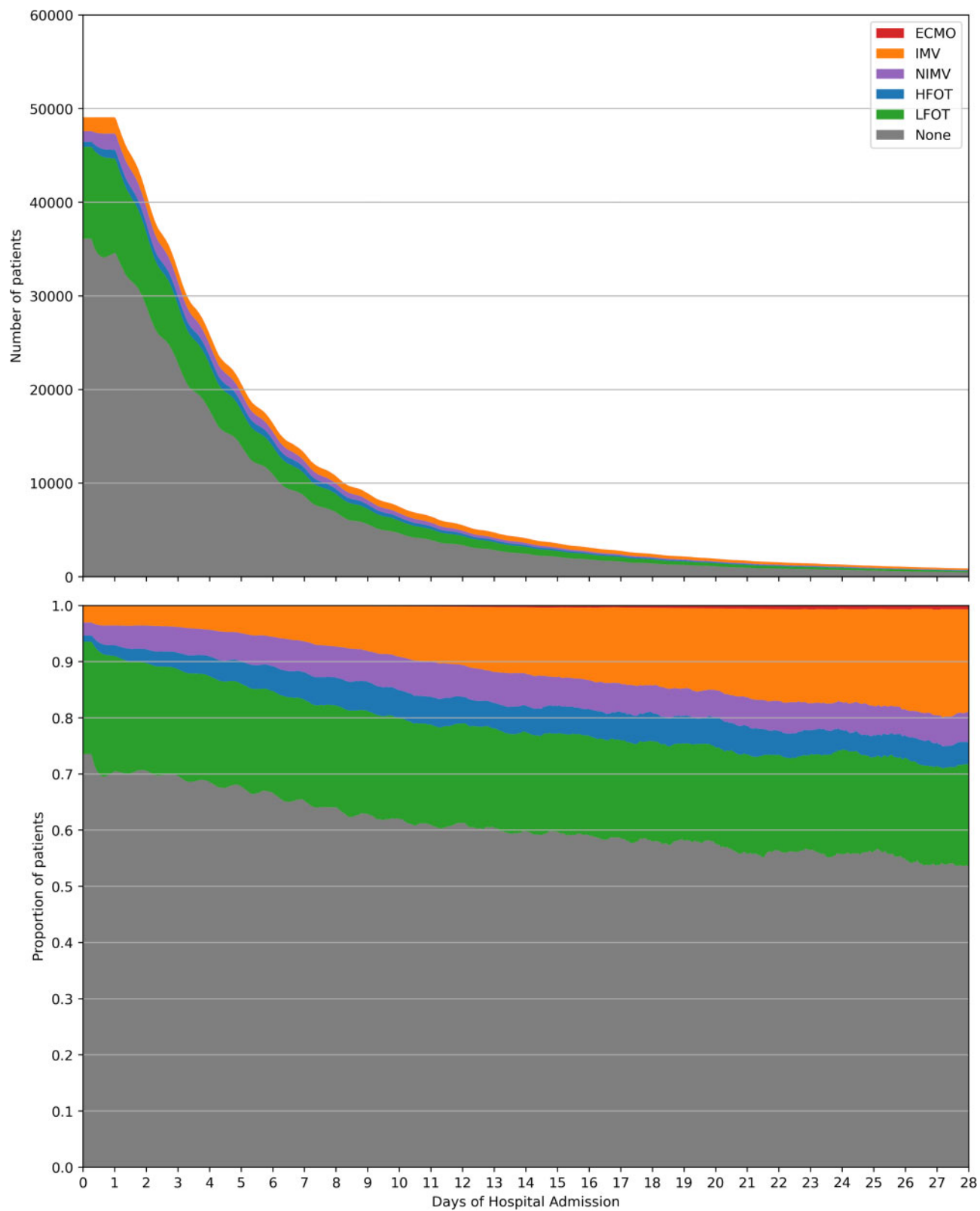
### Evaluation through in-hospital mortality prediction

A predictive modeling study was designed to evaluate the utility of the respiratory support information extracted through our heuristics on downstream analyses. The task was to predict in-hospital mortality within 30 days, at 24 h after hospital arrival for a COVID-19-tested, adult cohort presenting to an emergency department (ED); 121 *baseline* features were generated from demographic, laboratory, vital sign, and other clinical data extracted from the EHR. For most numeric measurements, the median value during the observation window was extracted, but for frequent measurements such as heart rate, other distributional statistics (25th quantile, 75th quantile, and interquartile range) were also extracted.

Ten additional, respiratory-support-derived features were generated using the *proposed* classification schema and heuristics which included duration of respiratory support per type, and the last respiratory support during the observation period. For comparison, we identified a small set of measurements *related* to respiratory status: fraction of inspired oxygen (FiO<sub>2</sub>) and oxygen flow rate. We also extracted the EHR-native representation of respiratory support

called oxygen delivery method (*O<sub>2</sub> Del Method*) which included ETT, CPAP, T-Piece, and so on. Lastly, we also considered a set of features based on the proposed classification, but using the *raw* time-stamped data prior to assembly into episodes. The feature sets including explicit respiratory support information—*O<sub>2</sub> Del Method*, *Raw*, and *Proposed*—all also include “Baseline” and “Related” features (Online Supplemental Materials, eTable 1, eFigure 1).

The compared algorithms were logistic regression (LogReg) and XGBoost (XGB). For LogReg, features were standardized and mean-imputed whereas for XGB, features were left as is. Hyperparameters such as regularization strength were optimized for log loss using the baseline features through 1000-iteration, 4-fold cross-validation (Online Supplemental Materials, eAppendix A). Once optimal hyperparameters were identified, five replicates of 2-fold cross-validation was performed to generate a distribution of performance metrics: area under receiver operating characteristic curve (AUROC), area under precision recall curve (AUPRC), and negative log loss.<sup>15</sup> The distributions of performance metrics were compared using the Wilcoxon signed-rank test (2-way, paired).<sup>16</sup> Feature importance was quantified using SHapley Additive exPlanations (SHAP) values for the XGBoost model, and coefficient values for the logistic regression model, both of which were aggregated over 100 bootstrap samples.<sup>17,18</sup>



**Figure 2.** Respiratory support utilization aligned at arrival. All patients were aligned at ED arrival, and their usage of respiratory support was plotted for the first 4 weeks. The top subplot shows the total number of patients utilizing each respiratory support method, whereas the bottom subplot shows the proportion. As expected, patients who have been in the hospital longer are more likely to be on higher levels of respiratory support. Abbreviations: ECMO: extracorporeal membrane oxygenation; HFOT: high-flow oxygen therapy; IMV: invasive mechanical ventilation; LFOT: low-flow oxygen therapy; NIMV: non-invasive mechanical ventilation.

**Table 2.** Cohort characteristics

Variable	Total (n = 45,908)	Outcome = in-hospital mortality within 30 days of index time		P <sup>a</sup>
		Yes (n = 1601, 3.5%)	No (n = 44,307, 96.5%)	
Age (years), median (IQR)	64.0 (51.0–76.0)	73.0 (62.0–83.0)	64.0 (50.0–76.0)	< 0.01*
Male, n (%)	22,638 (49.3%)	889 (55.5%)	21,749 (49.1%)	< 0.01*
Race, n (%)				< 0.01*
White	28,032 (61.1%)	933 (58.3%)	27,099 (61.2%)	0.021
Black	16,706 (36.4%)	576 (36.0%)	16,130 (36.4%)	0.747
Asian	340 (0.7%)	17 (1.1%)	323 (0.7%)	0.168
Other/unknown	830 (1.8%)	75 (4.7%)	755 (1.7%)	< 0.01*
BMI, median (IQR)	27.5 (23.2–33.4)	27.1 (22.7–32.2)	27.5 (23.2–33.4)	< 0.01*
COVID-19 positive, n (%)	8332 (18.1%)	502 (31.4%)	7830 (17.7%)	< 0.01*
Respiratory support duration during observation window (h), mean ± SD				
None	16.7 ± 10.4	7.5 ± 10.2	17.0 ± 10.2	< 0.01*
LFOT	5.31 ± 9.16	7.36 ± 10.05	5.24 ± 9.12	< 0.01*
HFOT	0.39 ± 2.70	1.86 ± 5.73	0.34 ± 2.50	< 0.01*
NIMV	0.76 ± 3.74	1.47 ± 5.23	0.73 ± 3.68	< 0.01*
IMV	0.85 ± 4.21	5.79 ± 9.73	0.67 ± 3.75	< 0.01*
ICU transfer, n (%)	10,311 (22.5%)	1321 (82.5%)	8990 (20.3%)	< 0.01*
Total LOS (h), median (IQR)	99.7 (59.0–172.7)	166.8 (83.4–313.6)	98.6 (58.2–170.1)	< 0.01*
In-hospital mortality, n (%)	1682 (3.7%)	1601 (100.0%)	81 (0.2%)	< 0.01*

*Abbreviations:* BMI: body mass index; COVID-19: Coronavirus disease 2019; ECMO: extracorporeal membrane oxygenation; HFOT: high-flow oxygen therapy; ICU: intensive care unit; IMV: invasive mechanical ventilation; IQR: interquartile range; LFOT: low-flow oxygen therapy; LOS: length of stay; NIMV: non-invasive mechanical ventilation; SD: standard deviation.

<sup>a</sup>Comparison of variables between those with and without the primary outcome of 30-day in-hospital mortality was performed using Mann-Whitney U test for continuous variables, and  $\chi^2$  for categorical variables. Statistical significance,

\* $P < 0.01$ , is indicated by.

## Statistical analysis

Variables were summarized using frequencies and proportions for categorical data and medians and interquartile ranges or means and standard deviations for continuous data. Statistical comparisons were performed using the Chi-square and Mann-Whitney U tests where appropriate unless specified otherwise. A  $P$ -value  $< 0.01$  was considered statistically significant. All resampling analyses, cross-validation and bootstrap, were performed using a fixed seed. All analysis and figure generation were performed with Python version 3.7.1 (Python Software Foundation, Beaverton, OR) using the following packages: scipy, numpy, pandas, matplotlib, sklearn, xgboost, and shap.<sup>17,19–24</sup>

## RESULTS

The severity-stratified respiratory support methods and the documentation serving as evidence for each method are listed in [Table 1](#). Examples of the full heuristic application process, along with documentation germane to respiratory status, can be found in [Figure 1](#). [Figure 2](#) shows respiratory support utilization over time, with patients temporally aligned at ED arrival.

Cohort characteristics and outcomes for the patient population used in this study can be seen in [Table 2](#). During the study period there were 45 908 hospitalizations lasting at least 24 h available for analysis. Of these, 1601 (3.5%) experienced in-hospital death within 30 days. Non-survivors were older, more likely to be male, more likely to be COVID-19 positive, and have a longer length of stay ([Table 2](#)).

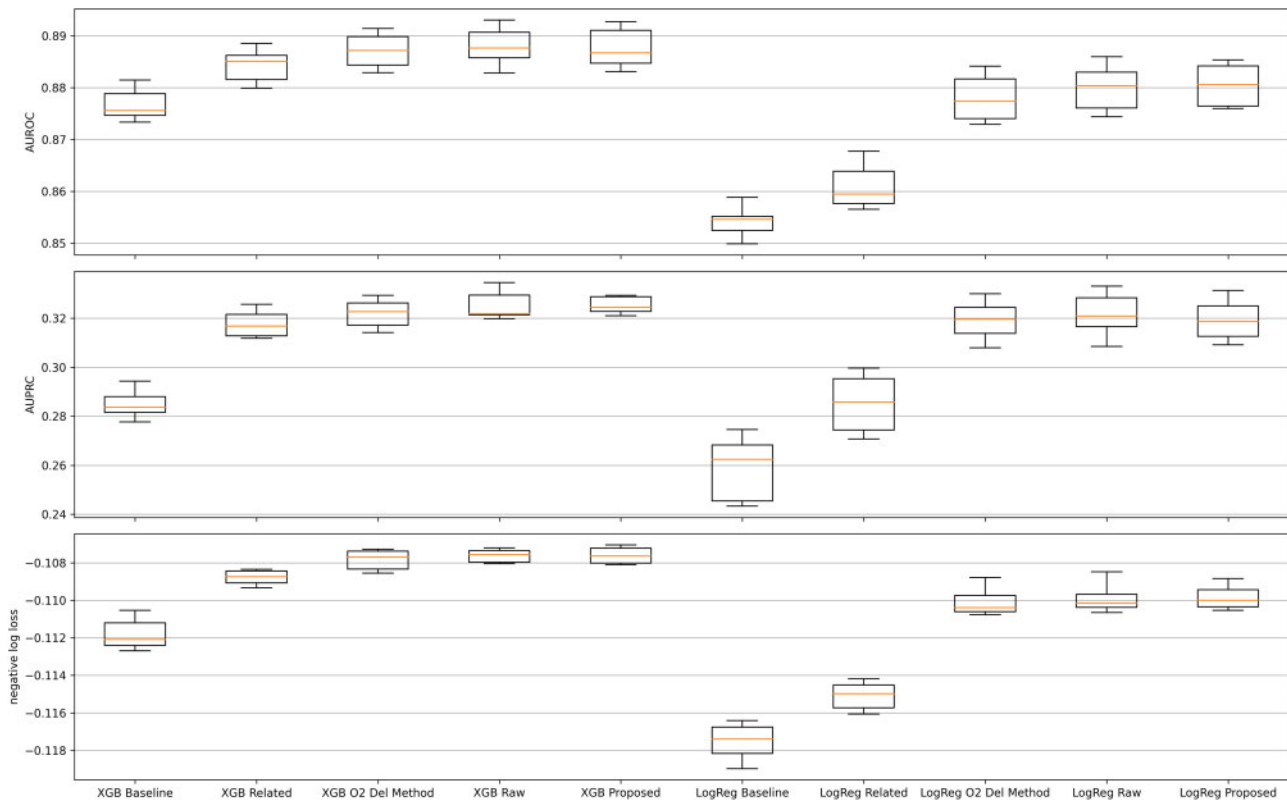
The optimized hyperparameters ([Online Supplemental Materials, eAppendix A](#)) were used for generating distributions of performance metrics through repeated cross-validation ([Figure 3](#)). For both XGB and LogReg, the addition of “Related” features significantly improved on “Baseline,” and the addition of “O<sub>2</sub> Del Method,” “Raw,” or “Proposed” features improved on “Related” across all 3 metrics: AUROC, AUPRC, and negative log loss ([Figure 3, Online Supplemental Materials, eTable 2, eTable 3](#)). However, “O<sub>2</sub> Del Method,” “Raw,” and “Proposed” rarely differed significantly, and when they did, the differences were very small as was the case for LogReg AUROC between “O<sub>2</sub> Del Method” and “Proposed” (0.887 [0.884–0.890] and 0.887 [0.885–0.891],  $P < 0.01$ , [Online Supplemental Materials, eTable 3](#)).

Six of the top 20 most impactful features for the LogReg model were respiratory-support-derived features, including last respiratory support, IMV, and LFOT duration ([Figure 4](#)). For the XGB model, respiratory-support-derived features ranked 10th (last respiratory support, IMV) and 17th (LFOT duration) ([Figure 5](#)).

## DISCUSSION

In this study, we (1) propose a preliminary, parsimonious, and pragmatic terminology system for respiratory support methods; (2) develop (meta-)heuristics for extraction of respiratory support information from EHR data; and (3) investigate the utility of the respiratory support information extracted through the proposed terminology system and heuristics via a mortality prediction study in a COVID-19-tested, ED-admit, adult cohort. The developed heuristic





**Figure 3.** In-hospital mortality prediction feature ablation performance comparison. Comparison of in-hospital mortality prediction performance for LogReg and XGB models with varying sets of features from 5-repeat, 2-fold cross-validation. “Baseline” includes demographics, common lab results, and vital signs from the EHR data. “Related” also includes O<sub>2</sub> flow rate and fraction of inspired oxygen. In addition, “O<sub>2</sub> Del Method” includes the EHR-native representation of respiratory support status, “Raw” includes data from the proposed approach prior to assembly into episodes, and “Proposed” includes features derived from respiratory support episodes based on the proposed approach. The center horizontal line represents median, box represents the interquartile range between 25th and 75th percentiles, and whiskers represent 2.5th and 97.5th percentiles. Abbreviations: AUPRC: area under precision recall curve; AUROC: area under receiver operating characteristic curve; LogReg: logistic regression; XGB: extreme gradient boosted trees model.

was successfully applied to EHR data to extract respiratory support episodes, which were then used for in-hospital mortality prediction as features, which were found to be among the most important features for both LogReg and XGB models.

Compared to models using demographics and commonly documented lab results and vital signs, the addition of respiratory-support-related information, FiO<sub>2</sub> and O<sub>2</sub> flow rate, significantly improved prediction performance for both XGB and LogReg across all measured performance metrics. Moreover, the additional inclusion of explicit respiratory support information further improved performance significantly, again for both model types and across all metrics.

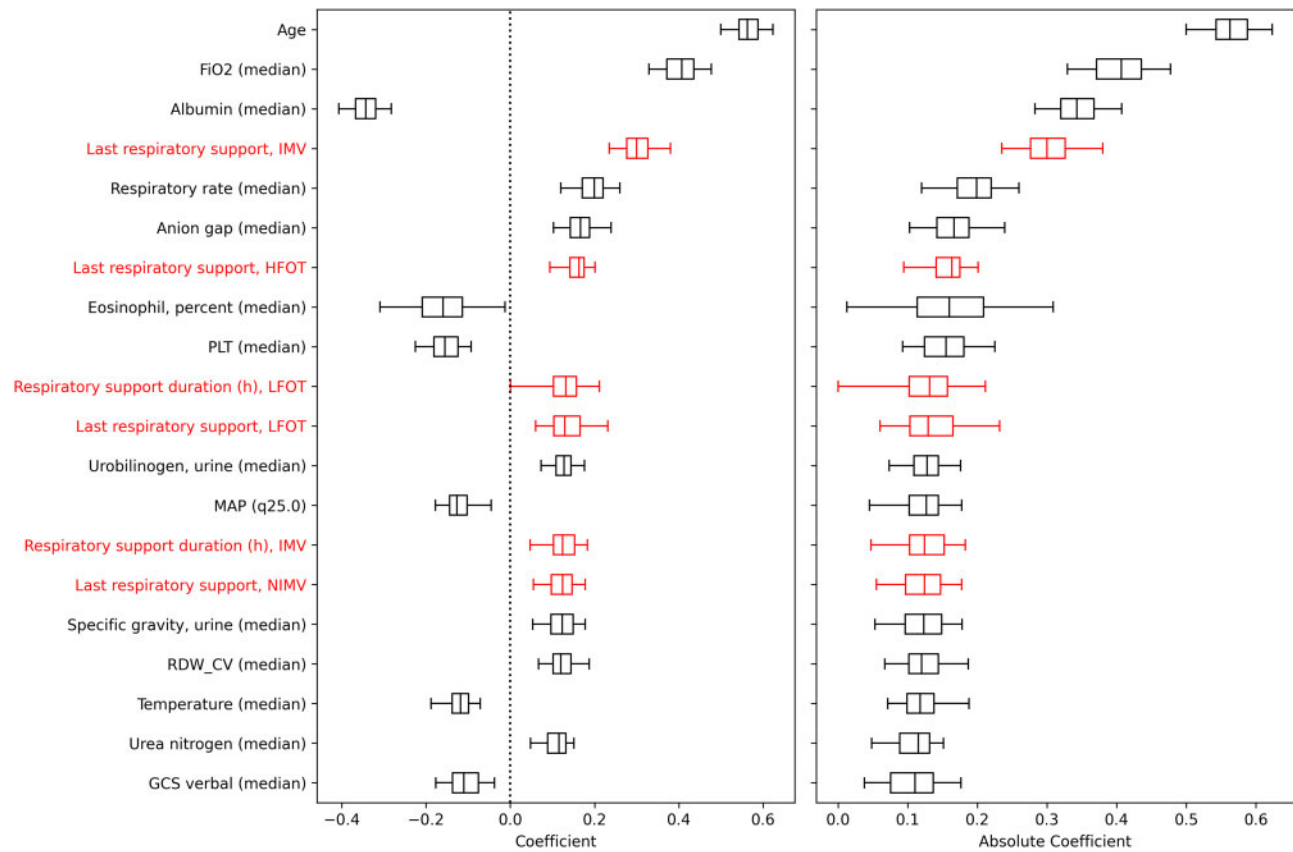
Because there are no other dataset-agnostic, full-severity-spanning classification, and heuristic system for respiratory support information found in literature, we compared our proposed approach against 2 other methods of explicit respiratory status representation: “O<sub>2</sub> Del Method” which used the EHR-native representation and “Raw” which uses the data elements for the proposed approach but prior to assembly into episodes (Online Supplemental Materials, eTable 2, eFigure 1). The proposed representation had commensurate performance to alternate representations of explicit respiratory support information, despite loss of both conceptual and temporal granularity resulting from the aggregation of heterogeneous timestamped raw data into the more human-understandable format of encounter-spanning series of episodes (Figure 3).

The increase in model performance associated with the addition of explicit respiratory support information in XGB was less than that of LogReg. We hypothesize that more complex models are able to infer respiratory support status or reconstitute information contained in respiratory status based on other features.

For the models using features from the proposed approach, IMV status and duration was an important predictor which is unsurprising—patients who are intubated are known to have higher rates of in-hospital mortality.<sup>25</sup> This result simply underscores the importance of leveraging respiratory support information, especially those of high severity, for understanding patient status. However, even features based on low-flow oxygen therapy status were among the most important features for both the XGB and LogReg models, indicating that lower severity respiratory support information is also critical for developing a complete clinical picture of patients.

In this study, respiratory support information was used as features for in-hospital mortality prediction. However, there are many other potential uses of respiratory support information, such as endpoint identification, patient sub-phenotype discovery, patient trajectory analytics, or characterization of patient cohorts.<sup>13,14,26</sup>

As documentation regarding respiratory support status varies across time and across sites, identification of timestamped documentation that serve as evidence for respiratory support cannot be generalized and must be specified in each study by researchers with



**Figure 4.** Logistic regression feature importance. The left subplot shows the top 20 most important features in the logistic regression model based on coefficient values aggregated over 100 bootstrap samples, and the right subplot shows the absolute coefficient values. For each feature, the center vertical line represents median, box represents the interquartile range between 25th and 75th percentiles, and whiskers represent 2.5th and 97.5th percentiles. Features based on respiratory support information are colored/shaded in red. Abbreviations: FiO<sub>2</sub>: fraction of inspired oxygen; GCS: Glasgow Coma Scale; HFOT: high-flow oxygen therapy; IMV: invasive mechanical ventilation; LFOT: low-flow oxygen therapy; MAP: mean arterial pressure; NIMV: non-invasive mechanical ventilation; PLT: platelet count; RDW\_CV: red blood cell distribution width coefficient of variation.

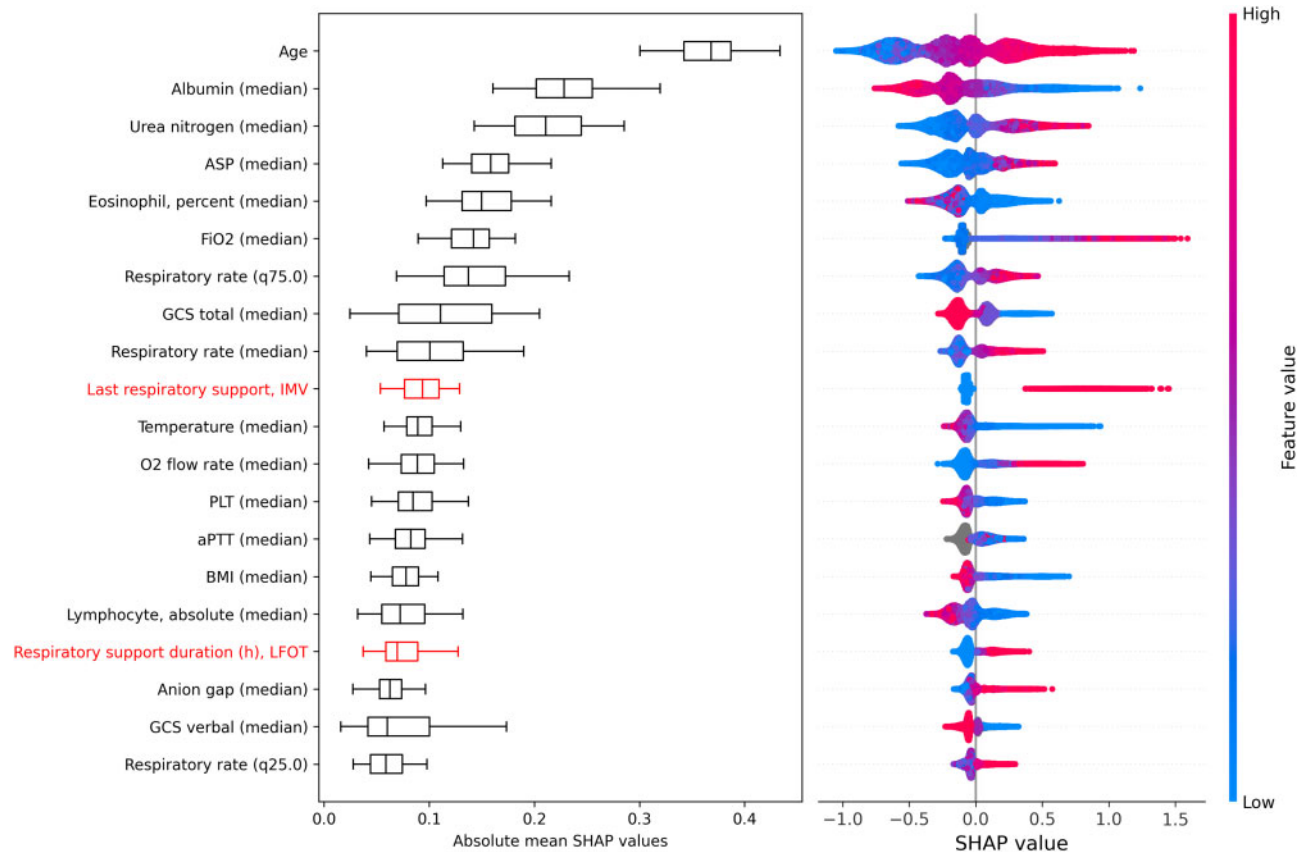
appropriate knowledge of local practice patterns contained within the dataset. Therefore, a (meta-)heuristic was developed which provides the structure for developing heuristics to establish episodes for any respiratory support method.

As is typical of studies using EHR data, this study suffers from missingness and inaccuracy of information. For example, we identified patients admitted through the ED with a recent positive COVID-19 test who were transferred to and remained in the ICU for several days, yet had no documentation of respiratory support throughout their entire stay. While heuristics can work with scattered, conflicting, and incomplete documentation, significant missingness will still result in unrealistic scenarios. There is trade-off in setting the MIN\_DURATION parameter—if it is too long then temporary/interim respiratory supports will be underrepresented; conversely, if it is too short then the heuristic will allow for unrealistically rapid oscillation among respiratory support methods. Additionally, tracheostomy status was considered orthogonal to the proposed system. NIMV is often used nightly for sleep apnea; thus, researchers utilizing the heuristic must decide whether to ignore those episodes based on their needs. Also, patients can be connected to a device for respiratory support, but not be actively using them (eg, delivering no or low-flow oxygen through a device capable of delivering high-flow oxygen), thus researchers must decide which is more important for their work: the occupation of the device or the

active use of the device. Respiratory support methods are ever-evolving—helmet NIV and high-flow nasal cannula, for instance, have only recently been used for adult patients, meaning that these terminology systems will also require regular revisiting and updating.<sup>27</sup>

## CONCLUSION

To facilitate generalizable and reproducible research, a terminology system was developed for standardized representation of respiratory support methods. (Meta-)heuristics were also developed to enable extraction of respiratory support episodes from EHR data, and transformation into encounter-spanning set of respiratory support trajectories. To demonstrate the utility of respiratory support information extracted through proposed methods, feature ablation and feature importance analyses were performed via an in-hospital mortality prediction study for COVID-19-tested, ED-admit, adult patients. The addition of features generated from the proposed approach significantly improved model performance. Further, those features were found to be among the most important for models. Finally, the proposed approach, which generated more interpretable and generalizable representations, despite the loss of conceptual and temporal granularity, had commensurate performance to alternate representations of explicit respiratory support information.



**Figure 5.** XGBoost SHAP feature importance. The left subplot shows the top 20 most important features in the XGB model based on absolute mean SHAP values aggregated over 100 bootstrap samples. The right subplot shows the individual SHAP value for the same top 20 features, for all encounters in the full dataset. For each feature, the center vertical line represents median, box represents the interquartile range between 25th and 75th percentiles, and whiskers represent 2.5th and 97.5th percentiles. Features based on respiratory support information are colored/shaded in red. Abbreviations: ASP: aspartate aminotransferase; BMI: body mass index; FiO<sub>2</sub>: fraction of inspired oxygen; GCS: Glasgow Coma Scale; IMV: Invasive Mechanical Ventilation; LFOT: low-flow oxygen therapy; PLT: platelet count.

## FUNDING

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

## AUTHOR CONTRIBUTIONS

SCY, MH, and APM conceived of the idea and planned the experiments. SCY performed the analysis and drafted the manuscript. MH and APM made substantial contributions to the language of the manuscript. AML, MHK, and PROP supervised the project for methodological rigor and veracity. All authors provided critical feedback and contributed to the final manuscript.

## SUPPLEMENTARY MATERIAL

[Supplementary material](#) is available at *Journal of the American Medical Informatics Association* online.

## CONFLICT OF INTEREST STATEMENT

None declared.

## DATA AVAILABILITY

The EHR data underlying this article cannot be shared publicly to protect the privacy of individuals included in the study, and thus is available only for authorized Washington University in St. Louis (WashU) investigators.

## REFERENCES

- Shelly MP, Nightingale P. Respiratory support. *BMJ* 1999; 318(7199): 1674–7.
- Vincent J-L, Moreno R, Takala J, *et al*. The SOFA (Sepsis-Related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Med* 1996; 22(7): 707–10.
- Yadav AS, Li Y-C, Bose S, *et al*. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *Lancet Digit Health* 2020; 2(10): e516–e25.
- Chute CG. Clinical classification and terminology: some history and current observations. *J Am Med Inform Assoc* 2000; 7(3): 298–303.
- McDonald CJ, Huff SM, Suico JG, *et al*. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin Chem* 2003; 49(4): 624–33.
- Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32 (Database issue): D267–70.
- Nishimura M. High-flow nasal cannula oxygen therapy in adults. *J Intensive Care* 2015; 3(1): 15–8.
- Xu Q, Wang T, Qin X, *et al*. Early awake prone position combined with high-flow nasal oxygen therapy in severe COVID-19: a case series. *Crit Care* 2020; 24(1): 1–3.
- Ang K, Green A, Ramaswamy K, *et al*. Preoxygenation using the Opti-flow™ system. *Br J Anaesth* 2017; 118(3): 463–4.
- Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. *Stud Health Technol Inform* 2006; 121: 279–90.



11. Johnson AE, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3(1): 1–9.
12. Cao H, Lee K, Ennett CM, *et al.*, eds. Heuristics to determine ventilation times of ICU patients from the MIMIC-II database. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology; 2010: IEEE.
13. 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(9): 1936–46.
14. Mody A, Lyons PG, Vazquez Guillamet C, *et al.* The clinical course of coronavirus disease 2019 in a US hospital system: a multistate analysis. *Am J Epidemiol* 2021; 190(4): 539–52.
15. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput* 1998; 10(7): 1895–923.
16. Demšar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 2006; 7: 1–30.
17. Lundberg SM, Erion G, Chen H, *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020; 2(1): 56–67.
18. Lundberg SM, Lee S-I. editors. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017.
19. Virtanen P, Gommers R, Oliphant TE, *et al.*; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020; 17(3): 261–72.
20. Harris CR, Millman KJ, van der Walt SJ, *et al.* Array programming with NumPy. *Nature* 2020; 585(7825): 357–62.
21. McKinney W, ed. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference; June 28–July 3, 2010; Austin, TX.
22. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng* 2007; 9(3): 90–5.
23. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–30.
24. Chen T, Guestrin C. editors. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016.
25. Wang Y, Lu X, Li Y, *et al.* Clinical course and outcomes of 344 intensive care patients with COVID-19. *Am J Respir Crit Care Med* 2020; 201(11): 1430–4.
26. Bobroske K, Larish C, Cattrell A, *et al.* The bird’s-eye view: a data-driven approach to understanding patient journeys from claims data. *J Am Med Inform Assoc* 2020; 27(7): 1037–45.
27. Patel BK, Wolfe KS, Pohlman AS, *et al.* Effect of noninvasive ventilation delivered by helmet vs face mask on the rate of endotracheal intubation in patients with acute respiratory distress syndrome: a randomized clinical trial. *JAMA* 2016; 315(22): 2435–41.