Unmasking Societal Biases in Respiratory Support for ICU Patients through Social Determinants of Health

Mira Moukheiber[1] , **Lama Moukheiber** [1] , **Dana Moukheiber** [1] and **Hyung-Chul Lee**[2,*]
[1]Massachusetts Institute of Technology
[2]Seoul National University College of Medicine, Seoul National University Hospital, Department of Anesthesiology and Pain Medicine
vital@snu.ac.kr

Abstract

In critical care settings, where precise and timely interventions are crucial for health outcomes, evaluating disparities in patient outcomes is important. Current approaches often fall short in comprehensively understanding and evaluating the impact of respiratory support interventions on individuals affected by social determinants of health. Attributes such as gender, race, and age are commonly assessed and essential, but provide only a partial view of the complexities faced by diverse populations. In this study, we focus on two clinically motivated tasks: prolonged mechanical ventilation and successful weaning. We also perform fairness audits on the models' predictions across demographic groups and social determinants of health to better understand the health inequities in respiratory interventions in the intensive care unit. We also release a temporal benchmark dataset, verified by clinical experts, to enable benchmarking of clinical respiratory intervention tasks.

1 Introduction

Critically-ill patients often find themselves in the intensive care unit (ICU) seeking specialized support for respiratory distress [Doyle *et al.*, 1995; Ware and Matthay, 2000]. Despite advances in supportive treatments, the in-hospital mortality rate remains 40% for conditions such as acute lung injury and acute respiratory distress syndrome [Rubenfeld *et al.*, 2005; Sweatt and Levitt, 2014]. Managing respiratory distress involves intricate treatment measures, including invasive mechanical ventilation [Esteban *et al.*, 2000], non-invasive mechanical ventilation [Esquinas *et al.*, 2017], and high-flow nasal cannula [Frat *et al.*, 2017]. However, existing recommendations and outcomes, especially regarding intubation and weaning procedures for ICU patients, remain controversial and poorly understood [Zuo *et al.*, 2020; Papoutsi *et al.*, 2021; Suo *et al.*, 2021; Wanis *et al.*, 2023; Kondrup *et al.*, 2023].

Health disparities are widespread within marginalized communities, particularly across respiratory diseases, acting as significant contributors to morbidity and mortality in the United States [Schraufnagel *et al.*, 2013; Moy *et al.*, 2017; Thakur *et al.*, 2014]. These communities, facing systemic barriers and social inequalities, bear a disproportionate burden of adverse health outcomes due to factors such as economic instability, limited access to education, and housing insecurity [Purnell *et al.*, 2016]. Recognizing and evaluating social determinants of health (SDOH) is important for addressing the complex factors that influence the quality of and access to healthcare [Holmes Fee *et al.*, 2023; Bundy *et al.*, 2023; Lua *et al.*, 2023; Marmot, 2005; Nakagawa *et al.*, 2023; Moukheiber *et al.*, 2024]. A comprehensive understanding of SDOH can offer insight into potential disparities that might be overlooked within studies focused solely on traditional attributes such as age, race, gender, and health insurance, making it important for the evaluation of algorithmic bias [Celi *et al.*, 2022; Nazer *et al.*, 2023].

Observational health data, derived from EHRs, presents a valuable resource with the potential to enhance healthcare. Although efforts have been made to establish benchmarks for EHR data [Harutyunyan *et al.*, 2019; Purushotham *et al.*, 2018; Wang *et al.*, 2020; Gupta *et al.*, 2022; Rocheteau *et al.*, 2021], these benchmarks primarily focus on conventional clinical prediction tasks, such as mortality and length-of-stay predictions. To the best of our knowledge, the current benchmark datasets lack dynamic aspects of pulmonary function, encompassing complex respiratory treatment strategies, ventilator settings, and pulmonary mechanics, along with other clinically-relevant variables for guiding decision-making. Furthermore, current ICU benchmark datasets often lack a link to SDOH, which limits the ability to fully understand and address the complexities influencing the recommendations for intubation and weaning in ICU patients. The recently released MIMIC-IV dataset, linked to SDOH features based on patient zip code [Yang *et al.*, 2023], enables detailed fairness assessments of SDOH dimensions. Therefore, we use MIMIC-IV to benchmark clinical respiratory intervention tasks for ICU patients.

In this work, we benchmark two time-dependent clinically-motivated prediction tasks, including successful weaning and prolonged mechanical ventilation. We further evaluate the differences in performance gaps across protected attributes, including age, gender, race, and English proficiency, as well as eight SDOH features. We also release a dataset with hourly

*Corresponding author

intervals to enable benchmarking of respiratory intervention tasks. This dataset is enriched with ventilation data and a wide range of other covariates, including demographics, lab results, measurements, illness severity scores, treatment interventions, and outcome variables. It covers 50,920 patients admitted to the ICU, with records collected over 90 days. This dataset can help address weaning delays and failures, optimize strategies for respiratory support, identify efficiencies in clinical practices, provide decision support to attending physicians regarding intubation decisions in the ICU, and facilitate time-series and reinforcement learning applications.

2 Methods

2.1 DataSet

Dataset Overview

We introduce a temporal benchmark for clinical respiratory interventions, a 90-day hourly ventilation dataset derived from MIMIC-IV version 2.2. MIMIC-IV is an open-access, de-identified database compiled from electronic health records of patients admitted to the ICU or Emergency Department at the Beth Israel Deaconess Medical Center in Boston between 2008 and 2019. Our temporal data includes confounding variables categorized into static and dynamic variables. Figure1 depicts hourly characteristics for a single patient's ICU stay over 30 days.

Cohort Selection

In the MIMIC-IV database, a patient can have multiple ICU stays over the years or experience transitions between different ICUs during the same hospital admission. To prevent data leakage and maintain data integrity, we choose the first ICU stay with respiratory support for each patient. This approach ensures that data used for modeling is independent and not influenced by information from subsequent stays. Additionally, patients with a do not resuscitate or do not intubate directives and those who were on invasive ventilation 24 hours before ICU admission are excluded, resulting in a total of 50,920 patients.

Data Extraction and Preprocessing

The majority of timestamps for time-varying variables in the raw MIMIC data are presented in the year, month, day, hour, minute, and second format, offering the potential to derive granular data for comprehensive medical analysis. The sporadic recording of multiple observations allows us to aggregate the data into hourly bins to improve the data density and analytical consistency. Our dataset spans the period of 0 to 2160 hours (equivalent to 90 days) following ICU admission for each subject.

Patient-level Static Variables

Static parameters extracted for patients, as outlined in Table 1, encompass demographic variables, comorbidity scores assessing neurological function (such as the Glasgow Coma Scale and its components: eye opening, verbal, and motor responses), as well as an evaluation of patient organ dysfunction (based on the maximum Sequential Organ Failure Assessment score) performed 24 hours after admission to the ICU.

Variable	Description
intime	ICU admission time
outtime	ICU discharge time
gender	patient gender
anchor year	patient shifted year
anchor age	patient age in anchor year
insurance	patient insurance type
language	English proficiency indicator
marital status	patient marital status
race	patient race
first_careunit	ICU type during first admission
pbw_kg	patient predicted body weight (kg)
height_inch	patient height (inches)
elixhauser_vanwalraven	Elixhauser-Van Walraven score
gcs	Glasgow Coma Scale (GCS) score
gcs_motor	GCS motor response component
gcs_verbal	GCS verbal response component
gcs_eyes	GCS eye-opening response component
gcs_unable	Endotracheal tube indicator
sofa_24_hours	Max 24-hour Sequential Organ Failure Assessment (SOFA) score

Table 1: Patient-level static variables.

Measurement Observations

The time-varying measurements in the data encompass ventilation settings, laboratory results, and vital signs. Ventilation settings and vital signs are extracted from the MIMIC chartevents table, while labs data are obtained from the MIMIC labevents table, each identified by their respective ItemIDs. To handle multiple values within a single hour for a subject, we aggregate the results by computing the median, as the median exhibits reduced sensitivity to noisy data. The labs are sourced from arterial blood gas (ABG) specimens, as arterial blood measurements are deemed to have greater clinical relevance and precision when evaluating parameters such as respiratory function, acid-base balance, and oxygenation status. Two parameters derived from ventilation settings are also presented: set_pc_draeger (set pressure for pressure-controlled ventilation from the Draeger ventilator) and set_pc (set pressure for pressure-controlled ventilation). Set_pc_draeger is calculated as the difference between the inspiratory pressure from the Draeger ventilator (pinsp_draeger) and the set peak inspiratory pressure (ppeak). Based on clinical knowledge, set_pc is populated with pcv_level (pressure controlled ventilation level) if present, pinsp_hamilton (inspiratory pressure from Hamilton ventilator) if pcv_level is absent, and set_pc_draeger (inspiratory pressure from Draeger ventilator) if both are absent. All variables related to ventilation parameters, vital signs, and labs, and their corresponding descriptions, are described in Table 2.

Treatment Interventions

Three respiratory support methods, including invasive ventilation (INV), non-invasive ventilation (NIV), and high-flow nasal cannula (HFNC), are presented as binary indicators per hour. The curation of these respiratory support variables is verified by clinical experts to ensure accuracy and reliability. In MIMIC, the procedureevents table identifies patients on INV or NIV during their ICU stay, while the chartevents table identifies patients on HFNC. INV and NIV in MIMIC have documented start and end times recorded by respiratory therapists, however, HFNC lacks a corresponding time inter-

Variable	Description
Ventilation parameters	
ppeak	peak Inspiratory pressure (cmH_2O)
set_peep	set peak inspiratory pressure (cmH_2O)
total_peep	total peak inspiratory pressure (cmH_2O)
rr	respiratory rate (insp/min)
set_rr	set respiratory rate (insp/min)
total_rr	total respiratory rate (insp/min)
set_tv	set tidal volume (mL)
total_tv	total tidal volume (mL)
set_fio2	set fraction of inspired oxygen
set_ie_ratio	set inspiratory-to-expiratory ratio
set_pc	set pressure for pressure controlled ventilation (cmH_2O)
_set_pc_draeger	set pressure from Draeger Ventilator (cmH_2O)
_pinsp_draeger	inspiratory pressure from Draeger Ventilator (cmH_2O)
_pinsp_hamilton	inspiratory pressure from Hamilton ventilator (cmH_2O)
_pcv_level	pressure controlled ventilation level (cmH_2O)
Labs	
calculated_bicarbonate	calculated bicarbonate, whole blood (mEq/L)
so2	oxygen saturation (%)
pCO2	partial pressure of carbon dioxide (mmHg)
pO2	partial pressure of oxygen (mmHg)
pH	pH
Vital Signs	
heart_rate	heart rate (bpm)
sbp	systolic arterial blood pressure (mmHg)
dbp	diastolic arterial blood pressure (mmHg)
mbp	mean arterial blood pressure (mmHg)
sbp_ni	systolic non-invasive blood pressure (mmHg)
dbp_ni	diastolic non-invasive blood pressure (mmHg)
mbp_ni	mean non-invasive blood pressure (mmHg)
temperature	temperature (°C)
spO2	oxygen saturation pulse oximetry (%)
glucose	blood glucose

Table 2: Measurement observations. "set" in ventilation settings refers to values set by healthcare professionals on the ventilator to suit the patients' respiratory needs. "_" refers to intermediate variables.

val; having only the time at which the measurement was observed. Therefore, we pre-process the data to establish a start time and end time for each HFNC event per ICU stay. In addition, multiple HFNC events could occur during a single ICU stay. Therefore, if the time gap between two consecutive HFNC events exceeded 24 hours, we treat them as separate events. For each HFNC event, the minimum and maximum time at which the HFNC is applied is used to obtain the HFNC start time and the end time. HFNC events with identical start and end times are excluded. In cases of overlapping mutually exclusive treatments, where patients are recorded to be on both non-invasive and invasive ventilation simultaneously, we prioritize the most invasive treatment strategy (INV > NIV > HFNC). The overlap of mutually exclusive treatments occurs due to the complexities involved in transitioning between ventilation therapies within the ICU, which often includes a series of procedures during the transition period. Furthermore, for short intervals (less than 6 hours) recorded between two different treatments, we attribute the gap to the less invasive treatment. This allows us to handle situations where the precise timing of treatments is unclear.

Additional binary indicators for interventions include vasopressor administration and continuous renal replacement therapy. Vasopressors are extracted from the MIMIC inputevents table and matched to the corresponding hour using their respective start and end times. A patient is classified as being on vasopressors if they received norepinephrine, epinephrine, dopamine, phenylephrine, or vasopressin. Information regarding continuous renal replacement therapy (CRRT) is extracted from the MIMIC chartevents table. Patients are identified as being on CRRT if they have a positive value for blood flow rate or fluid removal during dialysis. A summary of all treatment interventions is presented in Table 3.

Variable	Description
invasive	invasive ventilation indicator
noninvasive	non-invasive ventilation indicator
highflow	high-flow nasal cannula indicator
vasopressor	vasopressor treatment indicator
crrt	continuous renal replacement therapy indicator

Table 3: Treatment interventions.

Outcome Variables

The majority of the outcome variables are recorded as binary indicators at each hour, with one denoting the occurrence of the event. These include discharge outcome, ICU out-time outcome, death outcome, and sepsis. Discharge outcome and ICU out-time outcome indicate if a patient was discharged from the hospital or ICU respectively. The death outcome variable denotes whether a patient died at a specific hour. The date of death in MIMIC is derived from hospital and state records. In cases where both data sources are available, in-hospital mortality is preferentially used over state-linked data. The state-derived date of death includes only the date component, so a default time of midnight is used when converting the date to a timestamp. The data also includes a sepsis outcome variable that identifies whether a patient is septic according to the Sepsis-3 diagnostic criteria. Additionally, it contains the length of stay variable, which indicates the duration of a patient's ICU stay in fractional days. A summary of the outcome variables is presented in Table 4.

Variable	Description
discharge_outcome	hospital discharge indicator
icuouttime_outcome	ICU discharge indicator
death_outcome	death indicator
sepsis	presence of sepsis using sepsis 3 criteria
los	ICU length of stay (fractional days)

Table 4: Outcome variables.

2.2 Benchmark Tasks

We consider two clinically motivated prediction tasks for respiratory interventions in ICU settings: prolonged mechanical ventilation and successful weaning.

Task definition for prolonged mechanical ventilation: Prolonged mechanical ventilation can increase the caregiver burden and affect a patient's quality of life [Vali *et al.*, 2023; Sayed *et al.*, 2021]. We aim to predict prolonged mechanical ventilation using the first 24 hours of data in the ICU. Specifically, we define prolonged mechanical ventilation as the initial attempt to ventilate a patient for more than 14 days in the ICU.

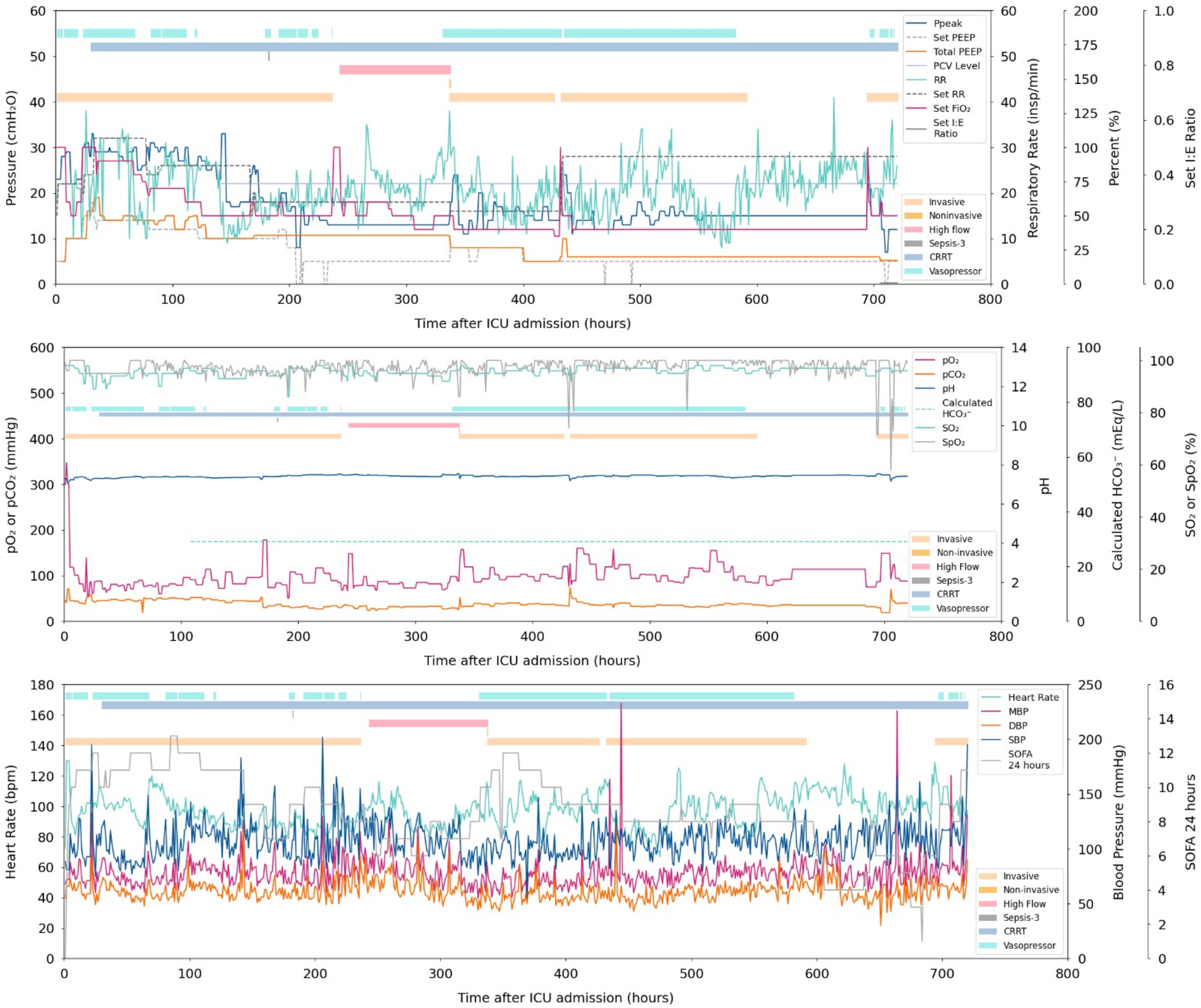

Figure 1: Visualization of time-varying covariates for a patient's stay, 30 days after ICU admission. The plots, listed from top to bottom, include ventilation parameters, laboratory results, and vital signs, respectively.

Task definition for prolonged successful weaning: Weaning has been studied in recent clinical trials [Pham *et al.*, 2023]. In this study, we use the first attempt to separate a patient from a mechanical ventilator. We aim to predict prolonged successful weaning using five days of ICU stay data. Specifically, we define successful weaning as no re-intubation or death within seven days of extubation.

The pre-processed patient cohorts for prolonged mechanical ventilation and successful weaning includes 4,930 and 2,358 cases, respectively. The numerical features for each task are normalized by min-max scaling. For each task, we split the data into 70% training, 10% validation, and 20% testing, while ensuring no patient overlap in the sets to avoid data leakage. In our hybrid sequence-based models, we combine continuous and static features to capture both the hourly dynamics of a patient's condition and the patient's individual characteristics, providing a comprehensive basis for our binary classification tasks on a stay level. For our hybrid fully connected networks which do not involve recurrent connections, we employ static features in conjunction with the median of the time-series features.

Model Architecture

In our proposed benchmark, we employ five types of machine learning models to address the aforementioned tasks. We specifically focus on deep learning-based methodologies, including sequence models and a multilayer perceptron (MLP) aiming to assess whether models that operate over time-steps can enhance overall model performance. The sequence models encompass Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Temporal Convolutional Neural Networks (TCN). For the sequence-based models, we apply a joint-fusion strategy to concatenate the hourly time-dependent variables with the static variables. Details of the sequence-based models are depicted in Figure 2. We fine-tune each model through an exhaustive hyperparameter search. The learning rate is initialized at 0.001 and decays by 5% for each epoch, with a batch size set to 512. The optimization algorithm used is the Adam Optimizer, and the loss function used is binary cross-entropy. We stop the model training when the validation loss does not improve over three consecutive epochs.

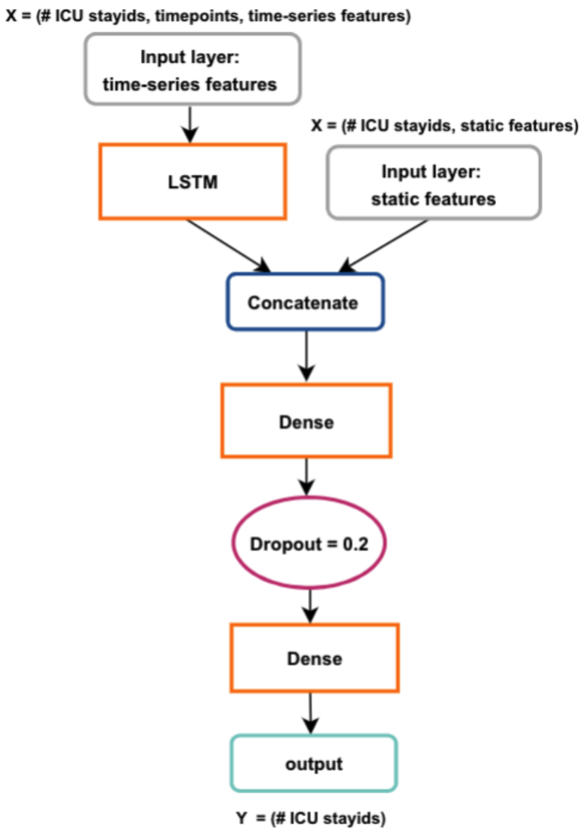

Figure 2: Model architecture of the sequence models. The architecture uses a joint-fusion strategy that concatenates hourly time-dependent features with static features and includes recurrent layers such as an LSTM, BiLSTM, or GRU, or neural networks like TCN.

Model Evaluation

We assess the models by evaluating their accuracy using the area under the receiver operating characteristic curve (AUROC). To perform binary classification on the predictions, we determine the optimal threshold value by selecting the threshold that maximizes the difference between true positive rate and false positive rate, and then we compare each prediction against this threshold. We report 95% confidence interval of the evaluation metrics, calculated by performing bootstrapping on the metric scores over 1000 iterations.

2.3 Fairness Audits Along SDOH & Demographic Attributes

We perform fairness audits by considering protected attributes, such as race, age, and gender, along with eight SDOH attributes. This provides deeper insights into the patient population beyond conventional demographic attributes. We utilize the MIMIC-IV census tract-level SDOH data to conduct fairness audits on our benchmark tasks [Yang *et al.*, 2023]. Our analysis includes investigating the differences in fairness across subgroups based on SDOH attributes, such as whether a patient resides in areas with high employment rates, has a high reliance on public assistance or food stamps, lives close to healthcare facilities, engages frequently in heavy drinking or smoking, has high student expenditure, resides in homes with high electricity heating and lives in areas with few deaths from firearms.

We assess the performance of downstream classifiers based on three definitions of fairness, including, demographic parity (parity gap), equality of opportunity for the positive class (recall gap), and equality of opportunity for the negative class (specificity gap) [Chen *et al.*, 2019]. We follow methods used in prior work to expand the demographic parity gap [Zhang *et al.*, 2020; Hashimoto *et al.*, 2018], and use a similar process to obtain the recall, and specificity gaps. These evaluations are conducted on the best-performing model for both tasks.

3 Results & Discussion

3.1 Benchmark Tasks

The AUROC for both prolonged mechanical ventilation and weaning are shown in Table 5. We found that the sequence-based model, the GRU (Hybrid) model, outperforms all other models on both binary prediction tasks.

Model	AUROC (↑)	
	Mechanical Ventilation	Successful Weaning
MLP	0.641 (0.638 - 0.643)	0.749 (0.747 - 0.751)
TCN (Hybrid)	0.747 (0.745 - 0.749)	0.743 (0.741 - 0.744)
LSTM (Hybrid)	0.770 (0.768 - 0.772)	0.764 (0.763 - 0.766)
BiLSTM (Hybrid)	0.775 (0.773 - 0.777)	0.752 (0.750 - 0.753)
GRU (Hybrid)	0.778 (0.776 - 0.780)	0.776 (0.774 - 0.778)

Table 5: Benchmark results for two clinically-motivated tasks: classifying mechanical ventilation lasting more than 14 days, using 24 hours of data, and successful weaning lasting more than 7 days, using 5 days of data. Scores are reported with 95% confidence intervals obtained through 1000 bootstrap samples.

3.2 Fairness Audits on Benchmark Tasks

We illustrate the differences in parity, recall, and specificity for demographic and social determinants of health attributes in the mechanical ventilation (Figure 3) and successful weaning tasks (Figure 4) using the best performing model (GRU). Recall indicates the proportion of actual positive instances that the model correctly identifies. It is particularly relevant in clinical settings where minimizing false negatives is crucial for timely effective patient diagnosis. To analyze variations in model performance among continuous SDOH attributes,

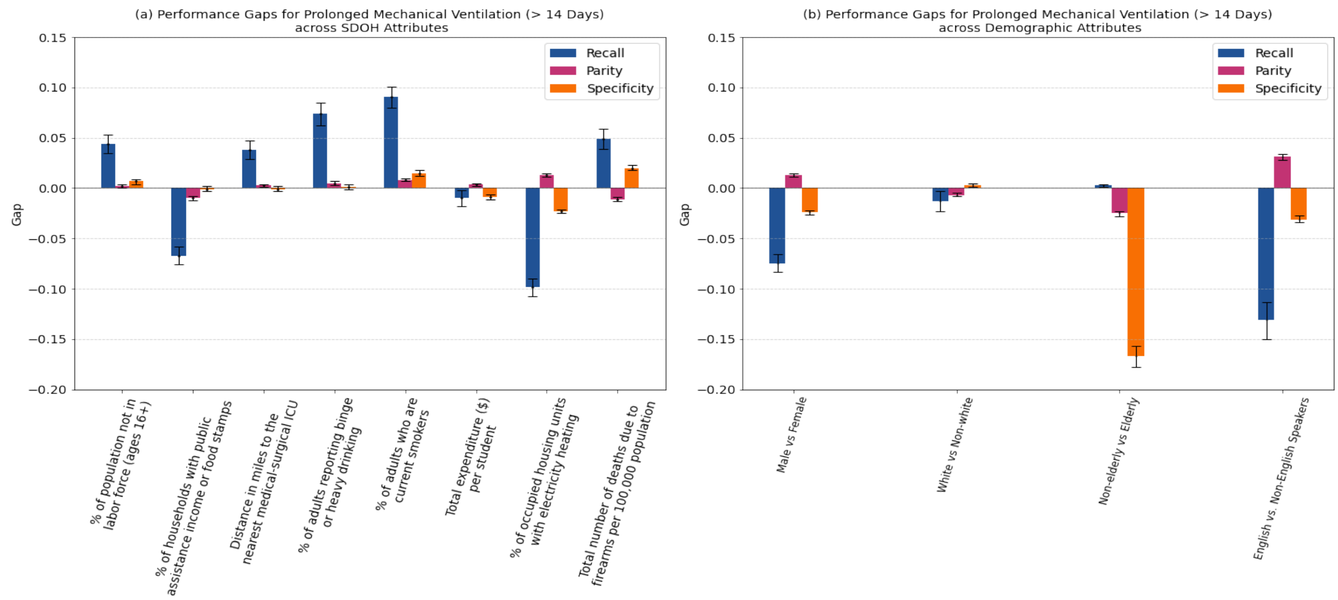

Figure 3: Performance gap measures for the prolonged mechanical ventilation task under the best model (GRU). A positive bar indicates the model favors one group over the other group. Error bars denote a 95% confidence interval obtained through 1000 bootstrap samples. a) Performance gap evaluation for SDOH attributes. b) Performance gap evaluation for demographic attributes.

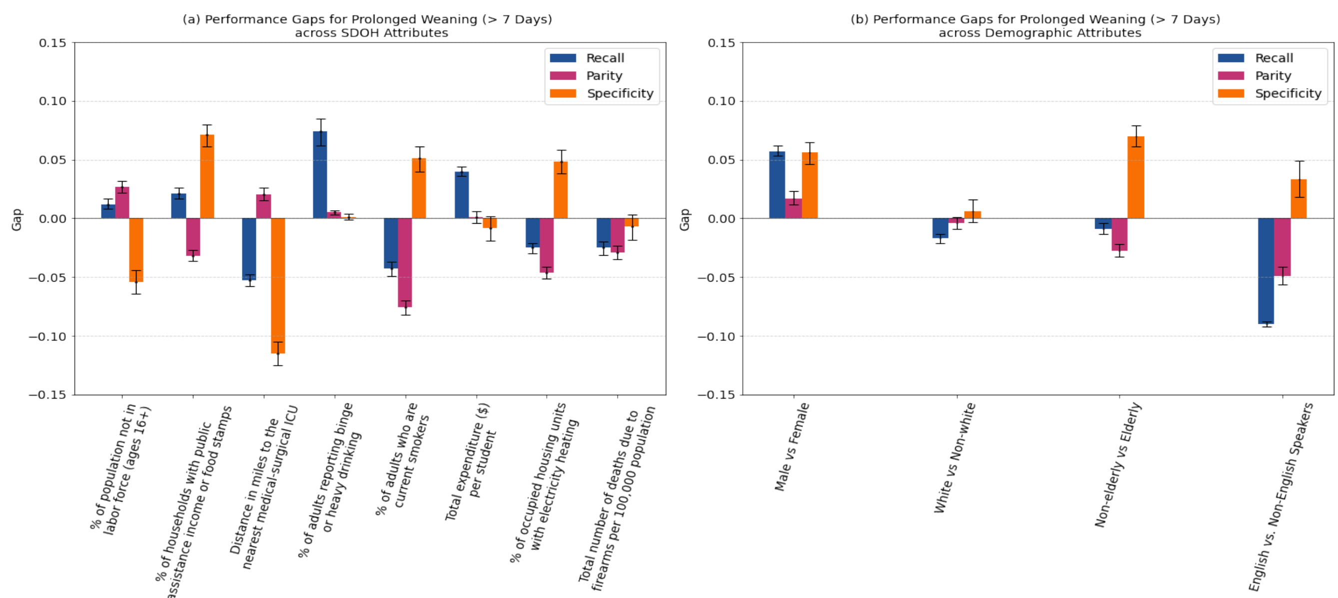

Figure 4: Performance gap measures for the successful prolonged weaning task under the best model (GRU). A positive bar indicates the model favors one group over the other group. Error bars denote a 95% confidence interval obtained through 1000 bootstrap samples. a) Performance gap evaluation for SDOH attributes. b) Performance gap evaluation for demographic attributes.

we discretize the attributes into two quantiles. A positive recall gap suggests that the model favors the low prevalence of the specified SDOH attribute over the high prevalence. For categorical variables like gender, race, age, and English proficiency, a positive recall gap indicates that the model favors males, whites, non-elderly individuals, or English speakers over their respective counterparts.

In Figure 3a, for the task of predicting prolonged mechanical ventilation the model favors individuals who reside in areas with high employment rate, have a high reliance on public assistance or food stamps, are close to a medical-surgical ICU, rarely engage in heavy drinking or smoking, have high student expenditure, reside in homes with high electricity heating, and live in areas with few deaths from firearms.

Additionally, as seen in Figure 3b the model favors certain demographic groups, including females, non-white individuals, younger individuals, and non-English speakers. On the other hand, as depicted in Figure 4a, for the task of predicting weaning, the model favors individuals who reside in areas with high employment rate, have a low reliance on public assistance or food stamps, are far from a medical-surgical ICU, rarely engage in heavy drinking, often smoke, have low student expenditure, reside in homes with high electricity heating, and live in areas with more deaths from firearms. Additionally, as seen in Figure 4b, the model favors certain demographic groups, including males, non-white individuals, elderly individuals, and non-English speakers.

The performance gaps illustrate the disparities in the model's predictive performance and the necessity for fairness auditing prior to model deployment. By assessing SDOH in addition to the previously studied traditional labels we hope to disentangle biases and uncover other hidden confounders and associations.

4 Conclusion

In critical care settings, it is important to carefully assess model biases across demographic and SDOH attributes before deployment. In this study, we benchmark two time-dependent tasks, including successful weaning and prolonged mechanical ventilation. Using different fairness definitions, we evaluate the differences in performance gaps for both tasks across demographic and SDOH attributes. Furthermore, we release an hourly dataset to support the benchmarking of respiratory intervention tasks. Our work aims to enable the development of machine learning models for timely interventions in critical care, emphasizing the consideration of social determinants to promote equitable healthcare access and improve patient outcomes.

Data Availability

The temporal dataset for respiratory support in critically ill patients is hosted on PhysioNet [Moody and Mark, 1996], an NIH-funded repository that is widely used to support biomedical research and education worldwide. It is available at this link, https://doi.org/10.13026/0d8j-2w14. The presented dataset consists of 50,920 distinct adult patients admitted to the ICU of Beth Israel Deaconess Medical Center (Boston, MA, USA) between 2008 and 2019. We extract static, time-varying, and outcome variables from MIMIC-IV in an hourly materialized view and store the content for each patient in a *.csv format named after the patient's unique identifier (subject ID).

Code Availability

We provide the GitHub repository at https://github.com/respiratory-support/respiratory-interventions which includes SQL scripts, offering tools for data management, querying, and analysis. Python scripts are also provided to demonstrate the application of the dataset in various clinical prediction tasks.

Acknowledgements

D.M., and L.M. are supported by grant NIH-R01-EB017205 from the National Institute of Health and by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH-R01-EB030362. D.M. is also supported by NIH National Library of Medicine under 75N97020C00013, and Massachusetts Life Sciences Center. H.C. is supported by the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (grant number: HI21C107409). The authors would like to thank Dr. Leo Celi, Dr. Kerollos Wanis, and Sicheng Hao for their valuable feedback.

References

[Bundy *et al.*, 2023] Joshua D Bundy, Katherine T Mills, Hua He, Thomas A LaVeist, Keith C Ferdinand, Jing Chen, and Jiang He. Social determinants of health and premature death among adults in the usa from 1999 to 2018: a national cohort study. *The Lancet Public Health*, 8(6):e422–e431, 2023.

[Celi *et al.*, 2022] Leo Anthony Celi, Jacqueline Cellini, Marie-Laure Charpignon, Edward Christopher Dee, Franck Dernoncourt, Rene Eber, William Greig Mitchell, Lama Moukheiber, Julian Schirmer, Julia Situ, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS Digital Health*, 1(3):e0000022, 2022.

[Chen *et al.*, 2019] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

[Doyle *et al.*, 1995] Ramona L Doyle, Nancy Szaflarski, Gunnard W Modin, Janine P Wiener-Kronish, and Michael A Matthay. Identification of patients with acute lung injury. predictors of mortality. *American journal of respiratory and critical care medicine*, 152(6):1818–1824, 1995.

[Esquinas *et al.*, 2017] Antonio M Esquinas, Maly Oron Benhamou, Alastair J Glossop, and Bushra Mina. Noninvasive mechanical ventilation in acute ventilatory failure: rationale and current applications. *Sleep Medicine Clinics*, 12(4):597–606, 2017.

[Esteban *et al.*, 2000] Andres Esteban, Antonio Anzueto, Inmaculada Alia, Federico Gordo, Carlos Apezteguia, Fernando Palizas, David Cide, Rosanne Goldwaser, Luis Soto, Guillermo Bugedo, et al. How is mechanical ventilation employed in the intensive care unit? an international utilization review. *American journal of respiratory and critical care medicine*, 161(5):1450–1458, 2000.

[Frat *et al.*, 2017] Jean-Pierre Frat, Rémi Coudroy, Nicolas Marjanovic, and Arnaud W Thille. High-flow nasal oxygen therapy and noninvasive ventilation in the management of acute hypoxemic respiratory failure. *Annals of translational medicine*, 5(14), 2017.

[Gupta *et al.*, 2022] Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. An extensive data processing pipeline for mimic-iv. In *Machine Learning for Health*, pages 311–325. PMLR, 2022.

[Harutyunyan *et al.*, 2019] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

[Hashimoto *et al.*, 2018] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[Holmes Fee *et al.*, 2023] Casey Holmes Fee, Rachel Scarlett Hicklen, Sidney Jean, Nebal Abu Hussein, Lama Moukheiber, Michelle Foronda de Lota, Mira Moukheiber, Dana Moukheiber, Leo Anthony Celi, and Irene Dankwa-Mullan. Strategies and solutions to address digital determinants of health (ddoh) across underinvested communities. *PLOS digital health*, 2(10):e0000314, 2023.

[Kondrup *et al.*, 2023] Flemming Kondrup, Thomas Jiralerspong, Elaine Lau, Nathan de Lara, Jacob Shkrob, My Duc Tran, Doina Precup, and Sumana Basu. Towards safe mechanical ventilation treatment using deep offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15696–15702, 2023.

[Lua *et al.*, 2023] Iracema Lua, Andrea F Silva, Nathalia S Guimarães, Laio Magno, Julia Pescarini, Rodrigo VR Anderle, Maria Yury Ichihara, Mauricio L Barreto, Carlos AS Teles Santos, Louisa Chenciner, et al. The effects of social determinants of health on acquired immune deficiency syndrome in a low-income population of brazil: a retrospective cohort study of 28.3 million individuals. *The Lancet Regional Health–Americas*, 24, 2023.

[Marmot, 2005] Michael Marmot. Social determinants of health inequalities. *The lancet*, 365(9464):1099–1104, 2005.

[Moody and Mark, 1996] George B Moody and Roger G Mark. A database to support development and evaluation of intelligent intensive care monitoring. In *Computers in Cardiology 1996*, pages 657–660. IEEE, 1996.

[Moukheiber *et al.*, 2024] Dana Moukheiber, Saurabh Mahindre, Lama Moukheiber, Mira Moukheiber, and Mingchen Gao. Looking beyond what you see: An empirical analysis on subgroup intersectional fairness for multi-label chest x-ray classification using social determinants of racial health inequities. *arXiv preprint arXiv:2403.18196*, 2024.

[Moy *et al.*, 2017] Ernest Moy, Macarena C Garcia, Brigham Bastian, Lauren M Rossen, Deborah D Ingram, Mark Faul, Greta M Massetti, Cheryll C Thomas, Yuling Hong, Paula W Yoon, et al. Leading causes of death in nonmetropolitan and metropolitan areas—united states, 1999–2014. *MMWR Surveillance Summaries*, 66(1):1, 2017.

[Nakagawa *et al.*, 2023] Keisuke Nakagawa, Lama Moukheiber, Leo A Celi, Malhar Patel, Faisal Mahmood, Dibson Gondim, Michael Hogarth, and Richard Levenson. Ai in pathology: what could possibly go wrong? In *Seminars in Diagnostic Pathology*, volume 40, pages 100–108. Elsevier, 2023.

[Nazer *et al.*, 2023] Lama H Nazer, Razan Zatarah, Shai Waldrip, Janny Xue Chen Ke, Mira Moukheiber, Ashish K Khanna, Rachel S Hicklen, Lama Moukheiber, Dana Moukheiber, Haobo Ma, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*, 2(6):e0000278, 2023.

[Papoutsi *et al.*, 2021] Eleni Papoutsi, Vassilis G Giannakoulis, Eleni Xourgia, Christina Routsi, Anastasia Kotanidou, and Ilias I Siempos. Effect of timing of intubation on clinical outcomes of critically ill patients with covid-19: a systematic review and meta-analysis of non-randomized cohort studies. *Critical Care*, 25:1–9, 2021.

[Pham *et al.*, 2023] Tài Pham, Leo Heunks, Giacomo Bellani, Fabiana Madotto, Irene Aragao, Gaëtan Beduneau, Ewan C Goligher, Giacomo Grasselli, Jon Henrik Laake, Jordi Mancebo, et al. Weaning from mechanical ventilation in intensive care units across 50 countries (wean safe): a multicentre, prospective, observational cohort study. *The Lancet Respiratory Medicine*, 11(5):465–476, 2023.

[Purnell *et al.*, 2016] Tanjala S Purnell, Elizabeth A Calhoun, Sherita H Golden, Jacqueline R Halladay, Jessica L Krok-Schoen, Bradley M Appelhans, and Lisa A Cooper. Achieving health equity: closing the gaps in health care disparities, interventions, and research. *Health Affairs*, 35(8):1410–1415, 2016.

[Purushotham *et al.*, 2018] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.

[Rocheteau *et al.*, 2021] Emma Rocheteau, Pietro Liò, and Stephanie Hyland. Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In *Proceedings of the conference on health, inference, and learning*, pages 58–68, 2021.

[Rubenfeld *et al.*, 2005] Gordon D Rubenfeld, Ellen Caldwell, Eve Peabody, Jim Weaver, Diane P Martin, Margaret Neff, Eric J Stern, and Leonard D Hudson. Incidence and outcomes of acute lung injury. *New England Journal of Medicine*, 353(16):1685–1693, 2005.

[Sayed *et al.*, 2021] Mohammed Sayed, David Riano, and Jesús Villar. Predicting duration of mechanical ventilation in acute respiratory distress syndrome using supervised machine learning. *Journal of Clinical Medicine*, 10(17):3824, 2021.

[Schraufnagel *et al.*, 2013] Dean E Schraufnagel, Francesco Blasi, Monica Kraft, Mina Gaga, Patricia W Finn, and

Klaus F Rabe. An official american thoracic society/european respiratory society policy statement: disparities in respiratory health. *American journal of respiratory and critical care medicine*, 188(7):865–871, 2013.

[Suo *et al.*, 2021] Daniel Suo, Naman Agarwal, Wenhan Xia, Xinyi Chen, Udaya Ghai, Alexander Yu, Paula Gradu, Karan Singh, Cyril Zhang, Edgar Minasyan, et al. Machine learning for mechanical ventilation control. *arXiv preprint arXiv:2102.06779*, 2021.

[Sweatt and Levitt, 2014] Andrew J Sweatt and Joseph E Levitt. Evolving epidemiology and definitions of the acute respiratory distress syndrome and early acute lung injury. *Clinics in chest medicine*, 35(4):609–624, 2014.

[Thakur *et al.*, 2014] Neeta Thakur, Meghan E McGarry, Sam S. Oh, Joshua M. Galanter, Patricia W Finn, Esteban G Burchard, and ATS Health Equality Committee. The lung corps' approach to reducing health disparities in respiratory disease. *Annals of the American Thoracic Society*, 11(4):655–660, 2014.

[Vali *et al.*, 2023] Mohebat Vali, Shahram Paydar, Mozhgan Seif, Golnar Sabetian, Ahmad Abujaber, and Haleh Ghaem. Prediction prolonged mechanical ventilation in trauma patients of the intensive care unit according to initial medical factors: a machine learning approach. *Scientific Reports*, 13(1):5925, 2023.

[Wang *et al.*, 2020] Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235, 2020.

[Wanis *et al.*, 2023] Kerollos Nashat Wanis, Arin L Madenci, Sicheng Hao, Mira Moukheiber, Lama Moukheiber, Dana Moukheiber, Sulaiman Moukheiber, Jessica G Young, and Leo Anthony Celi. Emulating target trials comparing early and delayed intubation strategies. *Chest*, 164(4):885–891, 2023.

[Ware and Matthay, 2000] Lorraine B Ware and Michael A Matthay. The acute respiratory distress syndrome. *New England Journal of Medicine*, 342(18):1334–1349, 2000.

[Yang *et al.*, 2023] Ming Ying Yang, Gloria Hyunjung Kwak, Tom Pollard, Leo Anthony Celi, and Marzyeh Ghassemi. Evaluating the impact of social determinants on health prediction. *arXiv preprint arXiv:2305.12622*, 2023.

[Zhang *et al.*, 2020] Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120, 2020.

[Zuo *et al.*, 2020] Mingzhang Zuo, Yuguang Huang, Wuhua Ma, Zhanggang Xue, Jiaqiang Zhang, Yahong Gong, and Lu Che. Expert recommendations for tracheal intubation in critically ill patients with noval coronavirus disease 2019. *Chinese Medical Sciences Journal*, 35(2):105–109, 2020.