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

Mira Moukheiber¹, Lama Moukheiber¹, Dana Moukheiber¹ and Hyung-Chul Lee²*

¹Massachusetts Institute of Technology
²Seoul 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 Matthy, 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 [Schaufnagel 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; Luo 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
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. Figure 1 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.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>intime</td>
<td>ICU admission time</td>
</tr>
<tr>
<td>outtime</td>
<td>ICU discharge time</td>
</tr>
<tr>
<td>gender</td>
<td>patient gender</td>
</tr>
<tr>
<td>anchor year</td>
<td>patient shifted year</td>
</tr>
<tr>
<td>anchor age</td>
<td>patient age in anchor year</td>
</tr>
<tr>
<td>insurance</td>
<td>patient insurance type</td>
</tr>
<tr>
<td>language</td>
<td>English proficiency indicator</td>
</tr>
<tr>
<td>marital status</td>
<td>patient marital status</td>
</tr>
<tr>
<td>race</td>
<td>patient race</td>
</tr>
<tr>
<td>first_careunit</td>
<td>ICU type during first admission</td>
</tr>
<tr>
<td>pbw_kg</td>
<td>patient predicted body weight (kg)</td>
</tr>
<tr>
<td>height_inch</td>
<td>patient height (inches)</td>
</tr>
<tr>
<td>elixaheuser_vanwalraven</td>
<td>Elixhauser-Van Walraven score</td>
</tr>
<tr>
<td>gcs</td>
<td>Glasgow Coma Scale (GCS) score</td>
</tr>
<tr>
<td>gcs_motor</td>
<td>GCS motor response component</td>
</tr>
<tr>
<td>gcs_verbal</td>
<td>GCS verbal response component</td>
</tr>
<tr>
<td>gcs_eyes</td>
<td>GCS eye-opening response component</td>
</tr>
<tr>
<td>gcs_unable</td>
<td>Endotracheal tube indicator</td>
</tr>
<tr>
<td>sofa_24hours</td>
<td>Max 24-hour Sequential Organ Failure Assessment (SOFA) score</td>
</tr>
</tbody>
</table>

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_pcv (set pressure for pressure-controlled ventilation from the Draeger ventilator) and set_pc (set pressure for pressure-controlled ventilation). Set_pcv is calculated as the difference between the inspiratory pressure from the Draeger ventilator (pinsp_draeger) and the peak inspiratory pressure (ppeak). Based on clinical knowledge, set_pcv 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-
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.

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.

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.
**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.

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUROC (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mechanical Ventilation</td>
</tr>
<tr>
<td>MLP</td>
<td>0.641 (0.638 - 0.643)</td>
</tr>
<tr>
<td>TCN (Hybrid)</td>
<td>0.747 (0.745 - 0.749)</td>
</tr>
<tr>
<td>LSTM (Hybrid)</td>
<td>0.770 (0.768 - 0.772)</td>
</tr>
<tr>
<td>BiLSTM (Hybrid)</td>
<td>0.775 (0.773 - 0.777)</td>
</tr>
<tr>
<td>GRU (Hybrid)</td>
<td>0.778 (0.776 - 0.780)</td>
</tr>
</tbody>
</table>

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,
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.

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.
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. Kerrollos Wanis, and Sicheng Hao for their valuable feedback.

References


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


