Data-Driven Prediction of Ventilation Duration in ICU Patients

Harold Haugen
Alec Pixton
Clarissa Benitez
Emmanuel Leonce

Context/Objective

Predicting Ventilation Duration to Improve ICU Resource Planning

Clinical Relevance: Early prediction of ventilation duration is critical for proactive ICU resource planning and informed patient management decisions.

Problem Framing: We divided our target variable into a binary classification of short-term ventilation (Mod_Under_Day) and long-term ventilation (Sig_Over_Day).

Objective: Develop and evaluate models that predict ventilation duration using pre-ventilation clinical features, enabling early identification of patients likely to require prolonged respiratory support.

Dataset Considerations: The data was stratified into non-invasive and invasive ventilation cases. However, the non-invasive subset was found to be non-representative. Thus, model comparisons focused on invasive cases vs. the full dataset, with the full dataset yielding more robust results.

Results

Key Results: The Random Forest model achieved the highest recall of 0.819 and XGBoost came in close second with a recall of 0.811 and demonstrating they effectively identified the severe cases (Sig_Over_Day).

Outcomes: Successfully developed a predictive model for ventilation duration and this included identifying key clinical features that significantly contribute to predicting prolonged ventilation needs.

Key Metrics: Recall was prioritized to ensure the model identifies as many severe cases (Sig_Over_Day) as possible.

In a clinical context, it is safer to capture all potential severe cases, even if it results in more false positives, rather than risk missing critical patients. High recall ensures early identification of at-risk patients, aiding ICU resource allocation and intervention planning.

Results: Main Takeaway

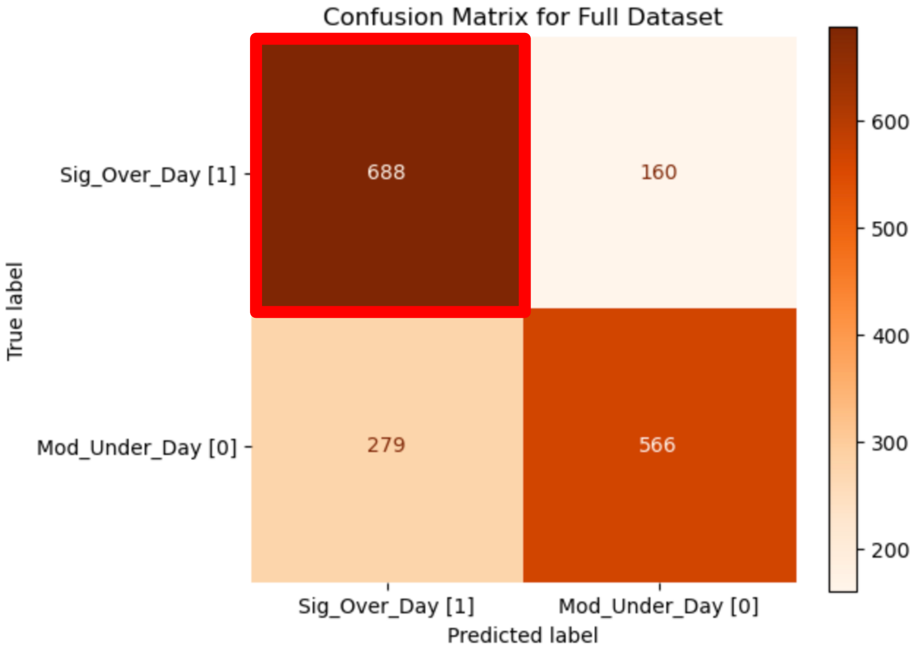
Including non-invasive ventilation improved recall for all models

Model	ROC-AUC	Recall	Specificity	Precision	Accuracy
XGBoost/Full	0.823	0.811	0.692	0.726	0.752
XGBoost/Invasive	0.821	0.803	0.708	0.738	0.756
Random Forest/Full	0.805	0.819	0.652	0.697	0.735
Random Forest/Invasive	0.811	0.785	0.682	0.710	0.733
Logistic Regression/Full	0.792	0.727	0.729	0.726	0.724
Logistic Regression/Invasive	0.792	0.725	0.723	0.726	0.725
Neural Network/Full	0.796	0.754	0.703	0.713	0.728
Neural Network/Invasive	0.547	0.351	0.700	0.464	0.552

Results: Confusion Matrix

Incorrectly predicting that patients will stay less than a day on a ventilator will cause more stress on the system

- Individual
- Family
- Nurses
- Doctors
- Hospital Administration

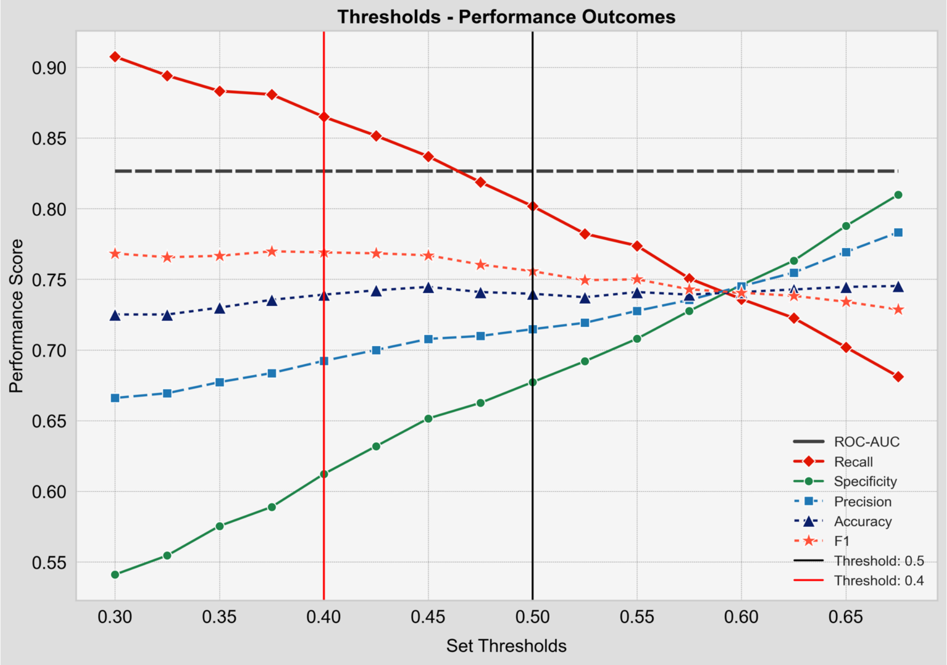


How does one day impact each of these?

Results: Threshold

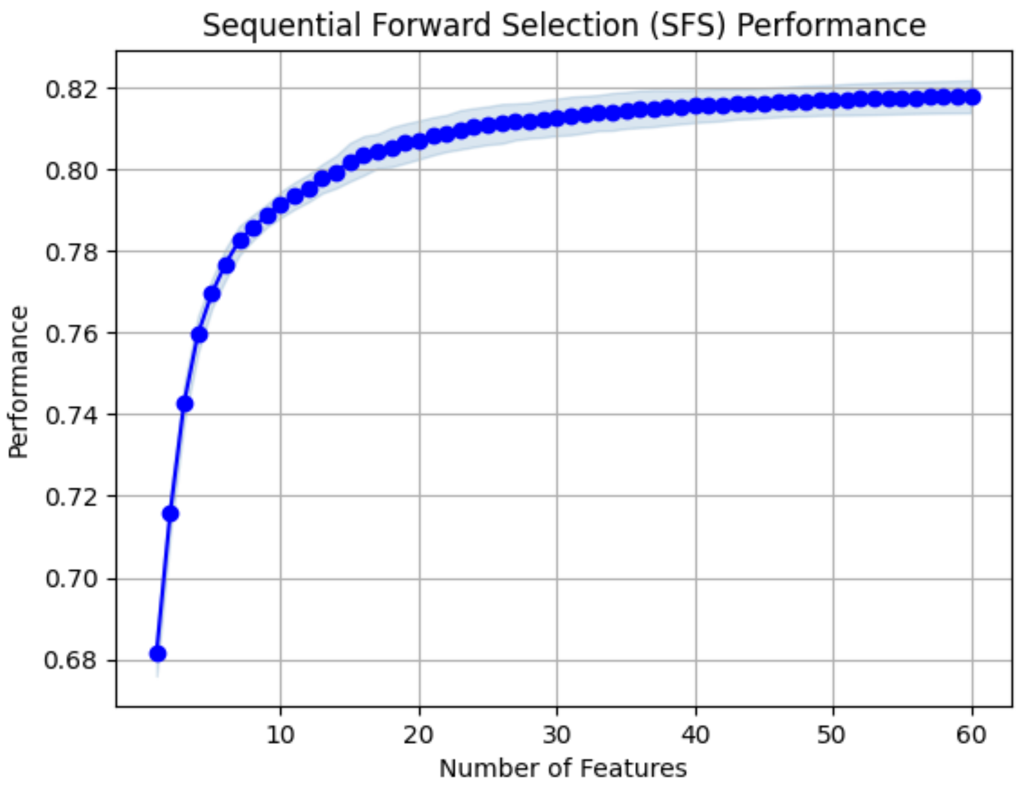
Adjusting the threshold can enable higher recall for better applicability

Would you prefer to be unexpectedly released from the hospital sooner than expected or have what was expected to be a short visit turn into a multi-day ordeal?



Results: Feature Selection

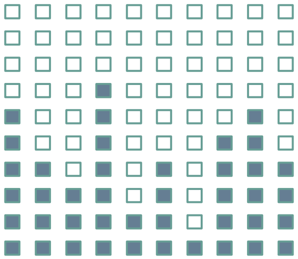
Increasing the number of included features provides diminishing returns in performance



- Full set vs Reduced set
- One-hot encoding
- Grouped vs Ungrouped
- 60th iteration took 8.6 min

If this model was deployed real-time, how many variables would we expect to collect from a patient?

Analysis



EDA

To uncover patterns, evaluate data quality, and identify potentially informative medical features contributing to predictive inference:

- Evaluating distributions of clinical data to detect skew, outliers, or group differences
- Identifying correlations or interactions between features and outcomes
- Forming hypotheses about which variables may hold predictive value

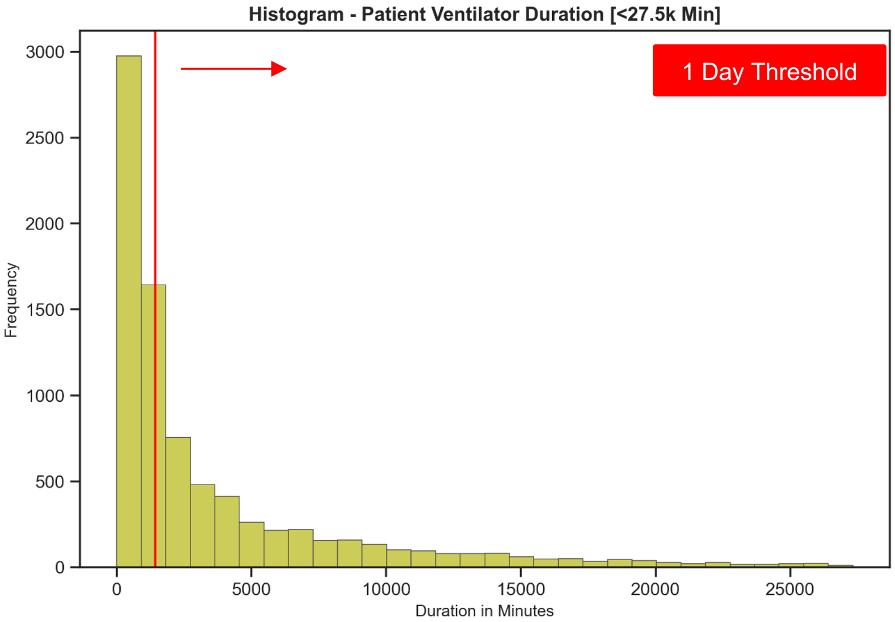
Feature Importance

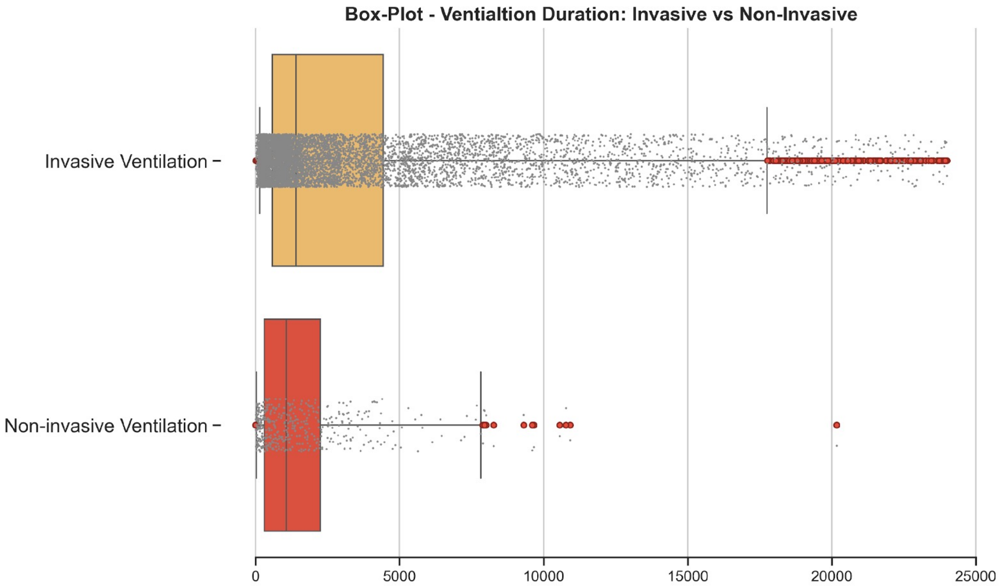
Analysis of Feature Importance will help identify which clinical variables most influence model prediction.

- Prioritizes the most relevant features for training
- Reduces model complexity
- Improves interpretability and computational efficiency
- Enhances generalization and clinical relevance

EDA

Confirmed: Skewed but Learnable Distribution for Predictive Modeling

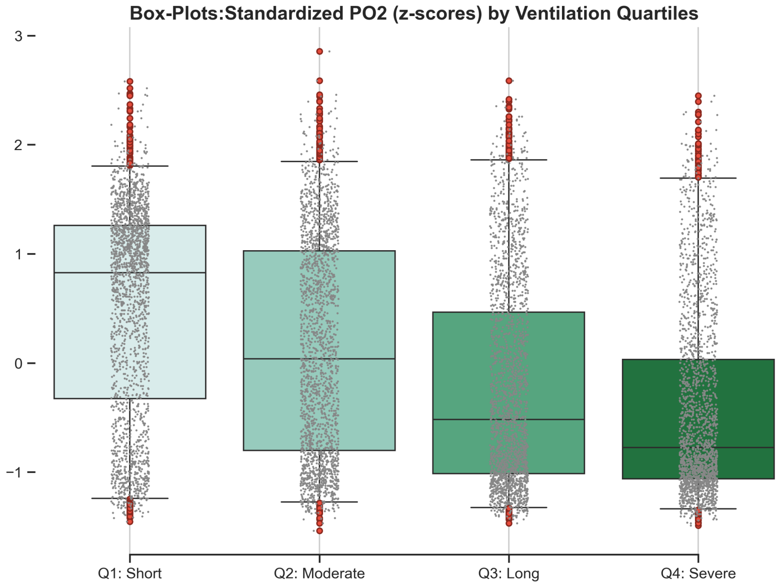


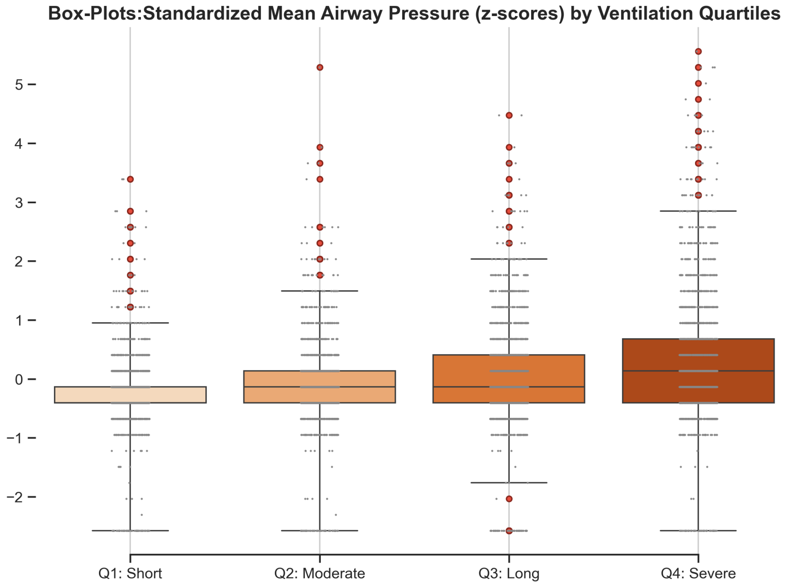


- **Does the distribution of ventilator duration offer sufficient variance & clinically meaningful signal / spread?**
- **Will the Ventilation Type Label instill leakage into the problem?**

EDA

Multiple Features Show Predictive Patterns, Including PO2 & Mean Airway Pressure





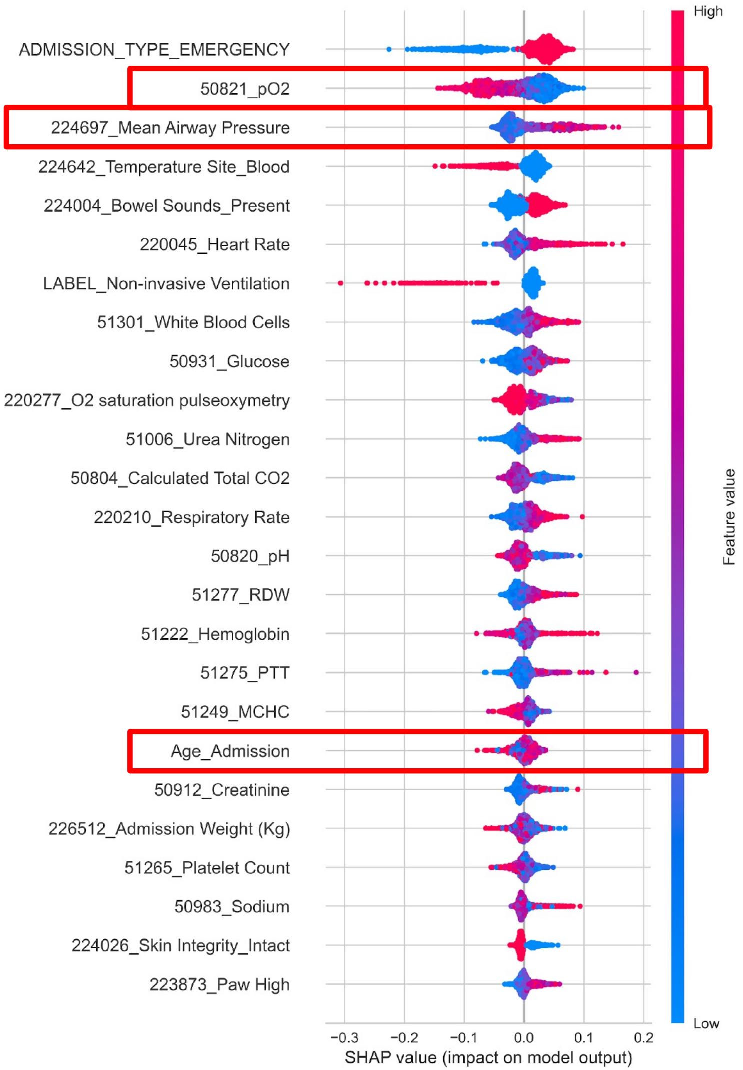
- **Are there Features that meaningfully correlate with the target and Show Predictive Power?**

Feature Importance

Core Features Identified - SHAP Importance Scoring

We confirmed that a series of Admission, Laboratory & Patient features meaningfully drive predication.

- **SHAP Values** – Indicate Magnitude / Push Towards Positive Class “> 1 Day Ventilated"
- **Model Interpretability / Trust and Transparency**
- **Model efficiency** – faster model training and lower computational cost w/ fewer features
- **Enhances feasibility for deployment** – by prioritizing a sub-set of inputs, real-time data collection is more practical
- **Improved Generalization** – avoid overfitting by eliminating noisy or irrelevant features
- **Enhanced Clinical Insight** for future research



Data Structure

MIMIC-III Data Schema for Scalable Research

Data Source

Mimic-III Clinical Database

- Source – intensive care unit (ICU) of Beth Israel Deaconess Medical Center (2001 to 2012)
- Data is de-identified in accordance with HIPPA using data cleaning and date shifting.
- DOBs where patient > 89 yrs – Date Shifted

Target Variable

Ventilation Duration ['VALUE']

- ITEMID 225792 – Invasive Ventilation
- ITEMID 225794 – Non-Invasive Ventilation

Features

73 Feature Variables from Tables

- Procedure Events
- Admissions
- Patients
- Laboratory Tests
- Charted Events / Observations

Data Structure

MIMIC-III Data Schema for Scalable Research

D_ITEMS	
ROW_ID	12,487
ITEMID	12,487
LINKSTO	Tables
LABEL	11,847 Uniq.
CATEGORY	95 Uniq.

- Research Identified 2 Labels Needed for Extraction
- Item IDs 225792 / 94

PROCEDUREEVENTS_MV	
ROW_ID	258,066
SUBJECT_ID	
HADM_ID	
ICUSTAY_ID	
STARTTIME	
ENDTIME	
ITEMID	12,160
VALUE	
** Other	

* ~12k Ventilation Events identified from Procedure Events

Unique Keys	
SUBJECT_ID	9,329 Unique Events
HADM_ID	
ICUSTAY_ID	

* Apply Data Assumptions / Cleaning

Ventilation Data Model	
ROW_ID	8,392
FEATURES	73 | 244 w/ Encoding

* Select Features merged for Final Data Model

PATIENT	
ROW_ID	46,520

ADMISSIONS	
ROW_ID	58,976

LABEVENTS	
ROW_ID	27.85M

CHARTEVENTS	
ROW_ID	330.7M

ICUSTAY	
ROW_ID	8,516

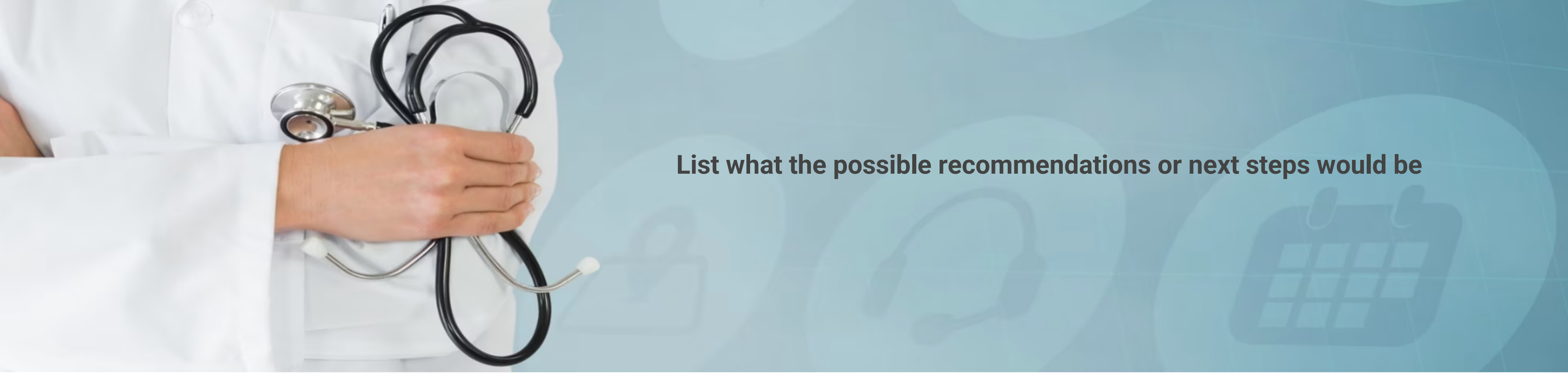
Data Structure

Data Assumptions

- **Unique Ventilation Events**: Defined using SUBJECT_ID and HADM_ID.
  - For multiple events per HADM, we kept the first hospital day's longest event to ensure clinical relevance and data availability.
- **First Occurrence Events**: Used the earliest lab and chart events per admission to maintain temporal consistency and reduce missing data.
- **Feature Inclusion Threshold**: Independent medical features with more than **~3%** missing ventilation Subjects were excluded to reduce noise and improve reliability, ensuring a balance between feature diversity and data completeness.
- **Representativeness**: Assumed that the filtered cohort remains representative of the overall ventilation population, with minimal introduction of bias.
- **Data Transformation**: Standard Scaling and One-Hot Encoding

Limitations

- **Data Limitations:** Dataset limited to MIMIC-III, a single-center ICU database (Beth Israel Deaconess Medical Center).
- **Data Cleaning:** Certain tables held moderate error levels, requiring manual cleaning and approximation.
- **Feature Exclusion / Missing Subjects:** Other potentially interesting features were not prevalent across all Subjects; requiring time for data Imputation.
- **Generalizability:** Since the data is from a single medical center and is de-identified, the findings may not apply to other healthcare systems.



Next Steps

Incorporate Time-Series Data: Evolve from single-day snapshots to models that track patient condition over time, enabling dynamic, longitudinal predictions.

Enable Cross-System Compatibility: Adapt the model to work with diverse hospital EHR systems and data schemas for broader adoption.

Embed Model into Clinical Workflow: Integrate predictions into triage and resource planning processes to support real-time clinical decision-making.

Summary Recap

- Early prediction of ventilation duration is critical for optimizing ICU resource allocation and improving patient care strategies.

- We developed a Random Forest model that classifies patients at risk of prolonged ventilation using clinical data available prior to intubation. The model achieved a recall of 0.819, effectively identifying severe cases.

- A key strength of the model is its reliance on routinely collected clinical measurements, such as pO2, Mean Airway Pressure, and Heart Rate, which are already recorded in real-time upon patient admission. This minimizes the need for additional testing, allowing the model to be seamlessly applied within existing clinical workflows.

Questions?