Supplemental Online Content for:

Early Identification of Hospitalized Patients with COVID-19 at Risk of Clinical Deterioration: Model Development and Multisite External Validation Study

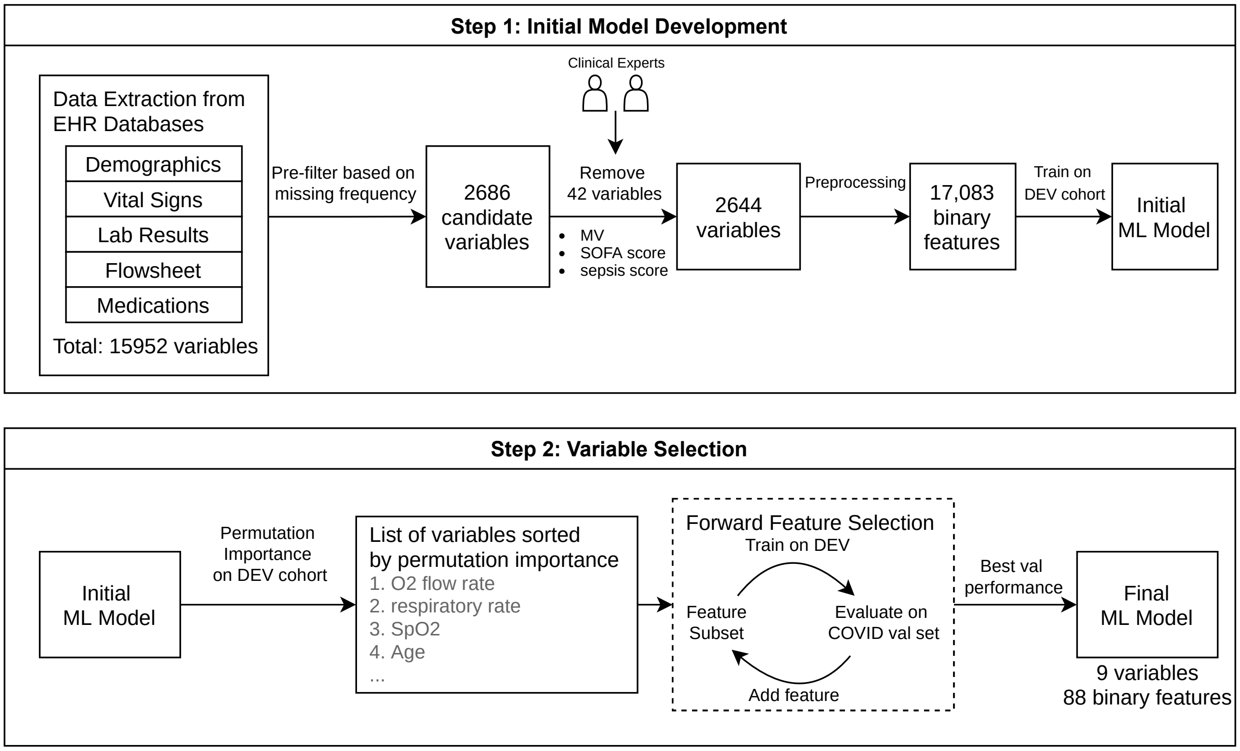
Fahad Kamran, Shengpu Tang, Erkin Otles, Dustin S. McEvoy, Sameh N. Saleh, Jen Gong, Benjamin Y. Li, Sayon Dutta, Xinran Liu, Richard J. Medford, Thomas S. Valley, Lauren R. West, Karandeep Singh, Seth Blumberg, John P. Donnelly, Erica S. Shenoy, John Z. Ayanian, Brahmajee K. Nallamothu, Michael W. Sjoding, Jenna Wiens

Table of Contents

eMethods 1. Additional Information on Model Development and Validation.

Variable Selection

Clinical knowledge and data-driven feature selection was used to reduce a set of 2,686 candidate variables from electronic health records (EHR) (including personal characteristics, laboratory results, and data recorded in nursing flowsheets; variables with a high level of missingness were excluded) to 9 variables. This process is summarized in the flowchart below.



DEV, development cohort; COVID val set, COVID validation set.

Step 1 - Initial Model Development: Data for the development cohort (2015-2019) were first extracted from the EHR and pertained to a total of 15,952 variables from 5 database tables. Examples of variables include: age, sex, heart rate, oxygen flow rate, white blood cell count (WBC) value, WBC flag, dose of ibuprofen, administration route of ibuprofen, etc. We applied a combination of pre-filtering heuristics to eliminate variables that are likely uninformative, primarily focusing on the level of missingness. Specifically, for laboratory results, medications, and nursing flowsheets, we inspected the missingness frequency for the set of variables and retained approximately 1,000 variables that are below a missingness threshold from each table (selected based on the missingness frequency distributions visualized via histograms). In addition, flowsheet variables with more than 12 categories were also removed (these are usually manual text entries or image URL links). This process led to a candidate set of 2,686 variables used for initial model training. After pulmonary critical care physicians and hospitalists reviewed this list, 42 variables (corresponding to 146 binary features) were further removed. Among these variables, 28 pertained to mechanical ventilation (e.g., documentation of positive end-expiratory pressure, which occurs in patients receiving mechanical ventilation). Such variables have the potential to 'leak' the outcome (i.e., indicate that the outcome had already occurred), rendering a prediction less meaningful. The remaining 14 variables pertained to SOFA and sepsis prediction scores. We removed these variables since they relied on existing (proprietary) deterioration indices or composite scores and are potentially unavailable across healthcare systems. Preprocessing this variable set using FIDDLE led to 17,083 binary features. The final feature set was used to train an initial ML model on the development cohort.

Step 2 - Variable Selection: After training the initial ML model, we ranked the input variables according to their permutation importance on the development cohort. Permutation importance quantifies how much the model relies on a variable for its predictions by measuring the performance drop when the values of that variable are shuffled across samples, effectively removing the information that variable provides. This process was repeated 1,000 times for every variable, and in the end, all variables were ranked from high to low by the average decrease in performance. Using the ranked list of variables, we applied forward feature selection, starting with the one variable with the highest importance and repeatedly adding the next best variable to the feature subset and retraining the model, until model performance no longer improved. Here, we used the COVID validation set consisting of 100 randomly selected hospital admissions from the development institution in 2020 (these were subsequently excluded from the internal validation results) to measure performance. This allowed us to derive a succinct model with 9 variables (corresponding to 88 features) while still maintaining good model performance. Though some features may not agree with clinical reasoning (i.e., choosing VBG pH over ABG pH), the data-driven approach for variable selection allowed us to choose variables that resulted in the best model performance.

Missing Data

If a particular variable was not collected within a 4-hour window, the group of binary features corresponding to that variable will all be set to 0. As an example, if a value of “Intermittent” was recorded for “Pulse Oximetry Type” during a particular 4-hour window, then the feature vector for this window will have “Pulse Oximetry Type: Continuous” set to 0 and “Pulse Oximetry Type: Intermittent” set to 1; if no value was recorded for the “Pulse Oximetry Type” in that window, then both features will have a value of 0 for that window. Consequently, this approach explicitly encodes variable missingness as a feature and allows the model to produce risk predictions every 4 hours, even when data elements are missing. This also allows us to forgo missing value imputation that makes certain assumptions about the variables, such as by using data from previous windows or by using the average value for that variable. For better context, in **eTable 3**, we evaluated the level of data missingness at each 4-hour interval in the internal validation cohort.

Sample Size Determination

In this retrospective cohort study, we did not pre-specify the sample size; instead, we used all available data to maximize the power and generalizability of our results.

Model Training

The goal of the primary prediction task was to identify high-risk patients who deteriorate quickly. Thus, we labeled a hospital admission based on whether or not the patient experienced the composite outcome within five days of hospital admission. We used all 4-hour windows from the time of the first vital sign up until (but not including) either i) 5 days after the first vital sign was measured or ii) the window in which the individual experienced the outcome or was discharged (whichever comes first). We randomly sampled one window per individual hospital admission to include in the training set, such that no individual was represented more than any other. We repeated this process and created 500 different training sets, leading to an ensemble of 500 regularized logistic regression models, whose outputs were averaged to create a final prediction. The model hyperparameter (L2 regularization strength) was selected using 5-fold cross-validation on the first model and applied to the remaining models in the ensemble.

Evaluation of Primary Use Case

Discriminative performance (receiver operating characteristics (ROC) and precision-recall (PR) curves) was evaluated at a per hospital admission level: we swept the decision threshold and identified individuals who exceeded that threshold prior to the endpoint (when outcome is met or when the 5-day mark is reached) as high risk and low risk otherwise. This approach has been used in past work and avoids biasing our evaluations to patient encounters with more windows [Henry et al. 2015, Oh et al. 2018, Singh et al. 2020]. Additionally, at inference time, to ensure the model is not biased by incomplete data, we removed all windows in which a complete 4-hour window of data was unavailable. Calibration performance was evaluated at a per window level (i.e., per prediction) as we want each prediction to be calibrated and reflect the probability of outcome.

Evaluation of Secondary Use Case

To evaluate models for the secondary use-case, (i.e., triaging low-risk patients), we consider a situation in which a triaging decision is made 48 hours after the patient's first vital sign is measured. Accordingly, we excluded patient hospital admissions that were no longer eligible for potential triaging at 48 hours (those who met the composite outcome or were discharged within 48h of the patient's first vital sign measurement). For each hospital admission, we make the triaging decision based on the average model prediction within the first 48 hours (excluding incomplete windows). A hospital admission's risk score is defined as their average model score of each complete window within the first 48 hours. To measure the number of hospital admissions we can correctly triage to lower acuity care, we calculated the maximum percentage of hospital admissions correctly flagged as low risk (i.e., those with the lowest average predicted score) where the negative predictive value (NPV) is greater than or equal to 0.95 (i.e., of the hospital admissions flagged as low risk, at least 95% will not meet the outcome during the hospital stay). Moreover, for these hospital admissions, we calculated the potential number of days saved, normalized by the total number of correct discharges, if the flagged individuals were discharged from the hospital at 48 hours into their stay. We repeated the procedure on 1,000 bootstrapped samples of each hospital's cohort and visualized the distributions of potential discharge proportions and potential bed days savings and reported the median values from the bootstrapped results.

Confidence Intervals

For all results, 95% confidence intervals (CIs) were generated using 1,000 bootstrapped samples of each validation cohort.

Personal Characteristic Subgroups

We considered the following subgroups defined by personal characteristics:

- Age groups are defined by pre-specified bins: 18-25, 26-45, 46-65, 66-85, >85
- Sex: Female, Male
- Race: Asian, Black, White, Other (which includes: American Indian or Alaskan, Native Hawaiian or Other Pacific Islander, Other, Unknown, Patient Refused, More than 1).
- Ethnicity: Hispanic or Latino, Non-Hispanic or Latino, Unknown

eMethods 2. Inclusion Criteria and Outcome Definitions for Internal and External Validation Cohorts. In alphabetical order, the external healthcare systems were Mass General Brigham (MGB), the University of California San Francisco Medical Center, and University of Texas Southwestern Medical Center. MGB included 10 hospitals: Brigham and Women's Faulkner Hospital, Brigham and Women's Hospital, Cooley Dickinson Hospital, Martha's Vineyard Hospital, Massachusetts General Hospital, McLean Hospital, Nantucket Cottage Hospital, Newton-Wellesley Hospital, North Shore Medical Center, and Wentworth-Douglass Hospital. Six sites with fewer than 100 cases that met the primary outcome were combined into a single cohort when performing evaluation, resulting in a total of 7 external validation cohorts. These medical centers represent both large academic medical centers and small to mid-size community hospitals in regions geographically distinct from the development institution (Midwest), including the Northeast, West, and South regions of the US. Institution-specific results were anonymized.

Inclusion criteria and outcome definitions were implemented by each institution individually, as hospital admission-level data were not shared across sites. An initial definition was developed by MM and then adapted by each individual institution to ensure that the same outcomes were captured accurately given differences in care processes and informatics infrastructure across institutions. Specific implementation details are summarized below.

Inclusion Criteria

- Michigan Medicine (MM)
    - **COVID Diagnosis**: To identify COVID-19, we included hospital admissions with either (i) a positive laboratory test or (ii) a recorded ICD-10 code for COVID-19 and the absence of a negative laboratory test.
    - **Respiratory Distress:** Adult inpatient hospital admissions in which the patient required supplemental oxygen.
- University of California, San Francisco (UCSF)
    - **COVID diagnosis:** Either "Detected" or "Indeterminate" COVID test result, or when patient flagged as having COVID from infection control status table.
    - **Respiratory Distress:** Any patient that had value for O2 device (that was not room air) OR (O2 flow rate > 0) OR (FiO2 > 21)
- University of Texas, Southwestern (UTSW)
    - **COVID diagnosis:** We included all COVID-19 infections associated with hospital encounters as retrieved from the COVID_19_HSP_INFECTIONS table. Patients are accessible in the table as part of the COVID-19 Hospital Infections registry where patients are added if they have an active or presumed COVID-19 infection flag during the admission.
    - **Respiratory Distress:** Includes all patients requiring supplemental oxygen during admission identified by flowsheet documentation of any oxygen device other than "room air", any ventilator settings, any O2 flow >0, or O2 concentration >21%.

- Mass General Brigham (MGB)
    - **COVID diagnosis:** We included hospital admissions where the patient had an active COVID-19 or CoV Presumed infection flag at some point during the admission. At MGB, COVID-19 infection flags are automatically added after a positive COVID-PCR test or positive BinaxNOW now antigen assay. CoV-Presumed is applied in the following scenarios: 1) symptoms and positive serological assay for SARS-CoV-2, 2) positive antigen assay when symptoms are documented, excluding the BinaxNOW assay; 3) PCR resulting as inconclusive, presumptive positive or NEG late signal (reported only at one institution on the Cepheid GeneXpert assay); 4) positive PCR or BinaxNOW assay in an individual who is between 91-180 days after initial diagnosis of COVID-19 or 5) at the discretion of Infection Control.
    - **Respiratory Distress:** Adult inpatient hospital admissions in which the patient required supplemental oxygen. Supplemental oxygen was defined as having flowsheet documentation of an oxygen device other than "None (Room Air)".

<u>Outcome Definition</u>

In general, MV and HHFNC are defined based on clinical events recorded in flowsheets; vasopressors are defined using keyword searches over medication administration records. While the MV and vasopressor definitions are mostly consistent, the HHFNC definition is not identical at each institution due to differences in workflows, though they all correspond to an elevated level of care. Specifically, at some institutions, we have an additional criterion of O2 flow rate ≥ 15L, because at these institutions, nasal cannula with low O2 flow rates were used on the floor but are recorded in the same way as nasal cannula with higher flow rates that are used in the ICU.

- MM
    - **IV vasopressors:** Vasopressors are defined by medication administration records (MAR); we performed a keyword search on the drug name of the MAR for the following well-recognized vasopressors: 'norepinephrine' (aka 'levophed'), 'epinephrine', 'dopamine', 'vasopressin', 'phenylephrine' (aka 'neo-synephrine', 'neosynephrine'), 'angiotensin', and further filtered administrations with route of 'IV' and notgiven = False.
    - **Mechanical Ventilation:** any of the following flowsheet event:
        - "UM IP R CMV START / STOP [Invasive Ventilation Start / Stop]" (313141) with value of "Start"
        - "UM IP R VENT MODE [Vent Mode]" (315640) with a few specific values
        - "UM ED R OXYGEN DEVICE [O2 Device]" 307923 with value 'Ventilator - Emergency Department' or 'Mechanical Ventilation - UH/CVC'
    - **HHFNC:** recorded flowsheet event of "UM ED R OXYGEN DEVICE [O2 Device]" (307923) with value 'Nasal Cannula - Heated High Flow'

- UCSF
  - **IV vasopressors:** Med admin route as (Intravenous, or continuous infusion, or continuous IV infusion, or central venous line induction) and following medications: Dobutamine, Dopamine, Ephedrine, Epinephrine, Milrinone, Norepinephrine, Phenylephrine, Vasopressin.
  - **Mechanical Ventilation:** After excluding patients who had MV on admission (included string "present_on_admission" in values related to intubation), first time where there was a value for "R RT VENT MODE" that was not null
  - **HHFNC:** Either Nasal Cannula or HFNC values for oxygen delivery device with flow rates > 15
- UTSW
  - **IV vasopressors:** Includes all MAR administration of the pressors below based on medication order ID only if route is "intravenous", medication was given (i.e., excludes the following MAR actions: 'Paused','Stopped','Canceled Entry','Held','HELD BY PROVIDER','Missed'), and rate is >0.
    - Vasopressin: '261095', '732983', '249321', '272111'
    - Norepinephrine: '240032', '240493', '12588', '732981'
    - Epinephrine: '3398', '250088', '244667', '3400', '266933', '735102', '250565', '250964', '732978'
    - Dobutamine: '118907','232425','232428','19051'
    - Milrinone: '31759', '231514'
    - Dopamine: '232499', '232498', '232500'
    - Phenylephrine: '240509', '7429', '246371', '732982', '240041','734102'
    - Ephedrine: '233968', '230498', '3382'
  - **Mechanical Ventilation:** Includes flowsheet documentation of ventilator mode ('UTSW R ED VENTILATOR MODE') or a ventilator FiO2 ('UTSW R ED VENTILATOR FIO2 (%)')
  - **HHFNC**: Includes flowsheet documentation of an oxygen device of "high-flow nasal cannula" with an O2 flow rate cutoff > 15.
- MGB
  - **IV vasopressors**: Defined as a documented MAR administration of a vasopressor with an associated MAR action indicating that the medication was given (i.e., excluding actions such as "missed" and "held"). Restricted to MAR actions with a documented route of "Intravenous" and a non-zero dose. Vasopressors were defined using pharmaceutical subclasses of Cardiovascular Sympathomimetic - Beta-Adrenergic Agonists, Antidiuretic and Vasopressor Hormones, Cardiovascular Sympathomimetics, and Renin-Angiotensin-Aldosterone System (RAAS) Hormones
  - **Mechanical Ventilation:** Defined as flowsheet documentation of a ventilator mode of 'AC/VC', 'AC/PC', 'SIMV/PC', 'ASV', 'AC/PRVC', 'PC-PSV', 'AC/VG', 'SIMV/PRVC', 'SIMV/VC', or 'HFJV'
  - **HHFNC**: Defined as flowsheet documentation of an oxygen device of "High Flow Nasal Cannula" or "High flow face mask". No additional O2 rate cutoffs were used.

eMethods 3. Documentation for External Validation.

Based on existing and new connections formed between different institutions and the relevant access to data each institution has, we identified the sites at which we can rapidly perform the external validation. We first provided a specification document to each institution that describes a unified format containing all information needed to perform the evaluation. Researchers at each institution performed their own cohort data extraction from EHR databases and outcome definitions and collated everything into a unified format. Model parameters for each of the 500 models along with the necessary code (including a standard feature processing procedure) were packaged into a transferable computer program by MM, which was sent to each institution. Researchers at each institution then ran the program on their own infrastructure and transferred back only model results; no identifiable information was shared. This procedure was done rapidly (within a month) and involved less risk of PHI-related issues compared to sharing raw patient data (which involves signing data use agreements with multiple institutions).

In the majority of these cases, these mappings were straightforward (e.g., vital signs such as respiratory rate were recorded in a consistent manner across institutions). However, in cases where variables could not be mapped exactly, we worked together towards reasonable mappings. For example, at one institution, head-of-bed position "less than 20 degrees" was sometimes documented, but this was not a possible head-of-bed variable category in the model. Therefore, these values were mapped to a head-of-bed position of "15 degrees" to be compatible with the preprocessing and model code.

We provided the following information to collaborators from external institutions to format the data, allowing for identical processing and model evaluation.

3.1 Instructions for Running Data Preprocessing

This information is also available in the public GitHub repository: https://github.com/MLD3/M-CURES.

Usage

- Refer to requirements.txt for the necessary pip packages.
- **preprocessing**: Run ./run.sh.
- **preprocessing_notebooks**: alternatively, run the notebooks in alphanumeric order.
- **evaluation**: Run the Eval.ipynb notebook.

Input

An example usage of the pipeline is provided with fake input data in preprocessing/sample_input and evaluation/sample_cohort.csv. The easiest way to use the code is to create local copies of preprocessing -> preprocessing_UM and evaluation -> evaluation_UM and replace the input files with real data. Please refer to the sample input files (and descriptions below) for formatting requirements.

Cohort specification

Provide the following files to specify the cohort and relevant windows on which the model is applied.

- windows_map.csv contains all 4h windows for all hosp_ids.
    - hosp_id column is the unique identifier for the encounter
    - window_id column is the index of 4h windows for the current encounter
    - ID column is "{hosp_id}-{window_id}"
- windows.csv has the same content as the ID column in windows_map.csv
- sample_cohort.csv is used by Evaluation_UseCase1.ipynb: predicting composite outcome that happens within the first 5 days. It has the same ID, hosp_id, and window_id columns as in windows_map.csv, and it contains an additional column y specifying the outcome label. The labels "y" for each window are defined as follows:
    - If a patient encounter experiences the outcome, then windows after the outcome window are not used for prediction and should not be included. Only windows before the outcome window are included and they have a label of 1.
    - If a patient does not have an outcome, then all of their windows have a label of 0, and we only include up to the first 30 windows (first 5 days).

Every encounter should have no more than 30 windows.

- sample_cohort_outcome_past_2days.csv is used by Evaluation_UseCase2.ipynb: predicting composite outcome that happens after 48h using the first 48h data. It has the same format as sample_cohort.csv, except it only contains encounters who have the outcome after two days, and the y label specifies if the outcome occurs *ever* (rather than within the first 5 days). Every encounter should have exactly 12 windows (48h worth of data).

Data

For details on the expected values of each variable, please refer to preprocessing/metadata/out_*/{discretization|feature_names}.json.

- demog.csv contains three columns:
    - age_value: numeric
    - sex_value: ['M', 'F']
    - race_value:
        - "African American"
        - "American Indian or Alaska Native"
        - "Asian"
        - "Caucasian"
        - "Native Hawaiian and Other Pacific Islander"
        - "Other"
        - "Patient Refused"
        - "Unknown"

The other input data files all have four columns: ['ID', 't', 'variable_name', 'variable_value'].

- The ID column specifies a 4h window of a specific encounter and should be contained in the windows_map.csv file.
- The t column is measured in minutes relative to the start of the current 4h window.

Below are the expected variable_names in each file:

- vitals.csv
    - heartrate
    - temperature
    - sbp
    - dbp
    - respiratoryrate
    - spo2
- flow.csv: (note the underscore prefix)
    - '_307928' for "O2 flow rate"
    - '_313030' for "Pulse Oximetry type"
        - "Intermittent"
        - "Continuous"
    - '_314689' for "BP: Patient Position"
        - "Lying"
        - "Sitting"
        - "Standing"
    - '_355444' for "Head of Bed Position"
        - "HOB at 15 degrees"
        - "HOB at 30 degrees"
        - "HOB at 45 degrees"
        - "HOB at 60 degrees"
        - "HOB at 90 degrees"
        - "HOB flat (medical condition)"
        - "Reverse Trendelenberg"
        - "other (see comments)"
- labs.csv
    - pH (Ven Blood Gas): '81723_value' and '81723_hilonormal_flag'
    - pCO2 (Art Blood Gas): '84066_value' and '84066_hilonormal_flag'
- meds.csv
    - currently none supported

3.2 Instructions for Running Model Evaluation

External collaborators were requested to put the model predictions into the following format, along with other relevant metadata useful for model evaluation. We then shared several Jupyter notebooks to take in this formatted table and produce summary information of model performance, e.g., points on the ROC curve, performance on personal characteristic subgroups, and confidence intervals. This information was then sent to the internal team and aggregated to produce the final figures and tables.

Each row should be an encounter, with a value for each of the following columns:

1. *hosp_id*: unique identifier for each hospital admission
2. *y_score_fourvar:* Maximum prediction across encounter for the four-variable model (excluding the 0th window)
3. *y_score_mcures*: Maximum prediction across encounter for the M-CURES model (excluding the 0th window)
4. *y_scores_four_lst*: All predicted scores for primary use-case for a particular encounter from the four-variable model (including the 0th window). Format: “[0.1, 0.2]”
5. *y_scores_mcures_lst*: All predicted scores for primary use-case for a particular encounter from the M-CURES model (Including the 0th window)
6. *y*: original label for encounter (for primary use-case, values in {0,1})
7. *admission_date:* Month and year of admission for encounter *(*format: ‘*(M)M/YY’*) - NOTE: please ensure that you do not provide the day
8. *outcome:* For a positive encounter (y=1), the first outcome experienced by the patient (options: ‘mortality’, ‘MV’, ‘HHFNC’, ‘IV’). Value should be set to None/np.nan/‘’ for negative encounters.
9. *outcome_time*: time of outcome or np.nan (in minutes) with respect to the beginning of the first window, for an outcome that happens any time within a hospital admission (not limited to first 5 days)
10. *final_time_min*: minimum of {outcome time, discharge time} (in minutes) with respect to the beginning of the first window
11. *age*: age of patient at admission (in integer years)
12. *sex:* sex of patient corresponding to the encounter (options: ‘M’, ‘F’)
13. *race:* race of patient corresponding to the encounter (options: "African American", "American Indian or Alaska Native", "Asian", "Caucasian", "Native Hawaiian and Other Pacific Islander", "Other", "Patient Refused", "Unknown", “More than 1”)
14. *ethnicity*: ethnicity of patient corresponding to the encounter (options: “Hispanic or Latino”, “Non-Hispanic or Latino”, “Patient Refused”, “Unknown”)

To get a table started with the first five pieces of information:

After the final code under the baseline header of Evaluation_1, you can get the four variable information with the following code:

```
"""""

y_score_lst = df_Yte_all.groupby(['hosp_id'])['y_score'].apply(list)
df1 = pd.DataFrame({'y_score_fourvar_lst': y_score_lst})
df2 = pd.DataFrame({'id': df_Yte_agg.index, 'y_score_fourvar': y_score, 'y': y_true})

mergedDf = df1.merge(df2, left_index=True, right_index=True)
""""
```

After the final code under the m-cures lite header, you can add the following lines of code:

```
"""
y_score_lst = df_Yte_all.groupby(['hosp_id'])['y_score'].apply(list)
df1 = pd.DataFrame({'y_score_mcures_lst': y_score_lst})
df2 = pd.DataFrame({'id': df_Yte_agg.index, 'y_score_mcures': y_score})

mergedDf2 = df1.merge(df2, left_index=True, right_index=True)
outcome_outputs = mergedDf.merge(mergedDf2, left_index = True, right_index = True)
"""
```

outcome_outputs then contains all of this information, you could then save this out.

Error Checking Descriptions (Cell 6 in the master notebook):

1. *Have all required columns?* Are the required fields present in the dataframe
2. *Age:*
    a. Is integer? Ensures age is an integer
    b. Min >= 18? Ensures there are no children in the dataset
    c. Max <= 90? Ensures that the oldest person is less than 90 years old (ok to fail this check)
3. *Only use allowed outcomes?* Ensures that only HHFNC, MV, mortality, and IV are used as outcomes for those with outcomes
4. *Race: Only use allowed race categories?* Ensures that only the allowed race categories are present in the dataframe
5. *Sex: Only used allowed sex categories?* Ensures that only the allowed sex categories are present in the dataframe
6. *Ethnicity: Only use allowed ethnicity categories?* Ensures that only the allowed ethnicity categories are present in the dataframe
7. *Score lists the same length?* Ensures that the length of the list of scores for the four variable model and the m-cures model are the same
8. *Score (eval1) lists the same length?* Ensures that the length of the list of scores for the four variable model and the m-cures model are the same for evaluation 1
9. *Is fourvar (eval1) max is max of list?* Ensures that the maximum score for an individual in the list of evaluation 1 scores is equal to the predicted score in the table for the four variable model
10. *Is MCURES (eval1) max is max of list?* Ensures that the maximum score for an individual in the list of evaluation 1 scores is equal to the predicted score in the table for the MCURES model
11. *Is len of fourvar equal to expected number of windows?* The number of windows present should be all windows up to (but not including) the final time-point (window of final_time_min) or 30, whichever is smaller. This check ensures this is true for the four variable list of scores.
12. *Is len of fourvar equal to expected number of windows?* The number of windows present should be all windows up to (but not including) the final time-point (window of final_time_min) or 30, whichever is smaller. This check ensures this is true for the four variable list of scores.

13. *Is len of fourvar equal to expected number of windows?* The number of windows present should be all windows up to (but not including) the final window (window of final_time_min) or 30, whichever is smaller. This check ensures this is true for the four variable list of scores.
14. *Is len of fourvar (eval1) equal to expected number of windows?* The number of windows present for evaluation 1 should be windows up from the 2nd (remove the first incomplete window) up to (but not including) the final window (window of final_time_min) or 30, whichever is smaller. This check ensures this is true for the four variable list of scores.
15. *Is len of MCURES equal to expected number of windows?* The number of windows present should be all windows up to (but not including) the final window (window of final_time_min) or 30, whichever is smaller. This check ensures this is true for the MCURES list of scores.
16. *Is len of MCURES (eval1) equal to expected number of windows?* The number of windows present for evaluation 1 should be windows up from the 2nd (remove the first incomplete window) up to (but not including) the final window (window of final_time_min) or 30, whichever is smaller. This check ensures this is true for the MCURES list of scores.

eTable 1. MCURES model weights (comma-separated file).

Please download the file from the website.

eTable 2. Characteristics of the development cohort and comparison with the internal validation cohort. Both cohorts are from Michigan Medicine. Statistically significant differences (at α=0.001 with a Bonferroni correction for multiple hypotheses) are denoted by *.

Characteristic	Development	Internal Validation	p-value
Number of patients	24,419	887	-
Number of hospital admissions	35,040	956	-
Median age in years [IQR]	63 [51-74]	64 [52–75]	-
Age Group (%) [18, 25] (25, 45] (45, 65] (65, 85] >85	 1,275 (3.6) 5,114 (14.6) 13,060 (37.3) 13,064 (37.3) 2,432 (6.9)	 17 (1.8) 129 (13.5) 374 (39.1) 365 (38.2) 70 (7.3)	0.02
Sex (%) Female Male	 16,877 (48.2) 18,163 (51.8)	 420 (43.9) 536 (56.1)	0.01
Race (%) White Black Asian Other/Unknown	 29,402 (83.9) 3,954 (11.3) 625 (1.8) 1,059 (3.0)	 649 (67.9) 187 (19.6) 30 (3.1) 90 (9.4)	<0.0001*
Ethnicity (%) Hispanic or Latino Not Hispanic or Latino Other/Unknown	 - - -	 34 (3.6) 883 (92.4) 39 (4.1)	-
Median LOS in hours [IQR]	97 [55–173]	138 [83–261]	-
Outcome ever (%) Death MV IV Vaso HHFNC	 963 (2.7) 2,341 (6.7) 1,320 (3.8) 1,858 (5.3)	 60 (6.3) 98 (10.3) 87 (9.1) 218 (22.4)	<0.0001*
Primary Outcome <= 5 days	3,757 (10.7)	206 (21.6)	<0.0001*
Reason for composite outcome (% of outcomes) Death MV IV Vaso HHFNC	 252 (6.7) 1,737 (46.2) 454 (12.1) 1,314 (35.0)	 5 (2.4) 20 (9.7) 9 (4.4) 172 (83.5)	<0.0001*

Acronyms: IQR, interquartile range; LOS, Length-of-Stay; MV, Mechanical Ventilation; IV, Intravenous Vasopressors, HHFNC, Heated High-Flow Nasal Cannula.

eTable 3. Extent of data missingness for each 4-hour interval in the internal validation cohort.

Variable	% 4h windows with missing data
Respiratory rate	17.0%
SpO2	16.3%
VBG pH	95.6%
ABG pCO2	99.6%
O2 flow rate	51.5%
Pulse oximetry type	62.7%
BP Patient position	40.6%

*Out of 19,997 4-hour windows in the internal validation cohort

eTable 4. P-values for pairwise comparisons of characteristics between the internal validation cohort and each external validation cohort. We applied chi-square tests for homogeneity to compare categorical personal characteristic variables. Every external validation cohort differed in at least one personal characteristic dimension. Statistically significant differences (at α=0.001 with a Bonferroni correction for multiple hypotheses) are denoted by *.

Characteristic	MM vs A	MM vs B	MM vs C	MM vs D	MM vs E	MM vs F	MM vs G
Sex	0.6	0.3	0.6	0.01	0.3	0.5	0.003
Age Group	0.02	0.009	3e-6*	0.09	5e-11	3e-21*	2e-5*
Race	2e-43*	3e-10*	1e-9*	4e-11*	2e-7	2e-16*	2e-70*
Ethnicity	2e-51*	1e-52*	2e-49*	4e-28*	1e-15	1e-14*	1e-45*
Has Outcome (Ever)	7e-36*	0.05	4e-14*	8e-14*	4e-8*	2e-12*	0.03
Has Primary Outcome	6e-9*	0.1	0.002	3e-5*	2e-5*	0.3	0.002

Acronyms: MM, Michigan Medicine.

eTable 5. P-values for pairwise comparisons of the reasons for meeting the composite outcome, between the internal validation cohort and the development cohort, as well as between the internal cohort and each external validation cohort. We applied chi-square tests for homogeneity to compare the reason for outcome. Statistically significant differences (at α=0.006 with a Bonferroni correction for multiple hypotheses) are denoted by *.

	MM vs DEV	MM vs A	MM vs B	MM vs C	MM vs D	MM vs E	MM vs F	MM vs G
Reason for Outcome	2e-41*	3e-30*	0.7	1e-11*	2e-15*	4e-7*	9e-12*	0.007

Acronyms: MM, Michigan Medicine; DEV, Development cohort.

eTable 6. Estimated 95% confidence intervals of the performance difference between the internal validation cohort and each external validation cohort. The difference is significant if the interval does not overlap with zero (denoted by *).

Performance Measure	MM vs A	MM vs B	MM vs C	MM vs D	MM vs E	MM vs F	MM vs G
Difference in AUROC	[-0.05, 0.03]	[-0.08, 0.02]	[-0.06, 0.04]	[-0.04, 0.08]	[-0.08, 0.01]	[-0.02, 0.09]	[-0.03, 0.09]
Difference in AUPR	[0.04, 0.23] *	[-0.13, 0.08]	[-0.01, 0.19]	[0.04, 0.25] *	[-0.01, 0.21]	[0.03, 0.24] *	[0.10, 0.32] *

Acronyms: MM, Michigan Medicine; AUROC, Area Under the Receiver Operating Characteristic; AUPR: Area Under the Precision Recall Curve.

eTable 7. Estimated 95% confidence intervals of model performance during a specific time period, within each validation cohort.

Time Period	MM	A	B	C	D	E	F	G
Mar ’20 – May ’20	0.87 (0.82, 0.92)	0.81 (0.77, 0.84)	0.93 (0.86, 0.98)	0.80 (0.74, 0.84)	0.78 (0.72, 0.84)	0.86 (0.81, 0.90)	0.78 (0.72, 0.83)	0.81 (0.69, 0.91)
Jun ’20 – Aug ’20	0.80 (0.60, 0.93)	0.74 (0.57, 0.87)	0.80 (0.71, 0.88)	0.72 (0.45, 0.93)	0.58 (0.34, 0.80)	0.89 (0.76, 0.98)	0.58 (0.31, 0.85)	0.75 (0.63, 0.85)
Sept ’20 – Nov ‘20	0.73 (0.66, 0.80)	0.85 (0.78, 0.91)	0.83 (0.77, 0.89)	0.86 (0.75, 0.94)	0.76 (0.59, 0.91)	0.86 (0.73, 0.95)	0.77 (0.64, 0.88)	0.71 (0.56, 0.84)
Dec ’20 – Feb ‘21	0.79 (0.73, 0.85)	0.82 (0.77, 0.88)	0.83 (0.79, 0.87)	0.83 (0.77, 0.89)	0.82 (0.74, 0.88)	0.81 (0.74, 0.86)	0.78 (0.69, 0.86)	0.81 (0.74, 0.88)

Acronyms: MM, Michigan Medicine.

eTable 8. Estimated 95% confidence intervals (99.8% CIs with Bonferroni correction) of the performance difference during a specific time period relative to overall performance, within each validation cohort. No difference is statistically significant (the intervals all overlap with zero).

Time Period	MM	A	B	C	D	E	F	G
Mar ’20 – May ’20	[-0.03, 0.16]	[-0.08, 0.06]	[-0.01, 0.18]	[-0.10, 0.05]	[-0.12, 0.11]	[-0.06, 0.10]	[-0.10, 0.10]	[-0.13, 0.22]
Jun ’20 – Aug ’20	[-0.34, 0.20]	[-0.36, 0.14]	[-0.19, 0.08]	[-0.55, 0.18]	[-0.60, 0.20]	[-0.22, 0.16]	[-0.71, 0.15]	[-0.23, 0.13]
Sept ’20 – Nov ‘20	[-0.18, 0.04]	[-0.08, 0.13]	[-0.12, 0.10]	[-0.11, 0.17]	[-0.39, 0.18]	[-0.20, 0.19]	[-0.19, 0.19]	[-0.30, 0.13]
Dec ’20 – Feb ‘21	[-0.13, 0.11]	[-0.10, 0.09]	[-0.08, 0.07]	[-0.09, 0.11]	[-0.09, 0.14]	[-0.14, 0.07]	[-0.15, 0.16]	[-0.09, 0.16]

Acronyms: MM, Michigan Medicine.

eTable 9. Estimated 95% confidence intervals of model performance on each personal characteristic subgroup, within each validation cohort. "N/A" denotes where evaluation was not performed because the subgroup consisted of fewer than 25 examples.

Subgroup	**MM**	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Sex:F	0.81 (0.74, 0.86)	0.83 (0.79, 0.87)	0.80 (0.75, 0.85)	0.82 (0.77, 0.88)	0.82 (0.76, 0.88)	0.85 (0.79, 0.90)	0.77 (0.70, 0.84)	0.75 (0.67, 0.82)
Sex:M	0.80 (0.75, 0.84)	0.80 (0.77, 0.84)	0.86 (0.82, 0.89)	0.81 (0.77, 0.85)	0.76 (0.70, 0.82)	0.83 (0.79, 0.88)	0.78 (0.72, 0.83)	0.79 (0.73, 0.85)
Age:17-25	N/A	0.84 (0.45, 1.00)	N/A	N/A	N/A	N/A	N/A	N/A
Age:25-45	0.77 (0.66, 0.87)	0.88 (0.83, 0.93)	0.82 (0.74, 0.90)	0.83 (0.72, 0.92)	0.86 (0.76, 0.94)	0.89 (0.77, 0.97)	0.88 (0.77, 0.96)	0.69 (0.48, 0.88)
Age:45-65	0.79 (0.72, 0.84)	0.79 (0.74, 0.83)	0.86 (0.82, 0.90)	0.85 (0.78, 0.90)	0.79 (0.71, 0.86)	0.83 (0.76, 0.89)	0.79 (0.70, 0.86)	0.76 (0.68, 0.83)
Age:65-85	0.84 (0.78, 0.88)	0.82 (0.77, 0.86)	0.81 (0.76, 0.86)	0.80 (0.74, 0.84)	0.76 (0.68, 0.82)	0.84 (0.78, 0.89)	0.79 (0.72, 0.85)	0.79 (0.70, 0.87)
Age:85+	0.86 (0.75, 0.95)	0.80 (0.70, 0.88)	0.82 (0.61, 0.98)	0.79 (0.68, 0.88)	0.79 (0.65, 0.91)	0.85 (0.78, 0.91)	0.70 (0.59, 0.81)	0.80 (0.67, 0.90)
Race:Asian	0.84 (0.54, 1.00)	0.84 (0.73, 0.93)	0.78 (0.53, 0.97)	0.83 (0.70, 0.93)	0.95 (0.84, 1.00)	N/A	0.93 (0.79, 1.00)	0.78 (0.69, 0.86)
Race:Black	0.85 (0.78, 0.91)	0.83 (0.73, 0.92)	0.83 (0.77, 0.89)	0.81 (0.69, 0.90)	0.84 (0.76, 0.91)	0.84 (0.71, 0.93)	0.71 (0.53, 0.89)	0.56 (0.30, 0.80)
Race:Other	0.82 (0.67, 0.93)	0.83 (0.79, 0.87)	0.87 (0.80, 0.92)	0.81 (0.68, 0.91)	0.86 (0.79, 0.91)	0.91 (0.85, 0.95)	0.65 (0.52, 0.76)	0.81 (0.73, 0.88)
Race:White	0.79 (0.74, 0.83)	0.80 (0.76, 0.84)	0.83 (0.79, 0.87)	0.82 (0.78, 0.86)	0.73 (0.65, 0.79)	0.83 (0.79, 0.87)	0.80 (0.75, 0.84)	0.78 (0.68, 0.87)
Ethnicity:Hispanic	0.77 (0.51, 0.95)	0.83 (0.79, 0.87)	0.82 (0.76, 0.88)	0.84 (0.77, 0.89)	0.82 (0.73, 0.89)	0.93 (0.88, 0.97)	0.74 (0.60, 0.87)	0.81 (0.73, 0.89)
Ethnicity:Non-Hispanic	0.81 (0.77, 0.84)	0.81 (0.78, 0.85)	0.84 (0.81, 0.88)	0.81 (0.76, 0.85)	0.77 (0.71, 0.82)	0.82 (0.77, 0.86)	0.79 (0.74, 0.83)	0.76 (0.69, 0.82)
Ethnicity:Unknown	0.82 (0.65, 0.95)	0.78 (0.67, 0.86)	0.85 (0.62, 1.00)	0.84 (0.60, 1.00)	N/A	0.94 (0.84, 0.99)	0.71 (0.52, 0.87)	N/A

Acronyms: MM, Michigan Medicine; F, Female; M, Male; AUROC, Area Under the Receiver Operating Characteristic.

eTable 10. Estimated 95% confidence intervals (99.8% CIs with Bonferroni correction) of the performance difference of each personal characteristic subgroup relative to overall performance, within each validation cohort. No subgroup is significantly different from overall performance in terms of AUROC. **"**N/A" denotes where evaluation was not performed because the subgroup consisted of fewer than 25 examples.

Subgroup	**MM**	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Sex:F	[-0.14, 0.11]	[-0.07, 0.10]	[-0.13, 0.06]	[-0.10, 0.10]	[-0.08, 0.14]	[-0.10, 0.12]	[-0.13, 0.12]	[-0.19, 0.10]
Sex:M	[-0.09, 0.08]	[-0.09, 0.07]	[-0.05, 0.10]	[-0.10, 0.07]	[-0.14, 0.08]	[-0.09, 0.08]	[-0.12, 0.11]	[-0.12, 0.13]
Age:17-25	N/A	[-0.61, 0.22]	N/A	N/A	N/A	N/A	N/A	N/A
Age:25-45	[-0.25, 0.11]	[-0.03, 0.14]	[-0.16, 0.12]	[-0.19, 0.15]	[-0.13, 0.21]	[-0.28, 0.17]	[-0.16, 0.23]	[-0.44, 0.26]
Age:45-65	[-0.15, 0.08]	[-0.13, 0.06]	[-0.05, 0.11]	[-0.09, 0.13]	[-0.20, 0.13]	[-0.14, 0.10]	[-0.15, 0.15]	[-0.16, 0.14]
Age:65-85	[-0.07, 0.13]	[-0.07, 0.09]	[-0.12, 0.06]	[-0.11, 0.08]	[-0.18, 0.11]	[-0.11, 0.11]	[-0.12, 0.13]	[-0.17, 0.15]
Age:85+	[-0.13, 0.20]	[-0.19, 0.10]	[-0.49, 0.18]	[-0.21, 0.14]	[-0.27, 0.20]	[-0.14, 0.13]	[-0.27, 0.10]	[-0.22, 0.18]
Race:Asian	[-0.58, 0.24]	[-0.16, 0.16]	[-0.63, 0.20]	[-0.19, 0.15]	[-0.01, 0.27]	N/A	[-0.10, 0.28]	[-0.16, 0.16]
Race:Black	[-0.09, 0.15]	[-0.12, 0.16]	[-0.12, 0.11]	[-0.23, 0.14]	[-0.09, 0.21]	[-0.27, 0.14]	[-0.41, 0.21]	[-0.63, 0.17]
Race:Other	[-0.32, 0.20]	[-0.10, 0.09]	[-0.10, 0.12]	[-0.32, 0.15]	[-0.09, 0.18]	[-0.03, 0.16]	[-0.33, 0.08]	[-0.14, 0.15]
Race:White	[-0.12, 0.07]	[-0.09, 0.07]	[-0.09, 0.08]	[-0.06, 0.10]	[-0.19, 0.07]	[-0.09, 0.07]	[-0.07, 0.12]	[-0.18, 0.15]
Ethnicity:Hispanic	[-0.60, 0.23]	[-0.06, 0.10]	[-0.11, 0.09]	[-0.08, 0.12]	[-0.12, 0.16]	[-0.02, 0.18]	[-0.29, 0.16]	[-0.12, 0.18]
Ethnicity:Non-Hispanic	[-0.08, 0.08]	[-0.06, 0.06]	[-0.06, 0.08]	[-0.10, 0.08]	[-0.14, 0.08]	[-0.11, 0.07]	[-0.10, 0.10]	[-0.14, 0.12]
Ethnicity:Unknown	[-0.39, 0.24]	[-0.20, 0.11]	[-0.39, 0.20]	[-0.46, 0.22]	N/A	[-0.10, 0.20]	[-0.40, 0.20]	N/A

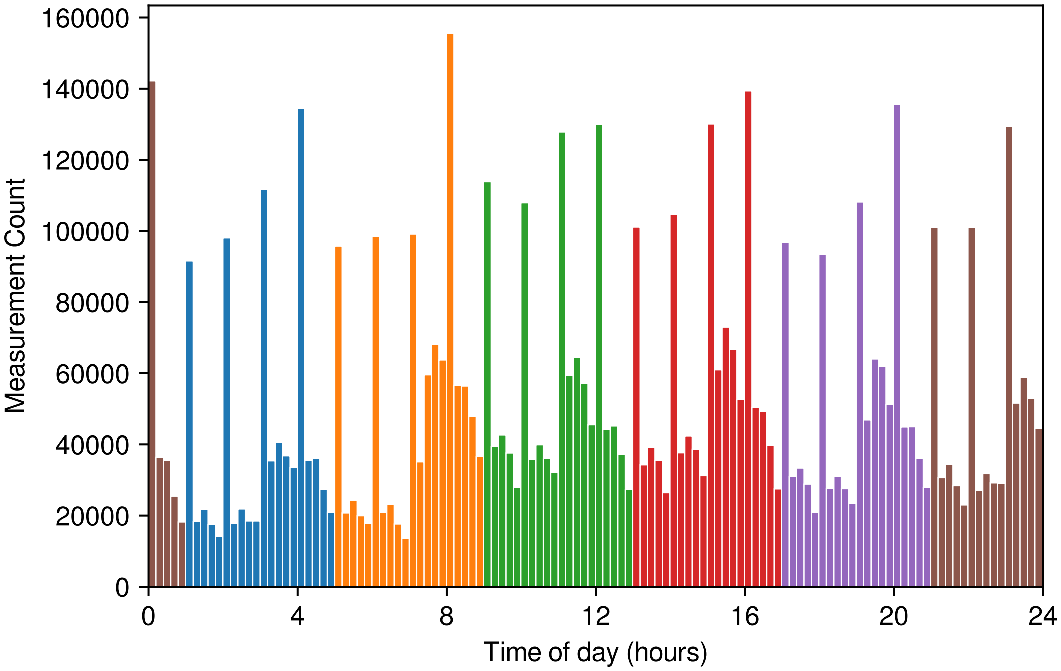
Acronyms: MM, Michigan Medicine; F, Female; M, Male; AUROC, Area Under the Receiver Operating Characteristic.

eTable 11. Estimated 95% confidence intervals (99.8% CIs with Bonferroni correction) of the performance difference between White and each other race subgroup, within each validation cohort. The difference is significant if the interval does not overlap with zero (denoted by *).

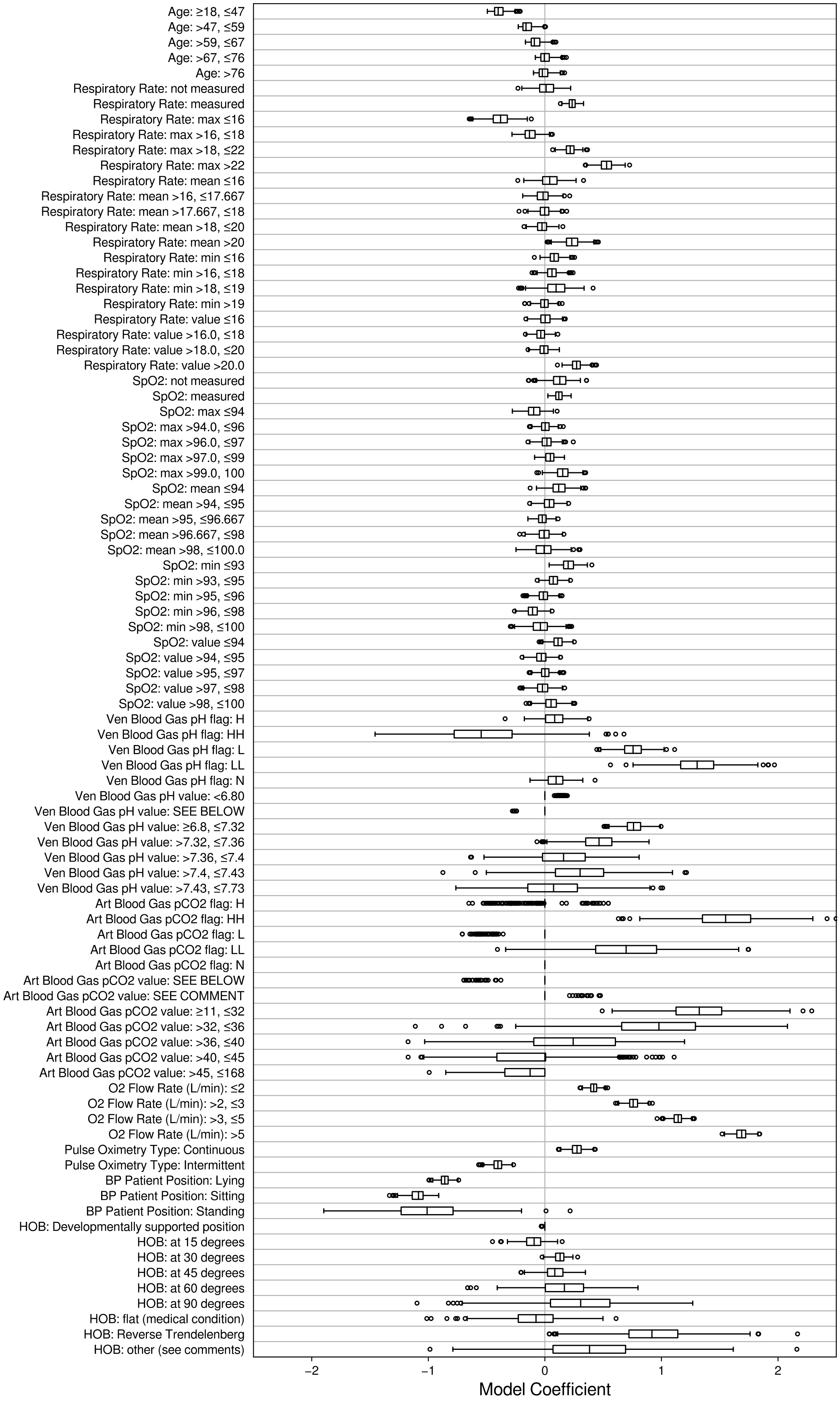
Comparison	**MM**	**A**	**B**	**C**	**D**	**E**	**F**	**G**
W-A	[-0.26, 0.53]	[-0.19, 0.13]	[-0.18, 0.45]	[-0.17, 0.19]	[-0.36, -0.04] *	N/A	[-0.25, 0.08]	[-0.21, 0.20]
W-B	[-0.17, 0.06]	[-0.18, 0.11]	[-0.10, 0.12]	[-0.12, 0.22]	[-0.26, 0.03]	[-0.13, 0.23]	[-0.17, 0.36]	[-0.17, 0.66]
W-O	[-0.20, 0.30]	[-0.12, 0.07]	[-0.13, 0.08]	[-0.14, 0.22]	[-0.28, 0.00]	[-0.17, 0.02]	[-0.02, 0.39]	[-0.22, 0.15]

Acronyms: MM, Michigan Medicine; A, Asian; B, Black; O, Other races; W, White.

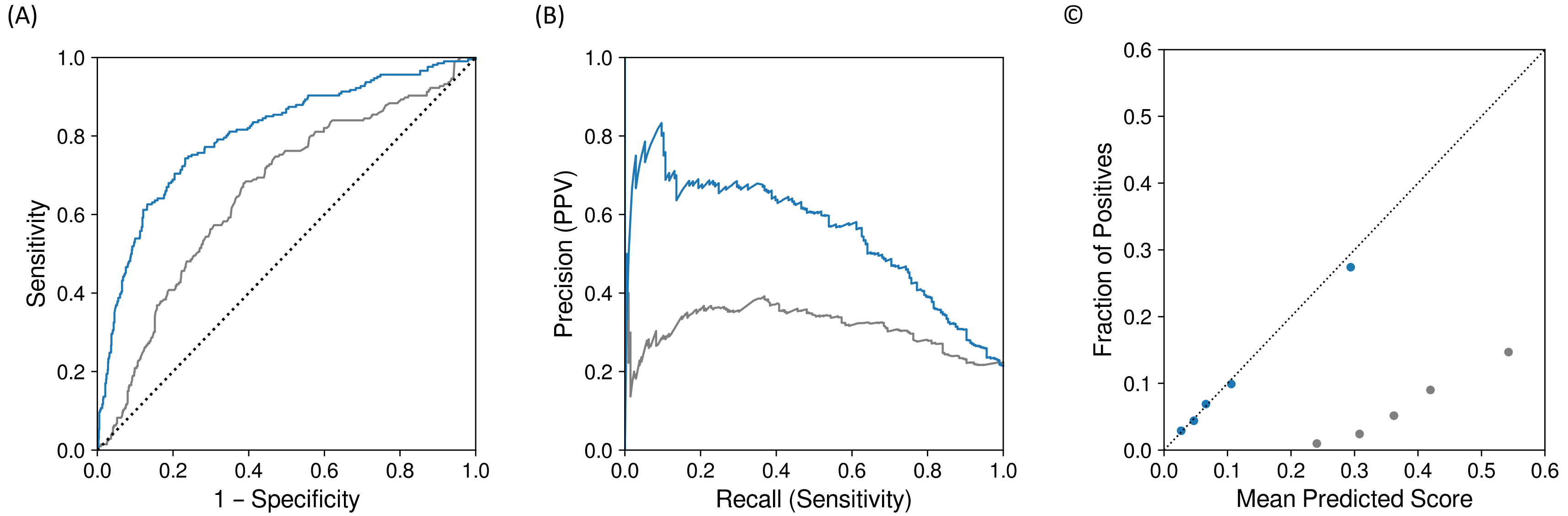
eFigure 1. Measurement frequency of patient heart rate throughout different times of the day, in the development cohort (Michigan Medicine, 2015-2019). Based on the empirical measurement frequency of important vital signs, we defined 4-hour time windows with respect to time points of a day at 1am, 5am, 9am, 1pm, 5pm, and 9pm. These time points correspond to right after the measurement "peaks" and were selected with the feasibility of real-time deployment of the system in mind.



eFigure 2. Visualization of weights of the 88 features over 500 regularized logistic regression models in the ensemble. A machine-readable version of all model parameters is provided in **eTable 1**.



eFigure 3. Model performance comparison of MCURES (shown in blue) and Epic Deterioration Index (shown in gray) on the MM internal validation cohort. We measure discriminative performance in (A) ROC curves and (B) PR curves. Model calibration is shown in (C) Reliability plots based on quintiles of predicted scores. Legend and results with 95% confidence intervals are summarized in (D). The MCURES model outperforms the Epic Deterioration Index in terms of both discriminative performance (the AUROC and AUPR scores) and calibration performance (the ECE score). ROC= receiver operating characteristics; PR=precision-recall; AUROC=area under the receiver operating characteristics curve; AUPR=area under the precision-recall curve; ECE=expected calibration error; EDI=Epic Deterioration Index.

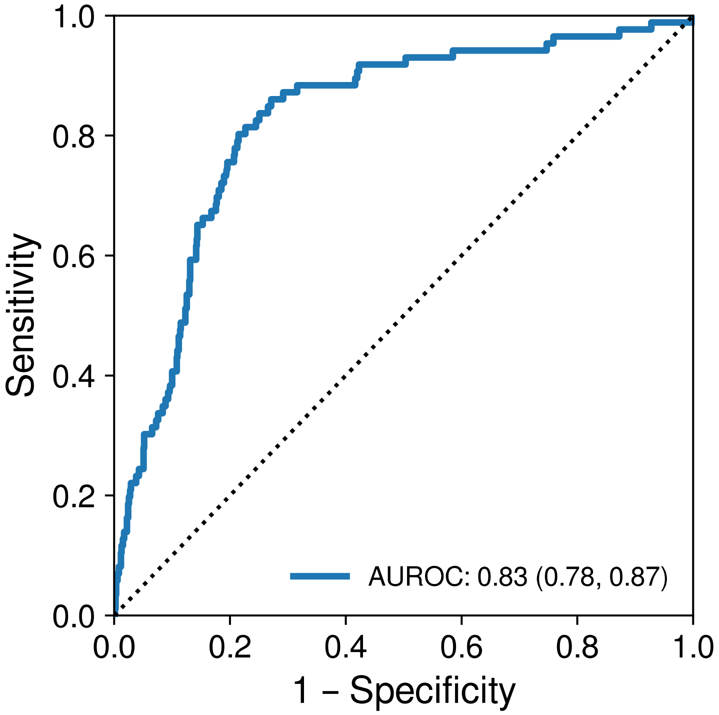


(D)

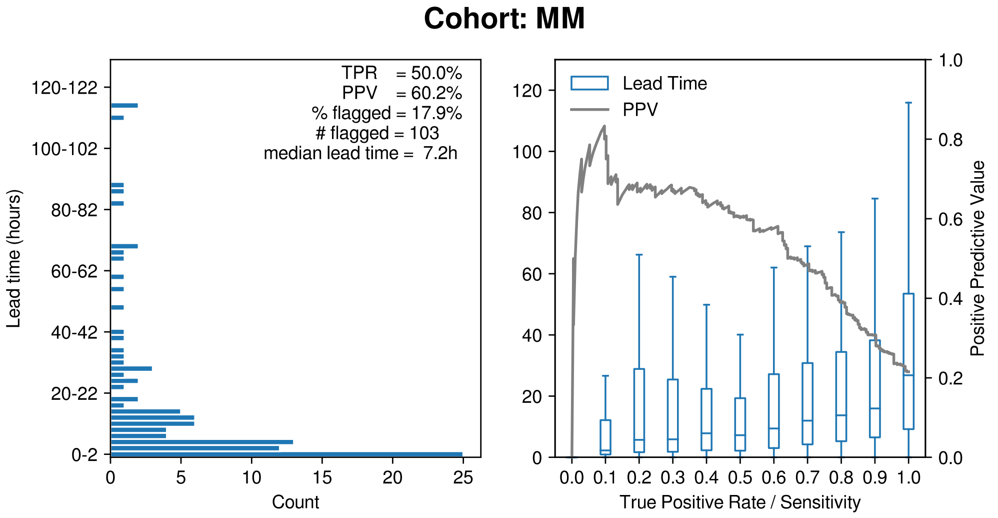
		AUROC	AUPR	ECE
●	MCURES	0.804 (0.770, 0.841)	0.549 (0.483, 0.631)	0.007 (0.003, 0.021)
●	EDI	0.657 (0.618, 0.701)	0.309 (0.264, 0.362)	0.310 (0.297, 0.322)

EDI, Epic Deterioration Index.

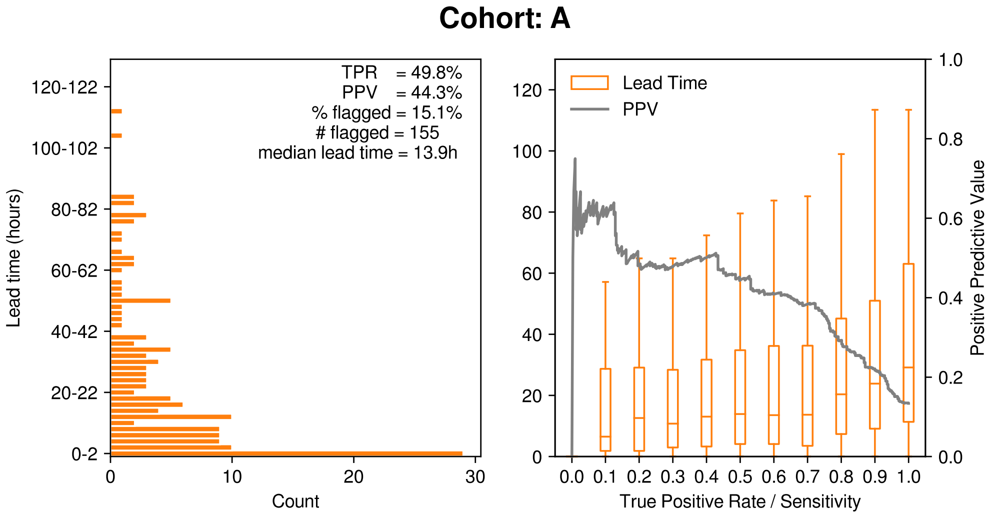
eFigure 4. Model performance with an alternative label definition on the MM internal validation cohort. On the internal evaluation cohort, we re-defined the outcome label to be the composite of in-hospital mortality, mechanical ventilation, and vasopressors (removing HHFNC as part of the outcome definition). This brings the outcome rate to 9% and the number of positive cases to 86 (out of 956 hospital admissions). Consequently, the performance of our model increases from AUROC of 0.80 (95% CI: 0.77, 0.84) to 0.83 (95% CI: 0.78, 0.87).



eFigure 5. Analysis of lead time on each validation cohort. On each validation cohort, we conducted an analysis of the “lead time”, i.e., how long in advance our model can flag a patient before he or she meets the deterioration outcome. We first considered a threshold resulting in a true positive rate (sensitivity) of 0.5, and visualized the distribution of lead time across all correctly flagged patients (true positives) in the left subplot. At this threshold, the model is able to flag patients well before the outcome occurs in many instances, with a median lead time of 7-18 hours at most institutions, and a positive predictive value of 40%-60%. In the right subplot, we swept over all score thresholds from 0 to 1 at 0.1 resolution, reporting the lead time distribution for true positives as boxplots. As the threshold is lowered, patients are flagged earlier in their hospital admission and the lead time increases. However, the positive predictive value decreases to random, or the incidence rate of the outcome. Note that, though we only considered using windows before the outcome during evaluation, the model prediction is only produced at the end of a 4-hour window. Hence, lead times can be arbitrarily close to 0 hours when an outcome occurs at the beginning of an omitted window, but data at the time of the outcome is never be used for making the prediction.

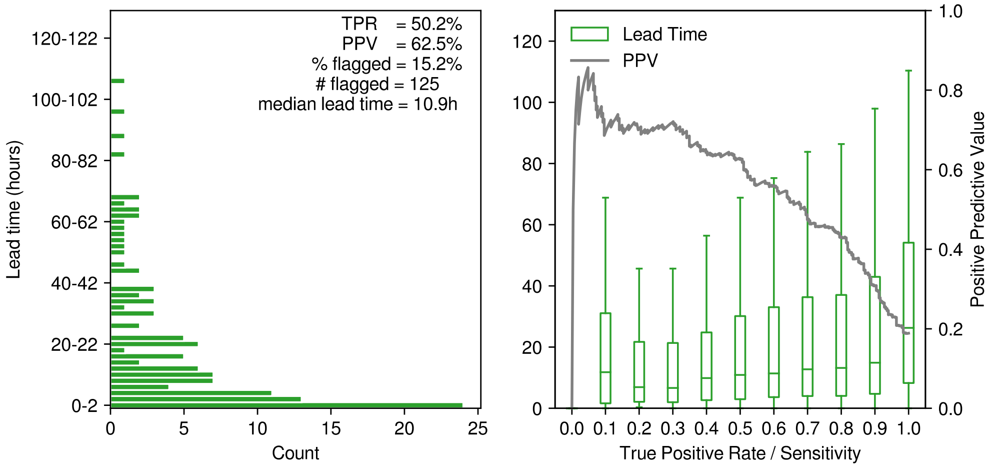


Cohort: MM

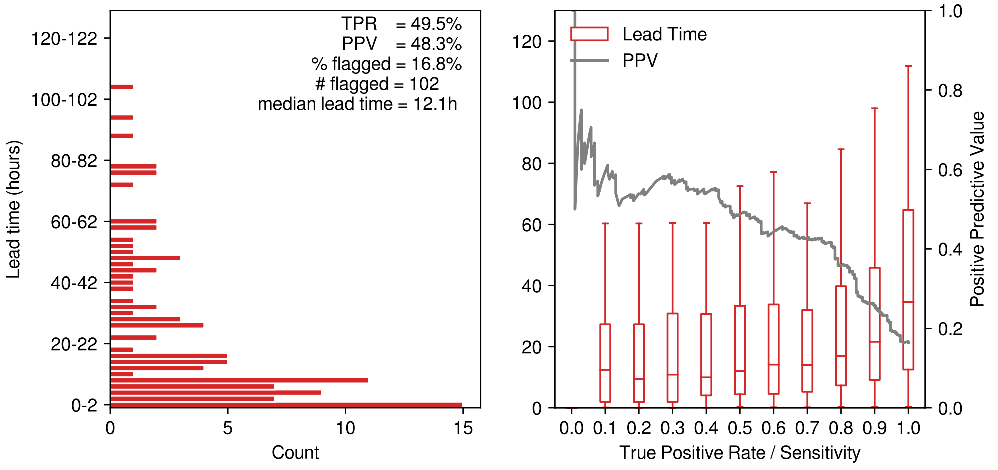


Cohort: A

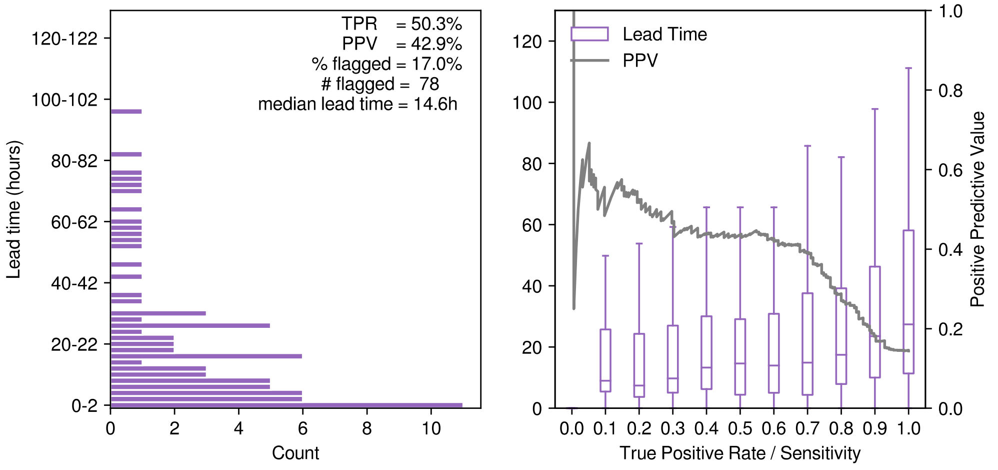
Cohort: B



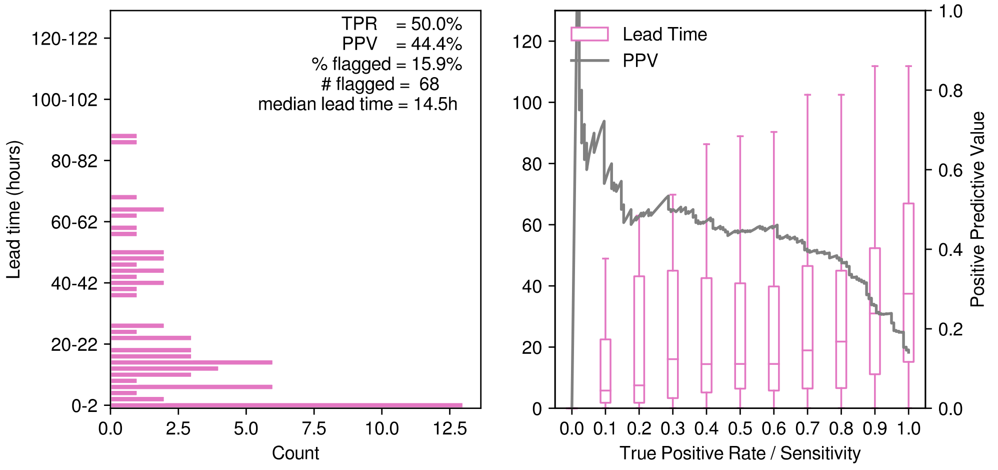
Cohort: C



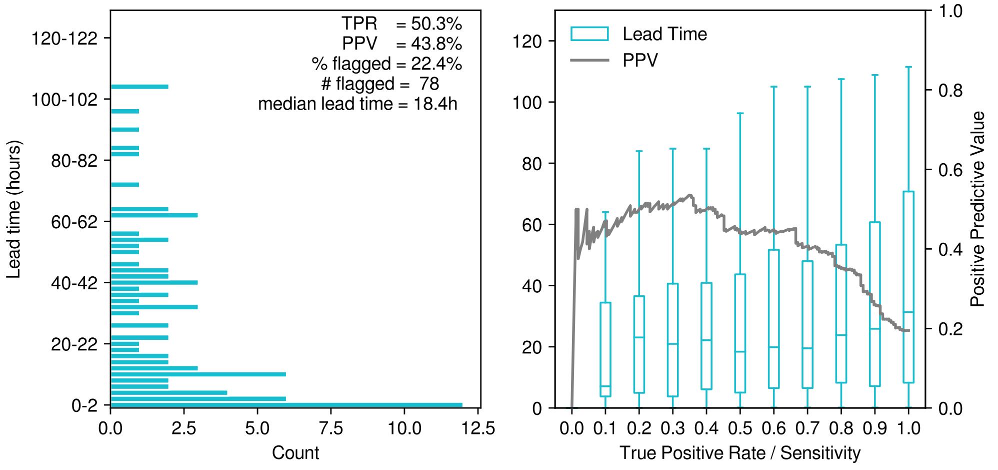
Cohort: D



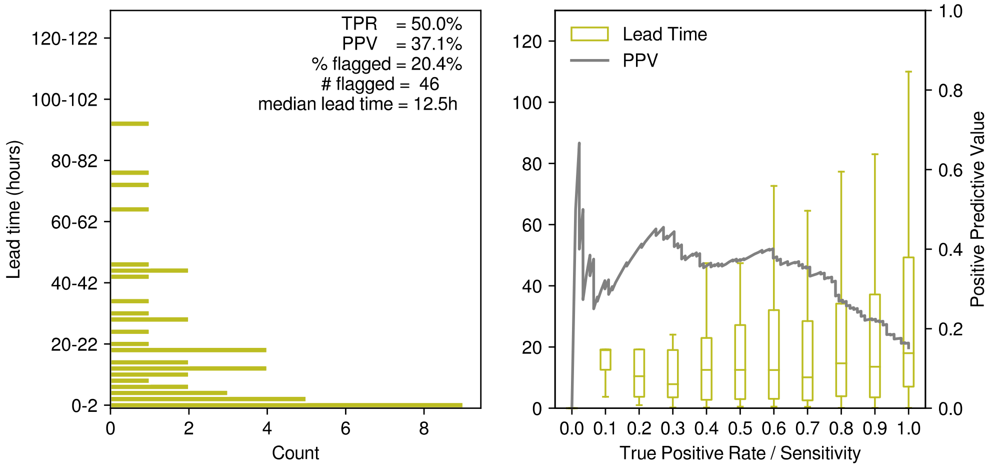
Cohort: E



Cohort: F



Cohort: G


TPR = 50.0%
PPV = 37.1%
% flagged = 20.4%
# flagged = 46
median lead time = 12.5h
Lead time (hours)
Count
Lead Time
PPV
True Positive Rate / Sensitivity
Positive Predictive Value