Short-Term ICU Capacity Stress Early Warning Using State-Level Hospital Data

Final Project Report

1. Executive Summary

Hospitals face significant operational risk when adult ICU utilization surges unexpectedly. This project developed an early-warning system that predicts whether California's adult ICU system will enter a high-stress state within the next 1–7 days, using upstream respiratory demand signals and recent ICU dynamics.

Two logistic regression models were trained and compared against a persistence-based heuristic baseline using expanding window time-series validation on the post-Omicron period (June 2022 – 2024). Key results from the test dataset:

Metric	Baseline	Model v1 (Signals)	Model v2 (Signals + Momentum + Smoothed Signals)
Recall	75.9%	94.8%	100%
Precision	75.9%	78.6%	76.3%
F1 Score	75.9%	85.9%	86.6%
AUC-ROC	0.852	0.945	0.966
Avg Lead Time	3.5 days	4.5 days	4.0 days

Both models substantially outperform the baseline across all metrics, with Model v1 providing the strongest early-warning lead time and Model v2 achieving perfect recall at the cost of slightly more false alarms.

Recommendations (see Section 9 for details):

1. Deploy Model v1 as the primary early-warning signal.
2. Calibrate the alert probability threshold to institutional risk tolerance.
3. Validate performance across additional surge seasons before full operational deployment.

2. Problem Context

Why ICU Stress Forecasting Matters

When adult ICU utilization spikes unexpectedly, hospitals face a cascade of operational failures: delayed admissions, deferred elective procedures, staff burnout, and degraded patient outcomes. Current surge response is predominantly reactive; administrators learn about capacity strain only after it has already materialized.

An early-warning system that reliably predicts stress 1–7 days in advance enables proactive staffing adjustments, elective procedure rescheduling, and inter-facility coordination before the system becomes overwhelmed.

How Stress Was Defined

High stress is defined as adult ICU bed utilization exceeding the **85th percentile** of its historical distribution. This threshold represents sustained operational strain, which is the point at which the system is consistently running at capacity levels seen only 15% of the time. It captures operationally meaningful stress rather than rare catastrophic overload, and produces sufficient positive examples for reliable model training.

Why Short-Term Forecasting

A 1–7 day forecast horizon aligns with actionable hospital decision timescales. Staffing changes, patient transfers, and surgical scheduling adjustments require days not hours or weeks of advance notice. This window is short enough for the signal to be actionable and long enough for respiratory demand patterns to carry predictive value.

Scope of This Report

The original project proposal described two components: (1) predicting near-term ICU capacity stress, and (2) conducting a counterfactual simulation of proactive capacity interventions triggered by early risk signals to evaluate their impact relative to reactive strategies. This report addresses component (1), that is the development and validation of a short-term stress forecasting model. The intervention simulation component is identified as future work in Section 9.

3. Data and Features

Data Source

The analysis uses the **HHS Protect Public Data Hub** dataset, filtered to **California state-level** daily reporting. This dataset includes adult ICU bed utilization, COVID and influenza admissions, inpatient burden, and related respiratory demand indicators.

Time Period

The analysis is restricted to the **post-Omicron regime** beginning **June 1, 2022**. The early pandemic period introduced extreme volatility, policy shocks, and structural changes in hospital operations that no longer reflect current ICU dynamics. The post-Omicron window provides a more stable and operationally relevant signal environment.

The dataset after filtering contains **697 daily observations** across 8 features. After applying label construction (7-day forward window), rolling feature engineering, and the expanding-window walk-forward split (365-day initial training window), the shared evaluation window contains **319 observations** (261 non-stress, 58 stress).

Features Used

Model v1 — Upstream Respiratory Signals Only:

Feature	Description

Feature	Description
`icu_patients_confirmed_influenza`	ICU patients with confirmed influenza
`total_patients_hospitalized_confirmed_influenza`	Total hospitalized influenza patients
`previous_day_admission_influenza_confirmed`	Prior-day confirmed influenza admissions
`previous_day_admission_adult_covid_confirmed_and_suspected`	Prior-day adult COVID admissions
`inpatient_beds_used_covid`	Inpatient beds occupied by COVID patients

All signal features were transformed using `log1p` to uncover structure in mid and low ranges of count variables.

`total_adult_patients_hospitalized_covid` was dropped due to multicollinearity with `inpatient_beds_used_covid` — the two features tracked nearly identically and reflect the same inpatient state rather than distinct demand signals.

Model v2 — Signals + Momentum + Smoothed Signals:

Extends Model v1 with:

- **7-day rolling means** of all five signal features (backward-looking, `min_periods=7`)
- **ICU utilization daily change** (`adult_icu_bed_utilization.diff()`), captures whether the system is accelerating toward the threshold

Stress Threshold

The 85th percentile of `adult_icu_bed_utilization` computed on the full post-Omicron dataset. Stress prevalence in the labeled dataset: **27.1%** of days preceded a stress event within 7 days.

Train/Test Split

An **expanding-window walk-forward** validation scheme was used. The initial training window is the **first 365 days**. For each subsequent day, the model is retrained on all data up to that point and evaluated on the next observation. This simulates real-world deployment where the model is continuously updated with new data.

4. Modeling Approach

Baseline Heuristic

The baseline uses a **rolling 7-day maximum** of ICU utilization. If the maximum utilization in the past 7 days exceeds the stress threshold, stress is predicted. This is a **persistence-based** heuristic which assumes that if utilization has been high recently, it will remain high. By design, it is reactive. Meaning it can only flag risk after utilization has already been elevated.

Logistic Regression Models

Both models use a `Pipeline` with `StandardScaler` and `LogisticRegression`:

- **Model v1** uses only upstream respiratory demand signals. The hypothesis: respiratory admissions and burden precede ICU utilization spikes, providing advance warning before ICU metrics themselves become elevated.

- **Model v2** extends v1 with 7-day smoothed signals and ICU momentum. The hypothesis: persistent respiratory inflow (smoothed signals) and system acceleration (momentum) improve early surge detection.

Forecast Horizon

The binary label `icu_stress_next_7d` answers: *will adult ICU utilization exceed the stress threshold at any point in the next 1–7 days?* The label is constructed by taking the maximum ICU utilization across days $t+1$ through $t+7$ and comparing it to the threshold.

Leakage Prevention

- Features use only data available on or before day t. Rolling windows are backward-looking (`center=False`).
- The label uses only future data ($t+1$ to $t+7$).
- Walk-forward validation ensures the model never trains on data from its evaluation period.
- Rows without a full 7-day forward window are excluded from labeling.

5. Model Evaluation (Test Dataset)

All metrics below are from the **apples-to-apples comparison** on the shared evaluation window, ensuring identical test indices across all three methods.

Performance Metrics

Metric	Baseline	Model v1	Model v2
Accuracy	91.2%	94.4%	94.4%
Precision	75.9%	78.6%	76.3%
Recall	75.9%	94.8%	100.0%
F1 Score	75.9%	85.9%	86.6%
AUC-ROC	0.852	0.945	0.966

Early-Warning Lead Time

Episode	Baseline	Model v1	Model v2

Episode	Baseline	Model v1	Model v2
Episode 1	0 days	2 days	1 day
Episode 2	7 days	7 days	7 days
Average	**3.5 days**	**4.5 days**	**4.0 days**

Classification Report (Model v2)

Class	Precision	Recall	F1 Score	Support
0 (No stress)	1.00	0.93	0.96	261
1 (Stress)	0.76	1.00	0.87	58
Overall Accuracy			**0.94**	**319**

Interpretation

Recall is the primary metric for an early-warning system. A missed stress event is the most costly error. Model v1 captures 94.8% of stress days; Model v2 captures 100%. The baseline misses approximately 1 in 4 stress days.

Precision is comparable across methods (76–79%). When an alert is raised, it is correct roughly three quarters of the time. The remaining alerts are false alarms (days where stress was predicted but did not materialize) within the 7-day window.

False positives carry operational cost (unnecessary staffing increases, deferred procedures) but are far less costly than false negatives (missed surges). Model v2 produces slightly more false positives than v1, consistent with its higher recall.

False negatives are the critical failure mode. The baseline produces the most missed stress events. Model v1 reduces false negatives substantially, and Model v2 eliminates them entirely in the evaluation window.

Stress prevalence in the full labeled dataset is 27.1%. In the evaluation window, the positive class (stress in next 7 days) constitutes 58 out of 319 observations (18.2%), making accuracy less informative than recall and precision.

6. Results Summary

Baseline vs Model v1 vs Model v2

The **persistence baseline** reacts to recent high utilization but provides no genuine anticipation. It misses roughly 1 in 4 stress periods and sometimes detects stress only on the same day it begins (0-day lead for Episode 1).

Model v1 (upstream signals only) demonstrates that respiratory demand signals carry predictive information about ICU stress. It achieves a major improvement in recall (94.8% vs 75.9%) while slightly improving precision, and provides the longest average early-warning lead time (4.5 days).

Model v2 (signals + momentum + smoothed signals) achieves perfect recall; no stress episodes were missed in the evaluation window. It achieves the highest AUC-ROC (0.966) but trades a small amount of precision for complete coverage. Its probability curve rises more gradually and in a more structured way prior to stress onset, suggesting that smoothed signals and momentum features capture meaningful temporal dynamics.

Key Tradeoffs

Recall vs Precision:

- Model v2 maximizes recall (100%) at the cost of slightly lower precision (76.3%)
- Model v1 provides a better balance of recall (94.8%) and precision (78.6%)
- The baseline underperforms both models on all metrics

Lead Time:

- Model v1 offers the strongest early signal (4.5 days on average)
- Model v2 remains strong but slightly less early (4.0 days)
- The baseline lags (3.5 days), with same-day detection on one episode

Early-Warning Behavior

Both models qualify as early-warning systems because their predicted stress probabilities rise several days before ICU utilization actually crosses the stress threshold. This confirms predictive rather than reactive behavior. The models detect upstream pressure before it manifests in ICU metrics. The baseline, by contrast, can only flag risk after utilization has already been elevated.

7. Limitations

Small Number of Stress Episodes

The post-Omicron evaluation window contains only **two stress episodes**. This creates a structurally fragile evaluation setting where performance estimates, particularly Model v2's perfect recall, may not generalize. Results should be interpreted as directional rather than definitive.

Threshold Sensitivity

The 85th percentile threshold is a single operating point. Different thresholds would change the number and duration of stress episodes, class balance, and model performance. The robustness of these results under alternative thresholds has not been evaluated.

State-Level Aggregation

The analysis uses California state-level data. ICU stress is fundamentally a facility-level or regional phenomenon. State-level aggregation may mask localized surges or dilute facility-specific patterns. Model performance at finer geographic granularity is unknown.

Short Evaluation Window

The expanding window validation begins after 365 training days, leaving approximately 11 months of evaluation data. This is a single seasonal cycle, insufficient to assess performance stability across multiple winter surge seasons or novel pathogen introductions.

Threshold Computed on Full Dataset

The stress threshold was computed on the full post-Omicron dataset rather than restricted to the training window. This is acceptable for iterative exploratory modeling but introduces a minor form of distribution leakage because the threshold incorporates information from future observations. In a strict deployment setting, the threshold should be computed using training data only and then locked.

8. Future Work

The following extensions would strengthen the system's operational readiness and address limitations identified in this analysis:

1. **Intervention simulation.** The original project proposal included a counterfactual evaluation component: comparing outcomes when proactive capacity adjustments are triggered by early risk signals versus reactive baseline strategies. This was not implemented in the current submission, which focuses on the forecasting model. Designing and evaluating intervention policies (e.g., temporary capacity increases triggered at a predicted risk threshold) remains the primary next step.

2. **Model interpretability (SHAP).** Applying SHAP analysis to identify which respiratory demand signals most strongly drive each prediction would validate that the model responds to clinically meaningful drivers rather than noise, and would provide operational stakeholders with interpretable reasoning behind each alert.

3. **Sub-state geographic granularity.** The current analysis uses California state-level data. Extending the model to regional or facility-level data would capture localized surges that state-level aggregation may mask.

4. **Multi-season validation.** The evaluation window covers a single seasonal cycle with two stress episodes. Running the model prospectively through additional winter surge seasons would provide stronger evidence of generalizability.

9. Recommendations

1. Deploy Model v1 as the Primary Early-Warning Signal

Hospital operations teams should adopt Model v1(the respiratory-signals-only logistic regression) as the primary tool for anticipating ICU stress events. In testing, this model correctly flagged 94.8% of stress days while maintaining 78.6% precision, and it provided the longest average advance warning at 4.5 days before stress onset. Because Model v1 relies entirely on upstream respiratory admissions and burden data rather than ICU utilization itself, it generates alerts before the ICU system shows visible strain exactly when early action is most valuable. Model v2 achieved perfect recall in the evaluation window, but with only two stress episodes available for testing, that result may not hold in future seasons. Model v1 offers a more reliable starting point for operational use.

2. Calibrate the Alert Threshold to Institutional Risk Tolerance

Before putting the model into practice, hospital teams should adjust the probability cutoff that determines when an alert is issued. The current system uses a default 0.5 cutoff. Meaning an alert fires whenever the model estimates a >50% chance of stress in the next 7 days. Lowering this cutoff (for example, to 0.3 or 0.4) would catch more true stress events at the cost of additional false alarms, while raising it would reduce unnecessary alerts but risk missing some surges. The right balance depends on each institution's tolerance for missed events versus extra preparedness actions. Because a missed ICU surge is far more costly than an unnecessary staffing adjustment, most hospitals should favor a lower cutoff that prioritizes recall. This calibration step requires no new modeling. It adjusts the operating point of the existing system.

3. Validate Performance Across Additional Seasons Before Full Deployment

The evaluation window in this project covers approximately 11 months and contains only two stress episodes. While the results are promising, they reflect a single seasonal cycle and a very small number of positive events. Before relying on this system for operational decisions, hospital teams should run the model prospectively through at least one additional winter surge season to confirm that its lead-time advantage and recall hold under different conditions. This validation period would also reveal whether the model performs consistently when respiratory pathogen patterns shift (for example, if influenza timing or COVID burden changes year to year). No changes to the model itself are required; only continued monitoring and comparison of predicted alerts against actual ICU outcomes.

All metrics reflect the test-period evaluation using expanding-window walk-forward validation on California state-level HHS Protect data, post-Omicron regime (June 2022 – 2024).