**Summary**

Accurately estimating infectious disease hospitalizations at small geographic scales remains a significant challenge. Most forecasting efforts to date have focused on national or state-level predictions, leaving a gap in localized decision support. Yet, county- and region-specific estimates are critical for effective resource allocation and targeted public health interventions. In this study, we developed a flexible, data-driven framework to distribute statewide projections of respiratory-related hospitalizations—including those from influenza, COVID-19, and RSV—across counties and regions in South Carolina.

We integrated hospitalization data from the South Carolina Revenue and Fiscal Affairs (RFA) office with electronic health record (EHR) encounter data from two major health systems in South Carolina—MUSC and Prisma Health. Separate models were developed for COVID-19, influenza, and RSV. Counties with fewer than 500 total encounters in 2023 were excluded from the COVID-19 and influenza models, and counties with fewer than 100 encounters were excluded from the RSV model. To avoid overcounting, repeat admissions occurring within 14 days were removed. Weekly hospitalization counts were smoothed using a peak-preserving moving average algorithm. Gradient-boosted tree models (XGBoost) were trained on observed hospitalizations from 2020 to 2023, using predictors such as geographic area, epidemiological week, year, and total EHR encounters. Model hyperparameters were optimized using Bayesian Optimization. The final models were applied to distribute CDC-projected statewide hospitalizations for 2024–2025 to counties and regions based on predicted risk. Model performance was assessed using Percent Agreement (PA) between smoothed observed and distributed hospitalization counts.

Across all three diseases, region-level models consistently produced higher agreement with observed hospitalizations than county-level models. For COVID-19, county-level predictions achieved a median PA of 0.82 (IQR: 0.71–0.91), while region-level predictions performed better with a median PA of 0.93 (IQR: 0.90–0.95). The county-level median PA for influenza was 0.75 (IQR: 0.53–0.90), compared to 0.82 (IQR: 0.70–0.91) at the region level. For RSV, county-level predictions yielded a median PA of 0.78 (IQR: 0.57–1.00), while region-level predictions reached a median PA of 0.89 (IQR: 0.83–0.95).

This study demonstrates the feasibility of distributing statewide projections of respiratory-related hospitalizations to smaller geographic units using a flexible, data-driven framework. By leveraging EHR encounter data and machine learning models, we achieved a strong agreement with observed hospitalization patterns—particularly at the regional level. These methods can support more localized public health planning and resource allocation for current and future respiratory disease outbreaks.

**Methods**

**Data Sources**

This study integrated data from three primary sources to distribute infectious disease hospitalizations at the regional and county level in South Carolina.

1. **Electronic Health Record (EHR) Data**: Weekly counts of infectious disease–related encounters were obtained from the Medical University of South Carolina (MUSC) and Prisma Health. These data were aggregated at the region and county level and represent clinical encounters attributed to COVID-19, influenza, or RSV.
2. **Hospitalization Records**: Administrative hospitalization data were extracted from the South Carolina Revenue and Fiscal Affairs (RFA) office, covering the period from 2021 to 2023. This dataset includes admission dates area identifiers. To mitigate overcounting, repeat admissions occurring within a 14-day window were removed.
3. **Statewide Forecasts**: CDC reported infectious disease hospitalization totals for the years 2024 and 2025 were incorporated to simulate future statewide burden. These forecasts were subsequently distributed to regions and counties using model-derived risk weights.

**Data Preprocessing and Integration**  
Hospital admission dates were converted to epidemiological weeks. Counties were assigned to one of four geographic regions: Upstate, Midlands, Pee Dee, and Lowcountry. Counties with fewer than a defined number of total admissions (500 for COVID-19 and Influenza; 100 for RSV) in 2023 were excluded to reduce sparsity and enhance model performance. Weekly hospitalization counts were then smoothed using a custom peak-preserving moving average algorithm, designed to retain critical epidemiological peaks while reducing random variability.

**Smoothing Algorithm**  
A two-step smoothing approach was applied. First, prominent hospitalization peaks were identified. Second, a windowed moving average was applied between peaks, preserving the integrity of peak values.

**Feature Engineering**  
Key predictive features included:

* Weekly statewide hospitalization projections
* Geographic region
* Week number
* EHR encounter counts

**Model Development**  
A gradient-boosted decision tree model (XGBoost) was trained to predict weekly regional hospitalization counts using the above features. The data were divided as follows:

* **Training set**: 2020–2022
* **Validation set**: January to June 2023
* **Test set**: July to December 2023
* **Forecasting set**: January 2024 to April 2025

**Hyperparameter Optimization**  
Model hyperparameters were tuned using Bayesian Optimization. The following parameters were optimized:

* Learning rate (eta)
* Maximum tree depth (max\_depth)
* Subsampling ratio (subsample)
* Column subsampling ratio (colsample\_bytree)
* Number of boosting rounds (nrounds)

The objective function minimized the root mean squared error (RMSE) on the validation set. Early stopping was used to avoid overfitting.

**Prediction and Evaluation**  
The final model was used to generate weekly region-level risk scores, which were converted into proportional weights for distributing statewide projections. Distributed predictions were then smoothed using the peak-preserving method described earlier.

Model performance was evaluated using Percent Agreement (PA).

**Results**  
The model demonstrated strong agreement between observed and predicted hospitalizations across all diseases. Region-level performance consistently exceeded that of county-level models:

* **COVID-19**: Region-level PA = 0.93 (IQR: 0.90–0.95), County-level PA = 0.82 (IQR: 0.71–0.91)
* **Influenza**: Region-level PA = 0.82 (IQR: 0.70–0.91), County-level PA = 0.75 (IQR: 0.53–0.90)
* **RSV**: Region-level PA = 0.89 (IQR: 0.83–0.95), County-level PA = 0.78 (IQR: 0.57–1.00)

These findings suggest that region-level models offer greater robustness and stability, likely due to reduced data sparsity and higher encounter volumes.











