Original Research Article

MDM Medical Decision Making

Medical Decision Making
1–15

DOI: 10.1177/0272989X251408845
journals.sagepub.com/home/mdm
Sage

Forecasting Local Surges in COVID-19 Hospitalizations through Adaptive Decision Tree Classifiers

Rachel E Murray-Watson, Xavier Guaracha, Alyssa Bilinski, and Reza Yaesoubi

Introduction. During the COVID-19 pandemic, many communities across the United States experienced surges in hospitalizations, which strained the local hospital capacity. Some risk metrics, such as the Center for Disease Control and Prevention's (CDC's) Community Levels, were developed to predict the impact of COVID-19 on the community-level health care system based on routine surveillance data. However, they had limited utility as they were not routinely updated based on accumulating data and were not directly linked to specific outcomes, such as surges in COVID-19 hospitalizations beyond local capacities. **Methods.** In this article, we evaluated decision tree classifiers developed in real time to predict surges in local hospitalizations due to COVID-19 between July 2020 and November 2022. These classifiers would have provided visually intuitive and interpretable decision rules and, by being updated weekly, would have responded to changes in the epidemic. We compared the performance of these classifiers with that of logistic regression and neural network models using various metrics, including the area under the receiver-operating characteristic curve (auROC) and the area under the precision-recall curve (auPRC). **Results.** Decision tree classifiers achieved an auROC of >80% for most pandemic weeks and outperformed the CDC's Community Levels in predicting high hospital occupancy. The auPRC, sensitivity, and specificity of the classifiers varied more substantially over time (between 20% and 100%) and in sync with pandemic waves. Decision tree classifiers demonstrated similar performance compared with logistic regression and neural network models while presenting more interpretable classification rules. **Conclusions.** Using routinely collected hospital surveillance data, decision tree classifiers can be adaptively updated to predict surges in local hospitalizations. However, the sensitivity and specificity of these classifiers could change markedly during different pandemic waves.

Highlights

- A major concern during the COVID-19 pandemic was the risk of exceeding local health care capacity due to COVID-19-related hospitalizations.
- To assess this risk and inform mitigating strategies, several risk assessment tools were developed during the pandemic. Many of these tools, however, did not predict local outcomes, were not updated as the pandemic progressed, and/or were not interpretable by decision makers.
- We propose an adaptive framework of decision tree classifiers to predict whether COVID-19–related hospital occupancy would exceed a given capacity threshold. These classifiers demonstrated reasonable and stable prediction performance over time. However, their sensitivity and specificity may change substantially over the course of pandemic waves.

Corresponding Author
Rachel E Murray-Watson, School of Public Health, Imperial College London, 90 Wood Lane, W12 0BZ; (rachel.murray-watson16@imperial.ac.uk)

Keywords
COVID-19, machine learning, decision tree, neural networks, logistic regression

Date received: November 19, 2024; accepted: October 27, 2025

When COVID-19 emerged in the United States in 2020, it proved an immediate threat to health care systems. Rapid surges in hospitalizations threatened to overwhelm the health care capacity and compromise the standard of care and patient health outcomes.[1] Even after vaccinations were introduced in late 2020, many communities remained at risk of intermittent surges in COVID-19 hospitalizations due to low rates of vaccination, waning infection- and vaccine-induced immunity, and seasonal changes in transmission dynamics.[2,3]

To assess the risk of surges in hospitalizations, significant efforts were made to predict the trajectory of the pandemic,[4] including the COVID-19 Forecast Hub,[5] the Scenario Modeling Hub,[6] and the IHME COVID-19 Forecast Model.[7] However, most of these predictions are made available at the state or national levels,[4] which reduces their utility in more local settings, where the pandemic trajectory may bear little resemblance to that of the nation or state. Hence, there was a need among local policy makers for tools that would convert the data collected by surveillance systems into meaningful predictions for their area.

One attempt at such a tool was the Center for Disease Control and Prevention's (CDC's) COVID-19 Community Levels[8] (replaced with the COVID-19 Hospital Admission Rate[9] and the COVID-19 County Check[10] in 2023), which was designed to indicate when there may be an upcoming strain on health care systems. The COVID-19 Community Levels were based on the number of new weekly COVID-19 cases, weekly hospital admissions due to COVID-19, and weekly inpatient beds occupied by COVID-19 patients and predicted whether the Community Level would be low, medium, or high. While these metrics were chosen for their correlation with future high mortality and hospital occupancy,[8] they did not directly predict any specific outcome of interest.[11] In addition, the thresholds for the Community Levels were chosen once and not updated as the pandemic progressed or new coronavirus strains became dominant. This limited the overall utility of the Community Levels tool in the long term. Previous work has attempted to address these issues by using regression models that are continuously updated to predict concrete outcomes (e.g., mortality level).[12] While logistic regression models are relatively interpretable, they do not provide the reasoning to understand why the model predicted a certain outcome (e.g., hospital capacity to be exceeded) for a given set of feature values.

In this article, we investigate whether accurate decision tree classifiers could be developed to predict local surges in COVID-19 hospitalizations. Decision tree classifiers are machine learning models that provide simple and interpretable classification rules to make predictions.[13,14] They resemble the threshold-based, flowchart-like structure of the CDC Community Levels, which makes them easy to use and interpret by local policy makers. Using decision tree classifiers, however, allows us to explicitly link the surveillance data to the outcome of interest, which here is to predict whether local hospital occupancy due to COVID-19 patients would exceed a set threshold. To develop these decision tree classifiers, we use data collected from US county surveillance systems and evaluate their performance during different waves of the pandemic, spanning from July 2020 to November 2022. Before this period, data were not routinely collected from each county in the United States. We also compare the performance of these decision tree classifiers with that of CDC's Community Levels and predictions provided by logistic regression and neural network models.

Method

Overview

From July 2020, several COVID-19 indicators (e.g., deaths, cases, hospital admissions) were collected

School of Public Health, Imperial College London (REM-W); Philip R. Lee Institute for Health Policy Studies, University of California San Francisco, San Francisco, CA, USA (XG, RY); Department of Health Services, Policy and Practice, Brown University School of Public Health, Providence, RI, USA (AB); and Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA (RY). The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under awards R21AI173746 and R01AI153351 to RY. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

through surveillance systems to monitor and predict trends in the pandemic.[15,16] Our goal is to use these indicators and metrics as features in decision tree classifiers to predict whether local hospital capacity is expected to exceed in the short term due to surges in COVID-19 hospitalizations.

We consider 4 groups of classifiers that differ based on the predictors (features) they use to predict whether local hospital capacity would be surpassed in 3 wk from the current week. These models are described in detail below. To develop and evaluate these models, we used data collected between July 15, 2020, and November 7, 2022, a 123-wk period. Before July 2020, data were not routinely collected and reported. In December 2022, there was a change in hospitalization reporting guidelines, and data were reported to the CDC's National Health Safety Network rather than to the Department of Health and Human Services. After this period, there were changes in the quality of the data being reported. Hence, we focused our analysis over the period July 15, 2020, to November 7, 2022.

For a given week t, we train decision tree classifiers using data collected between week 1 through week $t-1$. We then use the data collected in week t to predict the outcome in week $t+3$. For example, at the beginning of week $t=10$, we use the data collected through week 9 to develop models to predict whether hospital capacity would be exceeded in week 13. To evaluate how the performance of these classifiers changes throughout the pandemic (especially during the phases in which novel variants emerge), we repeat this procedure for every week $t \in \{5, 6, 7..., 120\}$.

Data

We obtained COVID-19 hospital admissions, occupancy, and intensive care unit (ICU) occupancy data from the Department of Health and Human Services[16] and the data on cases and deaths from *The New York Times*.[15] Only cases confirmed by a reverse-transcriptase polymerase chain reaction test were included in the definition of "case." We followed the procedures outlined in previous studies for data preprocessing, including the aggregation of weekly observations and the imputation of missing values.[11,12]

To account for patients leaving their county of residence to access health care, we aggregated data by health services areas (HSAs)[17] consistent with the CDC's Community Level calculations. We compiled data at the midpoint of each week. A total of 804 HSAs were included in the analysis; for each of the classifiers, we developed, $<1.5\%$ of the health service areas were omitted from the data set due to missing data.

Outcomes

The outcome of interest was whether the COVID-19–caused hospital occupancy would exceed 15 per 100,000 population in exactly 3 wk. The capacity threshold of 15 per 100,000 population is calculated in prior studies and falls in the middle of the CDC Community Level's "medium" risk assessment for the hospital admissions indicator[18–20] (Supplementary Figure S1). We chose the 3-wk period for consistency with the CDC's Community Levels.[8,12] However, we note that knowledge of hospital capacity in the preceding weeks is also useful to policy makers. Consequently, we considered whether this outcome occurs in any week during the interval $[t+1, t+3]$ as a sensitivity analysis. This time period skips the immediate next week but maintains a relatively short period between the training and target weeks.

In addition to the initial threshold of 15 cases per 100,000 population over a 3-wk period, we performed a sensitivity analysis using alternative threshold values and outcome periods.

Features

The prediction models we considered used all or a subset of the following features:

1. number of COVID-19 cases per 100,000 population in the last week,
2. number of COVID-19 deaths per 100,000 population in the last week,
3. number of COVID-19–related hospital admissions per 100,000 population in the last week,
4. number of hospital beds occupied by COVID-19 patients per 100,000 population in the last week,
5. number of ICU beds occupied by COVID-19 patients per 100,000 population in the last week,
6. portion of hospital beds occupied by COVID-19 patients in the last week,
7. change in each of the aforementioned metrics in the last week, and
8. whether the hospital occupancy by COVID-19 patients exceeded the selected capacity threshold in the last week.

Decision Tree and Benchmark Classifiers

We considered 4 broad categories of decision tree classifiers, which differ in the features they include to predict whether hospital capacity will be exceeded:

1. "naive" classifier that uses only a binary variable of whether the current hospital capacity threshold is exceeded;
2. "CDC optimized" classifier that uses the same features included in the CDC Community Levels (i.e., new weekly COVID-19 cases, hospital admissions, and the percentage of inpatient beds used by COVID-19 patients);
3. "reduced" classifier, which uses features related to hospital admissions, hospital and ICU occupancy, but not cases or deaths (because as of mid-2023, the county-level case and death data were no longer routinely reported[15]); and
4. "full" classifier that uses all features listed in the previous section.

We compared the performance of these classifiers with that of the CDC's Community Levels (Supplementary Figure S1), which were developed using all data collected between March 1, 2021, and January 24, 2022.[8] To measure the performance of Community Levels, we evaluated it over the period from March 3, 2022, to November 20, 2022, during which its use was encouraged by the CDC.

We also note that the Community Levels designate an area as "low," "medium," or "high" risk; hence, to evaluate its ability to predict surges in COVID-19 hospitalizations, we counted all weeks in which the set hospitalization threshold was exceeded as equivalent to a "high" community level and weeks in which it was below as equivalent to "medium/low."

We compared the performance of decision tree classifiers with logistic and neural network models that use the same features as the reduced and full models.

Model Development

For each week t between July 15, 2020, and November 7, 2022, we use the data collected from all HSAs from week 1 to week $t-1$ as a single training set to develop our classification trees. We used 10-fold spatial and temporal cross-validation to optimize the hyperparameters of each classifier for each week (see §S1 in the Supplementary Information for details). To ensure that the resulting decision trees were easy to interpret, we restricted the depth of the trees to less than 5 layers and then determined the optimal depth through hyperparameter tuning. This also ensures that the decision tree classifiers use the most relevant features to inform their predictions.

Since the data from larger populations are expected to be less noisy compared with data from smaller populations, we included instance weights in the model-fitting procedure, based on the HSA population. In addition, to account for class imbalance in the outcome, we trained the model using "balanced" class weights, with higher weights assigned to the minority class in the decision tree classifier's function.[21,22]

Although we used 123 wk of data, due to the 3-wk prediction task and the need for 1 wk of test data and 1 wk for initial training, we could train only 117 decision tree classifiers.

We developed our decision tree classifiers using the scikit-learn package in Python.[23]

Model Evaluation

In line with the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) recommendations,[24] model validation was performed using temporal validation, in which our dataset was partitioned into training and testing subsets based on time. To evaluate the performance of a classifier, we calculated the area under the receiver operating curve (auROC) and the area under the precision-recall curve (auPRC) based on data from all HSAs collected during the projection period. A classifier that provides 100% correct predictions has an auROC of 1, and a classifier that randomly guesses the outcomes has an auROC of 0.5. The auPRC metric is particularly useful when the dataset is imbalanced, which was the case for a considerable number of weeks during the pandemic. To investigate regional variation in the performance of our classifiers, we also present the area under the receiver-operating characteristic curve (auROC) scores for each HSA, which we calculated based on predictions made by each weekly classifier for a given HSA throughout the pandemic weeks.

The auROC and auPRC metrics are agnostic to the selected classification threshold and assumes that true/false positives and true/false negatives are equally desirable outcomes. In reality, a predicted surge in hospital capacity will cause different responses among policy makers, each with an associated cost. Under this assumption, the values of true and false predictions may differ. To account for any differences in preferences that policy makers may have, we additionally calculated the net benefit,[25] a metric that allows differences in the weighting of true and false positives, facilitated by an "exchange" parameter, ω. The net benefit function incorporating true-positive and false-positive rates is given by

$$NB_P(\omega, p) = TP(p) - \omega FP(p), \quad (1)$$

where $TP(p)$ is the true-positive rate and $FP(p)$ is the false-positive rate when the classification threshold p is selected (a classifier with classification threshold p predicts the hospital capacity would be exceeded if the estimated the probability of exceeding capacity is greater than p). A true-positive result means that the model has correctly predicted the surge in hospital capacity; hence, any action undertaken by the policy maker to avoid such a surge was justified and could have prevented loss in population health. In contrast, a false-positive result could waste resources if it leads to using mitigating strategies. Equation 1 allows policy makers to weigh the economic costs of unnecessary action ($\omega FP(p)$) with the health benefits of justified action ($TP(p)$). These outcomes, however, depend on the classification threshold p. To find the optimal classification threshold for a given ω, we solve

$$NB_P^*(\omega) = \max_p [TP(p) - \omega FP(p)],$$

using a grid search over values of p.

The above definition of net benefit does not consider the health and economic consequences of true and false negatives. True negatives mean that policy makers can avoid implementing costly interventions, and false negatives may risk overwhelming hospital capacity. To account for these factors, we also evaluate our classifiers using the following version of net benefit:

$$NB_{P,N}(\omega,p) = [TP(p) - FN(p)] - \omega[FP(p) - TN(p)]. \tag{2}$$

For a given ω, we use a grid search to find the optimal classification threshold that maximizes the above function.

To use these net benefit functions, we need to select the value of the penalty parameter ω. In practice, ω is determined according to the decision maker's preference over the consequences of true or false positives and true or false negatives. For example, in the net benefit function $NB_P(w,p)$, the penalty value ω could be set to 10 if the consequence of a false positive is 10 times more important than the consequence of a true positive for the decision maker. In the analyses presented below, we vary ω over the range 0 to 10.

Finally, we used Shapley Additive exPlanations (SHAP) values to evaluate the contribution of each feature to predictions.[26] SHAP values provide a way to allocate the contribution of each feature to a model's prediction by considering all possible combinations of feature values across instances.[27] It provides insight into changes in model performance that may occur when some features are excluded. We use the shap package in Python to calculate the SHAP values and present the SHAP summary plot.[27] In this plot, we present the SHAP values for each prediction period across all 117 reduced classifiers.

Results

Between July 15, 2020, and November 7, 2022, there was a substantial variation in the burden of COVID-19 across HSAs (Figure 1). Throughout this 123-wk period, on average, 78% of HSA weeks surpassed the 15 COVID-19 patients per 100,000 hospitalized threshold (dashed black curve in Figure 1C). COVID-19 cases, new hospital admissions, hospital beds, ICU beds occupied by COVID-19 patients, and the percentage of hospital beds occupied by COVID-19 patients were positively correlated with surpassing the hospitalization threshold (Table 1).

Across all decision tree classifiers, the auROC was always greater than 0.5, including for the naive classifiers, which used just 1 feature (Figure 2A). The full decision tree classifiers generally had the highest auROC, although the reduced decision tree classifiers were competitive despite not using features related to cases or deaths. The CDC optimized, reduced, and full decision tree classifiers had an auROC of >0.8 across all weeks (Figure 2A).

The full and reduced decision tree, logistic regression, and neural network models demonstrated comparative performance in terms of auROC and auPRC metrics (Figure 2). The auROC of full and reduced classifiers was relatively consistent across different waves of the pandemic (Figure 2, first column). However, the auPRC changed more substantially during pandemic waves. The highest auPRC scores were achieved when a high proportion of HSAs exceeded the 15 per 100,000 hospitalization threshold (Figure 2, second column). Across all models, around week 48, there was a decrease in the auPRC score, which coincided with when the omicron strain began circulating in the United States. There was additionally a decrease in auPRC before the peak during the omicron wave (around week 90) (Figure 2, second column).

For all classification methods (decision tree, logistic regression, and neural networks), the reduced models performed similarly to the full models based on auROC and auPRC (Figure 2). However, in contrast to full models, reduced models did not use features related to cases or deaths, making them more parsimonious. As such, for the remainder of our analysis, we focus on reduced classifiers. The sensitivity, specificity, and accuracy of reduced

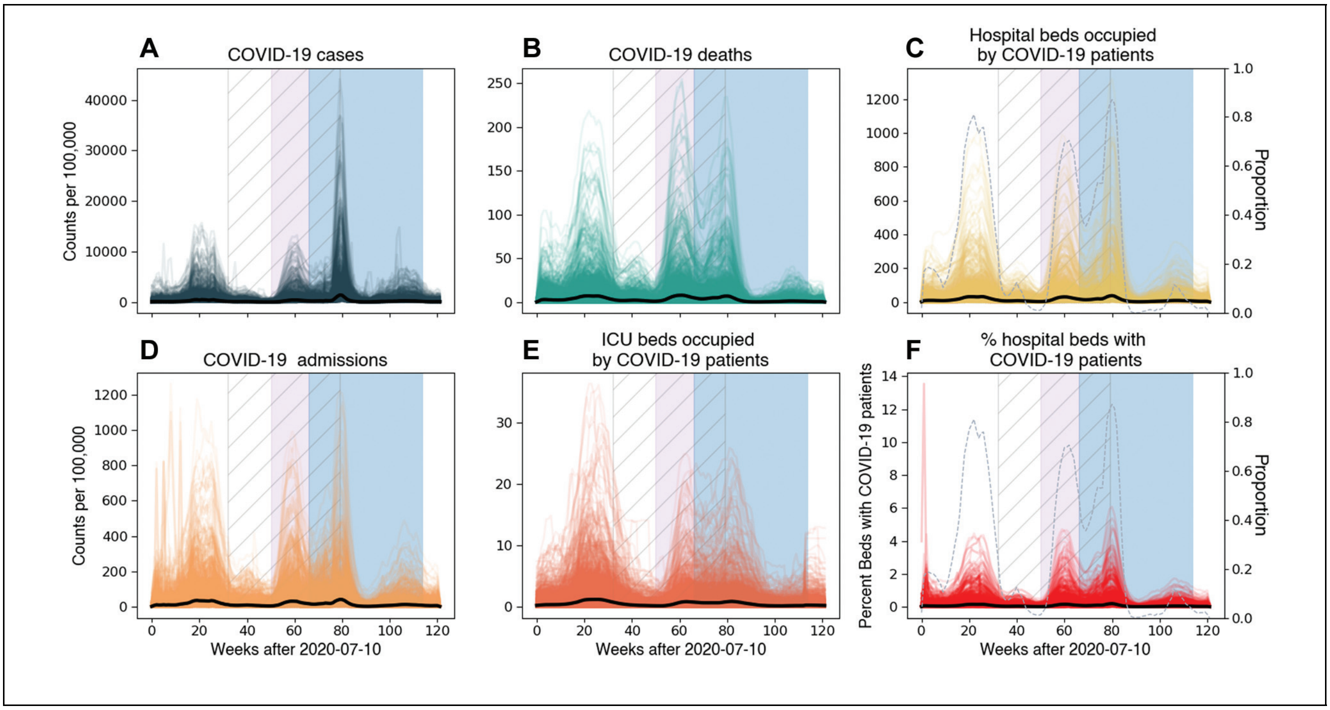

Figure 1 Weekly COVID-19 indicators between July 15, 2020, and November 7, 2022, reported by health services areas (HSAs). The purple and blue boxes show the period when the delta and omicron variants were dominant, respectively. The hatching shows the data used to develop the Centers for Disease Control and Prevention's Community Levels. The black curve shows the mean across all HSAs for the given indicator. The dashed curves in panels C and F represent the proportion of HSAs with hospital occupancy greater than 15 beds per 100,000 population. (A) COVID-19 cases. (B) COVID-19 deaths. (C) Hospital beds occupied by COVID-19 patients. (D) COVID-19 admissions. (E) Intensive care unit beds occupied by COVID-19 patients. (F) Percentage of hospital beds with COVID-19 patients.

Table 1 Mean and Standard Deviation of Weekly COVID-19 Observations Stratified by Whether the Hospital Occupancy Is Surpassed in 3 wks' Time or Not

	Surpassed		Not Surpassed	
Feature	**Mean**	**SD**	**Mean**	**SD**
Cases	143.6	290.4	53.4	122.9
Deaths	1.8	3.2	1.4	3.0
Hospital beds	37.9	87.3	1.4	4.3
Admissions	9.4	21.6	2.3	9.1
ICU beds	1.75	3.5	0.08	0.44
Percent beds occupied by COVID-19 patients	4.3	9.4	0.023	0.073
Change in cases	−1.4	169.0	2.5	88.1
Change in deaths	0.059	1.1	-0.16	1.3
Change in hospital beds	−0.098	41.6	0.016	4.3
Change in admissions	−0.19	14.7	0.36	9.1
Change in ICU beds	−0.0057	1.64	0.0076	0.38
Change in percentage beds occupied by COVID-19 patients	−0.00022	0.071	0.000024	0.0064

ICU, intensive care unit; SD, standard deviation.

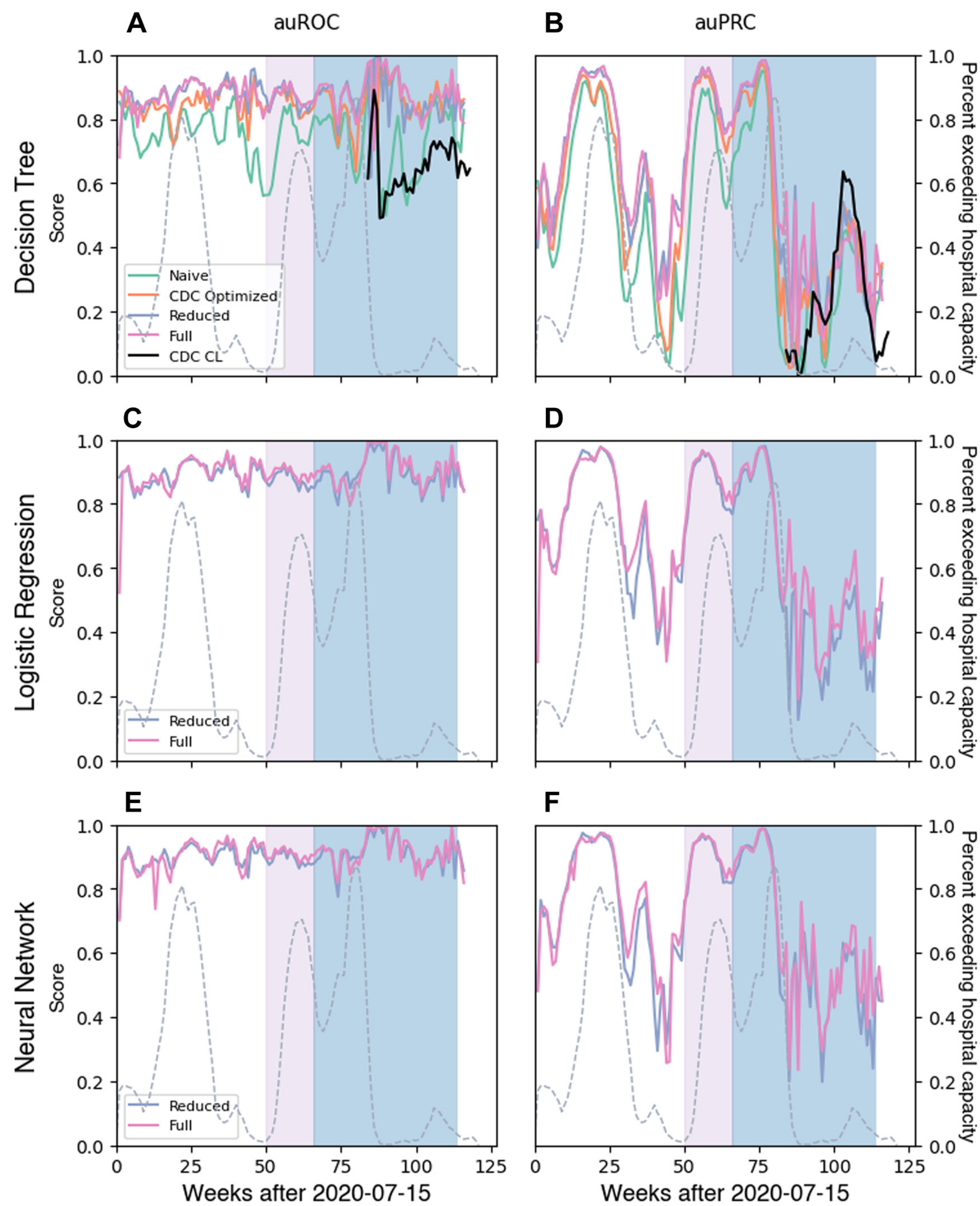

Figure 2 Performance of decision tree, logistic regression, and neural network classifiers to predict whether the COVID-19 hospital occupancy is expected to exceed 15 per 100,000 in exactly 3 wk. The purple and blue boxes show the period when, respectively, the delta and omicron variants were dominant. The gray dashed line shows the proportion of health services areas that exceed the hospitalization threshold of 15 per 100,000 population for a given week. (A) and (B) are the decision tree, (C) and (D) the logistic regression, and (E) and (F) the neural network. (A), (C), and (E) are the auROC and (B), (D), and (F) the auPRC.

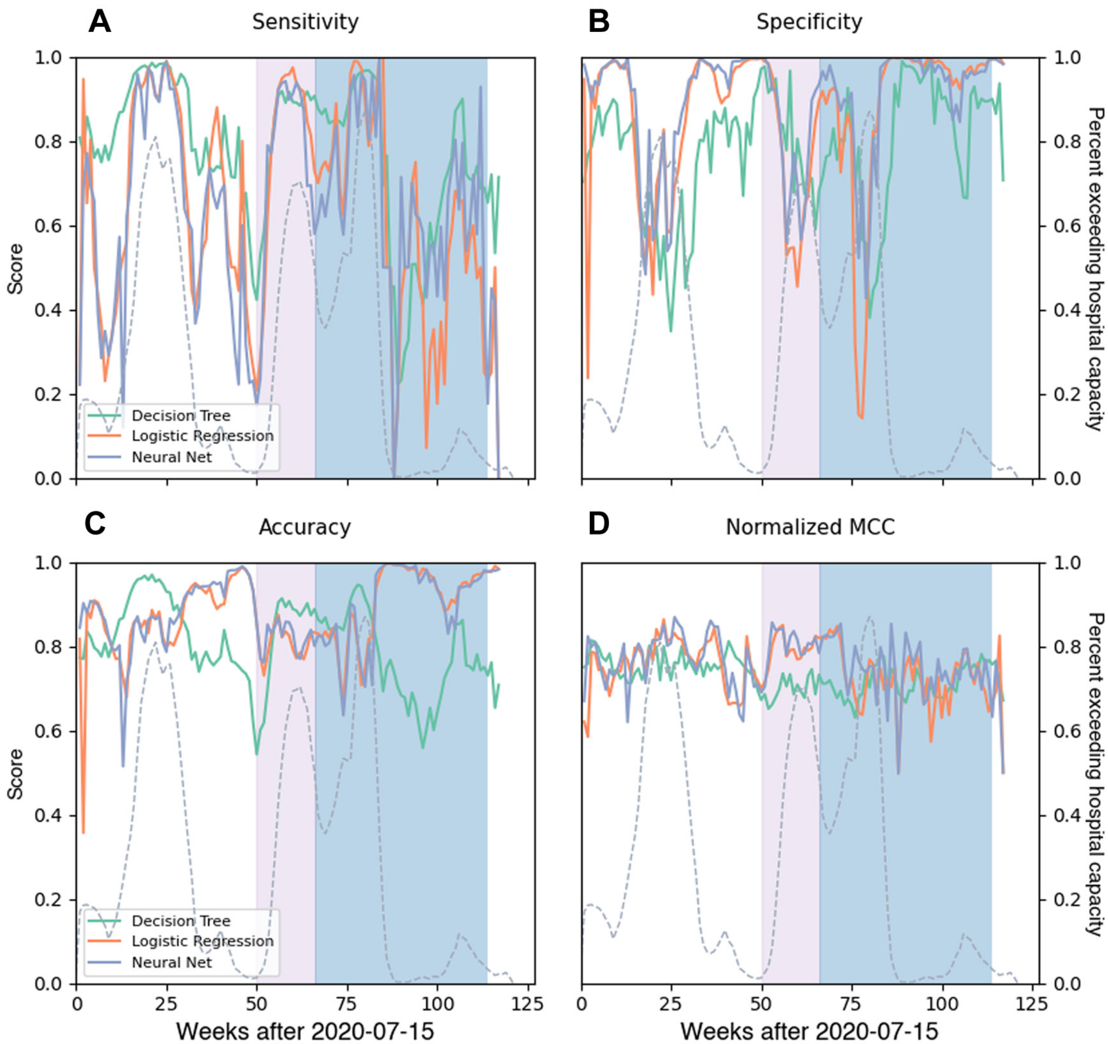

Figure 3 Sensitivity, specificity, accuracy, and the normalized Matthews correlation coefficient (MCC) for the reduced classification models. The purple and blue boxes show the period when, respectively, the delta and omicron variants were dominant. The gray dashed line shows the proportion of health services areas that exceed the hospitalization threshold of 15 per 100,000 population for a given week. (A) Sensitivity. (B) Specificity. (C) Accuracy. (D) Normalized MCC.

decision tree, logistic regression, and neural network models varied over pandemic weeks (Figure 3A–C). The sensitivity of all models coincided with the proportion of HSAs with exceeded hospital capacity, and it dropped substantially during weeks when less than 50% of HSAs exceeded the hospital capacity (Figure 3A). All models demonstrated similar and stable Matthews correlation coefficient over pandemic weeks (Figure 3D).

The accuracy of predictions varied across scenarios in which the continuation of a surge or nonsurge state is predicted as opposed to scenarios in which a change in the current state is predicted (Table 2). All classifiers achieved high accuracy (>85%) in predicting the continuation of a surge (Table 2, last row) and in predicting the continuation of a nonsurge state (first row). However, they performed worse in predicting changes in the surge status (second and third rows).

Based on the performance metrics we considered above, no single method demonstrated a clear advantage over the others. Hence, in the following, we focus on

Table 2 Accuracy of Prediction by the States in the Current Week and in 3 wk

Current State	State in 3 wk	CDC CL	DT Reduced	DT Full	LR Reduced	LR Full	NN Reduced	NN Full
Under	Under	82%	91%	90%	93%	93%	95%	96%
	Over	79%	75%	77%	71%	75%	66%	63%
Over	Under	44%	69%	71%	71%	76%	80%	80%
	Over	94%	89%	89%	88%	89%	84%	83%

CDC CL, Centers for Disease Control and Prevention's Community Levels; DT, decision tree classifiers; LR, logistic regression models; NN, neural network models.

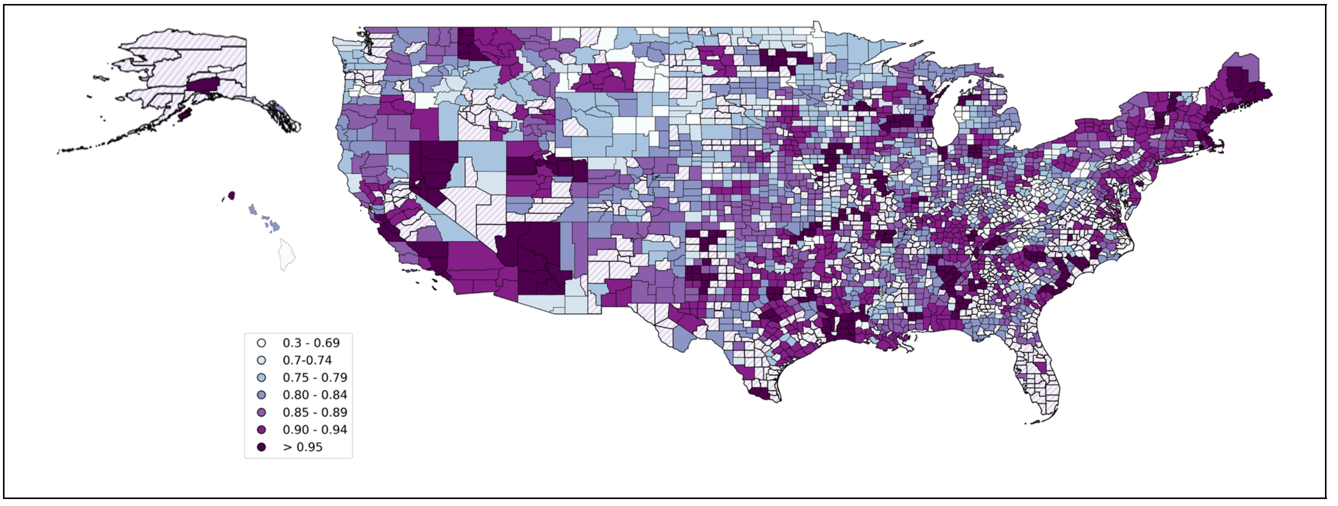

Figure 4 Performance of reduced decision tree classifiers across all counties. The area under the receiver-operating characteristic curve (auROC) was calculated by health services area using the predictions from all 117 reduced classifiers. The hatching indicates where there were no true-negative instances with which to calculate the auROC, and the auROC is recorded as "NA." See Supplementary Figure S2 for the spatial performance of classifiers that includes case and death data.

decision tree models as they are more interpretable and easier to use in practice.

For 76% of the counties, the auROC that could be calculated using the reduced decision tree classifiers throughout the study period exceeded 0.80 (Figure 4). However, for about 23% of counties, an auROC could not be calculated as across each 117 outcome weeks, the hospital beds occupied by COVID-19 patients either always or never exceeded the 15 per 100,000 threshold. Thus, there was no "true negative" with which to calculate the auROC. The full decision tree model presented similar spatial performance to the reduced model (Supplementary Figure S2).

The added benefit of the reduced decision tree classifiers compared with the naive decision tree classifiers varied over the pandemic weeks and depended on the selected penalty value (ω in equation 1) to model the tradeoff between the true-positive and false-positive rates (Figure 5A). For smaller values of ω, which represent scenarios in which the false positive is minimally penalized, the added benefit of using the reduced decision tree classifiers is unimportant. Reduced classifiers provide greater benefit when false positives are penalized more substantially and when the proportion of HSAs in which hospital capacity exceeds 15 per 100,000 is large (Figure 5A). However, the naive decision tree classifiers provide additional net benefit over the reduced decision tree classifiers when the proportion of HSAs exceeding the hospital occupancy threshold was low.

We observed similar behavior when the net benefit function $NB_{P,N}()$, which accounts for true- and false-positive rates and true- and false-negative rates, is used (Figure 5B). When this net benefit function is used, the reduced decision tree classifiers outperformed the naive decision tree classifiers for a larger number of weeks and penalty values.

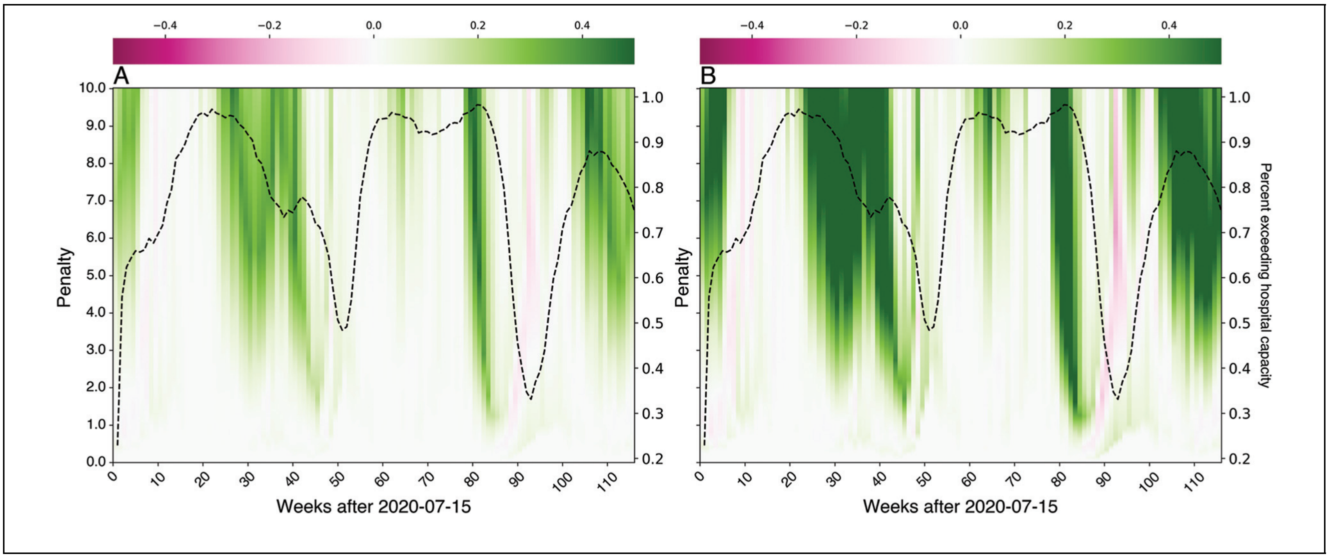

Figure 5 The net benefit of the reduced decision tree classifiers related to the naive classifiers. (A) Using the net benefit function $NB_P()$, which accounts for false-positive and true-positive rates and (B) using the net benefit function $NB_{P,N}()$, which accounts for true- and false-positive rates and true- and false-negative rates. In areas shaded green, the reduced decision tree classifiers outperform the naive classifier, while areas shaded pink indicate where the naive classifier performs better. The gray dashed line is the proportion of health services areas that exceed the 15 per 100,000 hospital capacity for a given week.

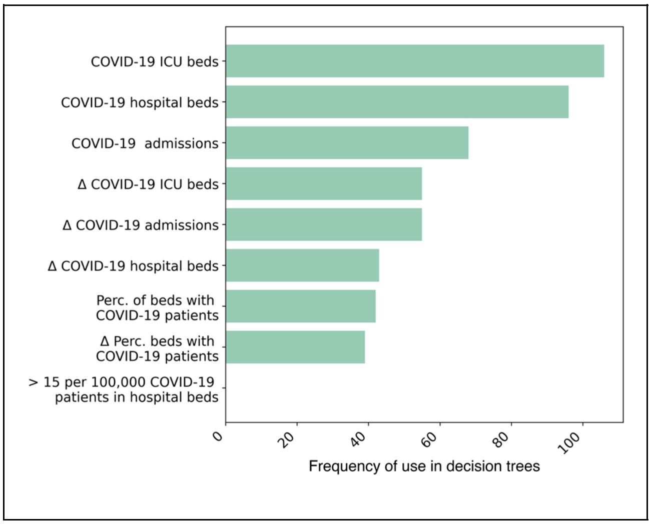

Figure 6 Frequency at which a feature was identified as important in 117 reduced decision tree classifiers created between July 15, 2020, and November 7, 2022. ICU, intensive care unit.

The change in the number of COVID-19 hospital admissions in the previous week, the number of beds occupied by COVID-19 patients, and the number of COVID-19 hospital admissions were selected as important features in more than 50% of 117 reduced decision tree classifiers between July 15, 2020, and November 7, 2022 (Figure 6). Whether the current hospital capacity exceeded the 15 per 100,000 threshold was not included as an important feature in any classifier, despite the good performance of the naive decision tree classifier, which used only this feature. However, this binary feature is highly correlated with other features such as the number of hospital beds by week and the percentage of beds occupied by COVID-19 patients (Table 1), which do appear as important features (Figure 6).

The SHAP values also indicate that COVID-19 admissions and ICU beds occupied by COVID-19 patients have a large influence on model predictions (Figure 7). The SHAP values broadly follow a trend in which high numbers of admissions and occupied hospital beds (indicated by the pink dots in Figure 7) increase the log odds of a positive prediction, whereas lower admissions decrease the log odds.

One major advantage of and the main motivation for using decision tree classifiers is that they are interpretable. To demonstrate, we present 3 decision tree classifiers developed for 3 different stages of the pandemic (Figure 8): the week of July 14, 2021, when the delta variant was circulating in the population but was not yet the dominant strain (panel A); the week of August 4, 2021, when the delta variant was dominant (panel B); and the final

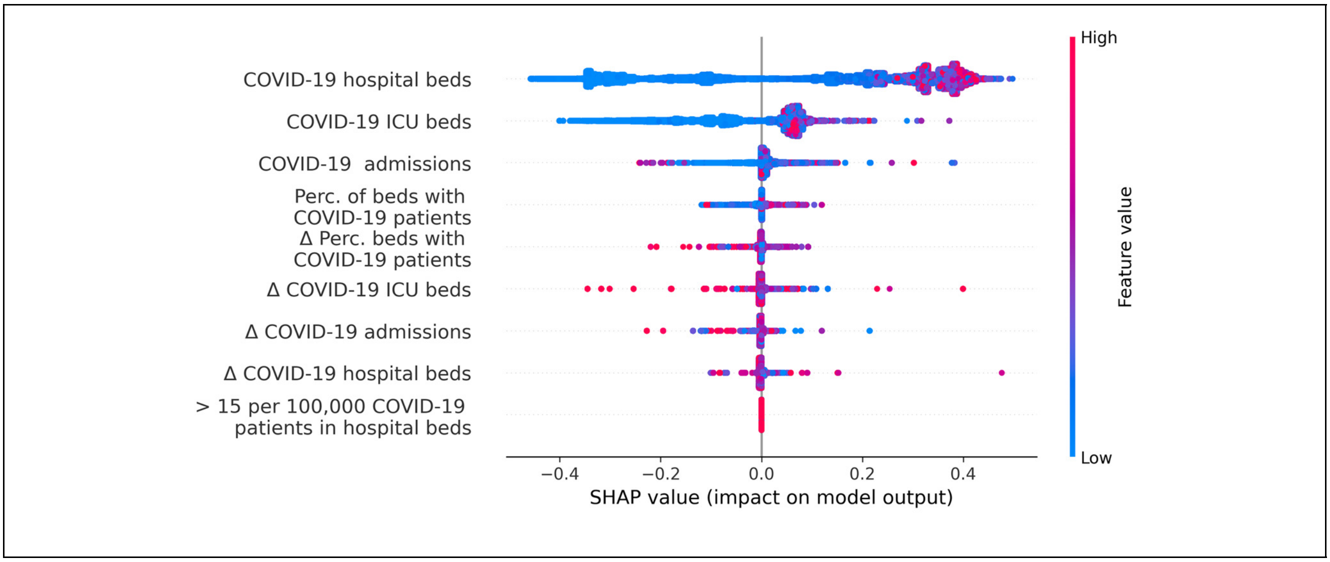

Figure 7 Shapley Additive exPlanations (SHAP) summary plot for the reduced decision tree classifiers. Each point represents the SHAP value for a feature and an instance (observations from a single health services area over a single week). The color of each point represents the value of the feature from low to high. Overlapping points are jittered vertically to display the distribution of SHAP values per feature. ICU, intensive care unit.

week in our study time period, that is, the week of November 20, 2022 (panel C). To illustrate how these decision tree classifiers could be used in practice by a local policy maker, we consider the scenario observed in an HSA in Maryland for July 14, 2021 (Table 3). Since hospital beds occupied by COVID-19 patients = 22.82, which is less than 28.835, the condition of the first decision node is satisfied. Hence, we check the condition of the second node, whether the number of hospital beds occupied by COVID-19 patients is ≤ 10.38, which is satisfied. Therefore, this classification decision tree predicts that the hospital capacity of 15 per 100,000 population is not expected to be exceeded in 3 wks' time.

We note that while the classification trees in Figure 8 are developed for different phases of the pandemic, they all identified the same features as important: 1) the number of hospital beds occupied by COVID-19 patients, 2) the number of ICU beds occupied by COVID-19 patients, and 3) the number of hospital admissions of COVID-19 patients. What is different between these classification trees is the classification thresholds and the depth of the tree; for example, the classification threshold at the first node of these trees is 28.835 (A), 26.346 (B), and 26.636 (C).

Our sensitivity analyses suggested that the decision tree classifiers considered here maintained their performance under various scenarios, representing when 1) the outcome of interest is whether the hospital capacity exceeds 15 per 100,000 over a 3-wk period, instead of at exactly 3 weeks from now (Supplementary Figure S3); 2) the hospital capacity is 10, 15, or 20 per 100,000 population (Supplementary Figure S4); 3) predictions are made over 2-, 3-, 4-, or 6-wk periods (Supplementary Figure S5); 4) the training dataset is limited to the past 4, 10, and 26 wk (Supplementary Figure S6); and 5) models are trained every 4 wk instead of every week (Supplementary Figure S7).

Discussion

We presented an adaptive framework for generating simple classification rules that predict whether COVID-19 hospitalizations would exceed local capacity. These classification rules are easy to communicate and use the latest data from surveillance systems that are available at the local level. To develop these classification rules from July 2020 to November 2022, we trained decision tree models on "expanding" datasets, in which all available data up to the target week of interest were used in the training procedure. This allowed for the model-training procedure to account for changes in pandemic trajectories due to factors not included as features in the model (e.g., vaccination rates, waning immunity, and easing nonpharmacologic interventions). These interpretable decision rules demonstrated similar performance compared with logistic regression and neural network models

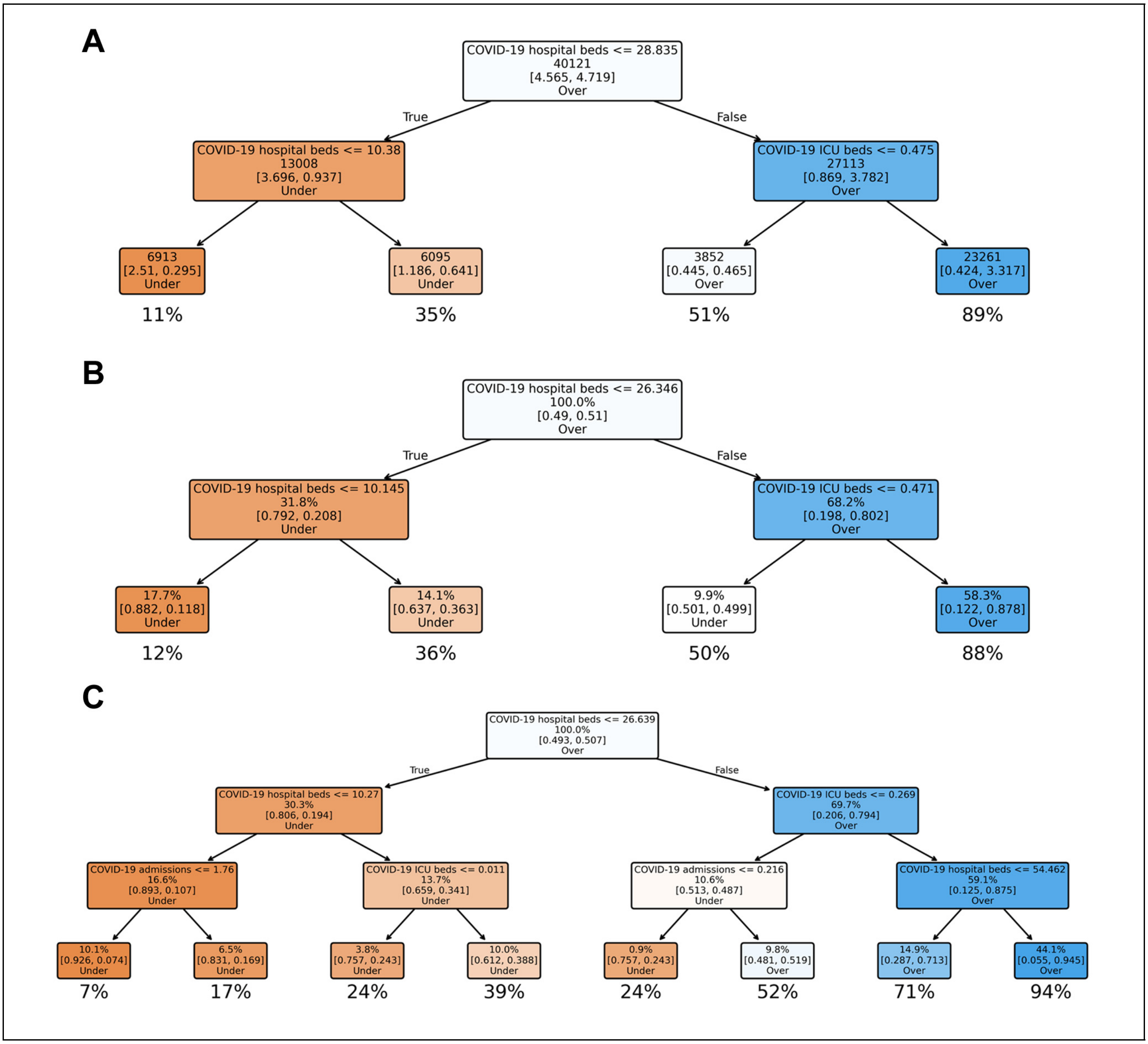

Figure 8 Decision tree classifiers developed at 3 different points during the pandemic. The week of July 14, 2021, when the delta variant was circulating in the population but was not yet the dominant strain (A); the week of August 4, 2021, when the delta variant was dominant (B); and the final week in our study time period, that is, the week of November 20, 2022 (C). The percentages represent the proportion of instances in the final leaves where hospital capacity was exceeded. ICU, intensive care unit.

based on various performance metrics such as auROC, auPRC, sensitivity, specificity, and accuracy (Figures 2 and 3).

Our classifiers were trained on a range of COVID-19 indicators that were routinely reported between July 2020 and November 2022, including hospital admissions and hospital occupancy. However, the source of case and death data used in this study for the "full" classifiers ceased being updated in mid-2023,[15] and other data sources are updated at only the state level[28] if at all.[29,30] Although this limits the training data on which the classifiers can be trained, we have shown that omitting death and case data does not significantly affect the performance of the characterized classification rules (Figure 2).

Table 3 Data for Allegany/Garrett Health Services Area in Maryland on July 14, 2021[a]

Value	
Hospital admissions	1.35
ICU beds	0.52
Hospital beds	22.82
Percent beds occupied by COVID-19 patients	1.5
Change in admissions	−1.81
Change in ICU beds	0.34
Change in hospital beds	5.65
Change in percentage beds occupied by COVID-19 patients	0.38
Currently exceed threshold capacity	Yes
Exceed capacity 3 wk later	Yes

ICU, intensive care unit.
[a]This HSA contains Allegany and Garrett counties in Maryland and Grant, Hardy, and Mineral Counties in West Virginia.

Indeed, the reduced classifiers presented reasonable performance, with an auROC >0.80 for most of the pandemic weeks.

Although all classifiers (decision tree, logistic regression, and neural network) had a stable auROC throughout the pandemic weeks, their auPRC, sensitivity, specificity, and accuracy varied more substantially with pandemic waves (Figure 3). Furthermore, the classifiers we studied demonstrated substantially different accuracy in predicting the continuation of a surge or a nonsurge state or predicting a change in the current state (Table 2). This suggests that different classifiers should be routinely evaluated as the pandemic unfolds, and during different phases of the pandemic, different classification methods might be selected to inform predictions.

We focused on providing predictions at the level of HSAs as this would have enabled local policy makers to detect potential surges in COVID-19 hospitalizations and to respond accordingly before the local hospital capacity is overwhelmed. One such previous attempt was the CDC's Community Levels framework. This framework, however, lacked a direct relationship with a specific outcome of interest[18] and was never updated after its development. Other studies have provided more systematic classification rules predicting specific outcomes (such as hospital capacity[20] or high mortality[12]). We extended this work by demonstrating that decision tree classifiers trained on surveillance data could predict whether COVID-19 hospitalizations may exceed the local capacity level with reasonable accuracy. These classification rules provided by these models are easy to use and interpret in practice (Figure 8).

The spread of SARS-CoV-2 varied substantially across different communities and geographic regions, depending on the implementation of nonpharmaceutical interventions,[31] local rates of vaccination,[32] and the emergence of novel strains, among other factors.[33–35] Despite the substantial heterogeneity in data reported by HSAs, our analysis suggests that our reduced decision tree classifiers maintained their auROC across US counties (Figure 4). We note that the spatial performance of classifiers was not dramatically improved when case and death data were used to develop classification rules (Supplementary Figure S2).

Two of the decision tree classifiers considered here (i.e., "CDC optimized" and "full") use case and death data. The available case data undercounted the actual number of infections, as only cases confirmed by a molecular laboratory test were included.[15] Limited testing availability,[36] particularly at the beginning of the pandemic, and asymptomatic transmission contributed to this undercount.[37] Similarly, death counts relied on the patient to be diagnosed according to guidelines specified by state and federal bodies,[38] which could have led to both the mis- and underdiagnosis of patients. However, our analysis shows that even without these COVID-19 indicators (as in our reduced classifier), reasonable predictive power can still be achieved (Figure 2). In the reduced decision tree classifiers, the admissions and change in admissions and occupied ICU beds were most frequently considered to be important in the decision tree classifiers (Figure 6). Similarly, these features had some of the largest explanatory effects on the auROC (Figure 7).

By using the net benefit framework, our approach allows for incorporating policy makers' preferences between different prediction outcomes (i.e., false negative and positive and true negative and positive). For example, predicting that there will be a surge in hospital occupancy when there will not actually be one (false positive) or predicting no surge when one is going to occur (false negative) have distinct health and economic costs; the former could have high economic costs, and the latter could lead to high adverse health outcomes. Our analysis shows that the classification rules identified by our approach present positive net benefit values, especially during the pandemic phases when the COVID-19 hospitalizations were rising or declining (Figure 5).

An important limitation of predictive models that are trained on only historical data is that historical data may not adequately capture the possible changes in pandemic trajectories due to future events. This is particularly true if novel strains emerge with characteristics that are substantially different from what is manifested in historical data. The use of simulated trajectories, which can be parameterized to allow for such variations, may help to make the predictions of the classifier more robust.[20] Although classification decision trees are easy to visualize and interpret, their structure could change throughout the pandemic when retrained frequently. Decision trees that change significantly over time may be more challenging to implement, as the reasons for the specific changes in the tree (e.g., changes in cutoff points, node ordering, tree depth) can be opaque to policy makers.

The proposed approach is flexible, allowing for several extensions. While we trained our models to predict a binary outcome, classification decision trees could also be developed to predict multiple outcomes, such as "low," "medium," and 'high" levels of COVID-19 hospital capacity, if thresholds to define these outcomes are available. Furthermore, we included only indicators related to the COVID-19 pandemic to predict surges in hospital occupancy. However, there is evidence of cocirculation of influenza, COVID-19, and respiratory syncytial virus infections during recent winters,[39,40] which threatens to overwhelm the health care system. Data from surveillance systems related to these infectious diseases could be incorporated to provide more robust predictions. Finally, while we focused on predicting surges in hospitalizations, other metrics of interest, such as ICU capacity or mortality, could also be considered depending on the availability and the quality of data related to these outcomes.

Ethical Considerations

Not applicable.

Consent to Participate

Not applicable.

Patient Consent

Not applicable.

Consent for Publication

Not applicable.

ORCID iDs

Rachel E Murray-Watson https://orcid.org/0000-0001-9079-5975
Xavier Guaracha https://orcid.org/0009-0000-6041-7834
Alyssa Bilinski https://orcid.org/0000-0001-9108-6660
Reza Yaesoubi https://orcid.org/0000-0002-9276-5750

Data Availability

All code and data used in this research are publicly available at https://github.com/RachelMurray-Watson/COVID_forecasting. The original data supporting this study were obtained from publicly accessible sources. COVID-19 hospital admissions, hospital occupancy, and ICU occupancy data were sourced from the US Department of Health and Human Services (2023). Data on confirmed COVID-19 cases and deaths were obtained from *The New York Times* (2021).
Full citations for these sources are:

- US Department of Health and Human Services. *COVID-19 Reported Patient Impact and Hospital Capacity Statistics*. 2023. Available at: https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u/about_data. [Accessed 7 August, 2023].
- Coronavirus (Covid-19) data in the United States. *The New York Times*. 2021. Available at: https://github.com/nytimes/covid-19-data.

Supplemental Material

Supplementary material for this article is available online at https://doi.org/10.1177/0272989X251408845

References

1. Kadri SS, Sun J, Lawandi A, et al. Association between caseload surge and covid-19 survival in 558 us hospitals, March to August 2020. *Ann Intern Med*. 2021;174(9):1240–51.
2. Murray CJL, Piot P. The potential future of the COVID-19 pandemic: will sars-cov-2 become a recurrent seasonal infection? *JAMA*. 2021;325(13):1249–50.
3. Saad-Roy CM, Wagner CE, Baker RE, et al. Immune life history, vaccination, and the dynamics of sars-cov-2 over the next 5 years. *Science*. 2020;370(6518):811–8.
4. Nixon K, Jindal S, Parker F, et al. An evaluation of prospective COVID-19 modelling studies in the usa: from data to science translation. *Lancet Digit Health*. 2022;4(10):e738–47.
5. Cramer EY, Huang Y, Wang Y, et al. The United States COVID-19 forecast hub dataset. *Sci Data*. 2022;9(1):462.
6. Lemaitre JC, Grantz KH, Kaminsky J, et al. A scenario modeling pipeline for COVID-19 emergency planning. *Sci Rep*. 2021;11(1):7534.
7. IHME COVID-19 Forecasting Team. Modeling COVID-19 scenarios for the United States. *Nat Med*. 2021;27(1):94–105.
8. Centers for Disease Control and Prevention. Science brief: indicators for monitoring COVID-19 community levels and making public health recommendations. In: *CDC COVID-19 Science Briefs*. Rockville (MD): Centers for Disease Control and Prevention; 2022. https://archive.cdc.gov/www_cdc_gov/coronavirus/2019-ncov/science/science-briefs/indicators-monitoring-community-levels.html

9. Scobie HM, Panaggio M, Binder AM, et al. Correlations and timeliness of COVID-19 surveillance data sources and indicators—United States, October 1, 2020–March 22, 2023. *Morb Mortal Wkly Rep*. 2023;72(19):529.
10. Centers for Disease Control and Prevention. COVID-19 county check. 2024. Available from: https://www.cdc.gov/TemplatePackage/contrib/widgets/covidcountycheck/ [Accessed 4 March, 2024].
11. Salomon JA, Bilinski A. Evaluating the performance of Centers for Disease Control and prevention COVID-19 community levels as leading indicators of COVID-19 mortality. *Ann Intern Med*. 2022;175(9):1240–9.
12. Bilinski AM, Salomon JA, Hatfield LA. Adaptive metrics for an evolving pandemic: a dynamic approach to area-level COVID-19 risk designations. *Proc Natl Acad Sci U S A*. 2023;120(32):e2302528120.
13. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2. New York: Springer; 2009.
14. Burkov A. *The Hundred-Page Machine Learning Book*. Vol. 1. Quebec City (QC): Andriy Burkov; 2019.
15. Coronavirus (Covid-19) data in the United States. *The New York Times*. 2021. Available from https://github.com/nytimes/covid-19-data.
16. US Department of Health and Human Services. COVID-19 reported patient impact and hospital capacity statistics. 2023. Available from: https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u/about_data. [Accessed 7 August, 2023].
17. Makuc DM, Haglund B, Ingram DD, Kleinman JC, Feldman JJ. Health service areas for the United States. *Vital Health Stat 2*. 1991;112:1–102.
18. Centers for Disease Control and Prevention. Indicators for monitoring COVID-19 community levels and COVID-19 and implementing COVID-19 prevention strategies. PowerPoint Presentation; 2022.
19. Moghadas SM, Shoukat A, Fitzpatrick MC, et al. Projecting hospital utilization during the COVID-19 outbreaks in the United States. *Proc Natl Acad Sci U S A*. 2020;117(16):9122–6.
20. Yaesoubi R, You S, Xi Q, et al. Generating simple classification rules to predict local surges in COVID-19 hospitalizations. *Health Care Manag Sci*. 2023;26(2):301–12.
21. King G, Zeng L. Logistic regression in rare events data. *Pol Anal*. 2001;9(2):137–63.
22. Scikit-learn Developers. Scikit-learn: machine learning in Python. 2021. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html. [Accessed 18 March, 2024].
23. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Machine Learn Res*. 2011;12:2825–30.
24. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1–73.
25. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6.
26. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable ai for trees. *Nat Mach Intell*. 2020;2(1):56–67.
27. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, et al. eds. *Advances in Neural Information Processing Systems 30*. Red Hook (NY): Curran Associates, Inc.; 2017. p 4765–74.
28. Centers for Disease Control and Prevention. COVID data tracker. Available from: https://covid.cdc.gov/covid-data-tracker [Accessed 2 November, 2023].
29. Johns Hopkins University. Johns Hopkins University and Medicine Coronavirus Resource Center. Available from: https://coronavirus.jhu.edu/ [Accessed 2 November, 2023].
30. USA Facts. US COVID-19 cases and deaths by state. Available from: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/ [Accessed 2 November, 2023].
31. Singh S, Shaikh M, Hauck K, Miraldo M. Impacts of introducing and lifting nonpharmaceutical interventions on covid-19 daily growth rate and compliance in the united states. *Proc Natl Acad Sci U S A*. 2021;118(12):e2021359118.
32. McLaughlin JM, Khan F, Pugh S, Swerdlow DL, Jodar L. County-level vaccination coverage and rates of covid-19 cases and deaths in the United States: an ecological analysis. *Lancet Reg Health Am*. 2022;9:100191.
33. Gandon S, Mackinnon MJ, Nee S, Read AF. Imperfect vaccines and the evolution of pathogen virulence. *Nature*. 2001;414(6865):751–6.
34. Ganslmeier M, Furceri D, Ostry JD. The impact of weather on COVID-19 pandemic. *Sci Rep*. 2021;11(1):22027.
35. Andersen LM, Harden SR, Sugg MM, Runkle JD, Lundquist TE. Analyzing the spatial determinants of local COVID-19 transmission in the United States. *Sci Total Environ*. 2021;754:142396.
36. Souch JM, Cossman JS. A commentary on rural-urban disparities in COVID-19 testing rates per 100,000 and risk factors. *J Rural Health*. 2021;37(1):188.
37. Rashki A, Dehkordi AH, Nemati R, Tavousi P. What does COVID-19 testing results really say? The real statistics concealed behind the accessible data. *J Med Virol*. 2021;93(10):5977–87.
38. Nguyen Q, Rivera J, Schechtman K, Glassman R, Mart M. Methods for counting COVID-19 deaths in US states and territories. *Int J Infect Dis*. 2022;116:S48.
39. Patel TA, Jain B, Raifman J. Revamping public health systems: lessons learned from the tripledemic. *Am J Prev Med*. 2024;66(1):185–8.
40. Furlow B. Triple-demic overwhelms paediatric units in us hospitals. *Lancet Child Adolesc Health*. 2023;7(2):86.

Supplementary Information

S1 Hyperparameter tuning and cross-validation

To avoid over-fitting and to ensure that the classifiers are robust to temporal and spatial variation in our dataset, we used a ten-fold spatial cross-validation approach (where the dataset is divided by HSAs into ten folds). For each week t, we applied this method to tune hyperparameters of each classifier. The ranges for the hyperparameters are shown in Table S1.

Table S1: Hyperparameters of classifiers considered in our analysis

Classifier	**Parameter**	**Values/Range**
Decision Tree	Criterion	gini, entropy
	Max depth of tree	2 - 5
Logistic Regression	Regularization	L1, L2
	Regularization strength	0.001, 0.01, 0.1, 1, 10, 100
	Solver	liblinear, saga
Neural Network	Neurons	64, 128, 255
	Learning rate	0.0001, 0.001, 0.01

S2 CDC Community Levels

New Cases (per 100,000 population in the last 7 days)	**Indicators**	**Low**	**Medium**	**High**
Fewer than 200	New COVID-19 admissions per 100,000 population (7-day total)	<10.0	10.0-19.9	≥20.0
	Percent of staffed inpatient beds occupied by COVID-19 patients (7-day average)	<10.0%	10.0-14.9%	≥15.0%
200 or more	New COVID-19 admissions per 100,000 population (7-day total)	NA	<10.0	≥10.0
	Percent of staffed inpatient beds occupied by COVID-19 patients (7-day average)	NA	<10.0%	≥10.0%

The COVID-19 community level is determined by the higher of the inpatient beds and new admissions indicators, based on the current level of new cases per 100,000 population in the past 7 days

Figure S1: **Criteria for establishing CDC Community Levels** [1].

S3 Sensitivity Analyses

S3.1 Change in prediction task

In the main text, we focused on predicting whether capacity will exceed a given threshold in week $t+3$. Here we present the evaluation metrics when the prediction task is to predict hospital capacity over a three-week period subsequent to t ($[t+1, t+3]$). The auROCs for the Full, CDC Optimized, and Reduced classifiers trained to predict this outcome were comparable with the prediction of capacity in week $t+3$ (comparing Figure S3 with Figure 2 in the main text).

S3.2 Change in hospitalization threshold

In addition to the hospital capacity threshold of 15 per 100,000 people that was used to generate our binary outcome in the main text, we also explored two other thresholds: 10 or 20 per 100,000 people. The models were trained in accordance with the procedure outlined in the main text, though now the feature and outcome relating to hospitalization capacity are replaced with a binary variable calculated based on the new threshold of interest. Overall, the performance did vary substantially between capacity thresholds and the auROC scores remained high (Figure S4).

S3.3 Change in duration of outcome period

In the main analysis, we predicted whether the hospital capacity would exceed 15 per 100,000 in three weeks' time. We additionally investigated three other periods: two (t+2), four (t+4), or six ($t+6$) weeks. Shorter outcome periods benefited the performance of the model (Figure S5), particularly when there was a decrease in the proportion of HSAs that exceeded capacity (around week 90).

S3.4 Limiting the size of the training dataset

The models in the main text have an "expanding" training set, which includes all available data up to the test week, t. We tested more limited training sets, namely restricting the training data to a four-week ($[t-5,\ t-1]$), ten-week period ($[t-11,\ t-1]$) and a twenty-six-week period ($[t-27,\ t-1]$). Expanding the training dataset did not significantly change the performance of the models considered here (Figure S6).

S3.5 Training models every four weeks

In this sensitivity analysis, we investigate the frequency at which the decision tree classifiers should be retrained with new, more recent data to ensure their continued accuracy. Rather than delivering predictions every week, we trained models to deliver predictions every four weeks using all data available up to t, and used it for prediction tasks over the next four weeks. The overall performance was comparable to scenarios where the prediction models were trained weekly (Figure S7).

S4 SI Figures

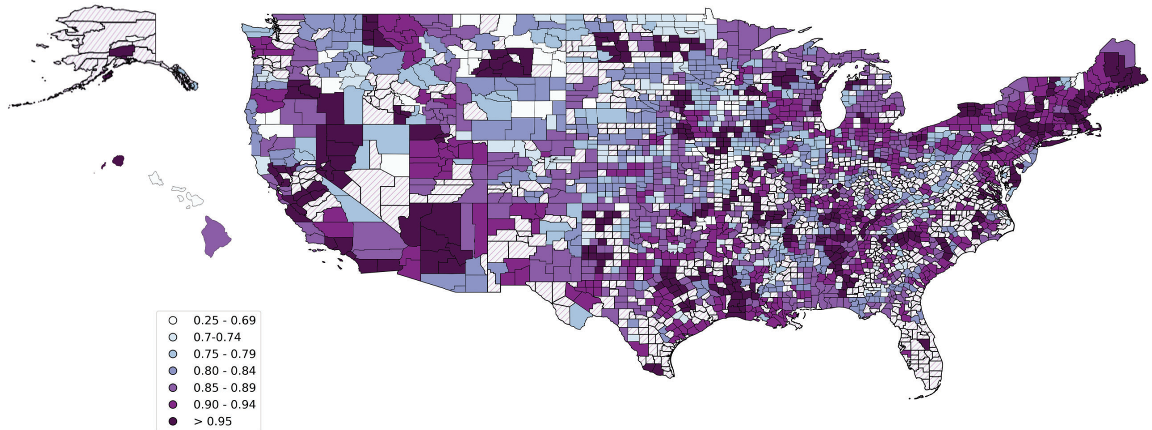

Figure S2: **Performance of the Full classifiers across all HSAs where COVID-19 death and case data are also used**. See caption of Figure 4 for additional information.

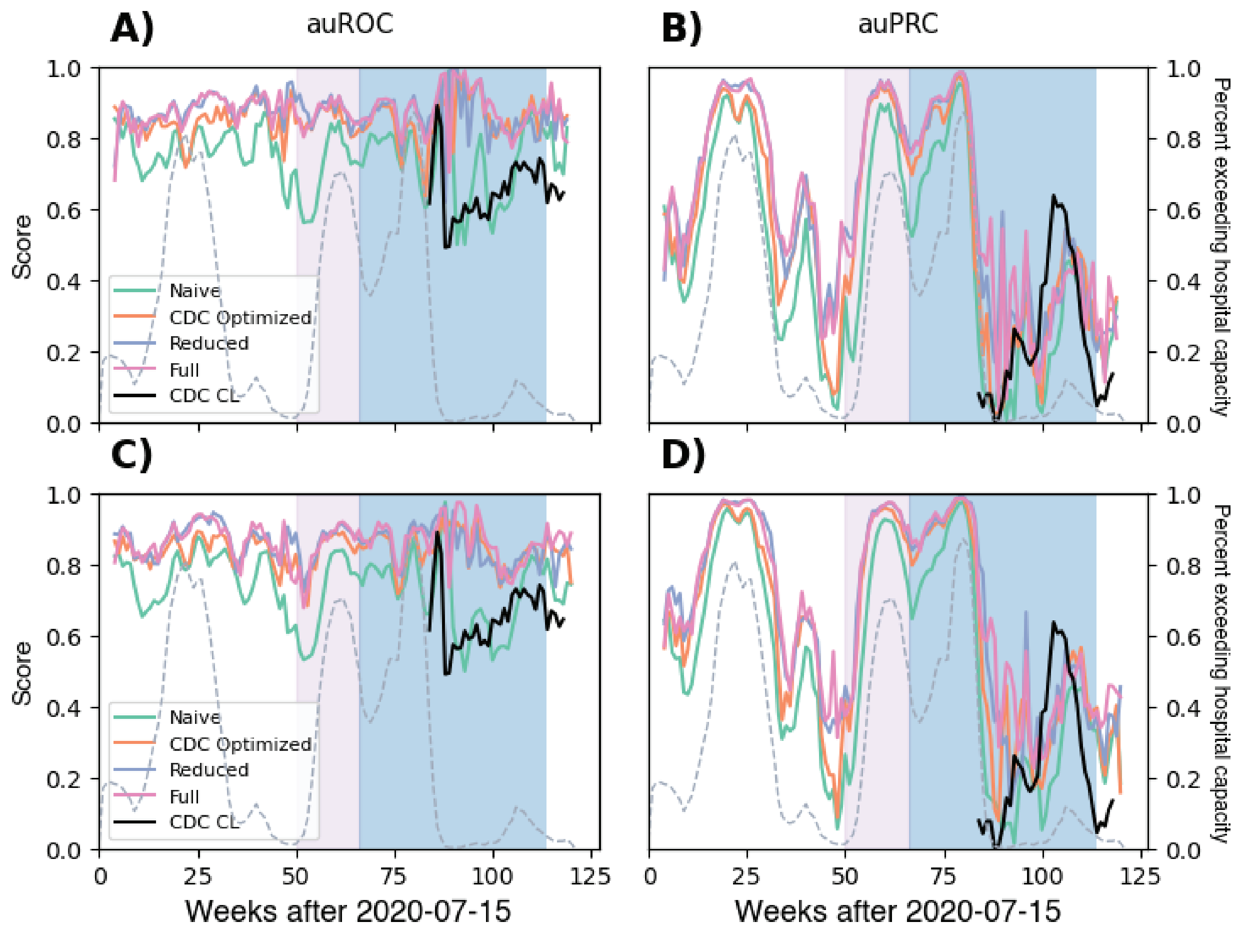

Figure S3: **Performance of decision tree classifiers when the outcome of interest is whether the hospital capacity exceeds 15 per 100,000 over a three-week period**. See the caption of Figure 2 for additional information. The first row represents the performance of models developed under the base scenario as presented in Figure 2.

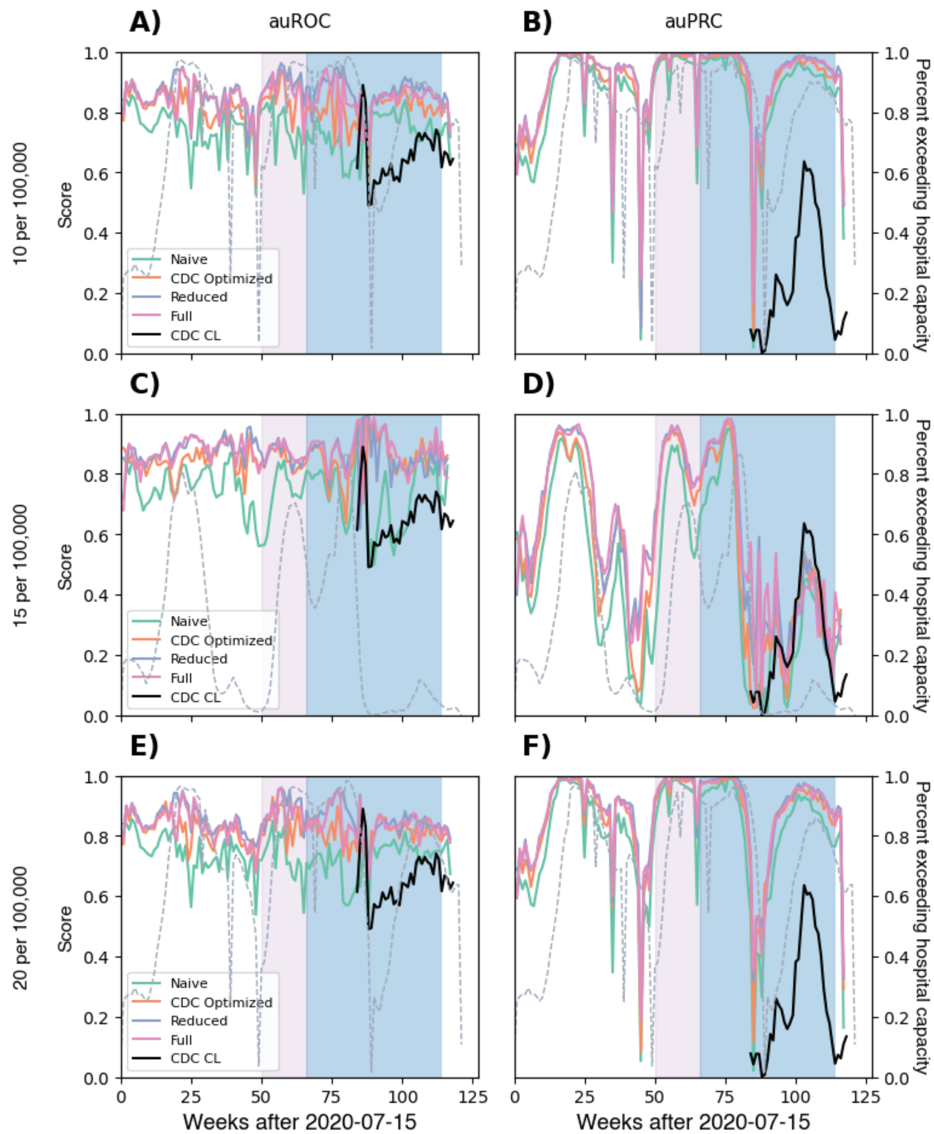

Figure S4: **Performance of decision tree classifiers when the outcome of interest is whether the hospital capacity exceeds either 10 or 20 per 100,000 during the outcome period.** See the caption of Figure 2 for additional information. The first row represents the performance of models developed under the base scenario as presented in Figure 2.

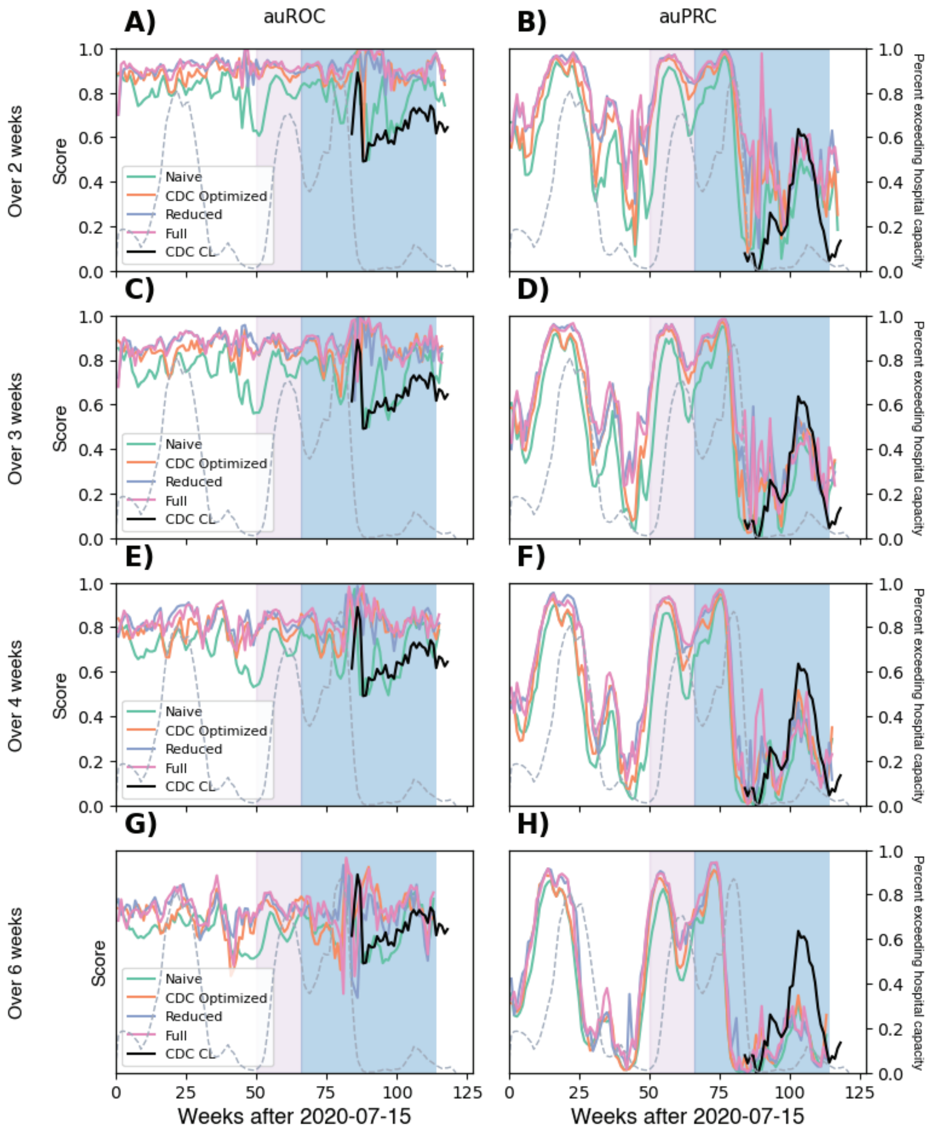

Figure S5: **Performance of full decision tree classifiers when the outcome of interest is whether the hospital capacity exceeds 15 per 100,000 people in over a 2, 3, 4, or 6 week period.** See the caption of Figure 2 for additional information. The second row represents the performance of models developed under the base scenario as presented in Figure 2.

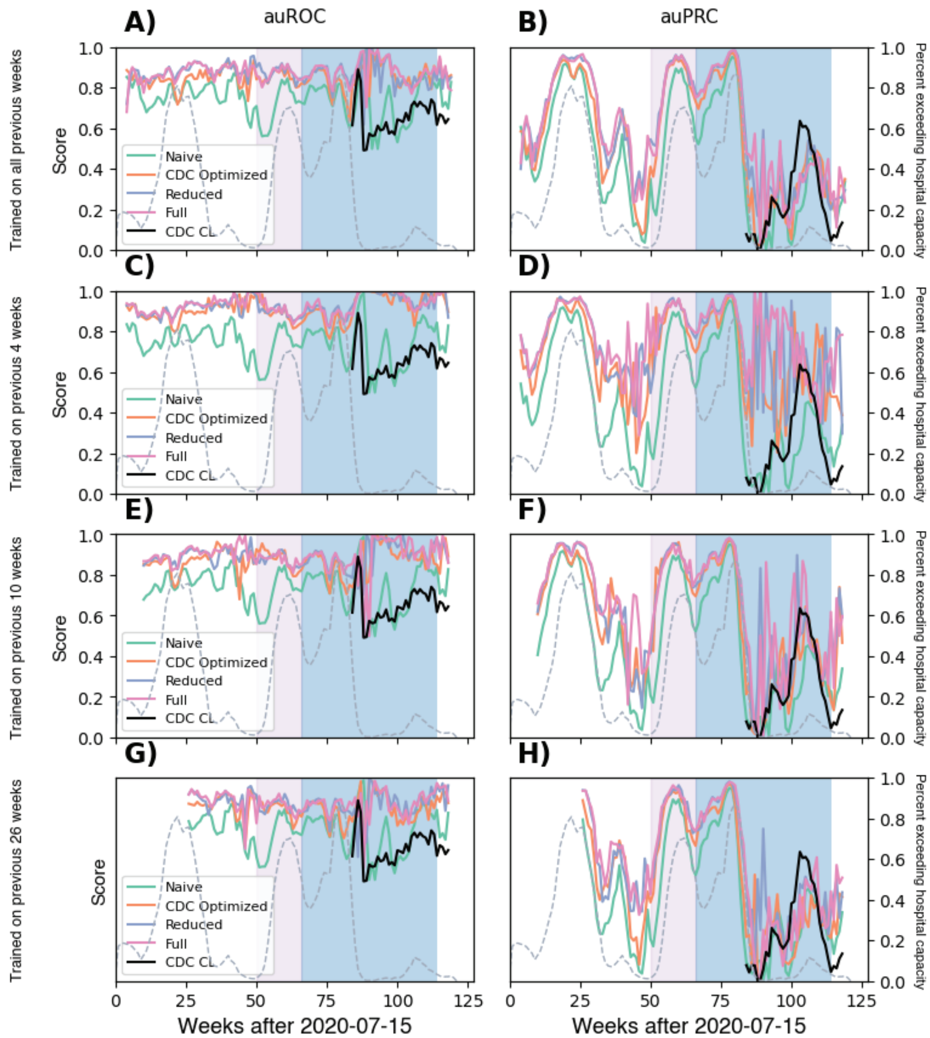

Figure S6: **Performance of full decision tree classifiers when the training set is either the previous 4, 10, or 26 weeks.** See the caption of Figure 2 for additional information. The first row represents the performance of models developed under the base scenario as presented in Figure 2.

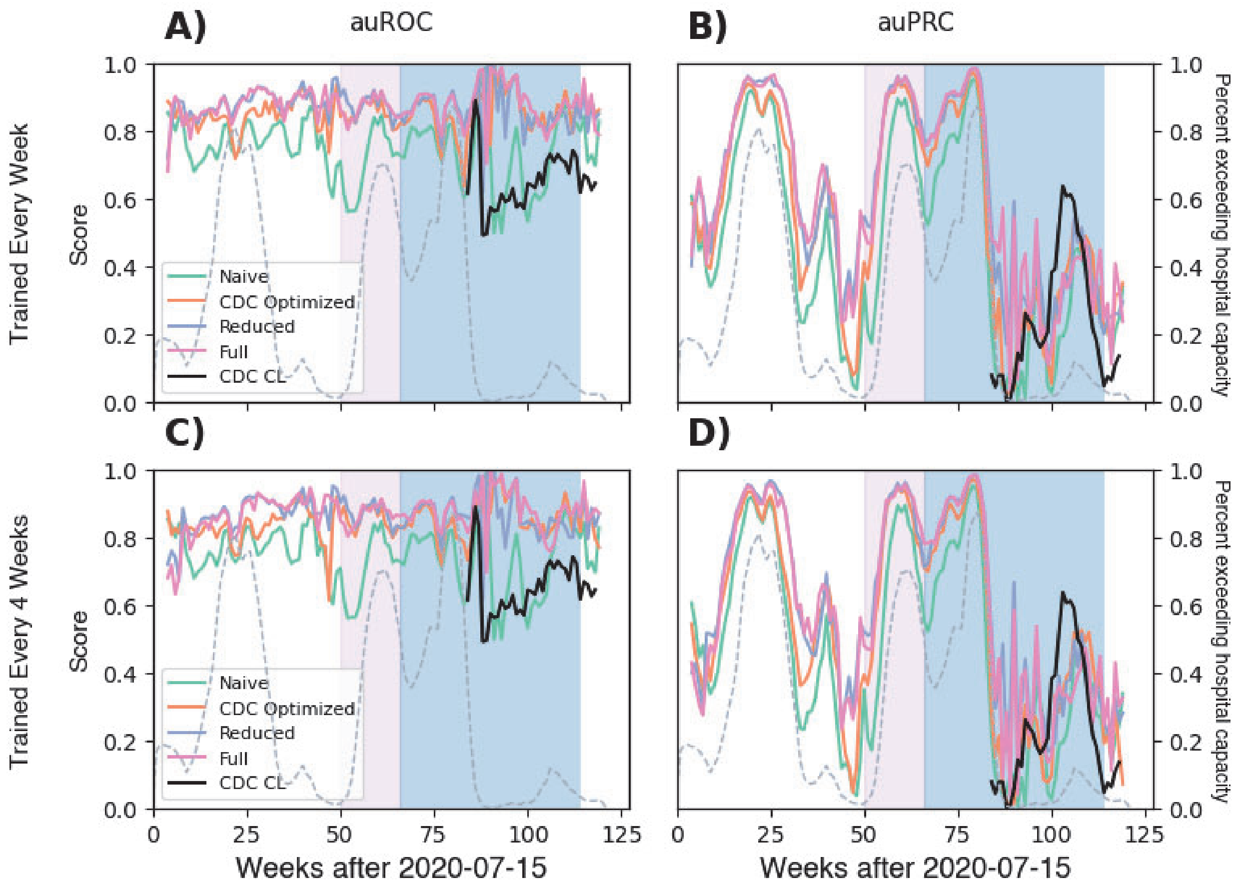

Figure S7: **Performance of decision tree classifiers when the model training set is only updated every four weeks.** See the caption of Figure 2 for additional information. The first row represents the performance of models developed under the base scenario as presented in Figure 2.

References

[1] Centers for Disease Control, Prevention, et al. Indicators for monitoring COVID-19 Community Levels and COVID-19 and implementing COVID-19 prevention strategies. *PowerPoint presentation, February*, 25, 2022.