Original Paper

Estimating COVID-19 Hospitalizations in the United States With Surveillance Data Using a Bayesian Hierarchical Model: Modeling Study

Alexia Couture[1], MPH; A Danielle Iuliano[1,2], PhD; Howard H Chang[3], PhD; Neha N Patel[1,4], MPH; Matthew Gilmer[1,5], MS; Molly Steele[1], PhD; Fiona P Havers[1,2], MD; Michael Whitaker[1], MPH; Carrie Reed[1], SCiD

[1]Centers for Disease Control and Prevention, Atlanta, GA, United States

[2]United States Public Health Service, Rockville, MD, United States

[3]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, United States

[4]Division of Health and Environment, Abt Associates, Atlanta, GA, United States

[5]General Dynamics Information Technology, Atlanta, GA, United States

Corresponding Author:
Alexia Couture, MPH
Centers for Disease Control and Prevention
1600 Clifton Road
Atlanta, GA, 30329
United States
Phone: 1 4044985984
Email: njh6@cdc.gov

Abstract

Background: In the United States, COVID-19 is a nationally notifiable disease, meaning cases and hospitalizations are reported by states to the Centers for Disease Control and Prevention (CDC). Identifying and reporting every case from every facility in the United States may not be feasible in the long term. Creating sustainable methods for estimating the burden of COVID-19 from established sentinel surveillance systems is becoming more important.

Objective: We aimed to provide a method leveraging surveillance data to create a long-term solution to estimate monthly rates of hospitalizations for COVID-19.

Methods: We estimated monthly hospitalization rates for COVID-19 from May 2020 through April 2021 for the 50 states using surveillance data from the COVID-19-Associated Hospitalization Surveillance Network (COVID-NET) and a Bayesian hierarchical model for extrapolation. Hospitalization rates were calculated from patients hospitalized with a lab-confirmed SARS-CoV-2 test during or within 14 days before admission. We created a model for 6 age groups (0-17, 18-49, 50-64, 65-74, 75-84, and ≥85 years) separately. We identified covariates from multiple data sources that varied by age, state, and month and performed covariate selection for each age group based on 2 methods, Least Absolute Shrinkage and Selection Operator (LASSO) and spike and slab selection methods. We validated our method by checking the sensitivity of model estimates to covariate selection and model extrapolation as well as comparing our results to external data.

Results: We estimated 3,583,100 (90% credible interval [CrI] 3,250,500-3,945,400) hospitalizations for a cumulative incidence of 1093.9 (992.4-1204.6) hospitalizations per 100,000 population with COVID-19 in the United States from May 2020 through April 2021. Cumulative incidence varied from 359 to 1856 per 100,000 between states. The age group with the highest cumulative incidence was those aged ≥85 years (5575.6; 90% CrI 5066.4-6133.7). The monthly hospitalization rate was highest in December (183.7; 90% CrI 154.3-217.4). Our monthly estimates by state showed variations in magnitudes of peak rates, number of peaks, and timing of peaks between states.

Conclusions: Our novel approach to estimate hospitalizations for COVID-19 has potential to provide sustainable estimates for monitoring COVID-19 burden as well as a flexible framework leveraging surveillance data.

(JMIR Public Health Surveill 2022;8(6):e34296) doi: 10.2196/34296

KEYWORDS

COVID-19; SARS-CoV-2; hospitalization; Bayesian; COVID-NET; extrapolation; hospital; estimation; prediction; United States; surveillance; data; model; modeling; hierarchical; rate; novel; framework; monitoring

Introduction

Monitoring disease burden and severity is a critical component of public health research, communication, and response. The current COVID-19 pandemic, which is caused by SARS-CoV-2, has been ongoing since early 2020 and presents novel challenges and barriers to monitoring due to the unique transmission, nature of the virus, and variety of symptom presentations. In the United States, COVID-19 cases, hospitalizations, and deaths are captured through the National Notifiable Disease Surveillance System (NNDSS) and death certificates reported to the National Vital Statistics System (NVSS) [1-3]. However, the hospitalization status of cases reported by states through the NNDSS is often incomplete and thus might inaccurately represent the burden of COVID-19 hospitalization in the United States. In addition, since July 15, 2020, hospitalizations known or suspected to be related to COVID-19 have been reported daily through the Department of Health and Human Services (HHS) Protect, known as the unified hospital time-series data [4]. This data collection is a burden on facilities that is likely unsustainable in the long term.

Current research and methods for estimating hospitalizations of COVID-19 are limited. In mid-2020, the Centers for Disease Control and Prevention (CDC) developed a multiplier method for estimating SARS-CoV-2 infections and hospitalizations for COVID-19 based on state- and territory-reported line-level case data [5]. To date, these COVID-19 burden estimates from this case-based multiplier model are calculated and published on the CDC's website [6]. Other papers have leveraged seroprevalence surveys to estimate SARS-CoV-2 infections and hospitalizations for COVID-19 [7,8]. These methods rely on data systems such as case reporting or wide-scale, special seroprevalence surveys that were initiated during the pandemic but might not exist in the future, as the pandemic winds down. Case count data and consistent, representative seroprevalence data may eventually be discontinued due to the pandemic slowing down and resources and attention going elsewhere, leaving a need for longer-term systems that can be sustained.

Since March 2020, the COVID-19-Associated Hospitalization Surveillance Network (COVID-NET) has collected data on laboratory-confirmed SARS-CoV-2–positive patients from a network of hospitals in 14 US states [9]. Although this sentinel surveillance system does not cover the entire United States, it is expected to continue monitoring rates of COVID-19 hospitalization even after the pandemic ends. The COVID-NET system was built off of the similar long-standing Influenza Hospitalization Surveillance Network (FluSurv-NET), which has been monitoring population-based rates of influenza hospitalization for almost 20 years [10]. Although the network does not currently make any further determination about the relationship between a positive SARS-CoV-2 test and the reason for hospitalization for each identified patient, this system and data are the best source available for the long-term surveillance of COVID-19 hospitalizations.

We created a method to utilize COVID-NET data to provide national and state-specific estimates of hospitalization to provide a long-term, sustainable framework to generate estimates of COVID-19 disease burden in the United States. The aim of this study was to estimate monthly COVID-19 hospitalization rates, defined as hospitalized patients with positive tests for SARS-CoV-2 infections, for all 50 states from May 2020 through April 2021. We adapted a Bayesian hierarchical model to estimate and extrapolate hospitalization rates, accounting for uncertainty and variability between states and across time.

Methods

COVID-NET Surveillance Hospitalization Data and Adjustments

We used COVID-19 hospitalization data from COVID-NET. The network identifies hospitalized patients with a positive SARS-CoV-2 test, including molecular assay and antigen detection, during hospitalization or within 14 days prior to hospitalization [9]. Hospitalization rates are calculated by the number of residents in a catchment area, defined as the area or population around the reporting hospital that the hospital potentially services, of the COVID-NET sites who are hospitalized with a confirmed, positive SARS-CoV-2 test divided by the total population within that defined catchment area. The network is made up of over 250 acute care hospitals representing 99 counties in 14 states: California, Colorado, Connecticut, Georgia, Iowa, Maryland, Michigan, Minnesota, New Mexico, New York, Ohio, Oregon, Tennessee, and Utah. Overall, the network covers about 10% of the United States population. For this analysis, case data were aggregated by month of hospitalization, state reporting, and the following 6 age groups: 0-17 years, 18-49 years, 50-64 years, 65-74 years, 75-84 years, and ≥85 years. Age groups were chosen based on available data age groupings as well as interest in breaking apart older age groups, which have been impacted more by severe COVID-19.

Recognizing that all hospital patients are unlikely to be tested for SARS-CoV-2 and, therefore, some true cases are not classified as COVID-19 patients, COVID-19 hospitalization rates are adjusted by weighting them for SARS-CoV-2 testing practices (ie, the probability of being tested for SARS-CoV-2 during their hospitalization). In addition, testing practices changed over the course of the pandemic. The probability of being tested was calculated from the IBM Watson Health Explorys electronic health record database (IBM Corporation), which includes more than 39 health system partners across the country. All states participating in COVID-NET, except Connecticut, used the same testing probabilities calculated from IBM Watson data, which were aggregated testing practices of all partners stratified by month and age group. The testing probabilities for these 13 states ranged from 0.28 to 0.67. Connecticut provided site-specific testing practice data through COVID-NET, which ranged from 0.32 to 1.00. Rates were also

adjusted to account for the SARS-CoV-2 assay sensitivity because, depending on the sensitivity of the assay, some patients could have false-negative test results (ie, would not be identified as a COVID-19 hospitalization). The assay sensitivity was assumed to be 0.885, which is the midpoint for the range found in a systemic review [11]. The adjusted hospitalization counts were used to calculate rates using COVID-NET catchment populations for each site. Due to the range in hospitalizations by age groups over time, 6 models were run, 1 for each age group:

$$Adjusted\ COVIDNET\ Count_{sm} = \frac{COVIDNET\ Count_{sm}}{Prob.of\ being\ tested_m * Testing\ sensitivity}$$

where s=1, ..., S for each COVID-NET state and m=1, ..., M for each month.

Covariate Data and Selection

To extrapolate COVID-19 hospitalization rates from COVID-NET sites to states not included in the COVID-NET network, we incorporated model covariates based on state, month, and age-specific demographic and epidemiological data. We used different data measures to account for differences between states with COVID-NET sites and those states without COVID-NET sites from multiple sources (Table 1). Including covariates in the model helps to quantify differences between age groups, months, and states and allows for the model to account for these differences when estimating how many COVID-19 hospitalizations have occurred. We considered both time-varying and time-invariant state-level covariates that captured other COVID-19 disease trends, population demographics, and population health indicators. For the time-varying covariates, we considered the percent of SARS-CoV-2 positive tests from commercial and public health laboratories, percent of all-cause deaths that were coded as COVID-19 deaths from the National Center for Health Statistics and NVSS, and the following hospital capacity variables: percent patients with COVID-19 out of all inpatients and percent intensive care unit (ICU) beds occupied out of all ICU beds [12-16]. We incorporated a 1-week lag to the percent positive COVID-19 tests to account for time between symptom onset and hospitalization and a 1-week lead to the percent of COVID-19 deaths out of all deaths to account for time between hospitalization and death. For the time-invariant covariates, we considered the percent Native American and percent Black American and the population prevalence of the following conditions or diseases from the Behavioral Risk Factor Surveillance System (BRFSS): obesity, heart disease, chronic obstructive pulmonary disease, diabetes, chronic kidney disease, and asthma [17,18]. Underlying medical and chronic conditions were found to be highly prevalent in hospitalized patients with COVID-19 and were therefore included as possible covariates [18]. Time- and age-varying data for population prevalence of underlying medical and chronic conditions were not available. Table 1 summarizes all of the variables that were considered as covariates. We used covariate selection methods to determine which of the possible covariates to include in the model. For the <18-year-old age group, only asthma was included as a possible covariate from the chronic conditions or diseases because of a lack of evidence that the prevalence of other chronic conditions or diseases affected COVID-19 hospitalization in that age group [19].

Table 1. Variables considered to be covariates in our Bayesian model to extrapolate COVID-19 hospitalizations for all 50 US states with stratification and source.

Variables	Stratified by	Source
Laboratory surveillance: SARS-CoV-2 % positive using rt-PCR[a] tests	Month, state, age	Commercial lab and public health lab data
Vital records death: % of all-cause deaths that were coded as COVID-19 deaths	Month, state, age	National Center for Health Statistics National Vital Surveillance System
Hospital capacity: % COVID patients out of all inpatients, % ICU[b] occupied out of all ICU beds	Month, state	HHS[c] Protect/National Center for Health Statistics
Race/ethnicity: % American Indian, % Black, % racial minority[d]	State, age	National Center for Health Statistics/National Vital Statistics System
Chronic conditions/diseases: % obesity, % heart disease, % COPD[e], % Diabetes, % CKD[f], % asthma	State	CDC[g] MMWR[h] Stacks/Behavioral Risk Factor Surveillance System

[a]rt-PCR: reverse transcription–polymerase chain reaction.

[b]ICU: intensive care unit

[c]HHS: Department of Health and Human Services.

[d]Racial minority was defined as non-White and non-Hispanic.

[e]COPD: chronic obstructive pulmonary disease.

[f]CKD: chronic kidney disease.

[g]CDC: Centers for Disease Control and Prevention.

[h]MMWR: Morbidity and Mortality Weekly Report.

Extreme values were detected for time-varying covariates and subsequently transformed using Winsorization (ie, minimized the influence of outliers by replacing them by the maximum or minimum values at a threshold of distribution percentiles) [20].

We used the adjusted COVID-NET hospitalization rates as the outcome to select covariates separately for each age group. Covariate selection methods assist with avoiding collinearity and ensuring that the most relevant and impactful covariates are included. Our method for covariate selection utilized Least Absolute Shrinkage and Selection Operator (LASSO) and spike and slab [21,22]. Covariates were included in the final model for the specific age group if they were selected by LASSO and then the model incorporated spike and slab selection. The LASSO chooses a subset of predictors by introducing an upper bound for the sum of squares and minimizing the errors present in the model. Spike and slab is a Bayesian approach in which we assigned priors to the regression coefficients to be zero or nonzero, which is where the name comes from. From that, the posterior distributions show a biseparation effect in the model coefficients—those that peak at zero and those significantly different from zero. Assumption for nonzero was high in the model due to LASSO selection being done first.

Bayesian Hierarchical Model and Extrapolation

We implemented a Bayesian hierarchical model for extrapolation adapted from a model to estimate global influenza burden rates [23]. Parameter estimation and inference were conducted under a fully Bayesian framework to better quantify uncertainties in predicted hospitalization rates, including those that are extrapolated to states without COVID-NET data.

We let A_{sm} denote the estimated, adjusted COVID-19 hospitalization count from the COVID-NET states during months from the pandemic, starting in May 2020, where s=1,..., S, and S=14 states in COVID-NET, m=1,..., M, and M=12 for each month included in the model (ie the observed data adjusted in section COVID-NET Surveillance Hospitalization Data and Adjustments). Because the observed hospitalization estimate is a count, we can view them as deriving from a Poisson probability [24]. This is used to account for the random variation from the observed data. Those estimated, adjusted COVID-19 hospitalization counts, along with the COVID-NET catchment populations and the selected covariates, were used as inputs into the following Bayesian hierarchical model:

Level 1: $A_{sm} \sim Pois(\theta_{sm} * Population_s/100{,}000)$

where A_{sm} = Adjusted COVIDNET Count_{sm} (the estimated hospitalization count for state and month from COVID-NET data), $Population_s$ is the catchment population for state s, and θ_{sm} is the unobserved true hospitalization rate.

Level 2: $\theta_{sm} \sim logN(\mu + \gamma_1 X_{1,sm} + ... + \gamma_k X_{k,sm}, \sigma^2)$

where X is the value of covariate i in state s at time m, k=1, ..., K, K = the number of selected covariates, and covariates are with mean 0 and variance 1.

Level 3: $\gamma_k \sim N(0, 1000000^{(1-gk)} * 0.001)$

$g_k \sim Bern(0.9)$

Priors: $\mu \sim N(0, 10^{-6})$

$\sigma^2 \sim Unif(0, 1000)$

where k=1, ..., K *and* K = the number of selected covariates.

Inference was carried out utilizing Markov chain Monte Carlo (MCMC) simulations with 20,000 iterations. The model outputs included samples from the posterior distribution of COVID-19–associated hospitalizations for each state and month. Using these samples, we calculated the median and 90% credible intervals (CrIs) for hospitalization counts, rounded to the hundreds due to MCMC errors, and used the state population by age group to calculate final hospitalization rates. To calculate overall age, age by month, age by state, and state by month hospitalizations and rates, we first summed the posterior samples. Since the median of sums does not equal the sum of medians, this led to slightly different total hospitalizations depending on which grouping was used to sum. For consistency, we calculated total hospitalizations from overall age medians, total monthly hospitalizations from age by month, and total state hospitalizations by age by state. We chose 20,000 iterations after starting with 2000 iterations and slowly increasing to obtain stable estimates that also minimized simulation error.

Validation and Comparison

We conducted sensitivity analyses to assess the effect of covariate selection and input data on the model. Multiple combinations of covariates were examined for each age group to assess how robust the hospitalization estimates were to covariate selection. To validate and test the sensitivity of the model, first, we compared how the model estimated hospitalizations for each COVID-NET state with the observed hospitalization rate from COVID-NET. In another sensitivity analysis, we dropped data from each COVID-NET state, one by one, and then compared the observed hospitalization rates to the extrapolated rates for each dropped state. Finally, we also compared our COVID-19 hospitalization estimates against other public estimates and databases, including COVID-19 hospitalization rates reported through Healthdata.gov (The Unified Hospital Timeseries data), the COVID Tracking project, and from the CDC's case-based multiplier model [5,6,25-28]. The Unified Hospital Timeseries data and COVID Tracking project are publicly available data sets providing state-aggregated data for COVID-19 hospitalizations over time. According to Healthdata.gov, the Unified Hospital Timeseries data had reliable counts of new hospitalizations with COVID-19 starting in the fall of 2020 when over 95% reporting from all hospitals reported by the HHS. The Unified Hospital Timeseries data are from reports at the facility level and do not account for nonresponse or missingness. The COVID Tracking Project compiled data taken directly from the websites of state or territory public health authorities but stopped and switched to reporting the Unified Hospital Timeseries on March 7, 2021. The CDC's case-based multiplier model estimates hospitalization in 2-month increments and by HHS regions, not by state. Our model output was aggregated appropriately for comparisons.

Ethical Statement

This activity was reviewed by the Centers for Disease Control and Prevention (CDC) and determined to be consistent with nonhuman participant research activity (#0900f3eb81da6749). Informed consent was waived, as data were deidentified and aggregated.

Results

The covariates selected for each age group varied (Multimedia Appendix 1). The SARS-CoV-2 percent positive, the percentage of inpatients with COVID-19 out of all inpatients, and the percentage of hospitalizations that were ICU admissions were selected for each of the age groups. The 18- to 49-year-old age group had the most covariates selected, and the <18-year-old age group had the fewest covariates selected.

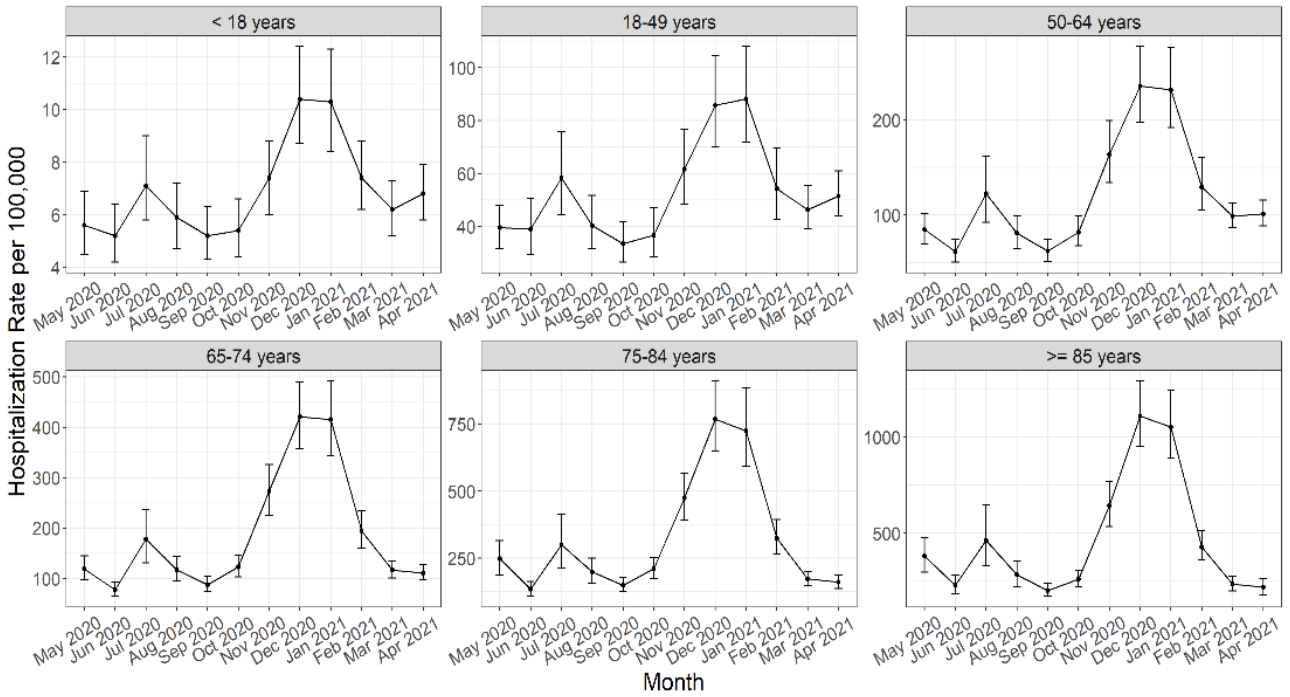
From May 2020 through April 2021 in the United States, we estimated there were 3,583,100 (90% CrI 3,250,500-3,945,400) hospitalizations representing a rate of 1093.9 (90% CrI 992.4-1204.6) hospitalizations per 100,000 population with COVID-19. The estimated rates varied by age group, state, and month. The highest rates of hospitalization were among those aged ≥85 years, with a rate of 5575.6 per 100,000 population (90% CrI 5066.4-6133.7), and the lowest hospitalization rate was for those <18 years of age, with a rate of 83.9 per 100,000 population (90% CrI 76.8-91.4). Table 2 summarizes the final estimated counts and rates of hospitalizations by age group from May 2020 through April 2021.

Hospitalization rates for all age groups peaked in either December 2020 or January 2021. Figure 1 shows the epidemiologic curves of hospitalizations over time by age group. During the study period, we observed the largest peak in hospitalization rates in December 2020 (183.7/100,000), followed by January 2021 (180.1/100,000). A second, smaller peak in COVID-19 hospitalizations was observed for all age groups in July 2020 (90.6/100,000). The lowest rate of hospitalization was observed across age groups in September 2020 (46.9/100,000). Following the peak in COVID-19 hospitalization rates during the winter months, COVID-19 hospitalizations declined until the month of April 2021 (Figure 1).

Table 2. Cumulative COVID-19 hospitalization count (median) and rate per 100,000 population and accompanying 90% credible intervals (CrIs) for each age group and overall from May 2020 through April 2021 for 50 US states from our Bayesian model output.

Age group	Hospitalization count	90% CrIs	Hospitalization rate per 100,000	90% CrI
<18 years	61,200	56,000-66,600	83.9	76.8-91.4
18-49 years	892,700	805,700-992,100	647.7	584.6-719.8
50-64 years	927,900	846,900-1,016,100	1477.1	1348.2-1617.6
65-74 years	709,800	645,200-776,500	2258.0	2052.5-2470.3
75-84 years	623,900	562,600-689,700	3912.7	3528.7-4325.7
≥85 years	367,600	334,000-404,400	5575.6	5066.4-6133.7
Total	3,583,100	3,250,500-3,945,400	1093.9	992.4-1204.6

Figure 1. COVID-19 hospitalization rates per 100,000 population and 90% credible intervals by age group over time from May 2020 through April 2021 for 50 US states from our Bayesian model output. The Y-axis limits are adjusted to the unique range for each age group (ie, they are not set to the same scale).



At a state level, cumulative hospitalization rates from May 2020 through April 2020 ranged from 359.3 (90% CrI 241.5-476.6) hospitalizations per 100,000 people in Vermont to 1855.6 (90% CrI 1184.3-2640.1) hospitalizations per 100,000 people in Nebraska. Figure 2 shows the overall cumulative hospitalization rate per 100,000 people from May 2020 to April 2021 for all states with a heat map (Figure 2A) and by bar graph (Figure 2B) to show the range of hospitalization burden across the country. COVID-NET states are well distributed throughout the highest to lowest rates by state.

Considering state-specific hospitalization rates over time, not all states had the same peaks or magnitudes of peaks. Figure 3 shows the epidemiological curves across the study period for the top 10 states with the highest upper 90% credible interval for cumulative hospitalization rates from May 2020 through April 2021. From these example states, we were able to observe differences in the time trends between states regarding the timing and number of peaks. States including Texas, Nevada, Alabama, Arizona, and Tennessee have 2 peaks; however, they differed by timing and magnitude of the peaks. In contrast, Nebraska, Kansas, Virginia, Missouri, and Oklahoma experienced only 1 major peak, which also differed by timing and magnitude. Hospitalization rates per 100,000 population from the final output model over time are provided in Figure 3.

To assess the sensitivity of the selected covariates, we ran the model using multiple combinations of the covariates, including those selected by the LASSO method alone and those by the spike and slab method alone. Hospitalization estimates did not vary greatly overall or by age depending on covariate combinations and were almost 100% consistent between LASSO alone, spike and slab alone, and when both were used, which are the covariates used in the final model for each age group. To validate the final model, we compared the observed COVID-NET hospitalization rates to the final model's estimated hospitalization rates. The rates are higher from the final model. However, the trends over time and by age group follow the observed, input rates (Multimedia Appendix 2). The supplementary images are a plot of each COVID-NET state comparing observed (input), estimated (final model), and extrapolated monthly hospitalization rate in the leave-one-state-out analysis, showing rates over time and by age group. Model median results for other states were mostly consistent whether the specific COVID-NET state was dropped or not. Almost all of the COVID-NET states' extrapolated estimates (ie, when dropped) had a 90% CrI that included the observed (input) estimate and estimated (final model) rate. The older age groups were more consistent and had more overlap between estimates than the younger age groups in the leave-one-state-out analysis. Finally, we compared our output with other hospitalization estimates and data for the final step of our sensitivity analysis. We compared our results with the Unified Hospital Timeseries data and data published on The COVID Tracking Project [25,26]. Figure 4 shows a comparison of hospitalization rate from each source over time. We also compared our results to the current published numbers from the CDC's case-based multiplier model (Multimedia Appendix 3) [5,6,27].

Figure 2. Cumulative COVID-19 hospitalization rate per 100,000 population by state from May 2020 through April 2021 in the United States from our Bayesian model output: (A) heat map of the United States of cumulative hospitalization rate per 100,000 population from May 2020 through April 2021 and (B) bar chart of cumulative hospitalization rate per 100,000 population from May 2020 through April 2021, with 90% credible intervals and states from COVID-19-Associated Hospitalization Surveillance Network (COVID-NET) in blue.

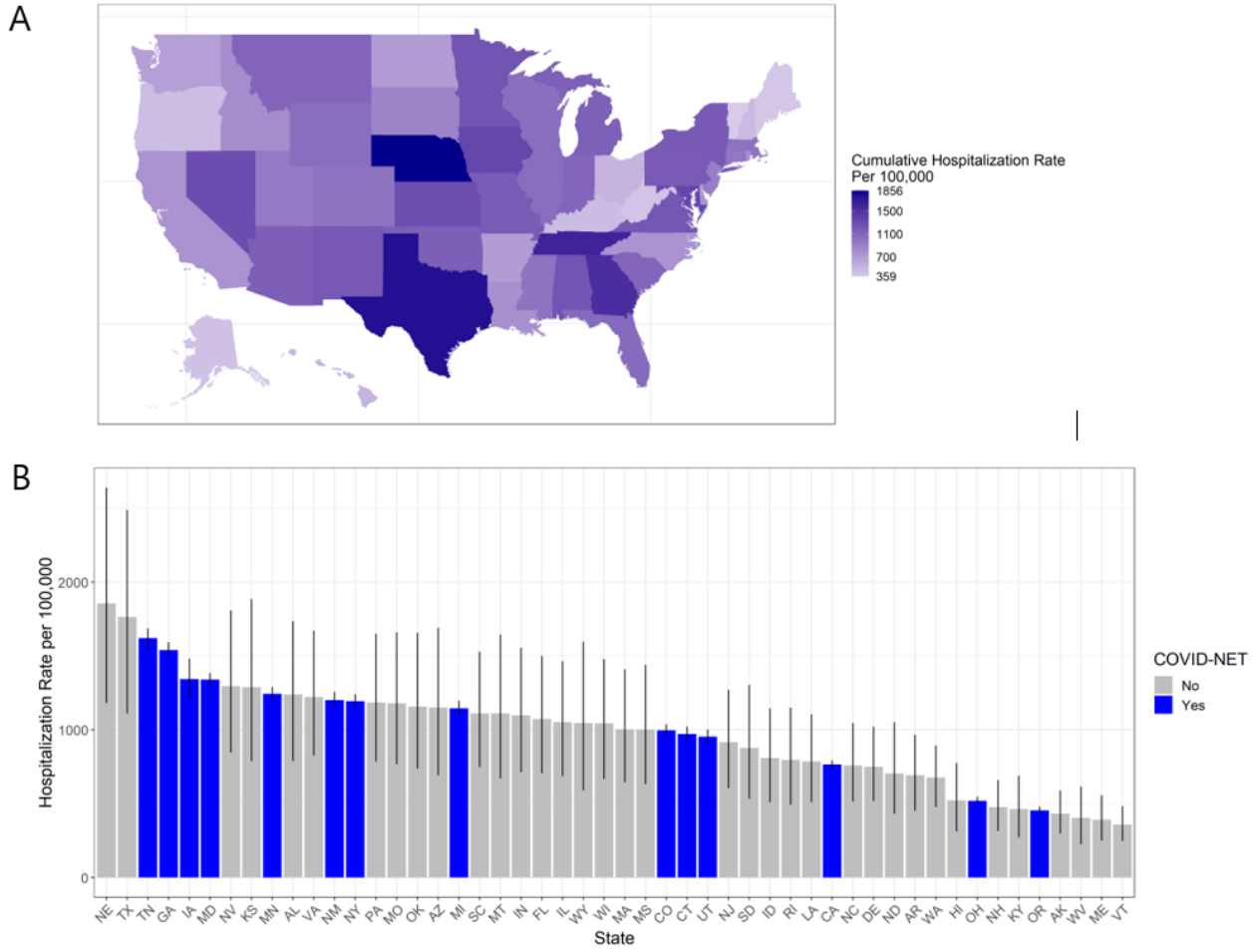


Figure 3. COVID-19 hospitalization rates per 100,000 population over time for the top 10 US states with the highest upper 90% credible interval for cumulative COVID-19 hospitalization rates from May 2020 through April 2021 from the Bayesian model output: (A) Nebraska, (B) Texas, (C) Kansas, (D) Nevada, (E) Alabama, (F) Arizona, (G) Tennessee, (H) Virginia, (I) Missouri, and (J) Oklahoma.

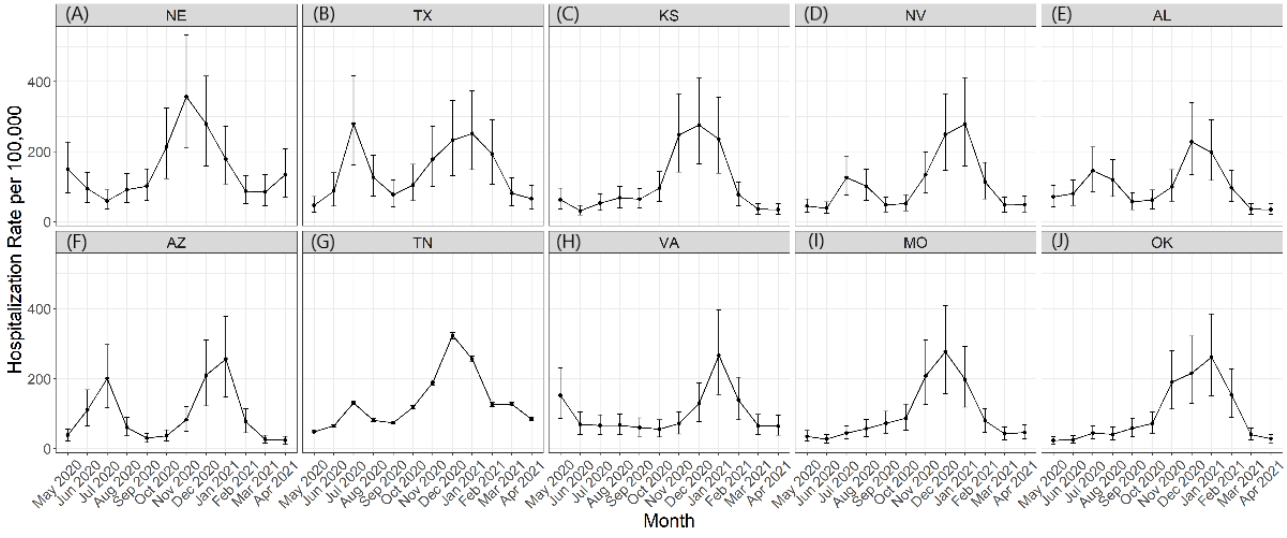
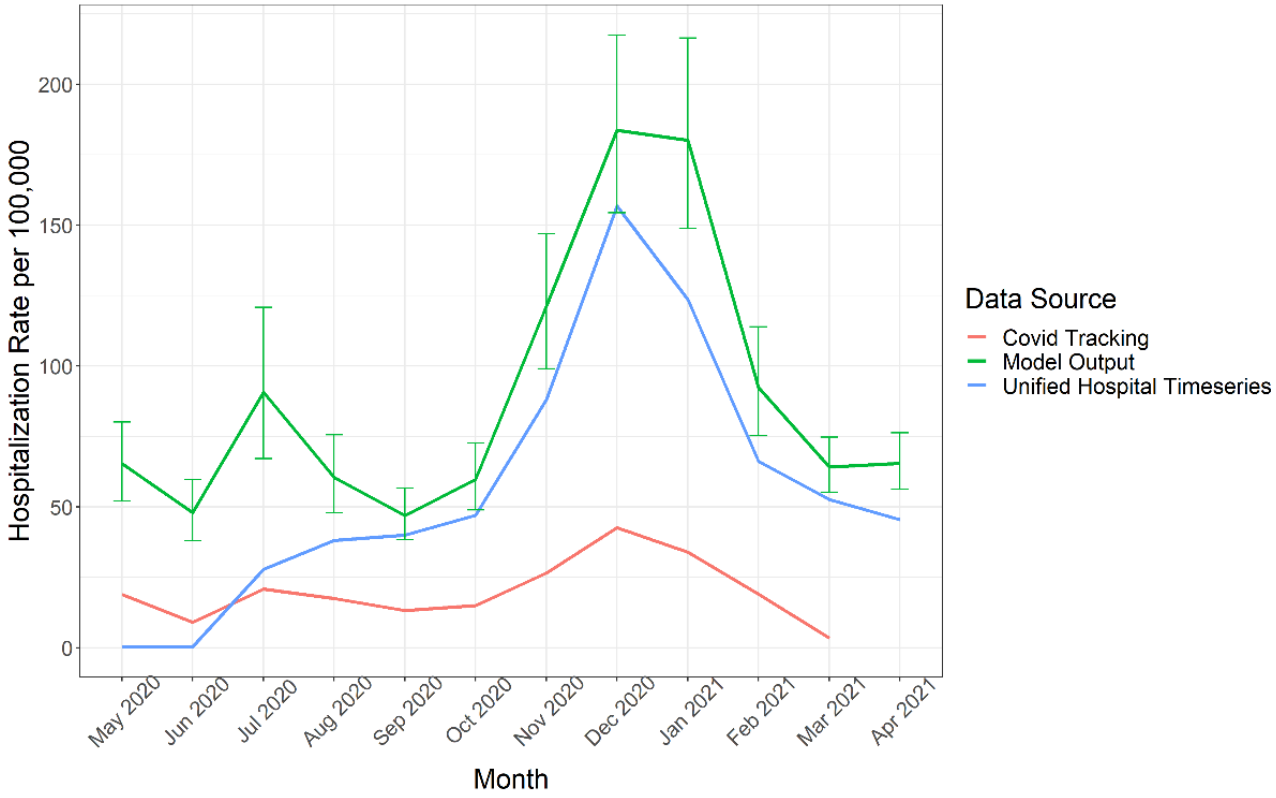


Figure 4. Comparison of COVID-19 hospitalization rates per 100,000 population over time from May 2020 through April 2021 in the United States from our Bayesian model output with 90% credible intervals, the Unified Hospital Timeseries data, and data from The COVID Tracking Project.



Discussion

Overall, our method estimated that 3,583,100 hospitalizations occurred in the United States from May 2020 through April 2021, with estimated rates varying by age group, state, and month. These estimates demonstrate the large burden of COVID-19 hospitalizations in the United States and provide visibility on variations in disease burden by age group, state, and time. As expected, the most severe burden of COVID-19 hospitalizations occurred among older age groups, specifically among people aged ≥65 years old. The largest peak in hospitalizations occurred in December 2020 and January 2021, aligning with the largest peak in reported case rates [28].

Our approach to estimating the burden of COVID-19 hospitalization using long-term surveillance data has several benefits. First, we designed our model to build on an existing system that was initially started to track hospitalizations for influenza and has expanded to capture other respiratory viruses including COVID-19. COVID-NET was built on a long-standing surveillance infrastructure that has been conducting surveillance for respiratory infections, including influenza and respiratory syncytial virus, for many years and is expected to continue monitoring COVID-19 hospitalization rates into the future [29]. Our model calculated estimates of state-level hospitalization rates by month and age group, rather than assuming the 14 COVID-NET sentinel sites are representative of the United States. Each US state has experienced the pandemic differently, and our models allow us to capture the variations in the number and magnitude of peaks and state-specific trends in hospitalization rates. Further, using covariates to extrapolate data from the COVID-NET sites to the rest of the United States provides useful information to understand state-level differences in hospitalization. The covariates add information to the input hospitalization rates to then create a better story for the states to which it extrapolates. This model helps preserve notable differences in the epidemiology of COVID-19 between states.

When we compared our model against the published Unified Hospital Timeseries and the COVID-Tracking Project, our COVID-19 hospitalization estimates were higher but showed the same trends and included the Unified Hospital Timeseries' rates in our 90% CrIs for a few months (Figure 4). We also compared our model to the case-based multiplier model. The CDC developed the case-based multiplier model using nationally notifiable COVID-19 case report data and assumptions for underdetection of confirmed cases, which is still being used to produce published burden estimates [5,6]. Our Bayesian model offers an alternative method of estimation by leveraging sentinel surveillance data if or when case report data become unreliable or unavailable. When we compared our model's output to the case-based multiplier model during time periods that overlapped, we found that our model generated more conservative estimates of hospitalization. Our model's output was lower than the estimates from the case-based multiplier model (Multimedia Appendix 3). From June 2020 to March 2021, our model estimated a cumulative incidence of 904.3 per 100,000 population whereas the case-based multiplier estimated 1345.3 per 100,000 population. When comparing estimates by age group, months, and HHS regions, specific differences are highlighted. Our model had much lower estimates of hospitalization rates per 100,000 for the 0- to 17-year-old age group (210.7 for the case-based multiplier model and 67.4 for ours) and ≥65-year-old age group (4401.7 for the case-based

multiplier model and 2800.8 for ours), while the other age groups were only slightly lower (Multimedia Appendix 3). In addition, our February through March estimate and HHS regions 2 and 9 were much lower. However, our model had higher estimates for a few HHS regions compared with the case-based multiplier estimates. Our method has several advantages over the case-based multiplier method. First, the case report data used were often incomplete for hospitalization status and relied on the imputation of hospitalization status. In our method, the input hospitalization data were from a surveillance system that actively identified laboratory-confirmed COVID-19 hospitalizations. This may account for the differences observed in the hospitalization estimates between the models. Imputation could lead to more hospitalizations than those counted from the surveillance system. For example, if those not missing in case data have a bias toward being hospitalized, then those with missing hospitalization status in the case data would also have a bias toward being hospitalized when imputed. A second difference between the methods was that the case-based multiplier method adjusted reported cases for factors that influenced case detection, including health care–seeking behaviors and testing practices at the HHS region level. Therefore, they adjusted and estimated at the HHS region level rather than the state level like our method. Estimating at the regional versus the state level may also explain differences in estimates.

The case-based multiplier model relies on COVID-19 being a nationally notifiable disease and continued case reporting by states and jurisdictions, which may not continue long term. In contrast, our method relies on routine sentinel surveillance data and allows for extrapolation to places without data. Both the case report data and seroprevalence data used by Angulo et al [7] as the basis for their national COVID-19 disease burden estimates were data sources created to inform the pandemic response, but it is unclear how long these data will continue to be collected.

Although we utilized this method for estimating state-level hospitalization rates for COVID-19 in the United States from May 2020 through April 2021, our method can be adapted for different outcomes or measures of interest both domestically and in international settings. The main components needed are reliable surveillance data in enough areas to have diversity in disease occurrence and covariates that help explain the variation between all areas of extrapolation. There are surveillance systems set up that do not have complete coverage. For example, this approach was adapted from an analysis using a Bayesian Hierarchical model to extrapolate influenza yearly rates by country [23]. This method provides an opportunity to leverage surveillance data and inform more accurate estimates of disease burden. Efforts to further expand the method to other levels of disease severity including infection, illness, or death are ongoing.

Our method also has some limitations. First and foremost, we are estimating hospitalizations with positive tests for SARS-CoV-2 infections, as the contributing surveillance data do not currently attribute whether patients were hospitalized due to complications caused by the infection. Even for hospitalizations that are incidental, like an elective surgery, the hospital still has to deal with cohorting and infection control for that person, which adds burden on the hospital. Second, since our goal was to use routine surveillance data, our time frame for estimates began in May 2020 in states where we believe the surveillance systems were established and providing stable data after being set up in the early months of the pandemic. Therefore, we cannot estimate cumulative hospitalizations since the start of the pandemic. Third, we assume that COVID-NET captures all patients who were tested for COVID-19 and had a positive result. Although we adjusted for testing practices (ie, those not tested), we could be underestimating hospitalizations if this assumption is not true and confirmed positives are not being reported. Fourth, we assumed that testing practices did not differ by states, except in Connecticut where testing practice data for COVID-NET sites were available. This assumption could result in either an over- or underestimation of hospitalizations. In addition, we assumed testing sensitivity for COVID-19 in COVID-NET was 0.885, which can lead to an over- or underestimation of hospitalizations depending on true sensitivity. We also did not adjust for false positives because the reported specificity for tests in COVID-NET is extremely high [11]. However, this could also lead to an overestimation of hospitalizations. Fifth, our method assumes that the COVID-NET sites are representative of the entire state. In some states, such as Maryland, COVID-NET includes all counties; in other states, such as Iowa, it includes only 1 county. Although the model accounted for uncertainty and variability between states, we are still limited by representativeness within a state between the COVID-NET site and the truth of the entire state. As a result, our model may be under- or overestimating hospitalizations at the state level for COVID-NET states depending on how well the particular catchment area reflects COVID-19 activity in the state. Sixth, our method assumes that COVID-NET states capture enough diversity across the nation to extrapolate data to all states, which may not be true. Although the 14 states from COVID-NET vary in many ways, we cannot be sure that they cover the variation in COVID-19 hospitalizations, including variations in things that may impact hospitalizations like mitigation strategies and vaccination rates. For example, we could not extrapolate to Washington DC or New York City appropriately due to the extreme variation between a state and a purely metropolitan city. Seventh, although the covariates are meant to inform the extrapolation, the covariates are limited by the quality, completeness, and availability of the data. There could be vital information around COVID-19 hospitalization rates that are missing, such as other chronic conditions, underlying risk factors in the population, mitigation measures, and vaccination rates. Although our model has time-varying covariates that describe the COVID-19 impact in each state, including percent positive, percent COVID-19 deaths, and hospital capacity covariates, vaccination rates were not included so we may be under- or overestimating age groups and states based on potential unaccounted variation from the correlation to vaccination rates. Another limitation is the wide CrIs. Median estimates from the model's output distributions of hospitalizations seem to be reasonable through our sensitivity, validation, and comparison analysis, but the 90% CrIs are wide for some of the states where extrapolation was carried out. This

limits the precision of true hospitalizations and inference of medians presented. Finally, since we ran a different model for each age group, we are limited in the interpretation of hospitalization estimates by month and state since combining models' outputs may underestimate variability and does not capture correlations between age groups. Although we calculated hospitalizations by month and state, combined variance is unknown, so CrIs may be wider than reported.

In conclusion, we estimated that about 4 million COVID-19 hospitalizations occurred in the United States from May 2020 through April 2021. As COVID-19 continues to circulate and cause illness, it will be important to develop a sustainable method to continue to estimate the disease burden of COVID-19 that can account for regional variation in timing and incidence of disease activity as well as changes in detection and reporting of COVID-19 and that utilizes ongoing surveillance data. With an unknown future of COVID-19, burden estimates will continue to be needed. Having a burden estimation method that uses a sentinel surveillance system ensures we will have the ability to create burden estimates despite changes in case data reporting. Knowing disease burden helps us understand vaccine-averted burden, post–COVID-19 conditions, and more important public health research. Our method leverages routine surveillance data that are expected to continue after the pandemic and a Bayesian hierarchical modeling approach as a novel way to continue estimating COVID-19 hospitalizations. The model offers an approach that will be useful not only to COVID-19 hospitalization estimations but also to other levels of the disease burden pyramid, including SARS-CoV-2 infections and COVID-19 deaths.

Acknowledgments

Funding for this work was supported by the Centers for Disease Control and Prevention (CDC; Atlanta, GA). The authors received no financial support for the research, authorship, or publication of these data.

Authors' Contributions

AC, ADI, HHC, NNP, MG, MS, and CR were involved in the study conceptualization. AC, ADI, NNP, MG, MW, and CR analyzed the primary and input data. AC and HHC coded the model, and AC was responsible for running all model analyses. AC, ADI, NNP, MG, MS, FPH, MW, and CR validated the data used in the analysis. AC and CR led the initial drafting of the manuscript. All authors contributed ideas for analysis and manuscript edits and subsequent drafts. All authors had access to the underlying data, have seen and approved the final manuscript, and were responsible for the decision to submit.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Covariates selected for each Bayesian model for extrapolation of COVID-19 hospitalizations for all 50 US states by age group out of all covariates. For the 0-17 years age group, only asthma was included as a possible covariate from the chronic conditions/diseases. CKD: chronic kidney disease, COPD: chronic obstructive pulmonary disease, ICU: intensive care unit, inpat: inpatients.
[DOCX File , 18 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Comparison of COVID-19 hospitalization rates per 100,000 population and 90% credible intervals (error bars) from our Bayesian model by age group for each US state in COVID-NET showing observed rate (COVID-NET input rate), estimated rate (final model), and extrapolated rate (dropped). Y-axis limits adjust to the unique minimum and maximum rate for each age group.
[DOCX File , 2181 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Comparison of COVID-19 hospitalization estimates between our Bayesian model and case-based multiplier model by age group, months, and HHS regions, including distribution of hospitalization for each group from June 2020 through March 2021.
[DOCX File , 21 KB-Multimedia Appendix 3]

References

1. Coronavirus Disease 2019 (COVID-19) 2020 Interim Case Definition, Approved August 5, 2020. Centers for Disease Control and Prevention. URL: https://ndc.services.cdc.gov/case-definitions/coronavirus-disease-2019-2020-08-05/ [accessed 2021-08-08]
2. Coronavirus Disease 2019 (COVID-19) 2020 Interim Case Definition, Approved April 5, 2020. Centers for Disease Control and Prevention. URL: https://ndc.services.cdc.gov/case-definitions/coronavirus-disease-2019-2020/ [accessed 2021-08-08]

3. COVID-19 Death Data and Resources. Centers for Disease Control and Prevention. 2021 Feb 05. URL: https://www.cdc.gov/nchs/nvss/covid-19.htm [accessed 2021-08-08]
4. Hospital Reporting. HHS Protect Public Data Hub. URL: https://protect-public.hhs.gov/pages/hospital-reporting [accessed 2021-08-08]
5. Reese H, Iuliano AD, Patel NN, Garg S, Kim L, Silk BJ, et al. Estimated Incidence of Coronavirus Disease 2019 (COVID-19) Illness and Hospitalization-United States, February-September 2020. Clin Infect Dis 2021 Jun 15;72(12):e1010-e1017 [FREE Full text] [doi: 10.1093/cid/ciaa1780] [Medline: 33237993]
6. Estimated COVID-19 Burden. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html [accessed 2022-02-11]
7. Angulo FJ, Finelli L, Swerdlow DL. Estimation of US SARS-CoV-2 Infections, Symptomatic Infections, Hospitalizations, and Deaths Using Seroprevalence Surveys. JAMA Netw Open 2021 Jan 04;4(1):e2033706 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.33706] [Medline: 33399860]
8. Hozé N, Paireau J, Lapidus N, Tran Kiem C, Salje H, Severi G, et al. Monitoring the proportion of the population infected by SARS-CoV-2 using age-stratified hospitalisation and serological data: a modelling study. The Lancet Public Health 2021 Jun;6(6):e408-e415. [doi: 10.1016/s2468-2667(21)00064-5]
9. Coronavirus Disease 2019 (COVID-19)-Associated Hospitalization Surveillance Network (COVID-NET). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covid-net/purpose-methods.html [accessed 2021-08-08]
10. Influenza Hospitalization Surveillance Network (FluSurv-NET). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/flu/weekly/influenza-hospitalization-surveillance.htm [accessed 2021-08-08]
11. Arevalo-Rodriguez I, Buitrago-Garcia D, Simancas-Racines D, Zambrano-Achig P, Del Campo R, Ciapponi A, et al. False-negative results of initial RT-PCR assays for COVID-19: A systematic review. PLoS One 2020;15(12):e0242958 [FREE Full text] [doi: 10.1371/journal.pone.0242958] [Medline: 33301459]
12. Syndromic Data Critical to COVID-19. National Syndromic Surveillance Program. URL: https://www.cdc.gov/nssp/index.html [accessed 2021-08-08]
13. National Vital Statistics System: Guidance for Certifying COVID-19 Deaths 2020. Centers for Disease Control and Prevention. 2020 Mar 04. URL: https://www.cdc.gov/nchs/data/nvss/coronavirus/Alert-1-Guidance-for-Certifying-COVID-19-Deaths.pdf [accessed 2021-08-08]
14. National Vital Statistics System: New ICD code introduced for COVID-19 deaths. Centers for Disease Control and Prevention. 2020 Mar 24. URL: https://www.cdc.gov/nchs/data/nvss/coronavirus/Alert-2-New-ICD-code-introduced-for-COVID-19-deaths.pdf [accessed 2021-08-08]
15. National Center for Health Statistics: Reporting and Coding Deaths Due to COVID-19. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/covid19/coding-and-reporting.htm [accessed 2022-05-26]
16. Hospital Utilization. HHS Protect Public Data Hub. URL: https://protect-public.hhs.gov/pages/hospital-utilization [accessed 2021-08-08]
17. CDC Wonder: bridged-race population estimates. Centers for Disease Control and Prevention. URL: https://wonder.cdc.gov/bridged-race-population.html [accessed 2021-08-08]
18. Razzaghi H, Wang Y, Lu H, Marshall KE, Dowling NF, Paz-Bailey G, et al. Estimated County-Level Prevalence of Selected Underlying Medical Conditions Associated with Increased Risk for Severe COVID-19 Illness - United States, 2018. MMWR Morb Mortal Wkly Rep 2020 Jul 24;69(29):945-950. [doi: 10.15585/mmwr.mm6929a1] [Medline: 32701937]
19. Clift AK, Coupland CAC, Keogh RH, Diaz-Ordaz K, Williamson E, Harrison EM, et al. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. BMJ 2020 Oct 20;371:m3731 [FREE Full text] [doi: 10.1136/bmj.m3731] [Medline: 33082154]
20. Ruppert D. Trimming and Winsorization. In: Balakrishnan N, Colton T, Everitt B, Piegorsch W, Ruggeri F, Teugels JL, editors. Wiley StatsRef: Statistics Reference Online. Hoboken, NJ: Wiley Blackwell; 2006.
21. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 2018 Dec 05;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
22. Ishwaran H, Rao JS. Spike and slab variable selection: Frequentist and Bayesian strategies. Ann Statist 2005 Apr 1;33(2):1. [doi: 10.1214/009053604000001147]
23. Iuliano A, Roguski K, Chang H, Muscatello D, Palekar R, Tempia S, Global Seasonal Influenza-associated Mortality Collaborator Network. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. Lancet 2018 Mar 31;391(10127):1285-1300 [FREE Full text] [doi: 10.1016/S0140-6736(17)33293-2] [Medline: 29248255]
24. Kochanek KD, Xu J, Murphy SL, Miniño AM, Kung H. Deaths: final data for 2009. Natl Vital Stat Rep 2011 Dec 29;60(3):1-116. [Medline: 24974587]
25. COVID-19 Reported Patient Impact and Hospital Capacity by State Timeseries. Healthdata.gov. URL: https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/g62h-syeh [accessed 2021-08-08]
26. The COVID Tracking Project. URL: https://covidtracking.com/ [accessed 2021-08-08]
27. Estimated COVID-19 Burden. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html [accessed 2021-08-08]

28. COVID Data Tracker. Centers for Disease Control and Prevention. URL: https://covid.cdc.gov/covid-data-tracker/ [accessed 2021-08-08]
29. Chaves SS, Lynfield R, Lindegren ML, Bresee J, Finelli L. The US Influenza Hospitalization Surveillance Network. Emerg Infect Dis 2015 Sep;21(9):1543-1550 [FREE Full text] [doi: 10.3201/eid2109.141912] [Medline: 26291121]

Abbreviations

BRFSS: Behavioral Risk Factor Surveillance System
CDC: Centers for Disease Control and Prevention
COPD: chronic obstructive pulmonary disease
COVID-NET: COVID-19-Associated Hospitalization Surveillance Network
CrI: credible interval
FluSurv-NET: Influenza Hospitalization Surveillance Network
HHS: Department of Health and Human Services
ICU: intensive care unit
LASSO: Least Absolute Shrinkage and Selection Operator
MCMC: Markov chain Monte Carlo
NNDSS: National Notifiable Disease Surveillance System
NVSS: National Vital Statistics System

Edited by T Sanchez; submitted 15.10.21; peer-reviewed by S Rigdon, N Fenton, J Fitzner; comments to author 09.12.21; revised version received 21.12.21; accepted 21.04.22; published 02.06.22

Please cite as:
Couture A, Iuliano AD, Chang HH, Patel NN, Gilmer M, Steele M, Havers FP, Whitaker M, Reed C
Estimating COVID-19 Hospitalizations in the United States With Surveillance Data Using a Bayesian Hierarchical Model: Modeling Study
JMIR Public Health Surveill 2022;8(6):e34296
URL: https://publichealth.jmir.org/2022/6/e34296
doi: 10.2196/34296
PMID: 35452402