Am. J. Trop. Med. Hyg., 104(4), 2021, pp. 1484–1492
doi:10.4269/ajtmh.20-1147

A Bayesian Mixture Model for Predicting the COVID-19 Related Mortality in the United States

Niko A. Kaciroti,[1,2]* Carey Lumeng,[3] Vikas Parekh,[4] and Matthew L. Boulton[4,5]

[1]*Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan;* [2]*Department of Pediatrics, Michigan Medicine, University of Michigan, Ann Arbor, Michigan;* [3]*Pediatrics-Pulmonary Medicine, Michigan Medicine, University of Michigan, Ann Arbor, Michigan;* [4]*Department of Internal Medicine, Michigan Medicine, University of Michigan, Ann Arbor, Michigan;* [5]*Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan*

Abstract. An outbreak of SARS-CoV-2 has led to a global pandemic affecting virtually every country. As of August 31, 2020, globally, there have been approximately 25,500,000 confirmed cases and 850,000 deaths; in the United States (50 states plus District of Columbia), there have been more than 6,000,000 confirmed cases and 183,000 deaths. We propose a Bayesian mixture model to predict and monitor COVID-19 mortality across the United States. The model captures skewed unimodal (prolonged recovery) or multimodal (multiple surges) curves. The results show that across all states, the first peak dates of mortality varied between April 4, 2020 for Alaska and June 18, 2020 for Arkansas. As of August 31, 2020, 31 states had a clear bimodal curve showing a strong second surge. The peak date for a second surge ranged from July 1, 2020 for Virginia to September 12, 2020 for Hawaii. The first peak for the United States occurred about April 16, 2020—dominated by New York and New Jersey—and a second peak on August 6, 2020—dominated by California, Texas, and Florida. Reliable models for predicting the COVID-19 pandemic are essential to informing resource allocation and intervention strategies. A Bayesian mixture model was able to more accurately predict the shape of the mortality curves across the United States than other models, including the timing of multiple peaks. However, given the dynamic nature of the pandemic, it is important that the results be updated regularly to identify and better monitor future waves, and characterize the epidemiology of the pandemic.

INTRODUCTION

An outbreak of SARS-CoV-2 has led to worldwide spread, affecting virtually every country globally with approximately 25,500,000 confirmed cases and more than 850,000 deaths as of August 31, 2020 based on WHO reporting.[1] In the United States alone (50 states plus District of Columbia), there have been more than 6,000,000 confirmed cases and 183,00 deaths as of the end of August. The pandemic has resulted in substantial disruptions to people's lives. At various points, more than 3 billion people throughout the world have been under various lockdown orders.[2] The application of these orders has varied widely across countries and within countries.[3] Public health countermeasures to interrupt and control transmission rely on predictive models and an understanding of disease dynamics.[4,5] Because disease is acquired through asymptomatic transmission in most of the cases, the epidemiology of the pathogen is not fully understood. It has been firmly established that it is readily transmitted via droplet nuclei, but airborne transmission also likely plays a role.[6,7] Understanding the transmission dynamics of the pandemic is crucial for informing decisions regarding resource allocation, in instituting control measures, and in assessing their effectiveness as a mitigation strategy.

The COVID-19 pandemic has spawned the development of a large number of predictive mathematical models.[8] Two commonly used approaches are based on transmission models[9–11] and curve-fitting models.[12] Transmission models simulate how quickly an infection can spread in a community that is immunologically naive, based on a number of initial assumptions. Although such models are useful, they are based on parameters that are hard to determine, and therefore are sensitive to initial values and assumptions. Consequently, the results can vary greatly (Imperial College model[10] versus Oxford model[11]) and substantially overestimate or underestimate the full extent of an outbreak.[13] Curve fitting models use available COVID-19 data to determine if trends exist, and project future disease trajectories by extrapolation (Institute for Health Metrics and Evaluation [IHME] COVID-19 health service utilization forecasting team[12]). Although such models can be useful for short-term prediction, their long-term projections will vary depending on the type of the curve used for extrapolation in addition to the impact of new and unforeseen factors, including, for example, the development of effective vaccines, virus mutations, or changes in government interventions including stay-at-home orders or prematurely phasing out of such orders.

We use data from the United States to inform and monitor COVID-19 mortality in 50 states and the District of Colombia using a Bayesian curve fitting model. Data on the number of confirmed cases are confounded by changes in test availability. We used mortality data as a more reliable measure for modeling pandemic progress over time. Our approach uses a Bayesian modeling framework[14] for modeling and predicting daily mortality, and subsequently deriving the cumulative mortality projections over time. The Bayesian framework allows for updating prior knowledge about the quantity of interest using the observed data and calculates posterior distribution for the quantity of interest. Bayesian inferences are derived from the posterior distributions of quantities of interest, which are used for projections and their corresponding credible intervals. In addition, the Bayesian framework provides computational power via the Markov chain Monte Carlo methodology to provide exact estimate of the quantity of interest, rather than using approximate optimization algorithms.

The Bayesian model is applied to COVID-19 mortality data in the United States, but can be used in a similar manner for predicting other COVID-19 measures, including the number of confirmed cases, the number of COVID-19–related hospitalizations, and healthcare utilizations.

MATERIALS AND METHODS

Curve fitting models are useful mathematical models for predicting the trajectory of pandemics over time.[8] A commonly used curve for such models is a bell-shaped curve

* Address correspondence to Niko A. Kaciroti, University of Michigan, 300 North Ingalls Bldg., 10-th Floor #1035NW, Ann Arbor, MI 48109. E-mail: nicola@med.umich.edu

defined by the Gaussian function: $f(t) = P*\exp(-[t-t_{peak}]^2/[2\sigma^2])$, where t_{peak} is the time to peak, P is the magnitude of the peak, and σ captures the width of the curve. Alternatively, $f(t)$ can be expressed as $f(t) = (P\sqrt{2\pi}\,\sigma)\frac{1}{\sqrt{2\pi}\sigma}\exp(-[t-t_{peak}]^2/[2\sigma^2]) = M*\phi(t;\ t_{peak}\ \sigma)$, where $M = P\sqrt{2\pi}\sigma$ is the overall expected mortality and $\phi(t; t_{peak,}\ \sigma)$ is the normal density function, which governs how the overall mortality M is distributed over time; that is, expected mortality at time t is $M^*\phi(t; t_{peak,}\ \sigma)$. The initial IHME model[12] uses a Gaussian sigmoidal function for forecasting the cumulative mortality data (which is equivalent to a Gaussian function for the daily mortality data). In the IHME model, the parameters of the curvature function are estimated based on a regression model with a normal distribution on the log-transform cumulative deaths as outcome.[12] For a curve fitting model to perform well, the shape of the curve chosen is important; it must have a sound fit with the observed data and an appropriate underlying theoretical justification. The curve defined by a Gaussian function has a symmetric shape around the peak, and future trends are forecasted by extrapolating the observed trend forcing the post-peak component of the trajectory to be symmetric with the pre-peak portion. These forecasts assume that factors related to the pandemic do not change over time, thus are suitable when the spread of the pandemic is relatively homogeneous with no mitigating measures of behavioral changes. Although the assumption of homogeneity over time might be reasonable over short time periods, it is unrealistic for long time periods and large areas, as changes in the pandemic progression will result in behavioral changes at the individual level, and policy and practice changes at the local, state, and federal levels (i.e., social distancing and stay-at-home orders). Such changes have the potential to alter, sometimes substantially, the natural arc of the disease course. To account for such changes, we propose a model where the curve for the daily mortality trajectory $\mu(t)$ at day t is expressed as a mixture of multiple homogeneous sub-curves. Each sub-curve can be seen as implicitly capturing a sub-epidemic or a specific trend in the trajectory during a given time period, which relates to underlying factors (known or unknown) for that span of time. This is particularly relevant for the COVID-19 pandemic in the United States, where, for example, changes in compliance with social distancing in late May early June 2020 resulted in a surge of cases in a large number of states.

Specifically, let the mortality at day t be comprised as $y(t) = \sum_{i=1}^{K} y_i(t)$, where $y_i(t)$ represents the number of deaths at day t attributed to a surge or sub-epidemic indexed by $i = 1,\ldots,K$, $K \geq 1$ is the number of sub-epidemics empirically identified, and $E(y_i[t]) = \mu_i(t)$. The trajectory $\mu_i(t)$ for surge i can be modeled by a homogeneous Gaussian function $\mu_i(t) = M_i^*\phi(t; t_{peak,i,}\ \sigma_i)$, where M_i is the overall mortality, $t_{peak,i}$ is the time to peak, and σ_i represents the width for surge i. The mortality curve for the whole pandemic $E(y[t]) = \mu(t)$ is then decomposed as:

$$\mu(t) = \sum_{i=1}^{K}\mu_i(t) = \sum_{i=1}^{K} M_i\phi\left(t, t_{peak,i}, \sigma_i\right).$$

Let $M = \sum_{i=1}^{K} M_i$ be the overall mortality of the pandemic, then $\pi_i = M_i/M$ is the proportion of all deaths attributed to surge i, with $M_i = M\,\pi_i$.

Therefore, we proposed the following mixture curve model for modeling the trajectory of daily COVID-19 mortality:

$$\mu(t) = M\sum_{i=1}^{K}\pi_i\phi\left(t, t_{peak,i}, \sigma_i\right),$$

where M is the overall mortality and $\phi(t; t_{peak,i}\ \sigma_i)$ represents the part of the curve related to some underlying factor(s), π_i is the proportion of total deaths that are attributed to such factor(s) indexed by i, and K is the number of sub-curves comprising the mixtures. The number of parameters identifying the curve is *3K: M,* $\pi_1 \ldots, \pi_{K-1,}$ $(t_{peak,1}\ \sigma_1), \ldots, (t_{peak,K}\ \sigma_K)$. The $\phi(t; t_{peak,i}\ \sigma_i)$ here is the Gaussian density function, defined by the location parameter $t_{peak,i}$, which represents the time from the first death to apex, and the scale parameter σ_i, which represents the spread of the curve. We propose a Bayesian approach for estimating the curve parameters and subsequently any other statistics of interest, with values for the parameters being drawn from their posterior distribution condition on the observed data. We assume the distribution for the observed mortality data $y(t)$ at time t to be negative binomial:

$$Y(t) \sim \text{negative-binomial}(N,\ p_t),$$

where N is the state's population size. A negative binomial distribution function is used for $Y(t)$, as it is appropriate for count data, with $E(Y[t]) = \mu(t)Np_t/(1-p_t)$. Bayesian Markov Chain Monte Carlo methods are used to make draws from the posterior distribution of the unknown parameters and derive future projections. Weakly informative prior distributions were used for all model parameters[15] (see Supplemental Material).

The number of mixtures K is selected based on choosing a parsimonious model having lower deviance information criteria. The resulting shape of the curve $\mu(t)$ can be multimodal or unimodal. For a multimodal curve, the proposed modeling identifies multiple surges (i.e., Sun Belt States, California, Florida, Texas). For unimodal curves, the proposed modeling can accommodate skewed long-tailed distributions,[16] where the curve comprises multiple sub-curves, but is dominated by one major surge or multiple surges occurring closely spaced in time (e.g., New York, New Jersey, Michigan).

RESULTS

In this section, we apply the proposed model to the COVID-19 mortality data in the United States (www.worldometers.info/coronavirus/). Because of variation in the number of death records by day of the week (particularly on weekends), we used a weekly 7-day moving average (± 3 days window) as a more reliable measure for daily mortality. All analyses were performed in SAS 9.4 (SAS Institute Inc., Cary, NC)[17] using *PROC MCMC*. We ran 500,000 iterations following 10,000 burn in and 1,000 thinnings to reduce high autocorrelation. The expected number of deaths for each day was derived based on the corresponding posterior distribution.

Based on the data available as of August 31, 2020, for most of the states, a mixture of $K = 2$ sub-curves showed a good and parsimonious fit of the data, with no substantial improvement for $K > 2$. For Arizona, Louisiana, Massachusetts, Michigan, New York, South Carolina, Tennessee, and Virginia, a mixture of $K = 3$ sub-curves improved the model fit, with no substantial improvement for $K > 3$. For the entire country, we derived the curve for daily mortality as the sum of estimates from each state plus the District of Columbia. Figure 1 displays the daily mortality curves for each state and District of Colombia and for the entire country. There is variability among curves in the context of their shape, magnitude, and timing of the

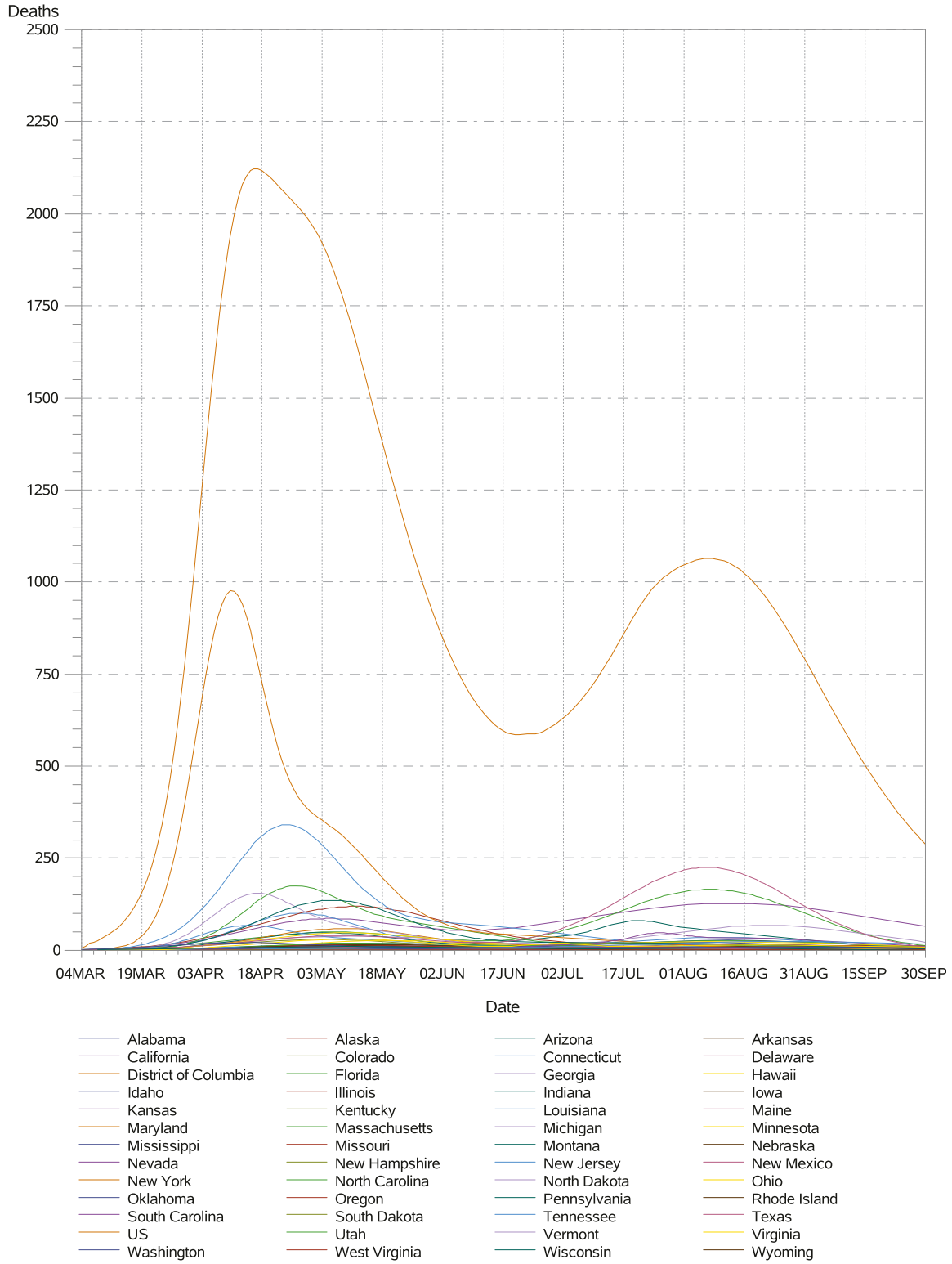

FIGURE 1. COVID-19 mortality curves for the United States and for each state.

pandemic. Daily mortality and cumulative mortality curves separately for each state and the entire nation are provided in Supplemental Figure 1. Most of the states demonstrate a bimodal curve with two major peaks. The first peak represents the initial dynamic of the pandemic, following the introduction of control measures in March, and the second or third peak captures the surge that occurred in many states after measures were phased out to varying degrees.

The estimated date of the peak for each state and the United States is shown in Table 1. The first peak date among states varied between April 4, 2020 for Alaska, and June 18, 2020 for Arkansas. Thirty-one states have a clear bimodal curve, with

TABLE 1
Projected COVID-19 mortality as of September 30, 2020 by state

State	September 30	Projected* (95%CI)	$Peak_1$	π_1 (%)	$Peak_2$	π_2 (%)	$Peak_3$	π_3 (%)
Alabama	2,540	2,478 (2,326–2,629) 2,291 (2,125–2,539)	May 5, 2020	25	August 1, 2020	75	–	–
Alaska	56	47 (40–59) 37 (34–46)	April 4, 2020	22	August 11, 2020	78	–	–
Arizona	5,650	5,513 (5,292–5,735) 5.902 (5,454–6,706)	May 7, 2020	17	July 19, 2020	10	July 28, 2020	73
Arkansas	1,369	1,215 (1,016–1,414) 1,040 (772–1,454)	June 18, 2020	30	September 12, 2020	70	–	–
California	15,900	15,906 (15,466–16,346) 17,448 (15,520–20,496)	May 2, 2020	23	August 12, 2020	77	–	–
Colorado	2,051	2,010 (1,955–2,093) 2,036 (1,967–2,169)	May 2, 2020	77	July 24, 2020	23	–	–
Connecticut	4,508	4,521 (4,468–4,604) 4,477 (4,460–5,512)	April 26, 2020	75	May 29, 2020	25	–	–
Delaware	636	620 (606–653) 618 (605–641)	May 20, 2020	94	June 25, 2020	6	–	–
District of Columbia	627	624 (609–654) 647 (637–663)	May 1, 2020	66	June 13, 2020	34	–	–
Florida	14,317	12,686 (12,367–13,006) 15,758 (13,107–20,248)	May 6, 2020	23	August 6, 2020	77	–	–
Georgia	7,021	6,992 (6,561–7,422) 7,469 (6,419–9,383)	May 10, 2020	39	August 21, 2020	61	–	–
Hawaii	136	136 (79–194) 61 (51–84)	April 13, 2020	11	September 12, 2020	89	–	–
Idaho	469	411 (372–459) 541 (408–778)	April 19, 2020	23	August 12, 2020	77	–	–
Illinois	8,916	8,719 (8,517–8,956) 9,044 (8,476–9,851)	May 10, 2020	74	August 11, 2020	26	–	–
Indiana	3,632	3,497 (3,367–3,628) 3,838 (3,616–4,183)	May 2, 2020	58	July 19, 2020	42	–	–
Iowa	1,346	1,315 (1,184–1,445) 1,542 (1,274–1,992)	May 15, 2020	50	August 22, 2020	50	–	–
Kansas	678	521 (472–575) 545 (507–620)	April 22, 2020	34	August 7, 2020	66	–	–
Kentucky	1,174	1,138 (1,042–1,235) 1,451 (1,102–2,089)	April 28, 2020	30	August 24, 2020	70	–	–
Louisiana	5,511	5,431 (5,223–5,639) 5,735 (5,331–6,334)	April 12, 2020	28	May 10, 2020	27	August 9, 2020	45
Maine	141	137 (133–154) 141 (134–154)	April 27, 2020	70	July 11, 2020	30	–	–
Maryland	3,949	3,879 (3,778–4,003) 3,880 (3,818–3,955)	May 6, 2020	69	July 11, 2020	31	–	–
Massachusetts	9,456	9,319 (9,101–9,538) 9,561 (9,320–9,929)	April 24, 2020	45	May 20, 2020	37	August 1, 2020	18
Michigan	7,083	7,013 (6,827–7,199) 7,175 (6,961–7,531)	April 15, 2020	48	May 9, 2020	33	August 12, 2020	19
Minnesota	2,089	2,048 (1,908–2,188) 2,273 (2,103–2,500)	May 18, 2020	71	August 22, 2020	29	–	–
Mississippi	2,969	2,940 (2,751–3,130) 2,969 (2,686–3,430)	May 12, 2020	31	August 12, 2020	69	–	–
Missouri	2,213	1,973 (1,849–2,097) 2,301 (1,693–3,810)	May 7, 2020	37	August 23, 2020	63	–	–
Montana	180	126 (111–150) 105 (96–123)	April 10, 2020	15	August 9, 2020	85	–	–
Nebraska	478	434 (404–483) 473 (452–501)	May 23, 2020	69	August 18, 2020	31	–	–
Nevada	1,600	1,579 (1,447–1,711) 2,050 (1,586–2,840)	April 28, 2020	28	August 13, 2020	72	–	–
New Hampshire	439	449 (432–478) 447 (436–458)	May 19, 2020	73	July 7, 2020	27	–	–
New Jersey	16,245	16,166 (16,077–16,320) 16,086 (16,010–16,177)	April 22, 2020	65	May 31, 2020	35	–	–
New Mexico	877	851 (791–919) 924 (857–1,016)	May 12, 2020	47	August 1, 2020	53	–	–
New York	33,246	33,201 (33,058–33,441) 33,125 (32,958–33,385)	April 8, 2020	52	April 29, 2020	36	June 7, 2020	12
North Carolina	3,532	3,233 (3,019–3,447) 3,412 (2,973–4,002)	May 13, 2020	36	August 13, 2020	64	–	–
North Dakota	246	164 (150–189) 208 (191–233)	May 10, 2020	45	August 13, 2020	55	–	–

(continued)

TABLE 1
Continued

State	September 30	Projected* (95%CI)	Peak$_1$	π_1 (%)	Peak$_2$	π_2 (%)	Peak$_3$	π_3 (%)
Ohio	4,821	4,624 (4,442–4,805) 4,912 (4,565–5,410)	May 8, 2020	51	August 8, 2020	49	–	–
Oklahoma	1,031	1,012 (887–1,137) 1,204 (954–1,640)	April 23, 2020	32	August 23, 2020	68	–	–
Oregon	559	564 (493–634) 590 (509–719)	April 19, 2020	23	August 15, 2020	77	–	–
Pennsylvania	8,224	7,993 (7,825–8,168) 8,515 (8,047–9,505)	May 4, 2020	70	July 8, 2020	30	–	–
Rhode Island	1,116	1,109 (1,055–1,193) 1,090 (1,065–1,132)	May 16, 2020	86	September 4, 2020	14	–	–
South Carolina	3,378	3,247 (2,936–3,558) 3,515 (3,083–4,160)	May 3, 2020	14	July 12, 2020	10	August 10, 2020	77
South Dakota	223	182 (169–206) 196 (183–212)	May 10, 2020	26	July 12, 2020	74	–	–
Tennessee	2,454	2,392 (2,160–2,623) 2,968 (2,143–4,467)	April 7, 2020	4	May 17, 2020	13	August 26, 2020	83
Texas	16,132	15,222 (13,433–19,820) 19,850 (16,450–24,306)	May 5, 2020	15	August 17, 2020	85	–	–
Utah	456	438 (414–479) 500 (452–565)	May 14, 2020	36	July 29, 2020	64	–	–
Vermont	58	62 (58–81) 59 (58–60)	April 8, 2020	78	July 30, 2020	22	–	–
Virginia	3,208	2,968 (2,652–3,574) 2,589 (2,481–2,768)	May 7, 2020	49	July 1, 2020	10	August 22, 2020	40
Washington	2,128	2,170 (2,051–2,290) 2,224 (2,129–2,350)	April 13, 2020	41	August 5, 2020	59	–	–
West Virginia	350	311 (237–400) 250 (224–284)	May 2, 2020	32	August 29, 2020	68	–	–
Wisconsin	1,327	1,215 (1,146–1,301) 1,334 (1,183–1,555)	May 9, 2020	67	August 10, 2020	33	–	–
Wyoming	50	45 (41–58) 34 (32–36)	May 12, 2020	57	August 14, 2020	43	–	–
The United States	206,796	200,839 (195,850–205,829) 215,441 (206,733–223,361)	April 16, 2020	44	August 6, 2020	56	–	–
Average bias† (%)	–		–	–	–	–	–	–
Bayesian model		5.8						
IHME model		10.6						
Median (IQR) bias†								
Bayesian		1.45% (2.8–8.0%)						
IHME model		1.75% (4.5–15.5%)						

* Projected mortality through September 30, 2020 is derived on August 31, 2020 for the Bayesian model (top estimate for each state and the United States), and on August 27, 2020 for the IHME model (bottom estimate for each state, District of Colombia, and the United States).
† Average and median (IQR) bias are derived using 52 projections: 50 states, District of Colombia, and for the United States.

the estimated second or third peak dates ranging between July 1, 2020 for Virginia and a projected date of September 12, 2020 for Hawaii. The first peak for the entire country occurred on approximately April 16, 2020—dominated by New York and New Jersey—with a second peak projected around August 6, 2020—dominated by California, Texas, and Florida. The projected overall mortality for September 30, 2020 (30 days' projection) is shown in Table 1, and a shorter term 15 days' projections for September 15 is shown in Table 2. The projected overall mortality through September 30, 2020 ranged between 45 (95% CI = [41–58]) for Wyoming, and 33,201 (95% CI = [33,059–33,441]) for New York, and 200,839 (95% CI = [195,850–205,829]) for the United States (50 states and District of Columbia). Data show that the institution of pandemic control measures had an impact that resulted flattening the curve. However, as control measures were relaxed, many states had a second surge; for example, California, Texas, and Florida are currently experiencing their second major outbreak. The proportion of two mixtures, π_1 and $\pi_2 = 1 - \pi_1$, are also shown in Table 1. Most of the states are characterized by the second outbreak, dominating the curve. For example, the majority of deaths through September 30, 2020 for California (π_2 = 77% versus π_1 = 23%), Florida (π_2 = 77% versus π_1 = 23%), and Texas (π_2 = 85% versus π_1 = 15%) are projected to occur during the second surge. For other states, the majority of deaths occurred during the first peak, with no new major surge projected through September 30, 2020, that is, New York, (π_1 = 52% and π_2 = 36%, versus π_3 = 12%), New Jersey (π_1 = 65% versus π_2 = 35%), Massachusetts (π_1 = 45% and π_2 = 37% versus π_3 = 18%), and Michigan (π_1 = 48% and π_2 = 33% versus π_3 = 19%). The majority of deaths for the entire country are projected during the second surge (π_2 = 56% versus π_1 = 44%); however, the first peak was more severe—more than 2,250 deaths/day—whereas the second surge has a lower peak—around 1,200 deaths/day—but is of longer duration.

Next, we evaluate the performance of the proposed Bayesian mixture model by comparing the projections based on the proposed Bayesian model to projections based on the widely used IHME hybrid model[18] (http://www.healthdata.org/covid/data-downloads) updated on August 27, 2020, representing the last update in August. Our projections are derived on August 31, 2020; however, following the revision of our

TABLE 2
Projected COVID-19 mortality as of September 15, 2020 by state

State	September 15	Projected* (95% CI)	$Peak_1$	π_1 (%)	$Peak_2$	π_2 (%)	$Peak_3$	π_3 (%)
Alabama	2,387	2,375 (2,261–2,490)	May 5, 2020	25	August 1, 2020	75	–	–
		2,198 (2,097–2,333)						
Alaska	44	44 (40–54)	April 4, 2020	22	August 11, 2020	78	–	–
		36 (33–41)						
Arizona	5,344	5,377 (5,2051–5,548)	May 7, 2020	17	July 19, 2020	10	July 28, 2020	73
		5.532 (5,262–5,948)						
Arkansas	1,010	1,021 (926–1,117)	June 18, 2020	30	September 12, 2020	70	–	–
		881 (745–1,065)						
California	14,615	14,750 (14,459–15,041)	May 2, 2020	23	August 12, 2020	77	–	–
		15,186 (14,243–16,667)						
Colorado	1,996	1,990 (1,955–2,070)	May 2, 2020	77	July 24, 2020	23	–	–
		1,987 (1,951–2,053)						
Connecticut	4,485	4,521 (4,468–4,604)	April 26, 2020	75	May 29, 2020	25	–	–
		4,452 (4,443–4,470)						
Delaware	618	619 (606–662)	May 20, 2020	94	June 25, 2020	6	–	–
		608 (601–620)						
District of Columbia	627	623 (609–653)	May 1, 2020	66	June 13, 2020	34	–	–
		626 (622–632)						
Florida	12,788	12,344 (12,073–12,615)	May 6, 2020	23	August 6, 2020	77	–	–
		13,615 (12,247–15,857)						
Georgia	6,398	6,531 (6,287–6,776)	May 10, 2020	39	August 21, 2020	61	–	–
		6,450 (5,954–7,318)						
Hawaii	100	105 (79–135)	April 13, 2020	11	September 12, 2020	89	–	–
		59 (50–74)						
Idaho	423	401 (372–443)	April 19, 2020	23	August 12, 2020	77	–	–
		439 (376–535)						
Illinois	8,564	8,527 (8,327–8,726)	May 10, 2020	74	August 11, 2020	26	–	–
		8,471 (8,220–8,779)						
Indiana	3,460	3,422 (3,332–3,543)	May 2, 2020	58	July 19, 2020	42	–	–
		3,569 (3,464–3,714)						
Iowa	1,234	1,238 (1,154–1,321)	May 15, 2020	50	August 22, 2020	50	–	–
		1,298 (1,184–1,475)						
Kansas	560	492 (472–538)	April 22, 2020	34	August 7, 2020	66	–	–
		495 (476–531)						
Kentucky	1,074	1,050 (978–1,121)	April 28, 2020	30	August 24, 2020	70	–	–
		1,155 (1,008–1,375)						
Louisiana	5,278	5,281 (5,119–5,443)	April 12, 2020	28	May 10, 2020	27	August 9, 2020	45
		5,333 (5,125–5,613)						
Maine	137	136 (133–153)	April 27, 2020	70	July 11, 2020	30	–	–
		135 (132–141)						
Maryland	3,849	3,839 (3,778–3,959)	May 6, 2020	69	July 11, 2020	31	–	–
		3,811 (3,778–3,849)						
Massachusetts	9,225	9,206 (9,077–9,407)	April 24, 2020	45	May 20, 2020	37	August 1, 2020	18
		9,296 (9,180–9,455)						
Michigan	6,932	6,903 (6,791–7,074)	April 15, 2020	48	May 9, 2020	33	August 12, 2020	19
		6,932 (6,836–7,077)						
Minnesota	1,979	1,984 (1,889–2,085)	May 18, 2020	71	August 22, 2020	29	–	–
		2,045 (1,978–2,130)						
Mississippi	2,734	2,777 (2,649–2,906)	May 12, 2020	31	August 12, 2020	69	–	–
		2,723 (2,561–2,967)						
Missouri	1,866	1,831 (1,739–1,922)	May 7, 2020	37	August 23, 2020	63	–	–
		1,860 (1,609–2,358)						
Montana	140	120 (111–140)	April 10, 2020	15	August 9, 2020	85	–	–
		102 (95–113)						
Nebraska	436	422 (404–460)	May 23, 2020	69	August 18, 2020	31	–	–
		432 (421–445)						
Nevada	1,482	1,492 (1,397–1,587)	April 28, 2020	28	August 13, 2020	72	–	–
		1,665 (1,446–2,003)						
New Hampshire	438	448 (432–477)	May 19, 2020	73	July 7, 2020	27	–	–
		437 (431–442)						
New Jersey	16,166	16,164 (16,077–16,319)	April 22, 2020	65	May 31, 2020	35	–	–
		16,038 (15,992–16,086)						
New Mexico	830	824 (791–883)	May 12, 2020	47	August 1, 2020	53	–	–
		848 (816–888)						
New York	33,141	33,192 (33,059–33,432)	April 8, 2020	52	April 29, 2020	36	June 7, 2020	12
		33,011 (32,916–33,141)						
North Carolina	3,127	3,056 (2,921–3,192)	May 13, 2020	36	August 13, 2020	64	–	–
		3,064 (2,838–3,346)						
North Dakota	172	158 (150–178)	May 10, 2020	45	August 13, 2020	55	–	–

(continued)

TABLE 2
Continued

State	September 15	Projected* (95% CI)	$Peak_1$	π_1 (%)	$Peak_2$	π_2 (%)	$Peak_3$	π_3 (%)
		175 (167–186)						
Ohio	4,511	4,442 (4,297–4,586)	May 8, 2020	51	August 8, 2020	49	–	–
		4,531 (4,356–4,763)						
Oklahoma	912	927 (847–1,007)	April 23, 2020	32	August 23, 2020	68	–	–
		988 (874–1,175)						
Oregon	519	520 (470–573)	April 19, 2020	23	August 15, 2020	77	–	–
		518 (478–576)						
Pennsylvania	7,961	7,921 (7,825–8,092)	May 4, 2020	70	July 8, 2020	30	–	–
		8,059 (7,849–8,441)						
Rhode Island	1,090	1,085 (1,055–1,148)	May 16, 2020	86	September 4, 2020	14	–	–
		1,061 (1,049–1,077)						
South Carolina	3,098	3,074 (2,894–3,254)	May 3, 2020	14	July 12, 2020	10	August 10, 2020	77
		3,146 (2,909–3,465)						
South Dakota	184	176 (169–199)	May 10, 2020	26	July 12, 2020	74	–	–
		182 (175–191)						
Tennessee	2,127	2,121 (1,955–2,247)	April 7, 2020	4	May 17, 2020	13	August 26, 2020	83
		2,323 (1,932–2,917)						
Texas	14,717	14,487 (13,433–16,418)	May 5, 2020	15	August 17, 2020	85	–	–
		16,321 (14,616–18,419)						
Utah	436	431 (414–468)	May 14, 2020	36	July 29, 2020	64	–	–
		454 (431–486)						
Vermont	58	61 (58–78)	April 8, 2020	78	July 30, 2020	22	–	–
		58 (58–59)						
Virginia	2,839	2,810 (2,652–3,038)	May 7, 2020	49	July 1, 2020	10	August 22, 2020	40
		2,540 (2,471–2,643)						
Washington	2,015	2,070 (1,974–2,167)	April 13, 2020	41	August 5, 2020	59	–	–
		2,079 (2,033–2,143)						
West Virginia	280	276 (237–327)	May 2, 2020	32	August 29, 2020	68	–	–
		221 (207–239)						
Wisconsin	1,220	1,222 (1,146–1,267)	May 9, 2020	67	August 10, 2020	33	–	–
		1,229 (1,149–1,314)						
Wyoming	46	44 (41–55)	May 12, 2020	57	August 14, 2020	43	–	–
		34 (32–36)						
The United States	195,660	194,904 (192,696–201,995)	April 16, 2020	44	August 6, 2020	56	–	–
		198,702 (194,462–202,382)						
Average bias† (%)	–	–	–	–	–	–	–	–
Bayesian model		2						
IHME model		5.5						
Median (IQR) bias†		–						
Bayesian model		0.4% (1.0–2.25%)						
IHME model		0.8% (1.65–6.0%)						

* Projected mortality through September 15, 2020 is derived on August 31, 2020 for the Bayesian model (top estimate for each state and the United States), and on August 27, 2020 for the IHME model (bottom estimate for each state, District of Columbia, and for the United States).
† Average and median (IQR) bias are derived based on 52 projections: 50 states, District of Colombia, and for the United States.

article, the mortality data as of September 30, 2020 have become available; therefore, we include these data in Tables 1 and 2. Consequently, we also evaluate the performance of each projection based on the Bayesian model and IHME model by calculating the bias and the mean square error (mean square error $[\text{MSE}] \approx \text{bias}^2 + [\{\text{upper bound} - \text{lower bound}\}/4]^2$). The results from both models are similar, with the Bayesian model having lower bias and MSE in 35 of 52 projections for September 15, 2020, and lower bias in 35 and lower MSE in 38 of 52 projections for September 30, 2020. The average (median, IQR) bias for September 15 and September 30 projections were 2% (median = 0.4%, IQR = 1–2.25%) and 5.8% (median = 1.45%, IQR = 2.8–7.96%), respectively, based on the Bayesian model versus 5.5% (median = 0.8%, IQR = 1.65–6.0%) and 10.6% (median = 1.75%, IQR = 4.5–15.5%) based on the IHME hybrid model.

DISCUSSIONS

The novel coronavirus, SARS-CoV-2, has caused an unprecedented global public health crisis, with the pandemic spreading to virtually every country worldwide in less than a year and accompanied by overwhelming levels of related morbidity and mortality. Predictive models continue to have a fundamental role to play in estimating the future burden of disease and in informing the allocation of critical laboratory, medical, and public health resources needed to successfully interrupt and eventually control the pandemic. We propose a Bayesian mixture model, which can capture multiple surges or sub-epidemics attributed to a number of different underlying factors, including the introduction and phasing out of control measures.

As of August 31, 2020, a combination of two or three subcurves provided a parsimonious good fit for modeling daily mortality curve among all states in the United States through September 30, 2020. The results showed a second surge for some states and a prolonged recovery for others. For many states (e.g., Arizona, California, Florida, and Texas), most of the cases occurred in the second or third peak characterized by a major surge starting in late July. Other states experienced only a single major peak, but the distribution of mortality was skewed with a long tail end of the distribution (e.g., New York,

New Jersey, and Michigan). Importantly, the mixture modeling approach accommodates the fit of both multimodal and unimodal skewed distribution as shown in Supplemental Figure 1. The shapes of the mortality curves reveal that even for states that have successfully lowered mortality relative to its peak, it remains consistently greater than zero with a long tail or is even increasing. This is an indication of how challenging it will be to eradicate the pandemic or to reduce the risk of new surges if control measures are phased out too quickly.

There is limited information to inform how a post-peak world will appear. At this point, we lack sufficient data on numerous parameters, including duration of immunity, the degree of public compliance with social distancing over time, and the political and governmental response to COVID-19, among others.[8] A gradual and data-driven relaxation of restrictions accompanied by continuous monitoring is necessary to avert an exponential increase in the cumulative number of cases.[19] Alterations in social mixing patterns and increased contact among susceptible individuals will clearly result in ongoing challenges to achieving control of the pandemic.[20]

Our monthly predictions run through September 30, 2020, at which point the number of projected deaths is low for many states. However, as children return to school, lockdown orders expire, social distancing behaviors are relaxed, and individuals engage in greater social mixing including traveling during the holidays; there is likely to be a prolongation of transmission potentially accompanied by new surges and an overall increase in COVID-19 mortality. This also serves to underscore the importance of regularly updating model projections using an appropriate number of mixtures to capture new surges as they occur. Mathematical models can play a key role in better understanding the course of the pandemic. However, it is also important to be familiar with their underlying assumptions, strengths, and limitations. Given the dynamic and rapidly changing nature of the pandemic, any long-term projections will be sensitive to unforeseen changes. As such, these models are most reliable at shorter term monthly projections, and for monitoring trends, which inform planning for optimal management and distribution of resources, and evaluating the impact of control measures on the pandemic. Conversely, long-term projections for number of cases or deaths are sensitive to even small daily changes as these can translate into larger cumulative changes. This does not necessarily speak to model shortcomings as much as it confirms the dynamic nature of the disease transmission and changes in factors related to it. Specifically, in the last several months (after this article was submitted), two vaccines against COVID-19 were developed by Pfizer and Moderna. Although both vaccines are highly effective, there are many logistic challenges to achieve a high rate of vaccination. In addition, new strains of the virus are occurring, and it is hard to know what the new strains will look like in months from now or how resistant they will be to the vaccines.

The proposed Bayesian mixture model is an effective tool for monitoring the pandemic over time and consequently provides monthly projections. Such model can and should be used in a rolling bases as new data come in. Whereas updating estimates using additional data is helpful, constant changes of the model used for prediction introduce the risk of overfitting the observed data, and potentially give rise to inconsistent projections. Any update of a model should be guided by theory that may include using different numbers of mixtures based on the data or the occurrence of new factors affecting the pandemic that may result in new surges.

The quality of the model prediction will also depend on the availability of data. In the early stages of an epidemic, or even the early post-peak phases, data are often limited, with a weak data signal relative to the noise. Consequently, any model projections will be more sensitive to initial assumptions or prior information. Our Bayesian approach can accommodate different levels of prior knowledge and uncertainty into the model, such as information from other countries by introducing informative prior distributions. In general, using weakly informative priors is preferred,[16] as they have low impact in early projections that quickly fade away while more data become available, and in return, they improve model convergence. For the current modeling, we did not use informative priors in any of the model parameters.

In summary, Bayesian mixture models are useful for monitoring and predicting COVID-19–related mortality in the United States or globally. These models are particularly helpful for identifying multiple surges and forecasting trajectories of skewed and multimodal curves. The results for the United States based on data as of August 31, 2020 showed that many states are experiencing a second surge, which for many is of greater magnitude than the first. Our model was able to more accurately characterize the actual bimodal shape of the pandemic mortality curves through September 30 for many states or unimodal but skewed curves reflecting the prolonged recovery for other states like New York.

We are running our model regularly using the most updated data; the model performs well and is able to capture the new surge (after August 31, 2020) by increasing the number of mixtures.

Identifying and monitoring the dynamic or multiple surges is important to understanding why such sub-epidemics occurred, and to inform future policy and practice decisions to more effectively prevent them. Moreover, providing regular pandemic forecasts is needed to guide the introduction or phasing out of programmatic interventions intended to control transmission in addition to providing an evidence-based decision-making for optimal resource allocation to address feature health needs.

Received September 5, 2020. Accepted for publication February 9, 2021.

Published online February 19, 2021.

Note: Supplemental material appears at www.ajtmh.org.

Acknowledgments: Publication charges for this article were waived due to the ongoing pandemic of COVID-19.

Authors' addresses: Niko A. Kaciroti, Carey Lumeng, Vikas Parekh, and Matthew L. Boulton, University of Michigan, Ann Arbor, MI, E-mails: nicola@med.umich.edu, clumeng@umich.edu, viparekh@med.umich.edu, and mboulton@umich.edu.

REFERENCES

1. World Health Organization, 2020. *Coronavirus Disease (COVID-19) Pandemic*. Geneva, Switzerland: WHO. Available at: www.who.int/emergencies/diseases/novel-coronavirus-2019. Accessed September 1, 2020.
2. Buchholz K, 2020. *Infographic: What Share of the World Population is Already on COVID-19 Lockdown?* Statista Infographics. Available at:

http://www.statista.com/chart/21240/enforced-covid-19-lockdowns-by-people-affected-per-country/. Accessed August 20, 2020.
3. Secon H, 2020. *An Interactive Map of the US Cities and States Still Under Lockdown — and Those that Are Reopening*. Business Insider. Available at: http://www.businessinsider.com/us-map-stay-at-home-orders-lockdowns-2020-3. Accessed June 7, 2020.
4. Pan A et al., 2020. Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China. *JAMA 323:* 1915–1923.
5. Harapan H, Itoh N, Yufika A, Winardi W, Keam S, Te H, Megawati D, Hayati Z, Wagner AL, Mudatsir M, 2020. Coronavirus disease 2019 (COVID-19): a literature review. *J Infect Public Health 13:* 667–663.
6. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J, 2020. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science 368:* 489–493.
7. Bi Q et al., 2020. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *Lancet Infect Dis 20:* 911–919.
8. Jewell NP, Lewnard JA, Jewell BL, 2020. Predictive mathematical models of the COVID-19 pandemic underlying principles and value of projections. *JAMA 323:* 1893–1894.
9. Tolles J, Luong T, 2020. Modeling epidemics with compartmental. *JAMA 323:* 2515–2516.
10. Ferguson MN et al., 2020. On behalf of the Imperial College COVID-19 response Team. *Report 9: Impact of Non-pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand*. Imperial College COVID-19 Response Team, London.
11. Lourenço J, Robert P, Ghafari M, Kraemer M, Thompson C, Simmonds P, Klenerman P, Gupta S, 2020. Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic. *medRxiv*. doi: 10.1101/2020.03.24.20042291.
12. IHME COVID-19 health service utilization forecasting team, Murray CJL, 2020. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. *medRxiv*. doi: 10.1101/2020.04.21.200074732.
13. Butler D, 2014. Models overestimate Ebola cases. *Nature 515:* 18.
14. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB, 2013. *Bayesian Data Analysis*, 3rd Edition. Boca Raton, FL: Chapman & Hall/CRC Text in Statistical Science.
15. Gelman A, Simson D, Betancourt M, 2017. The prior can often only Be understood in the context of the likelihood. *Entropy 19:* 555.
16. Schork NJ, Schork MA, 1988. Skewness and mixture of normal distributions. *Commun Stat Theor Methods 17:* 3951–3969.
17. SAS Institute Inc, 2013. *Base SAS 9.4 Utilities*. Cary, NC: SAS Institute Inc.
18. IHME COVID-19 Forecasting Team, Hay SI, 2021. COVID-19 scenarios for the United States. *Nature Medicine 27:* 94–105.
19. Leung K, Wu JT, Liu D, Leung G, 2020. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. *Lancet 395:* 1382–1393.
20. Zhang J et al., 2020. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science 368:* 1481–1486.