Quantitative assessment of the effects of resource optimization and ICU admission policy on COVID-19 mortalities

Lang Zeng,[1,*] Ying-Qi Zeng,[2,*] Ming Tang,[1,3,†] Ying Liu,[4] Zonghua Liu,[1] and Ying-Cheng Lai[5,6,‡]

[1]*School of Physics and Electronic Science, East China Normal University, Shanghai 200241, China*
[2]*School of Communication and Electronic Engineering, East China Normal University, Shanghai 200241, China*
[3]*Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China*
[4]*School of Computer Science, Southwest Petroleum University, Chengdu 610500, People's Republic of China*
[5]*School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona 85287, USA*
[6]*Department of Physics, Arizona State University, Tempe, Arizona 85287, USA*

(Received 25 November 2021; accepted 23 May 2022; published 16 September 2022)

It is evident that increasing the intensive-care-unit (ICU) capacity and giving priority to admitting and treating patients will reduce the number of COVID-19 deaths, but the quantitative assessment of these measures has remained inadequate. We develop a comprehensive, non-Markovian state transition model, which is validated through the accurate prediction of the daily death toll for two epicenters: Wuhan, China and Lombardy, Italy. The model enables prediction of COVID-19 deaths in various scenarios. For example, if appropriate treatment priorities had been used, the death toll in Wuhan and Lombardy would have been reduced by about 10% and 7%, respectively. The strategy depends on the epidemic scale and is more effective in countries with a younger population structure. Analyses of data from China, South Korea, Italy, and Spain suggest that countries with less per capita ICU medical resources should implement this strategy in the early stage of the pandemic to reduce mortalities. We emphasize that the results of this paper should be interpreted purely from a scientific and a quantitative-analysis point of view. No ethical implications are intended and meaningful.

DOI: 10.1103/PhysRevResearch.4.033209

I. INTRODUCTION

The continuous and recently accelerated spreading of COVID-19 in more than 145 different countries and regions of the world has placed an unprecedented burden on the corresponding healthcare systems. As of December 1, 2020, there have been more than 63.3 million confirmed cases with over 1.4 million deaths reported, and the daily number of deaths [1] has exceeded 8000. To mobilize the available medical resources to the maximum degree to reduce the COVID-19 fatalities has become the top priority in many hospitals and healthcare facilities. In a hospital, there are two types of care resources for inpatients: General Ward (GW) and Intensive Care Unit (ICU) beds, with the latter currently serving mostly COVID-19 patients. Increasing the ICU beds would undoubtedly reduce the COVID-19 fatalities. When a hospital is overwhelmed with COVID-19 patients so that the ICU beds are in a serious shortage, a selective ICU admission policy must be invoked to allow specific groups of patients to receive the ICU care. In this regard, from a sheer mathematical point of view, exercising priority will reduce the fatality rate. Stretching the ICU resources to their limit and implementing selective ICU admission policy are becoming inevitable and even absolutely necessary in many parts of the world as the COVID-19 cases have continued to skyrocket in recent months.

It is intuitively evident that enhancing ICU capabilities and selectively admitting patients into ICU can reduce the COVID-19 mortalities, but we lack a modeling framework to quantitatively assess and characterize their effects. The main goal of this paper is to address this issue that is critically important to maximizing the usage and effectiveness of the limited medical resources to minimize COVID-19 deaths.

There have been intense modeling efforts on early warning, prevention and control of the COVID-19 pandemic [2–11]. For example, Kraemer *et al.* [2] found that travel restriction in the early stage of the COVID-19 outbreak can effectively prevent the infection imported from known sources. Once cases begin to spread in the community, the contribution of newly imported cases tends to diminish, requiring a set of control measures including travel restrictions, detection, tracking, and isolation to mitigate the pandemic. Kissler *et al.* [3] established a SARS-CoV-2 transmission model with seasonal variations, immune duration, and cross immunization, where the peak of SARS-CoV-2 infection is assumed to be low in spring and summer and a larger epidemic outbreak

*These authors contributed equally to this work.
†tangminghan007@gmail.com
‡ying-cheng.lai@asu.edu

can occur in autumn and winter. A finding was that, with respect to imposing one time social distancing, intermittent social alienation measures can prevent the overload of public health resources. Hao *et al.* [4] found that the new coronavirus has two characteristics: high infectivity and concealment. If 87% of the infected individuals are not detected, without any prevention and control measures, after 14 consecutive days of zero confirmed cases, the probability of a second wave of epidemics will be 32%. If only 53% of the infected are undetected, the probability of an epidemic rebound will drop to 6%. Premature removal of prevention and control measures will greatly increase the possibility of a second outbreak of the epidemic. Long *et al.* [5–7] developed a time-delay transmission model that can accurately simulate and predict the epidemic development in countries or regions, and the model enables evaluation of the impact of virus detection and social distancing intensity on the epidemic, as well as an accurate estimate of the time zero point of the epidemic in various countries and regions. In particular, it was found [7] that community-based transmission occurred in Europe and in the United States in early January 2020. The study of Flaxman *et al.* [8] revealed that the outbreak interventions implemented in 11 European countries have reduced about 1.3 million deaths and these measures are enough to bring down the basic reproduction number R_0 to less than one. Dehning *et al.* [9] studied the spread of COVID-19 in Germany by combining epidemiological modeling with Bayesian inference, and found that the change point of the effective growth rate of new infection is closely related to the time point when the intervention measures were imposed.

There have also been recent studies on the interplay between the medical resources and COVID-19 mortality. Particularly, in comparison with the common influenza virus, a higher proportion of the patients infected with SARS-CoV-2 need to be hospitalized. The aggravated growth in the number of infected individuals in recent months has stretched the capacities of the medical systems in many countries and regions to their limit. A shortage of medical resources, such as the respiratory support devices, will result in a large number of severely ill patients not getting timely and effective treatment, exposing them to a greater risk of mortality. Ferguson *et al.* [12,13] evaluated the demand for medical resources of the nonpharmaceutical interventions type, predicting that, even when the most effective mitigation measures are implemented, the numbers of hospital GW and ICU beds in the UK still need to be expanded by more than eight times to meet the needs of patient care. Miller *et al.* [14] used the age-specific mortality and demographic data to predict the cumulative COVID-19 cases and medical resource burden in different regions of the United States, pointing out that medical resources are relatively scarce in remote areas. When the medical system is overloaded, the nursing standard for the patients will be compromised, which would affect the treatment outcome. Without proper and rigorous care, the condition of critically ill patients with COVID-19 will further deteriorate, and the shortage of ICU facilities will aggravate the high mortality rate of such population [15,16]. These studies [12–16] made evident the potential significant impact of the medical resource availability on the COVID-19 mortality.

In this paper, we develop a comprehensive state transition model to predict the COVID-19 death and its evolution over time for a regional healthcare system with limited medical resources and selective ICU admission policy. The typical systems are a city or a region, e.g., Wuhan city in China or the Lombard region of Italy. Realistic time delays associated with various state transitions are fully incorporated into the model, rendering it non-Markovian to better describe the real world situations. The model is based on and validated with empirical data such as the number of confirmed cases, clinical data, demographics, and the amount of medical resources, and it enables such questions to be addressed as if the ICU resources were deployed certain days earlier or if the ICU capacity was doubled or tripled, how much reduction in the death toll would be achieved? The findings of this paper can be best described with concrete numbers. For example, for Wuhan city, if the ICU resources had been deployed a week in advance or if the number of ICU beds had been doubled, the death toll would have been reduced by 5% or 13%, respectively. For Lombardy, the corresponding numbers are similar: 3% or 14%. For both Wuhan and Lombardy, tripling the ICU capacity would have resulted in a 21% reduction in the mortalities. With respect to selective ICU admission policy, prioritizing certain groups of patients would have reduced the death toll in Wuhan and Lombardy by about 10% and 7%, respectively. As illustrated by the exemplary numbers, our model provides a framework to predict COVID-19 deaths in arbitrary scenarios. Because the model has been fully validated with real data, we expect the model predictions to be reliable and accurate. A further analysis of the data from China, South Korea, Italy and Spain indicates that the age-selective admission strategy is more effective in countries with a younger age structure, while the countries with less per capita ICU medical resources should implement the selective admission strategy in the early stage of the epidemic to suppress mortalities.

It is emphasized that the results of this paper should be interpreted purely from a scientific and a quantitative-analysis point of view. No ethical implications are intended and may be regarded as meaningful.

II. MODEL

Our COVID-19 patient admission model under limited medical resources has ten dynamical states, as shown in Fig. 1. It describes the basic transition process of a diagnosed individual's state from onset to recovery or death. Time delays associated with the various state transitions are taken into account, rendering non-Markovian the dynamical evolution. In particular, morbidity (M) state denotes that a patient is infected with COVID-19, state F means that the patient needs to be hospitalized but is currently under home isolation or centralized isolation, critical (C) state describes that the patient needs ICU support treatment but is currently isolated, General Ward (G) and Intensive Care Unit (U) states indicate an inpatient occupying a GW and an ICU bed, respectively, a patient in state W currently occupies a GW bed but needs ICU treatment, a patient in the G_U state is transferred out of ICU who does not need to occupy a GW bed and is isolated in a provisional square cabin bed, a patient in the M_X state is discharged from the hospital with lessened symptoms or was

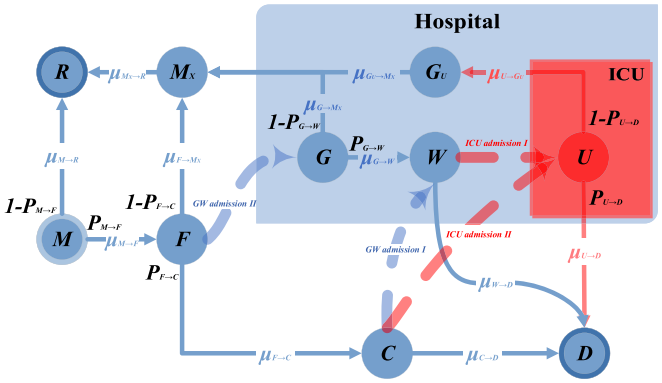

FIG. 1. Schematic diagram of the proposed COVID-19 patient admission model under limited medical resources. Each node represents a possible state of the patients. The specific definition of each state and the transition paths among them (light blue and light red arrows with time representing the average delay of the transition) are described in detail in the text. The various quantities $P_{\cdot\to\cdot}$ denote the fractions of the state transition paths, and the time delays follow a normal distribution. The light red dotted arrows indicate the two admission paths of ICU, where ICU admissions I and II have the first and second priorities, respectively. The light blue dotted arrows indicate the two treatment paths of GW, which have a lower priority than ICU admission: GW admissions I and II have the third and fourth priorities, respectively.

very ill but now has lessened symptoms, and the recovered (R) and dead (D) states represent two clinical outcomes: recovery and death.

The medical resources can be divided into two categories: GW and ICU resources denoted as Q_G and Q_U, respectively, which can be measured, e.g., by the corresponding numbers of beds. That the resources are limited is modeled as follows. When the GW beds are all used up, new patients will no longer be admitted. When all ICU beds have been occupied, patients who require ICU care either continue to wait at GW (W state) or continue to wait outside the hospital (C state). A common hospital admission policy is "first-come first-serve" (FCFS), which gives priority to patients who have had a long waiting time. The admission policies of GW and ICU thus consist of four admission paths with different priorities: GW admission I/II and ICU admission I/II, where ICU admission I for the W-state patients has the highest priority, followed by ICU admission II for the C-state patients, GW admission I for the remaining C-state patients in GW, and GW admission II for the F-state patients. The priorities define the execution order of the corresponding paths. For the M-state patients, two different transitions can occur: some patients deteriorating into the F state so as to require hospitalization and others slowly recovering into the R state. For patients in the F state, there are three types of transitions: (i) when the GW resources are available, the patients are admitted to the hospital through GW admission II, (ii) a proportion of the remaining F-state patients deteriorate to C state, requiring ICU treatment, and (iii) the rest are cured so they switch to the M_X state. Patients in state C enter ICU through ICU admission II to transition into U state, or through GW admission I to enter GW to become W state, while the rest switch to D state. For G-state patients, some switch to M_X state and no longer occupy GW resources, while others deteriorate to the W state, requiring ICU care. W-state patients will enter ICU through ICU admission I, and the remaining will be in the D state. Some of the U-state patients occupying ICU resources will transition to the G_U state, and the others will switch to the D state. For patients in the G_U state, the symptoms are relieved after a period time and they enter the M_X state. Patients in the M_X state recover to the R state after a period of time.

Let $\triangle Q_G$ and $\triangle Q_U$ denote the changes in the GW and ICU resources in each time step, respectively. A decrease in the GW resources is the sum of the number of patients admitted through the GW admission I/II pathway, and an increase is the sum of the number of ICU admission I, the number of patients in G(W) state cured (or died), and the amount of newly deployed GW resources. Likewise, a decrease in ICU resources is the sum of the number of patients admitted through ICU admission I/II pathway, and an increase is the sum of the number of patients cured (or died) in state U and any newly deployed ICU resources. The state dynamical evolution is described by a set of difference equations with time delays characterizing the state ages (Appendix A).

III. RESULTS

We validate our ten-state model by simulating the trend of the death tolls in Wuhan and Lombardy, using the daily number of confirmed cases, clinical and demographic data. The model then enables us to assess the overall effects of varying the timing of resource deployment and its amount on the COVID-19 mortalities. At a more detailed level, we divide the patients into several age groups and calculate the death-toll trend in each age group. This allows us to assess the impact of resource allocation scheme of different age groups on the number of deaths, and to obtain the optimal ICU admission strategy in terms of the age structure and outbreak scale with limited medical resources.

A. Impact of limited medical resource deployment

1. Model validation

To simulate the trend of the death tolls in Wuhan and Lombardy by using the the second-order difference equations in our modeling framework, three types of data are required: time delays associated with state transitions, patient morbidity data, and local medical resource deployment data. The details of these data are described in Appendixes B which include information such as the average number of days in the transition delay from M to F state and the average mortality rate of patients in ICU. Most of the data were obtained through references and official reports [17–19]. As detailed in Appendix B, the values of some model parameters need to be estimated in an optimal way, e.g., at the 95% confidence level, through model simulation and empirical data such as the average fraction of the patients switching from G to W state, denoted as $|P|_{G\to W}$, and that from F to C state, denoted as $|P|_{F\to C}$. We use the cumulative days of insufficient GW and ICU resources to quantify the level of stress on the healthcare system, defined as the GW overload days O_G and ICU overload days O_U (see Appendix C - medical system stress level indicators).

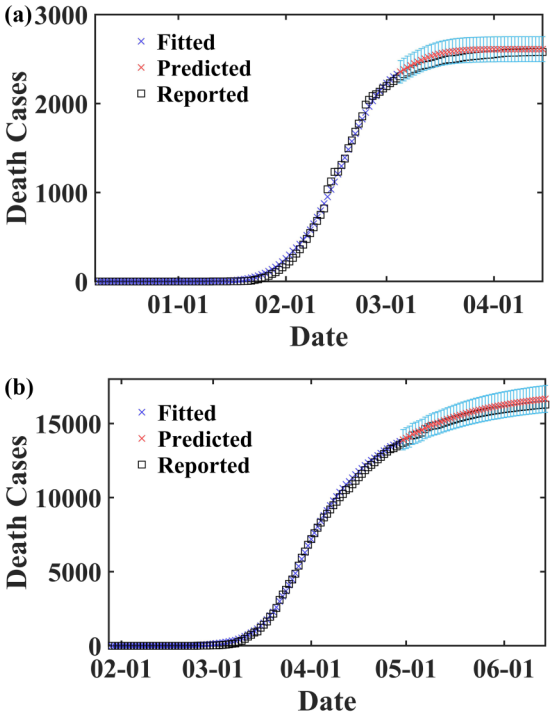

FIG. 2. Simulation results of the daily evolution of death tolls for the two COVID-19 epicenters. [(a) and (b)] Empirical data and simulation of cumulative death toll with time in Wuhan and Lombardy, respectively. The open black squares represent the empirical data. The blue and red crosses represent, respectively, the simulated and predicted data, where the latter are obtained by using the model parameter values estimated from the training data. For each data point, the 95% confidence interval is indicated. There is a good agreement between the predicted daily number of deaths and the actual data, thereby validating our ten-state model.

To validate the model, we first obtain the optimal estimates of the model parameters from an incomplete training data set, i.e., the first half of the epidemic time series (see Appendix B: Parameter estimation), and then test if the model is capable of predicting the second half of the time series. Specifically, the empirical data of cumulative death toll with time in Wuhan (from December 8, 2019 to March 4, 2020) and Lombardy (from January 28 to April 30, 2020) are set as the training data. With the optimal parameter values and empirical data as inputs to the model, we predict the cumulative death toll over time afterwards in Wuhan and Lombardy for the time periods from March 4, 2020 to April 15, 2020 and from April 30, 2020 to June 14, 2020, respectively. The results are shown in Fig. 2, where the red crosses represent the predicted numbers of deaths. It can be seen that the model predicted trends of the cumulative death toll agrees with the actual data quite well, thereby validating the model. The model also gives that the values of O_G are 14 and 17 days for Wuhan and Lombardy, respectively, with the corresponding values of O_U as 49 and 74 days.

2. *Deployment of medical resources*

Since the onset of COVID-19 pandemic, many countries (most notably the USA) and regions have missed the best time window to control the spreading. At present, there is a skyrocketing increase in the demand for ICU beds and medical resources such as ventilators and other special medical devices. Many hospitals and healthcare facilities have reached or will soon reach the limit of their operating capacities.

Optimizing the usage of medical resources by deploying them as early as possible and augmenting them as much as possible are key to reducing the mortality rate. Let DT and RI denote the two key parameters: the deployment time and the resource input, where $\mathrm{DT} = 0$ and $\mathrm{RI} = 1$ represent the actual deployment time and the available normalized amount of resources, respectively. If the resources are deployed seven days ahead of the actual time, we have $\mathrm{DT} = -7$. Likewise, if the resources are doubled (e.g., twice as many ICU beds as in the actual case), we have $\mathrm{RI} = 2$ (see Appendix C for COVID-19 special medical resource deployment plans). A virtue of our modeling framework lies in its ability to provide a quantitative picture of the dependence of the mortalities on the two key parameters.

To uncover the impact of deploying medical resources on the patient mortality rate in a concrete way, we consider the following three scenarios: (i) varying the GW resources only, (ii) varying ICU resources only, and (iii) varying both resources simultaneously. Some representative results are presented below (more results in Appendix D).

Figure 3(a) shows, for Wuhan, the impact of varying GW resource deployment on the number of deaths for fixed ICU resource deployment, where the red hexagon indicates the true death toll under the actual resource deployment. If the GW resources had been deployed one week in advance or if its amount had been doubled, the death toll would have been reduced by 10%. While remarkable, this 10% figure is about the maximum reduction that can be achieved by varying the GW resource deployment. For example, deploying the GW resources earlier than one week or an increase in its investment over three times would not reduce the death toll further. The maximum reduction in the deaths that can be achieved is determined approximately by the contour of the actual death toll—the curve connecting the white circles Fig. 3(a). These results indicate that deploying the GW resources earlier and/or augmenting them have only limited effects on reducing the deaths.

In contrast, deploying the ICU resources earlier and/or augmenting them can be more effective at reducing the deaths in Wuhan, as shown in Fig. 3(b), where the GW resource deployment is fixed. If the ICU resources had been deployed one week ahead of the actual time or if the ICU resource amount had been doubled, the death toll would have been decreased by 5% and 13%, respectively. While these figures are similar to those that would have been achieved with the same adjustment in the GW resources [Fig. 3(a)], deploying the ICU beds earlier or increasing their number can have a much more significant effect on the number of deaths. For example, if the number of ICU beds had been tripled, the death toll would have been reduced by about 20% in the most optimistic scenarios. A significant reduction of about 30% in the deaths is achieved for $\mathrm{DT} = -14$ and $\mathrm{RI} = 2.5$, which is close to the maximum reduction with infinite resources. The corresponding ICU overload days would have been decreased to

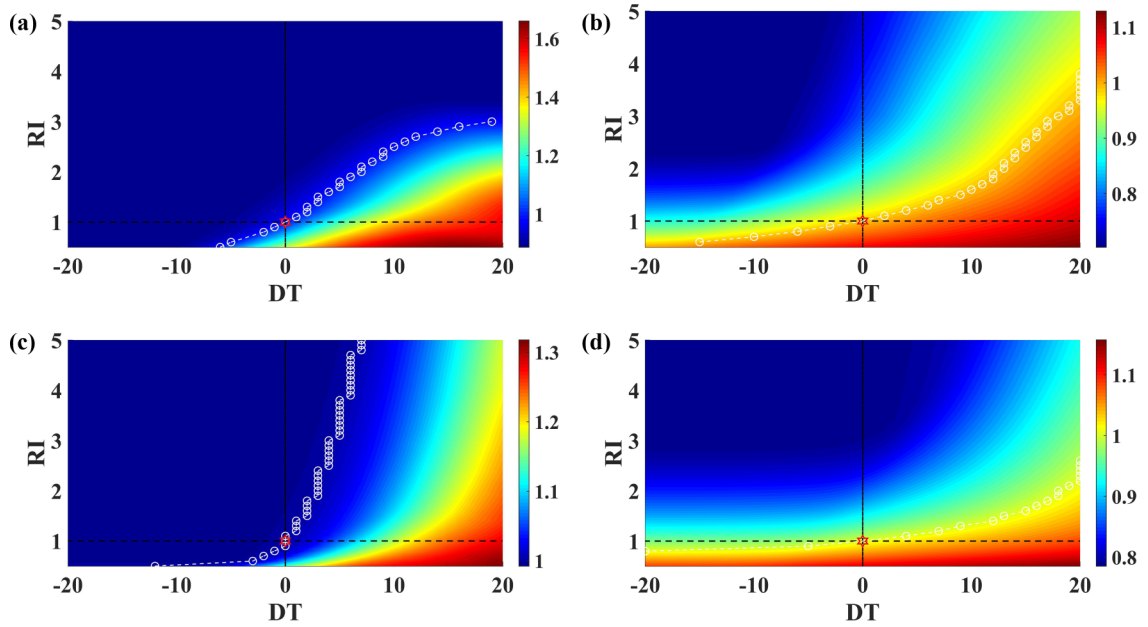

FIG. 3. Impact of timing of medical resource deployment and resource input on COVID-19 mortalities. Each panel shows the ratio of the number of deaths to the actual data (color coded) in the parameter plane of deployment time and resource input: (a) and (b) are for Wuhan, China, and (c) and (d) are for Lombardy region, Italy. In (a) and (c), the ICU resource amount is fixed and the resource that varies is GW. In (b) and (d), the GW resource amount is fixed and the resources that vary are ICU. The red hexagon indicates the true death toll under the actual resource deployment, and the white circles mark the contour of the actual death toll in the parameter plane.

less than seven days (comparing with the actual 49 days for Wuhan). Similar results are obtained for Lombardy. We find that adjusting the GW resource deployment would have only limited effects in reducing the death toll. For example, as shown in Fig. 3(c), if the GW resources had been deployed one week or two weeks earlier or if the number of GW beds had been doubled or tripled, the reduction in the deaths would have been only about 1%. This indicates that the deployment of the GW resources in Lombardy was timely and its amount was appropriate, a fact that can also be seen from the value of the GW overload days O_G: to make it less than a week, the resources would need to be deployed only five days earlier or the number of GW beds would need to be only 1.4 times higher. In contrast, deploying the ICU resources earlier or augmenting them would have been much more effective at reducing the deaths. For example, the number of ICU beds had been tripled in an ideal scenario, the death toll would likely have been reduced by about 20%, as shown in Fig. 3(d).

In fact, if many doctors and nurses get ill, the effective hospital resources will be decreased. A decrease in RI would lead to an increase in the number of deaths, as shown in Fig. 3. Appendix C presents results on the impact of varying the GW and ICU resource deployment simultaneously on COVID-19 death toll for both Wuhan and Lombardy.

B. Admission strategy based on age groups

To improve the efficiency of treatment and to reduce the mortality rate, it is essential and imperative to allocate the available resources as reasonably as possible because they are not unlimited. For COVID-19, the hospitalization and mortality rates of the elderly patients with underlying medical conditions are higher than those of other patients [20]. A simple and common strategy is to divide the patients into distinct age groups. Our modeling framework provides a rigorous way to calculate the mortalities for different age groups.

1. Age groups

Individuals in any age group have the possibility of getting infected by COVID-19 with certain death risk [21]. Currently available data indicate that, in most countries, the risk of hospitalization and death increases with age [22,23]. A direct manifestation is that the fraction of hospitalized patients and ICU mortality vary among the age groups.

For Wuhan, the available data divide patients into three groups: [0-69] years old, [70-79] years old, and [80+] years old. A similar division scheme applies to Lombardy. We use a linear regression to obtain the estimates of the fractions of various state transitions for each age group: $P_{M\to F}$, $P_{G\to W}$, $P_{F\to C}$, and $P_{U\to D}$. (Appendix D provides more details.)

We input the parameters associated with each age group into the model and calculate the death tolls as a function of time for Wuhan and Lombardy. Figure 4 shows that the mortalities obtained from the simulation agree well with those of the actual data. Figures 4(a) and 4(c) show that the death toll of patients over 80 years old in Wuhan is the lowest among the age groups, while it is the highest in Lombardy: about 59%. There are two possible reasons for this difference. First, the aging populations in the two regions are different: the fraction of people over 65 years old is 14.06% in Wuhan and 20% in Lombardy [24,25]. The elderly have weak autoimmunity and most of them have preexisting, underlying diseases, resulting

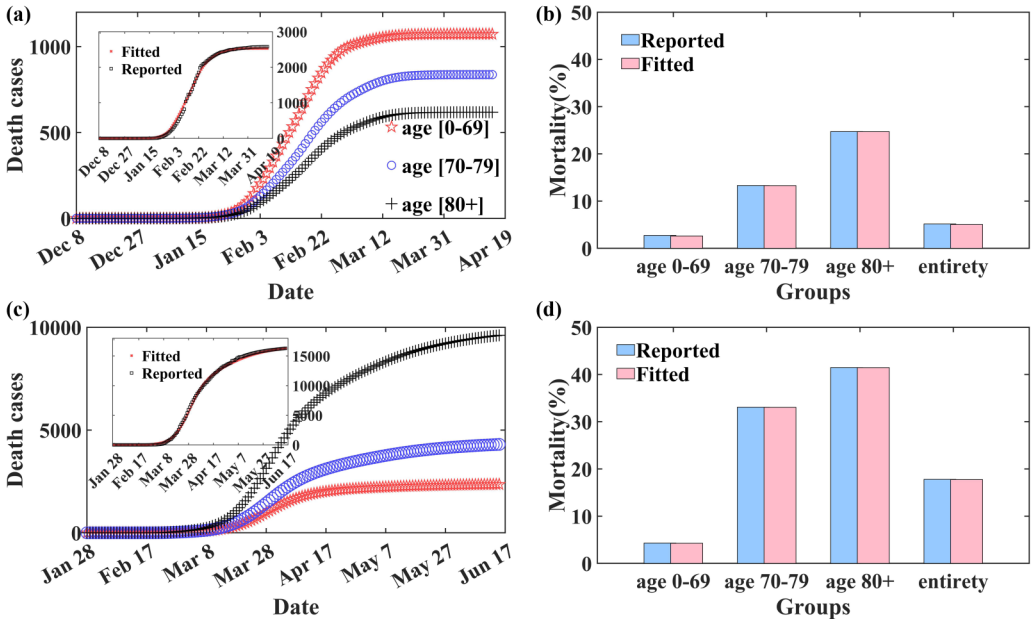

FIG. 4. Evolution of death toll with time for different age groups. (a) Simulated evolution trend of the death toll of the three age groups in Wuhan; (b) the actual mortality rate of each age group in Wuhan in comparison with model prediction; (c,d) the corresponding results for Lombardy. The insets in (a) and (c) display the actual daily death toll and model predictions for Wuhan and Lombardy, respectively. In panels (b) and (d), the light blue and pink columns represent the actual and simulated mortality rate of each group, respectively.

in an excessively high death rate in Lombardy. The second reason is that strict containment policies protect most of the elderly against infections in Wuhan. Nonetheless, Figs. 4(b) and 4(d) show that the mortality rate of patients over 70 years old is higher than that of patients under 70 for both Wuhan and Lombardy. Especially, in Lombardy, the mortality rate of patients over 80 years old reached 41%.

When the population is divided into nine age groups, our model generates essentially the same phenomenon that the risk of death increases with age (Appendix E).

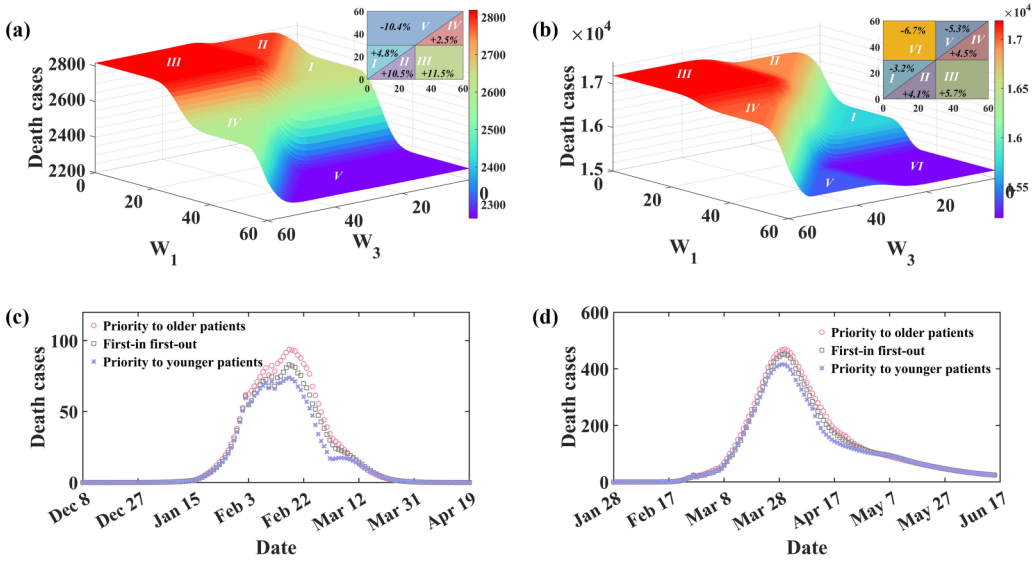

FIG. 5. Dependence of death toll on weights. [(a) and (b)] Model generated death toll and its dependence on w_1 and w_3 for fixed $w_2 = 30$ for Wuhan and Lombardy, respectively, where the blue and purple colors correspond to relatively fewer deaths. In the regions marked by Roman numerals, the death toll changes slowly with weights. The insets show the distribution of the number of deaths for different weight combinations, where the numerical values represent the differences in the number of deaths than that from the FCFS treatment strategy. [(c) and (d)] Daily number of new deaths under different admission strategies for Wuhan and Lombardy, respectively, where red circles, gray boxes, and blue-purple forks denote the results from the three strategies: preferential treatment for elderly patients, FCFS, and preference for young patients, respectively. Giving priority to young patients results in the lowest daily death curve.

2. Group weighting strategy

People of all ages have certain risk of being infected with COVID-19. Generally, the patients without any underlying medical conditions have a low mortality rate. Setting priority of admission for appropriate patients can reduce the number of deaths. In Wuhan, the ICU resources were more scarce than GW resources, so making ICU admission policy dependent on age is especially important for suppressing the mortality rate. That is, the conventional FCFS admission policy is not suitable for COVID-19 under limited ICU resources. To set the priority of admission for different age groups, we shift the state age of each group by a priority weight. We then change the admission order of different age groups with the goal to find an optimal set of weights that minimizes the number of deaths. In particular, we set the priority weight of the i_{th} age group as w_i and record τ_i as the state age of the this age group. After weighting, the new state age τ_i' becomes

$$\tau_i' = \tau_i + w_i. \tag{1}$$

The state age of patients in the i_{th} age group is increased by w_i days. The patients, after incorporating the weighted state age, are admitted according to FCFS. It is worth noting that this group weighting strategy does not change the order of treatment (i.e., FCFS) for patients in the same group.

To carry out the optimization procedure, we fix the priority weight of the second age group, i.e., the [70, 80) age group, to be $w_2 = 30$. As the state age of each group does not exceed 15, we implement different admission strategies for the age groups by changing the value of w_1 and w_3 (see Appendix E for a specific method).

Figure 5 shows the effect of the state age of groups 1 and 3 on the number of deaths when the weight of group 2 is fixed at constant 30. The larger the state age, the higher priority of admission is. When adjusting the weights of the first and third groups, their corresponding state ages will vary and so the admission order of each age group will change accordingly. Different orders of treatment will lead to different death tolls. Figure 5 reveals, in the parameter plane of (w_1, w_3), five distinct death toll regions for Wuhan. For Lombardy, there are six such regions, as shown in Fig. 5(b). The weight range in region V in Fig. 5(a) corresponds to the ranges in regions V and VI in Fig. 5(b), due to the relatively small fraction of patients in the third group in Wuhan. When patients in the first group are given the priority, whether the priority of treatment for patients in the third group is higher than that of the second group has little impact on the total number of deaths in region V of Fig. 5(a). While the patients in the third group in Lombardy have a large proportion, if the priority weights of both the first and the second age groups are fixed, the priority of treatment for patients in the third group will obviously affect the total number of deaths. As the patients in the third group have a higher mortality rate in ICU, the number of deaths in region V with the high priority for these patients is higher than that in region VI with the low priority.

A striking result in Fig. 5 is the occurrence of plateaued regions, where a change in both w_1 and w_3 values does not alter the death toll. To understand the emergence of the plateaus, we take area III in Fig. 5(a) as an example. As can be seen from Fig. 14 in Appendix F, we first fix the priority weight

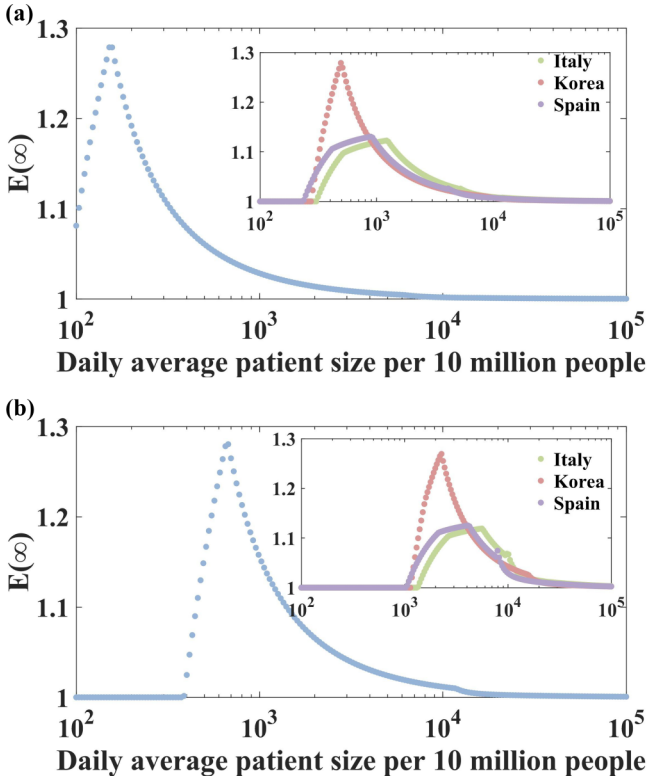

FIG. 6. Asymptotic optimization efficiency in China, Korea, Italy, and Spain. Shown is the steady-state efficiency $E(\infty)$ vs the cases per ten million people. (a) The result for China based on the parameter values from Wuhan. Inset shows the corresponding results for Italy, South Korea, and Spain. (b) The result for China based on the parameter values from Lombardy and the results for the other three countries are displayed in inset.

of the second age group, whose state age range is [30, 42]. If we give the highest priority to the elderly, i.e., the state age of the third group is increased by $w_3 \in [40, 60]$, the state age of the third group will be in the range [40, 68]. Likewise, if the state age of the first group is increased by $w_1 \in [0, 15]$, the state age of the first group will be in the range [0, 35]. The older the state age, the earlier the treatment. This suggests that, in region III, regardless of the variations in w_1 and w_3, the priority for admission is given to group 3, followed by group 2, and finally to group 1. As a result, there is a plateau in area III where the death toll changes little.

In area III in Fig. 5, the weighting order is $w_3 > w_2 > w_1$ and the largest death tolls are observed in Wuhan and Lombardy, as characterized by an increase of 11% and 5.7%, respectively, in comparison with the FCFS case. In area V in Fig. 5(a), we have $w_1 > w_3 > w_2$, and the number of deaths in Wuhan is the lowest as represented by a decrease of 10.4% relative to that associated with the FCFS strategy. In the same area V for Lombardy, the corresponding decrease in the death toll is by 5.3%, as shown in Fig. 5(b). In area VI in Fig. 5(b), the priority ordering is $w_1 > w_2 > w_3$ and there is a decrease of 6.7% in the death toll. The implication of these results is that giving the treatment priority to young patients will reduce the number of deaths but giving priority to elderly patients will increase the death toll. Due to the resource limit, this strategy

TABLE I. State transition time delay parameter setting.

Parameter(delay)		Estimate/assumption	Definition	Justification
$\mu_{M\to R}$	Wuhan/Lombard	14d	Average delay from M state to report recovery	[1]
$\mu_{M\to F}$	Wuhan/Lombard	5d(std=6.67)	Average delay from M to F state	[17]
$\mu_{F\to M_X}$	Wuhan/Lombard	8d(std=4.4477)	Average delay from F to M_X state	[12]
$\mu_{F\to C}$	Wuhan/Lombard	2d(std=3.7064)	Average delay from F to C state	[17,29]
$\mu_{G\to M_X}$	Wuhan/Lombard	8d(std=4.4477)	Average delay from G to M_X state	[12]
$\mu_{G\to W}$	Wuhan/Lombard	3d(std=3.7064)	Average delay from G to W state	[17]
$\mu_{U\to G_U}$	Wuhan/Lombard	8d(std=4.4477)	Average delay from U to G_U state	[18]
$\mu_{U\to D}$	Wuhan/Lombard	7d(std=5.93)	Average delay from U to D state	[18,19]
$\mu_{C\to D}$	Wuhan/Lombard	1d(std=2.2239)	Average delay from C to D state	Assume
$\mu_{W\to D}$	Wuhan/Lombard	3d(std=3.7064)	Average delay from W to D state	Assume
$\mu_{G_U\to M_X}$	Wuhan/Lombard	8d(std=4.4477)	Average delay from G_U to M_X state	[12]
$\mu_{M_X\to R}$	Wuhan/Lombard	14d	Average delay from M_X state to report recovery	[1]

is intuitively reasonable and has in fact been commonly practiced in many hospitals and healthcare facilities. Our results provide a validation at a quantitative level.

Figures 5(c) and 5(d) demonstrate the evolution of new deaths over time in Wuhan and Lombardy, respectively, under different admission strategies. In both cases, there is no significant difference in the number of deaths in the early stages of the epidemic under three treatment strategies. When the epidemic has lasted for a period of time and the death toll increases sharply, giving priority to young patients can significantly reduce the number of deaths. At the end of the epidemic, the rapid decrease in the number of patients needing ICU treatment again makes the differences among the three treatment strategies diminish. Taken together, these results verify that, under limited medical resources, during the rapidly increasing phase of COVID-19 infection, admitting and treating appropriate patients are necessary to reduce the final death toll. It is emphasized that in implementing the strategy, the age of a patient should not be taken as the only criterion for judging whether he/she is an appropriate patient. Many factors of the patient such as symptoms and underlying medical conditions should be taken into account.

3. *Optimization efficiency*

When the ICU resources are in a serious shortage, medical staff have to face the hard choice of adopting the strategy of giving priority to treating appropriate patients with a higher survival probability. The strategy's effectiveness notwithstanding, it has serious implications for medical ethics. It is thus worth evaluating this admission/treatment strategy further. To this end, we exploit our model to study more extensively the impacts of different strategies for four countries: China, South Korea, Italy, and Spain. For each country, we collect information about the numbers of GW and ICU beds per 100 000 population and about the age distribution of patients diagnosed with COVID-19. For example, in China and Italy, the numbers of ICU beds per 100 000 individuals are 3.6 and 12.5, respectively. (A detailed display of the information can be found in Appendix F). We define the time-dependent optimization efficiency as

$$E(t) = \frac{V_Q(t)}{V_Y(t)}, \tag{2}$$

where $V_Q(t)$ and $V_Y(t)$, respectively, are the numbers of new deaths per day with the FCFS and the age-selective strategy. In the asymptotic time limit $t \to \infty$ when the system has reached a steady state, $E(\infty)$ characterizes the final optimization efficiency. Since information about the average state transition fractions for patients in all four countries is not available, we use the model parameter values for Wuhan and Lombard to calculate the value of $E(\infty)$ for four countries versus the average daily patient size. Note that, in this case, we assume that the number of new cases per day, $M(0,t)$, is the same for all t.

Based on the per capita medical resource data of the four countries as well as the age distribution data of confirmed patients in each country, by making use of the relevant model parameter values for Wuhan and Lombard, we calculate $E(\infty)$ of the four countries versus the daily average patient size. We assume that up to 50% of the medical resources can be deployed to treating patients with COVID-19. Since the average state transition fractions for patients in all four countries are not available, here we respectively use for reference the model parameter values for Wuhan and Lombard, which

TABLE II. Parameter setting of state transition fraction.

Parameter(delay)		Estimate/assumption	Definition	Justification
$P_{M\to F}$	Wuhan/Lombard	see Tables S3.1 and S3.2	Transition fraction from M to F state	
$P_{F\to C}$	Wuhan	23.26% (21.24%, 28.55%)	Transition fraction from F to C state	Fitting according to the reported death data
	Lombard	56.45% (53.33%, 65.23%)		
$P_{G\to W}$	Wuhan	20.76% (18.15%, 24.56%)	Transition fraction from G to W state	
	Lombard	41.22% (39.24%, 55.43%)		
$P_{U\to D}$	Wuhan/Lombard	61.5%	Transition fraction from U to D state	[18,30]

TABLE III. Setting of transition fraction from M to F state in five stages in Wuhan.

Date	Dec 8, 2019–Jan 9, 2020	Jan 10–Jan 22	Jan 23–Feb 1	Feb 2–Feb 16	Feb 17–
$P_{M\to F}$(Wuhan)	53.10%	35.10%	23.50%	15.90%	10.30%
Justification			[31]		

does not qualitatively affect the impacts of age structures and ICU resources per capita on $E(\infty)$.

As shown in Fig. 6, when the number of patients is below 100, the value of $E(\infty)$ for the four countries is equal to one. In this case, there is no need to implement the age-selective admission strategy. When the number of patients exceeds a critical value, $E(\infty)$ for all four countries increases rapidly as the number of patients increases through the critical value and then reaches a peak value. In fact, for a wide range of variation of the patient size, the values of $E(\infty)$ are high. Note that the peak values of $E(\infty)$ for China and South Korea are slightly below 1.3, but those for Italy and Spain are lower than 1.2 with a relatively slow increasing rate to reach the peak value. A plausible reason lies in the difference in the age distribution of patients diagnosed with COVID-19 in the eastern and western countries. In particular, in China and South Korea, less than 12% of the patients are over 70 years old, while those in Italy and Spain account for 37% and 34%, respectively (Appendix F). For a small average daily patient size, $E(\infty) = 1$ means that ICU beds are fully sufficient, but $E(\infty) > 1$ signifies a shortage of ICU beds. If the daily average patient size is large, we have $E(\infty) = 1$, which means that critically short ICU beds cannot meet the requirement of any age group. In this case, the age-selective admission strategy has no effect on the decrease of death toll. Figure 6(a) reveals that, with the per capita medical resources of South Korea, Italy, and Spain, the need for ICU of 1,000 cases per tens of millions of people per day can be met for a long time, while the capacity of China may be less than half of the capacity of those countries. Indeed, the number of ICU beds per capita in China is 3.6 per 100,000 persons, while the corresponding numbers in Italy, South Korea, and Spain are 12.5, 10.6, and 9.7, respectively (Appendix F). We conclude that, while the age-selective admission strategy is more effective in countries with a younger population structure, countries with less per capita ICU medical resources need to implement this strategy, particularly in the early stage of the pandemic when the number of patients is relatively small.

IV. DISCUSSION

With the resurgence of COVID-19 cases in many countries and regions, a severe shortage of medical resources for treating the disease is inevitable, potentially leading to a significant death toll. To study the impact of limited medical resources on the patient mortality at a quantitative level can provide insights into developing optimal resource allocation schemes to reduce the number of deaths. Based on the COVID-19 data, medical resources and other relevant information, we have developed an admission treatment model subject to limited medical resources, which enables a quantitative and systematic assessment of the mortality rate associated with various resource allocation scenarios. Using empirical data from Wuhan and Lombardy, we validate the model by demonstrating that it can reproduce accurately the evolution of daily death toll in both places. The validated model is then used to assess the impact of different scenarios of medical deployment (including deployment time and resource investment) on the number of deaths. In general, the ICU resources have a significant impact on the mortality rate, and it is intuitively evident that deploying the ICU resources earlier or augmenting them will reduce the death toll. A virtue of our model is that it enables an accurate quantification of such intuitive expectations. For example, we find that, if the deployment of ICU resources had been made one week earlier or if they had been doubled, the death toll in Wuhan would have been reduced by 5% or 13%, and that in Lombardy by 3% and 14%, respectively. If the number of ICU beds had been tripled, the death toll in both Wuhan and Lombardy would have been reduced by about 20% in the most optimistic scenario. Likewise, our model is fully capable of assessing the impacts of GW resources on COVID-19 mortalities. In comparison with the ICU resources, the GW resources play a less role in reducing the death toll.

The morbidity and mortality of patients infected with COVID-19 increase with age. We have studied the role of age selective admission and treatment strategies in the death toll. In general, giving admission priorities to patients in different age groups would result in different mortality rates. In particular, if priorities had been given to the younger age groups, the number of deaths in Wuhan and Lombardy would have been reduced by 10.4% and 6.7%, respectively, in comparison with that from the normal FCFS strategy. In contrast, if priority had been given to the elderly patients, the number of deaths would have increased by 11.5% and 5.7% for the two places, respectively.

The optimal admission strategies also depend on the scale of COVID-19 outbreak, the age structure of patients in the general population, and the per capita medical resources. We

TABLE IV. Setting of transition fraction from M to F state in four stages in Lombardy.

Date	Jan 28, 2020–Feb 19, 2020	Feb 20–Feb 25	Feb 26–Mar 20 1	Mar 21–
$P_{M\to F}$(Lombardy)	63.00%	61.00%	56.00%	32.21%(29.33%, 34.14%)
Justification		[32]		Fitting according to the reported death data

have quantified these effects by defining the steady-state optimization efficiency $E(\infty)$ and calculated this quantity for four countries: China, South Korea, Italy, and Spain. We conclude that the age-selective admission strategy is more effective in countries with a younger age structure, while the countries with less per capita ICU medical resources may need to implement this admission strategy in the early stage of the epidemic when the number of patients is relatively small.

Our model provides a general evaluation framework for assessing, at a quantitative level, the necessary medical resource deployment and admission strategy. It can be used to predict and articulate, under limited medical resources, optimal scenarios with respect to resource deployment and hospital admission/treatment strategies to minimize the death toll for future outbreaks of infectious diseases. In addition, our model can be used to quantitatively evaluate the number of overload days of medical resources and age-specific mortality over time in cases where empirical data are not available. Our model is based on the assumption that certain groups of patients have a lower mortality rate. In fact, with timely and comprehensive treatment, the vast majority of the patients without underlying medical conditions can successfully recover. It should be emphasized that the ages of patients should not be taken as the only criterion for judging whether they should be treated preferentially. The flexibility of our model allows for further development in more realistic scenarios to explain phenomena such as relapses from lessened symptoms to serious illness and the longer time required for older patients to recover in intensive care.

ACKNOWLEDGMENTS

We thank Zhai Zhengmeng and Long Yongshang for correcting the procedure, Kang Jie for collecting data, and Lin Zhaohua and Han Lilei for discussing the experimental design. This work was supported by the National Natural Science Foundation of China (Grant Nos. 82161148012, 11975099, 11675056, 61802321, and 11835003). The work at Arizona State University was supported by the Office of Naval Research through Grant No. N00014-21-1-2323.

APPENDIX A: EQUATIONS GOVERNING NON-MARKOVIAN DYNAMICAL EVOLUTION OF STATES

Epidemic spreading in the real world depends on human behaviors and the event occurrences cannot be simply described as a Poisson random process. In general, the event time is not exponentially distributed, which is characteristic of non-Markovian dynamics [26–28]. To accurately capture the non-Markovian nature of the dynamical evolution of the states, we model the underlying process in terms of difference equations over infinitesimal intervals in both time and delay. In the following, we derive, one by one, the difference equations for all ten states in our model.

1. M state

The evolution equation of the number of M-state patients with state age τ at time t is

$$M(\tau+d\tau;t+dt)=[1-\omega_{M\to F}(\tau)d\tau]\frac{P_{M\to F}\int_{\tau}^{+\infty}f_{M\to F}(\tau')d\tau'}{P_{M\to F}\int_{\tau}^{+\infty}f_{M\to F}(\tau')d\tau'+(1-P_{M\to F})\int_{\tau}^{+\infty}f_{M\to R}(\tau')d\tau'}M(\tau;t)$$
$$+[1-\omega_{M\to R}(\tau)d\tau]\frac{(1-P_{M\to F})\int_{\tau}^{+\infty}f_{M\to R}(\tau')d\tau'}{P_{M\to F}\int_{\tau}^{+\infty}f_{M\to F}(\tau')d\tau'+(1-P_{M\to F})\int_{\tau}^{+\infty}f_{M\to R}(\tau')d\tau'}M(\tau;t), \tag{A1}$$

where $M(\tau;t)$ is the density function of M-state patients with state age τ at time t, so the number of nodes in the M state with the state age in the interval $(\tau,\tau+d\tau)$ is $M(\tau;t)d\tau$. A node in the M state enters the F state (R state) at the conditional rate $\omega_{M\to F}(\tau)[\omega_{M\to R}(\tau)]$, which are given by

$$\omega_{M\to F}(\tau)=\frac{f_{M\to F}(\tau)}{\int_{\tau}^{+\infty}f_{M\to F}(\tau')d\tau'},\quad \omega_{M\to R}(\tau)=\frac{f_{M\to R}(\tau)}{\int_{\tau}^{+\infty}f_{M\to R}(\tau')d\tau'},$$

where $f_{M\to F}(\tau)$ and $f_{M\to R}(\tau)$ represent the probabilities of M-state individuals transferring to F and R states within the state age interval $(t,t+dt)$, respectively, and $\int_{\tau}^{+\infty}f_{M\to F}(\tau')d\tau'$ and $\int_{\tau}^{+\infty}f_{M\to R}(\tau')d\tau'$ are the probabilities that M-state individuals have not moved to F state and R state before state age τ, respectively. Since the M state can transition to two different states: $M\to F$ and $M\to R$, the two processes will compete with each other, so the remaining probabilities are

$$\frac{P_{M\to F}\int_{\tau}^{+\infty}f_{M\to F}(\tau')d\tau'}{P_{M\to F}\int_{\tau}^{+\infty}f_{M\to F}(\tau')d\tau'+(1-P_{M\to F})\int_{\tau}^{+\infty}f_{M\to R}(\tau')d\tau'}\quad\text{and}\quad\frac{(1-P_{M\to F})\int_{\tau}^{+\infty}f_{M\to R}(\tau')d\tau'}{P_{M\to F}\int_{\tau}^{+\infty}f_{M\to F}(\tau')d\tau'+(1-P_{M\to F})\int_{\tau}^{+\infty}f_{M\to R}(\tau')d\tau'}, \tag{A2}$$

respectively.

The newly increased patients enter the M state. Suppose that $Z_M(t)$ new patients are added within $(t,t+dt)$, we have

$$M(0;t+dt)=Z_M(t). \tag{A3}$$

The state age of the newly added M-state patients is zero.

2. F state

The evolution equation of the number of F-state patients with state age τ at time t is

$$F(\tau + d\tau; t + dt) = [1 - \omega_{F\to C}(\tau)d\tau]\frac{P_{F\to C}\int_{\tau}^{+\infty} f_{F\to C}(\tau')d\tau'}{P_{F\to C}\int_{\tau}^{+\infty} f_{F\to C}(\tau')d\tau' + (1-P_{F\to C})\int_{\tau}^{+\infty} f_{F\to M_X}(\tau')d\tau'}F(\tau;t)$$
$$+ [1 - \omega_{F\to M_X}(\tau)d\tau]\frac{(1-P_{F\to C})\int_{\tau}^{+\infty} f_{F\to M_X}(\tau')d\tau'}{P_{F\to C}\int_{\tau}^{+\infty} f_{F\to C}(\tau')d\tau' + (1-P_{F\to C})\int_{\tau}^{+\infty} f_{F\to M_X}(\tau')d\tau'}F(\tau;t), \tag{A4}$$

with notations similar in their meanings to those of the M-state equation.

After admission of patients in the C state, the remaining available GW resources are $\Delta Q'_G(t)$. There are two cases: (i) the new resources available are sufficient to accommodate all F-state patients; and (ii) new available resources can only accept some or none of the F-state patients. In these two cases, patients in the F state are admitted according to the strategy of FCFS (first-come first-serve). The number of F-state patients changing to the G state in $(t, t + dt)$ is denoted as $\triangle_{F\to G}(t)$.

In the first case where the GW resources are sufficient to accommodate all current F-state patients, if

$$\Delta Q'_G(t)(t) \geqslant \int_0^{+\infty} F(\tau + d\tau; t + dt)d\tau,$$

then all F-state individuals will change to the G state. We have $F(\tau + d\tau; t + dt) = 0$ for all $\tau \in [0,+\infty)$, and

$$\triangle_{F\to G}(t) = \int_0^{+\infty} F(\tau + d\tau; t + dt)d\tau.$$

In the second case where the GW resources can hold only some or none of the current F-state patients, if

$$\Delta Q'_G(t)(t) = \int_{\tau_c}^{+\infty} F(\tau + d\tau, t + dt)d\tau,$$

then some F-state individuals will change to the G state. We have $F(\tau + d\tau; t + dt) = 0$ for $\tau \in [\tau_c, +\infty)$, and

$$\triangle_{F\to G}(t) = \Delta Q'_G(t)(t).$$

When the patients in the M state enter into the F state, the initial state age is set to be zero. The number of newly added F-state patients is given by

$$F(0; t + dt) = \int_0^{+\infty} \omega_{M\to F}(\tau)\frac{P_{M\to F}\int_{\tau}^{+\infty} f_{M\to F}(\tau')d\tau'}{P_{M\to F}\int_{\tau}^{+\infty} f_{M\to F}(\tau')d\tau' + (1-P_{M\to F})\int_{\tau}^{+\infty} f_{M\to R}(\tau')d\tau'}M(\tau;t)d\tau. \tag{A5}$$

3. C state

The evolution equation of the number of C-state patients with state age τ at time t is

$$C(\tau + d\tau; t + dt) = [1 - \omega_{C\to D}(\tau)d\tau]C(\tau;t). \tag{A6}$$

To get the number of C-state patients admitted to the hospital, we note that there are two admission paths: (a) through ICU admission II to the U state; (b) through GW admission I to the W state.

ICU admission II path. After admission of W-state patients, the remaining available ICU resources are $\Delta Q'_U(t)$. There are two cases: (i) the newly available resources are sufficient to accommodate all the C-state patients and (ii) the newly available resources are able to accommodate some or none of the C-state patients. The number of C-state patients who switch to the U state in $(t, t + dt)$ time interval is denoted as $\triangle_{C\to U}(t)$.

In case (i), if

$$\Delta Q'_U(t) \geqslant \int_0^{+\infty} C(\tau + d\tau; t + dt)d\tau,$$

all C-state patients will change to the U state. We have $C(\tau + d\tau; t + dt) = 0$ for $\tau \in [0,+\infty)$, and

$$\triangle_{C\to U}(t) = \int_0^{+\infty} C(\tau + d\tau; t + dt)d\tau.$$

In case (ii), if

$$\Delta Q'_U(t) = \int_{\tau_c}^{+\infty} C(\tau + d\tau; t + dt)d\tau,$$

and $0 \leqslant \tau_c \leqslant +\infty$, then the C-state individuals whose state ages are greater than τ_c will enter the U state. We have $C(\tau + d\tau, t + dt) = 0$ for $\tau \in [\tau_c, +\infty)$ and

$$\triangle_{C\to U}\ (t) = \Delta Q'_U(t).$$

The remaining C-state patients can switch to the W state through the GW admission I pathway.

GW admission I path. Considering that the remaining C-state patients are in critical conditions, when the ICU resources are unavailable to them, they will be admitted to GW and change into the W state with the highest priority. The available resources within the time interval $(t, t + dt)$ are denoted as $\Delta Q_G(t)$. There are two cases: (i) the available GW resources are enough to accommodate all the remaining C-state patients, and (ii) the available GW resources can accommodate some or none of the remaining C-state patients. The number of patients in the C state who change to the W state in $(t, t + dt)$ is denoted as $\triangle_{C\to W}\ (t)$.

In case (i), we have

$$\Delta Q_G(t) \geqslant \int_0^{+\infty} C(\tau + d\tau; t + dt)d\tau.$$

The amount of the remaining available resources through F-state patients for GW admission II is

$$\Delta Q'_G(t) = \Delta Q_G(t) - \int_0^{+\infty} C(\tau + d\tau; t + dt)d\tau.$$

Individuals in the C state will change to the W state, so we have $C(\tau + d\tau; t + dt) = 0$ for $\tau \in [0,+\infty)$, and

$$\triangle_{C\to W}\ (t) = \int_0^{+\infty} C(\tau + d\tau; t + dt)d\tau.$$

In case (ii), we have

$$\Delta Q_G(t) = \int_{\tau_c}^{+\infty} C(\tau + d\tau; t + dt)d\tau.$$

The amount of the remaining available resources for GW admission II is $\Delta Q'_G(t) = 0$. The C-state individuals with state ages greater than τ_c will enter the W state. We have $C(\tau + d\tau, t + dt) = 0$ for $\tau \in [\tau_c, +\infty)$ and

$$\triangle_{C\to W}\ (t) = \Delta Q_G(t).$$

When the F-state patients enter the C state, the initial state age is set to be zero. The number of newly added C-state patients is

$$C(0; t + dt) = \int_0^{+\infty} \omega_{F\to C}(\tau)\frac{P_{F\to C}\int_\tau^{+\infty} f_{F\to C}(\tau)'d\tau'}{P_{F\to C}\int_\tau^{+\infty} f_{F\to C}(\tau)'d\tau' + [1 - P_{F\to C}]\int_\tau^{+\infty} f_{F\to R}(\tau)'d\tau'}F(\tau; t)d\tau. \tag{A7}$$

4. G state

The evolution equation of the number of G-state patients with state age τ at time t is

$$\begin{aligned} G(\tau + d\tau; t + dt) = {} & [1 - \omega_{G\to W}(\tau)d\tau]\frac{P_{G\to W}\int_\tau^{+\infty} f_{G\to W}(\tau')d\tau'}{P_{G\to W}\int_\tau^{+\infty} f_{G\to W}(\tau')d\tau' + (1 - P_{G\to W})\int_\tau^{+\infty} f_{G\to M_X}(\tau')d\tau'}G(\tau; t) \\ & + [1 - \omega_{G\to M_X}(\tau)d\tau]\frac{(1 - P_{G\to W})\int_\tau^{+\infty} f_{G\to M_X}(\tau')d\tau'}{P_{G\to W}\int_\tau^{+\infty} f_{G\to W}(\tau')d\tau' + (1 - P_{G\to W})\int_\tau^{+\infty} f_{G\to M_X}(\tau')d\tau'}G(\tau; t). \end{aligned} \tag{A8}$$

The number of newly added G-state patients is

$$G(0; t + dt) = \triangle_{F\to G}\ (t). \tag{A9}$$

5. W state

The evolution equation of the number of W-state patients with state age τ at time t is

$$W(\tau + d\tau; t + dt) = (1 - \omega_{W\to D}(\tau)d\tau)W(\tau; t). \tag{A10}$$

To have the number of W-state patients admitted to the hospital, we denote the available ICU resources at time t as $\Delta Q_U(t)$. There are two cases: (i) there are sufficient resources to accommodate all current W-state patients and (ii) the available resources can accommodate some or none of the W-state patients. The number of W-state patients who change to the U state in $(t, t + dt)$ is $\triangle_{W\to U}\ (t)$.

In case (i), if

$$\Delta Q_U(t) \geqslant \int_0^{+\infty} W(\tau + d\tau; t + dt)d\tau,$$

the amount of the remaining available resources is

$$\Delta Q'_U(t) = \Delta Q_U(t) - \int_0^{+\infty} W(\tau + d\tau; t + dt)d\tau.$$

We have

$$W(\tau + d\tau; t + dt) = 0 \text{ for } \tau \in [0, +\infty), \quad \triangle_{W\to U}(t) = \int_0^{+\infty} W(\tau + d\tau; t + dt)d\tau. \tag{A11}$$

In case (ii), if

$$\Delta Q_U(t) = \int_{\tau_c}^{+\infty} W(\tau + d\tau, t + dt)d\tau$$

and $0 \leqslant \tau_c < +\infty$, the amount of the remaining available resources is $\Delta Q'_U(t) = 0$. We get

$$W(\tau + d\tau; t + dt) = 0 \text{ for } \tau \in [\tau_c, +\infty), \triangle_{W\to U}(t) = \Delta Q_U(t). \tag{A12}$$

The sources of the W-state patients are the G-state patients whose conditions have deteriorated and the C-state patients admitted through GW admission I. The number of newly added W-state patients is

$$W(0; t + dt) = \triangle_{C\to W}(t) + \int_0^{+\infty} \omega_{G\to W}(\tau)\frac{P_{G\to W}\int_\tau^\infty f_{G\to W}(\tau')d\tau'}{P_{G\to W}\int_\tau^\infty f_{G\to W}(\tau')d\tau' + (1 - P_{G\to W})\int_\tau^\infty f_{G\to M_X}(\tau')d\tau'}G(\tau; t)d\tau. \tag{A13}$$

6. U state

The evolution equation of the number of U-state patients with state age τ at time t is

$$\begin{aligned} U(\tau + d\tau; t + dt) = {} & [1 - \omega_{U\to D}(\tau)d\tau]\frac{P_{U\to D}\int_\tau^{+\infty} f_{U\to D}(\tau')d\tau'}{P_{U\to D}\int_\tau^{+\infty} f_{U\to D}(\tau')d\tau' + (1 - P_{U\to D})\int_\tau^{+\infty} f_{U\to G_U}(\tau')d\tau'}U(\tau; t) \\ & + [1 - \omega_{U\to G_U}(\tau)d\tau]\frac{(1 - P_{U\to D})\int_\tau^{+\infty} f_{U\to G_U}(\tau')d\tau'}{P_{U\to D}\int_\tau^{+\infty} f_{U\to D}(\tau')d\tau' + (1 - P_{U\to D})\int_\tau^{+\infty} f_{U\to G_U}(\tau')d\tau'}U(\tau; t). \end{aligned} \tag{A14}$$

Patients in the U state are from the W state and the admitted C-state patients. The number of newly added U-state patients is

$$U(0; t + dt) = \triangle_{W\to U}(t) + \triangle_{C\to U}(t). \tag{A15}$$

7. G_U state

The evolution equation of the number of G_U-state patients with state age τ at time t is

$$G_U(\tau + d\tau; t + dt) = \big[1 - \omega_{G_U\to M_X}(\tau)d\tau\big]G_U(\tau; t). \tag{A16}$$

The number of newly added G_U-state patients is

$$G_U(0; t + dt) = \int_0^{+\infty} \frac{(1 - P_{U\to D})\omega_{U\to G_U}(\tau)\int_\tau^{+\infty} f_{U\to G_U}(\tau')d\tau'}{P_{U\to D}\int_\tau^{+\infty} f_{U\to D}(\tau')d\tau' + (1 - P_{U\to D})\int_\tau^{+\infty} f_{U\to G_U}(\tau')d\tau'}U(\tau, t)d\tau. \tag{A17}$$

8. M_X state

The evolution equation of the number of M_X-state patients with state age τ at time t is

$$M_X(\tau + d\tau; t + dt) = \big[1 - \omega_{M_X\to R}(\tau)d\tau\big]M_X(\tau; t). \tag{A18}$$

The sources of M_X state patients are three: patients from the F, G, and G_U states. The number of newly added M_X-state patients is given by

$$\begin{aligned} M_X(0, t + dt) = {} & \int_0^{+\infty} \omega_{G_U\to M_X}(\tau)G_U(\tau; t)d\tau \\ & + \int_0^{+\infty} \frac{(1 - P_{G\to W})\int_\tau^{+\infty} f_{G\to M_X}(\tau')d\tau'}{P_{G\to W}\int_\tau^{+\infty} f_{G\to W}(\tau')d\tau' + (1 - P_{G\to W})\int_\tau^{+\infty} f_{G\to M_X}(\tau')d\tau'}\omega_{G\to M_X}G(\tau; t)d\tau \end{aligned}$$

$$+\int_0^{+\infty}\frac{(1-P_{F\to C})\int_\tau^{+\infty}f_{F\to M_X}(\tau')d\tau'}{P_{F\to C}\int_\tau^{+\infty}f_{F\to C}(\tau')d\tau'+(1-P_{F\to C})\int_\tau^{+\infty}f_{F\to M_X}(\tau')d\tau'}\omega_{F\to M_X}F(\tau;t)d\tau. \tag{A19}$$

9. R state

The R-state patients come from the M and M_X states. We have

$$R(t+dt)=R(t)+\int_0^{+\infty}\omega_{M_X\to R}M_X(\tau;t)d\tau + \int_0^{+\infty}\frac{(1-P_{M\to F})\int_\tau^{+\infty}f_{M\to R}(\tau')d\tau'}{P_{M\to F}\int_\tau^{+\infty}f_{M\to F}(\tau')d\tau'+(1-P_{M\to F})\int_\tau^{+\infty}f_{M\to R}(\tau')d\tau'}\omega_{M\to R}M(\tau;t)d\tau. \tag{A20}$$

10. D state

The D-state patients come from C, W, and U states. We have

$$D(t+dt)=D(t)+\int_0^{+\infty}\omega_{C\to D}C(\tau;t)d\tau+\int_0^{+\infty}\omega_{W\to D}W(\tau;t)d\tau + \int_0^{+\infty}\frac{P_{U\to D}\int_\tau^{+\infty}f_{U\to D}(\tau')d\tau'}{P_{U\to D}\int_\tau^{+\infty}f_{U\to D}(\tau')d\tau'+(1-P_{U\to D})\int_\tau^{+\infty}f_{U\to G_U}(\tau')d\tau'}\omega_{U\to D}(\tau)U(\tau;t)d\tau. \tag{A21}$$

11. Calculation of available ICU resources $\triangle Q_U(t)$

The available ICU resources are those that have been newly added, those that are released when patients go from the U to the D state or the G_U-state patients, and the resources consumed by the W-state and C-state patients. We have

$$\Delta Q_U(t+dt)=\Delta Q_U(t)+Z_U(t)-\Delta_{W\to U}(t)-\Delta_{C\to U}(t) + \int_0^{+\infty}\frac{P_{U\to D}\int_\tau^{+\infty}f_{U\to D}(\tau')d\tau'}{P_{U\to D}\int_\tau^{+\infty}f_{U\to D}(\tau')d\tau'+(1-P_{U\to D})\int_\tau^{+\infty}f_{U\to G_U}(\tau')d\tau'}\omega_{U\to D}(\tau)U(\tau;t)d\tau + \int_0^{+\infty}\frac{(1-P_{U\to D})\int_\tau^{+\infty}f_{U\to G_U}(\tau')d\tau'}{P_{U\to D}\int_\tau^{+\infty}f_{U\to D}(\tau')d\tau'+(1-P_{U\to D})\int_\tau^{+\infty}f_{U\to G_U}(\tau')d\tau'}\omega_{U\to G_U}(\tau)U(\tau;t)d\tau. \tag{A22}$$

12. Calculation of available GW resources $\Delta Q_G(t)$

The available GW resources are those that have been newly added, those released when patients switch from the G to the M_x state and from the W to the D state or the U state, and the resources consumed by the C-state and F-state patients. We have

$$\Delta Q_G(t+dt)=\Delta Q_G(t)+Z_G(t)-\Delta_{C\to W}(t)-\Delta_{F\to G}(t)+\int_0^{+\infty}W(\tau;t)\omega_{W\to D}(\tau)d\tau + \int_0^{+\infty}\frac{(1-P_{G\to W})\int_\tau^{+\infty}f_{G\to M_X}(\tau')d\tau'}{P_{G\to W}\int_\tau^{+\infty}f_{G\to W}(\tau')d\tau'+(1-P_{G\to W})\int_\tau^{+\infty}f_{G\to M_X}(\tau')d\tau'}\omega_{G\to M_X}(\tau)G(\tau;t)d\tau. \tag{A23}$$

APPENDIX B: PARAMETER ESTIMATION

1. Average delay and fraction of state transition

As shown in Table I, we assume that the time delay of patients switching from state M to F follows the normal distribution with the average of five days [17]. The distributions of the time delays associated with other state transitions are also assumed to be normal. The average time delays from the U to the D and G_u states are set as seven days [18,19] and eight days [18], respectively, and that from the G to the W state is three days [17]. In our model setting, the clinical symptoms of the F-state and G-state patients are identical, and the difference lies in whether the patients are admitted (i.e., occupying GW beds), and the same rule applies to the C-state and W-state patients.

The average time delay from the F to the C state is two days [17,29], and that from the G and F states to the M_X state is eight days [12]. The average time delays from the W and C states to the D state are assumed to be three days and one day, respectively.

As shown in Table II, we set the average mortality rate of ICU patients as 61.5% [18,30]. During the epidemic, due to the different testing methods and conditions, the fraction of M-state patients in the F state is different in different time periods. According to the analysis of 32583 cases of laboratory-confirmed patients in Wuhan and reconstruction

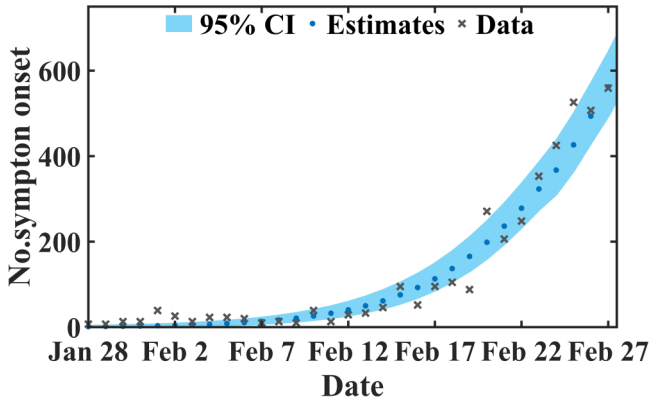

FIG. 7. Validation of our backtracking method. The validation method is based on the data of early laboratory-confirmed cases in Lombardy area (January 28–February 27, 2020). The incidence data of patients are estimated from the laboratory reported confirmed cases and the distribution of the time delay between the onset of the disease and the reported confirmation. The blue dots and black crosses represent, respectively, the incidence estimation curve and the retrospective survey data of laboratory-confirmed cases in the literature [37]. The light blue shaded region represents the 95% confidence interval of the estimation.

of the epidemic trend [4,31], we divide the epidemic process in terms of the fraction of the transition from the M to the F state into five stages. As shown in Table III, the time periods and the transition fractions of the five stages are [31]: from 8 December 2019 to 10 January 2020 with the fraction $|P|_{M\to F}(1) = 53.10\%$; from January 11 to January 22 with $|P|_{M\to F}(2) = 35.10\%$; from January 23 to February 1 with $|P|_{M\to F}(3) = 23.50\%$; from February 2 to February 16 with $|P|_{M\to F}(4) = 15.90\%$; and after February 17 with $|P|_{M\to F}(5) = 10.30\%$ [31].

For Lombardy, the early epidemic can be divided into three stages [32]. We add a new time point: March 21, 2020, leading to a division into four stages. After (including) this date, the intervention measures in Lombardy and Italy as a whole reached maximum [33]. As shown in Table IV, the time periods and the fractions associated with the four stages are from 28 January 2020 to 19 February 2020 with $|P|_{M\to F}(1) = 63\%$; from February 20 to February 25 with $|P|_{M\to F}(2) = 61\%$; from February 26 to March 20 with $|P|_{M\to F}(3) = 56\%$ [32]; after March 21 with $|P|_{M\to F}(4)$=32.21%.

TABLE V. Actual medical resource deployment plan for Wuhan, China [34].

Date	Beds(ICU+General ward)
2020/2/1	842
2020/2/2	842
2020/2/3	1962
2020/2/4	2234
2020/2/5	2264
2020/2/6	2515
2020/2/7	2657
2020/2/8	2911
2020/2/9	3263
2020/2/10	4307
2020/2/11	4461
2020/2/12	5873
2020/2/13	6078
2020/2/14	6636
2020/2/15	6926
2020/2/16	7035
2020/2/17	7067
2020/2/18	7225
2020/2/19	7296
2020/2/20	7560
2020/2/21	7560
2020/2/22	7844
2020/2/23	7844
2020/2/24	7936
2020/2/25	8194

TABLE VI. Actual medical resource deployment in Lombardy, Italy [38].

Date	Beds(ICU)	Beds(General ward)
2020/2/24	19	76
2020/2/25	25	79
2020/2/26	25	79
2020/2/27	41	172
2020/2/28	47	235
2020/2/29	80	256
2020/3/1	106	406
2020/3/2	127	478
2020/3/3	167	698
2020/3/4	209	877
2020/3/5	244	1169
2020/3/6	309	1622
2020/3/7	359	1661
2020/3/8	399	2217
2020/3/9	440	2802
2020/3/10	466	3319
2020/3/11	560	3852
2020/3/12	605	4247
2020/3/13	650	4435
2020/3/14	732	4898
2020/3/15	767	5500
2020/3/16	823	6171
2020/3/17	879	6953
2020/3/18	924	7285
2020/3/19	1006	7387
2020/3/20	1050	7735
2020/3/21	1093	8258
2020/3/22	1142	9439
2020/3/23	1183	9266
2020/3/24	1194	9711
2020/3/25	1236	10 026
2020/3/26	1263	10 681
2020/3/27	1292	11 137
2020/3/28	1319	11 152
2020/3/29	1328	11 613
2020/3/30	1330	11 815
2020/3/31	1334	11 883
2020/4/1	1342	12 009
2020/4/2	1351	12 009
2020/4/3	1381	12 009

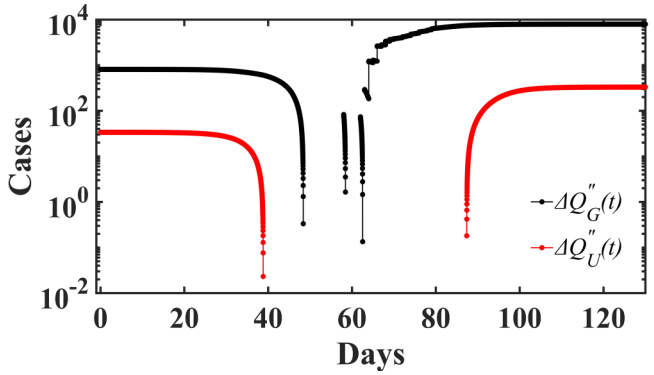

FIG. 8. Remaining resources of GW and ICU in Wuhan after two admission processes. The ordinate is on a logarithmic scale, and the red and black dots indicate that the amounts of remaining resources of Wuhan ICU and GW at time t are $\Delta Q_U(t)$ and $\Delta Q_G(t)$, respectively. The missing areas of the red and blue dots represent the period of $\Delta Q_U = 0$ and $\Delta Q_G = 0$, respectively.

By using the weighted least-squares method, we obtain the optimal estimates of the parameters. In particular, for the Wuhan scenario, we obtain the optimal set of parameters $\Xi = (|P|_{F\to C}, |P|_{G\to W})$ by minimizing the sum of the weighted difference squares between the reported death curve [23,34] $D(t_i)$ and the model fitting value $F(t_i, \Xi)$ $(i = 0, 1, \ldots, n-1)$, where

$$\widehat{\Xi} = \text{argmin} \sum_i w_{t_i}[F(t_i, \Xi) - D(t-i)]^2. \tag{B1}$$

To quantify the uncertainties of parameter estimation, we resort to the general bootstrap method [35,36] and then obtain the 95% confidence interval of the estimated parameter values. As shown in Table II, the optimal values and 95% confidence interval of the average transition fractions from G to W state and from F to C state in Wuhan are $|P|^*_{F\to C}$=23.26% (21.24%, 28.55%) and $|P|^*_{G\to W}$=20.76% (18.15%, 24.56%), respectively.

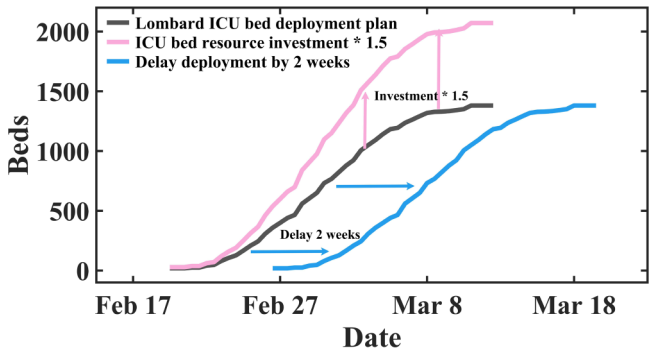

FIG. 9. Effects of varying ICU deployment plan in Lombardy, Italy. The black curve shows Lombardy's actual ICU medical resource deployment plan, as displayed in Table IV. The blue and red curves show the two cases where Lombardy's government would delay the deployment of ICU resources by two weeks (DT = 14 and RI = 1) and the ICU resource is expanded by a factor of 1.5 (DT = 0 and RI = 1.5), respectively.

For Lombardy, the required optimal parameters are $\Xi = (|P|_{M\to F}(4), |P|_{F\to C}, |P|_{G\to W})$. As shown in Tables II and IV, the optimal values and the 95% confidence intervals of the average transition fractions in Lombardy are $|P|^*_{M\to F}$=32.21%(29.33%, 34.14%), $|P|^*_{F\to C}$=56.45% (53.33%, 65.23%, and $|P|^*_{G\to W}$=41.22% (39.24%, 55.43%, respectively.

2. Incidence dates estimated from the confirmed data

According to the average time delay from the onset date to the diagnosis date of laboratory-confirmed cases in the literature [31,32], we determine the distributions of state transition delay from onset to diagnosis for Wuhan and Lombardy using a backtracking method based on non-Markov processes that we have developed, where the incidence dates are deduced from the confirmed cases. We designate a new state, the J state that occurs after M state, to denote that a patient has been confirmed. The deduction process from J to M state is described by the following difference equations:

$$\begin{aligned} J(\tau + d\tau; t - dt) &= [1 - \omega_{J\to M}]J(\tau; t)d\tau, \\ Z_M(t - dt) &= \int_0^{+\infty} J(\tau; t)\omega_{J\to M}(\tau)d\tau, \\ J(0; t - dt) &= Z_J(t), \end{aligned} \tag{B2}$$

where $\omega_{J\to M}(\tau)$ is the conditional rate of transition from J to M states with state age τ, with the specific form

$$\omega_{J\to M}(\tau) = \frac{f_{J\to M}(\tau)}{\int_\tau^{+\infty} f_{J\to M}(\tau')d\tau'},$$

with $f_{J\to M}(\tau)$ being the probability density function of the transition time delay from J to M state. For Wuhan and Lombardy, the state transition delays are assumed to follow the Gamma distribution with the mean value of 9.5 and 7.3 days, respectively. The quantities $Z_M(t)$ and $Z_J(t)$ are the numbers of new patients in the M and J states in the time interval $(t, t + dt)$, respectively, where the former is our estimated value and the latter is the newly reported, daily confirmed data. We validate the accuracy of this method using data

TABLE VII. Comparison of different medical resource deployment plans (simultaneously adjusting GW and ICU resources).

	Simulated death toll	
	Wuhan	Lombard
The actual deployment	2553	16363
Sufficient resources	Decrease by 33%	Decrease by 22%
Deploy 7 days in advance	Decrease by 14%	Decrease by 3.4%
Deploy 14 days in advance	Decrease by 18%	Decrease by 4%
7 days delay in deployment	Increase by 36%	Increase by 14%
14 days delay in deployment	Increase by 57%	Increase by 28%
Resource input *0.5	Increase by 43%	Increase by 26%
Resource input *2	Decrease by 22%	Decrease by 15%
Resource input *5	Decrease by 32%	Decrease by 22%
Deploy 7 days in advance and invest resources *2	Decrease by 30%	Decrease by 17%

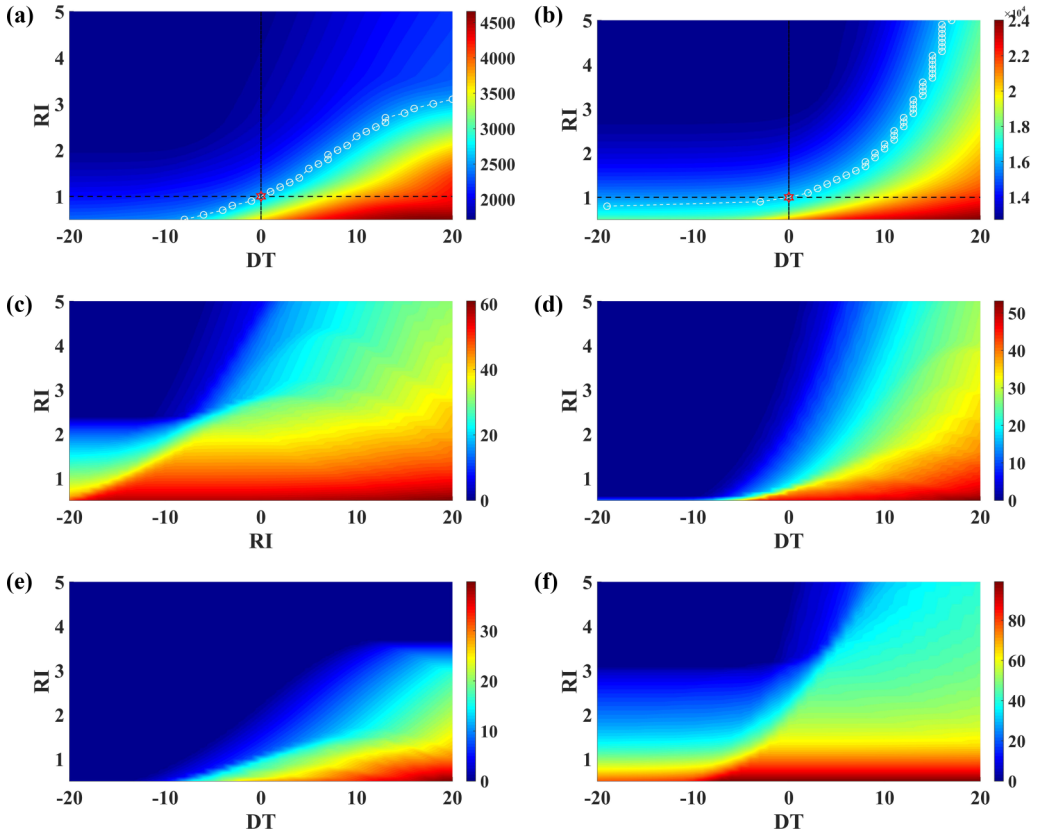

FIG. 10. Changes in the number of deaths and overload days for simultaneous GW and ICU resource deployment for Wuhan and Lombardy. [(a) and (b)] Numbers of deaths in Wuhan and Lombardy, respectively. [(c) and (d)] GW ward overload days in Wuhan and Lombardy, respectively. [(e) and (f)] ICU ward overload days in Wuhan and Lombardy, respectively.

from early laboratory-confirmed cases in Lombardy (January 28-February 27, 2020) [37], as shown in Fig. 7.

There are 50 008 laboratory/clinically confirmed cases in Wuhan, and 93,901 such cases in Lombardy. We use the same backtracking method to trace the onset time of the nonlaboratory confirmed cases. The incidence dates of all confirmed cases is April 15 for Wuhan and June 30 for Lombardy.

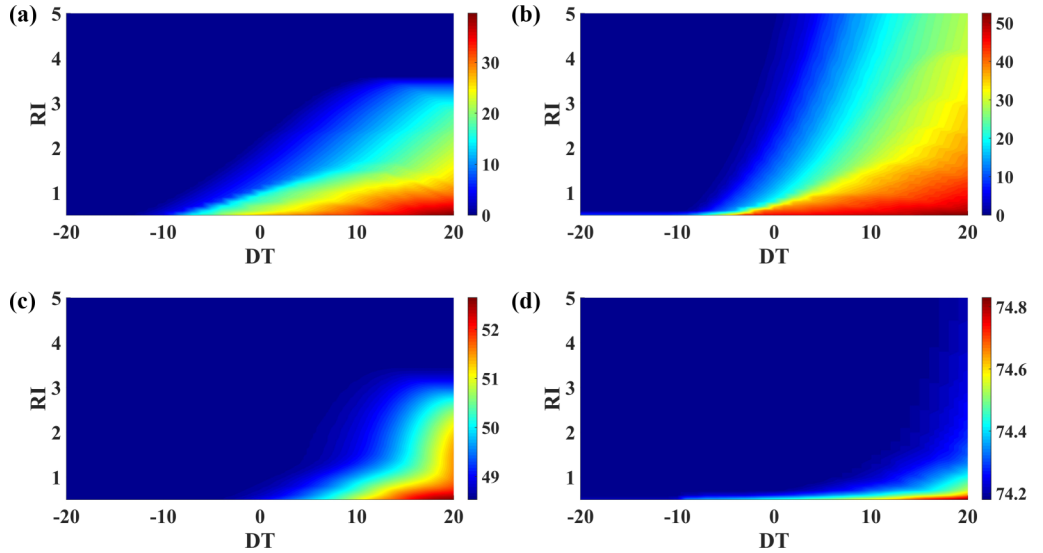

FIG. 11. Change in overload days for fixed ICU resource deployment and varying GW resource deployment. [(a) and (b)] Overload days of GW in Wuhan and Lombardy, respectively. [(c) and (d)] Overload days of ICU in Wuhan and Lombardy, respectively.

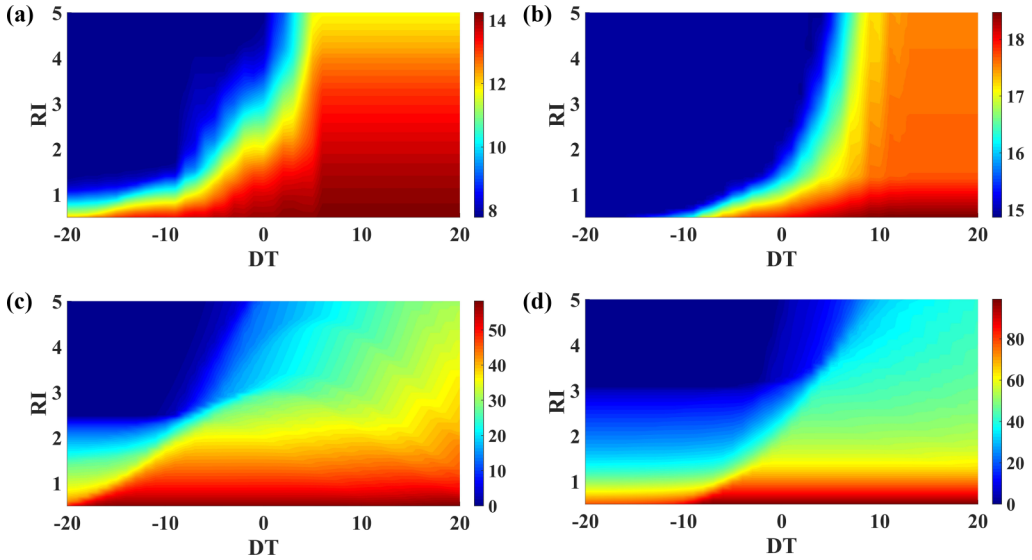

FIG. 12. Change in overload days for fixed GW resource deployment and varying ICU resource deployment. [(a) and (b)] Overload days of GW in Wuhan and Lombardy, respectively. [(c) and (d)] Overload days of ICU in Wuhan and Lombardy, respectively.

3. Actual medical resource deployment

In response to an emerging public health event, the government usually adopts the policy of gradual deployment of medical resources to designated hospitals. The dedicated medical resources in an area typically slowly increase with time.

As shown in Tables V and VI, we collect medical resource deployment data dedicated to COVID-19 patients in Wuhan and Lombardy, which include GW and ICU beds [34,38]. For Wuhan, we count the number of open beds in critical hospitals. As the official reports do not distinguish GW beds from ICU beds, we assume that ICU beds account for 4% of the total beds [39].

APPENDIX C: PARAMETERS OF MEDICAL RESOURCE

1. Health care system stress metrics

In addition to the final number of deaths, the load of local special medical resources for COVID-19 patients is also an important indicator to measure the stress of the medical system. We simulate and obtain the full-load working days of GW and ICU in Wuhan and Lombardy during the recovery stage, denoted as O_G and O_U, respectively. When the GW or ICU is under full load, no new patients can be admitted.

In Fig. 8, the remaining ICU resources after ICU admission I/II at time t is denoted as $\triangle\, Q_U(t)$, and the corresponding quantity for GW is $\triangle\, Q_G(t)$, which are given by

$$O_U = \int_{T_s}^{T_e} \delta[\triangle Q_U(t)]dt \quad \text{and} \quad O_G = \int_{T_s}^{T_e} \delta[\triangle Q_G(t)]dt, \tag{C1}$$

where $\delta(*)$ is the Dirac-δ function, the integration interval is the whole recovery stage, T_s and T_e represent the starting and ending time of the recovery stage, respectively. For example, T_s in Wuhan is December 8, 2019 and T_e is April 16, 2020. The time interval is set to be $dt = 0.01$ day.

2. Deployment plan of dedicated medical resources for COVID-19

The deployment plan of local officially dedicated medical resources is determined by two key factors: the deployment time DT and resource input RI, where DT is the time for local authorities to start the deployment of the dedicated medical resources and RI represents the financial and personnel investment of the dedicated medical resources as characterized by the open GW and ICU beds.

Figure 9 shows the effects of varying ICU deployment plan in Lombardy, where the black curve represents Lombardy's current dedicated ICU deployment plan. The actual deployment time serves as the benchmark DT = 0 and the actual resource input is RI = 1. The left blue arrow indicates the scenario where the Lombard authorities had delayed the deployment time by two weeks (DT = 14 and RI = 1), and the red up arrow corresponds to the scenario where the official resource input had been increased by 50% (DT = 0 and RI = 1.5).

3. Effects of varying medical resource deployment on death toll

We study the impacts of varying both GW and ICU resources on the number of deaths. As shown in Table VII and Fig. 10, if resources had been deployed one week in advance, the death toll in Wuhan and Lombardy would have been reduced by 14% and 3.4%, respectively. If the input of resources had been doubled, the death toll would have been suppressed by 22% and 15%, respectively. If the deployment had been one week ahead of the actual time and the input of resources had been doubled, the death toll would have been reduced by 30% and 17% in Wuhan and Lombardy, respectively, indicating that varying the two factors simultaneously can be more effective at reducing the death toll. Figure 11 demonstrates the change in the ward overload days when GW deployment is modified while ICU deployment is unchanged. Figure 12 displays the change in the ward overload days when

TABLE VIII. Reference transition fraction matrix among states and age distribution of patients in M state.

Parameter(Reference matrix)		Estimate/assumption	Definition	Justification
$R_{M\to F}$	Wuhan	[0.0953 0.2430 0.2730]	Reference transformation matrix from M to F state	[12]
	Lombard	[0.0827 0.2430 0.2730]		
$R_{F\to C/G\to W}$	Wuhan	[0.0982 0.1419 0.2393]	Reference transformation matrix from F/G to C/W state	Fitting according to the reported death data fitting
	Lombard	[0.1249 0.2240 0.2393]		
$R_{U\to D}$	Wuhan/Lombard	[0.2297 0.7727 0.8571]	Reference transformation matrix from U to D state	[19]
D_M	Wuhan	[0.8240 0.1260 0.0500]	Age group distribution matrix of M-state patients	[21]
	Lombard	[0.6050 0.1420 0.2530]		

GW deployment is unchanged but the ICU deployment plan is modified.

APPENDIX D: ESTIMATION OF STATE TRANSITION FRACTIONS FOR DIFFERENT AGE GROUPS

In the main text, we divide the patients in Wuhan and Lombardy into three age groups: [0-69], [70-79] and [80+], at intervals of actual mortality rate 10%. The transition fractions from M to F state, from G to W state, from F state to C state and the ICU mortality rates are different for different age groups. We articulate a linear regression method to estimate the state transition fractions for different age groups. Take as an example the transition from M to F state. We first introduce the matrix of average transition fraction, as (detailed in Appendix B)

$$|P|_{M\to F} = \begin{pmatrix} |p|_1^{M\to F} \\ \vdots \\ |p|_1^{M\to F} \end{pmatrix}, \tag{D1}$$

where $|p|_i^{M\to F}$ represents the average transition fraction from M to F state at stage i. The age distribution of patients in the M state at different stages is

$$D_M = \begin{pmatrix} d_{1,1}^M & \cdots & d_{1,y}^M \\ \vdots & \ddots & \vdots \\ d_{s,1}^M & \cdots & d_{s,y}^M \end{pmatrix}, \tag{D2}$$

where $d_{i,j}^M$ is the fraction of patients in age group j among all patients at stage i. We then define the matrix of average transition fraction from M sto F state in different age groups as

$$R_{M\to F} = \left[r_1^{M\to F} \cdots r_y^{M\to F}\right], \tag{D3}$$

where $r_j^{M\to F}$ is the average transition fraction from M to F state in the jth age group during the whole recovery period.

The parameters can be obtained empirically. We make a linear transformation of matrix $R_{M\to F}$ to obtain the estimation of the fraction of the M-state patients to F state whose weighted average transition fraction is $|P|_{M\to F}$. The linear

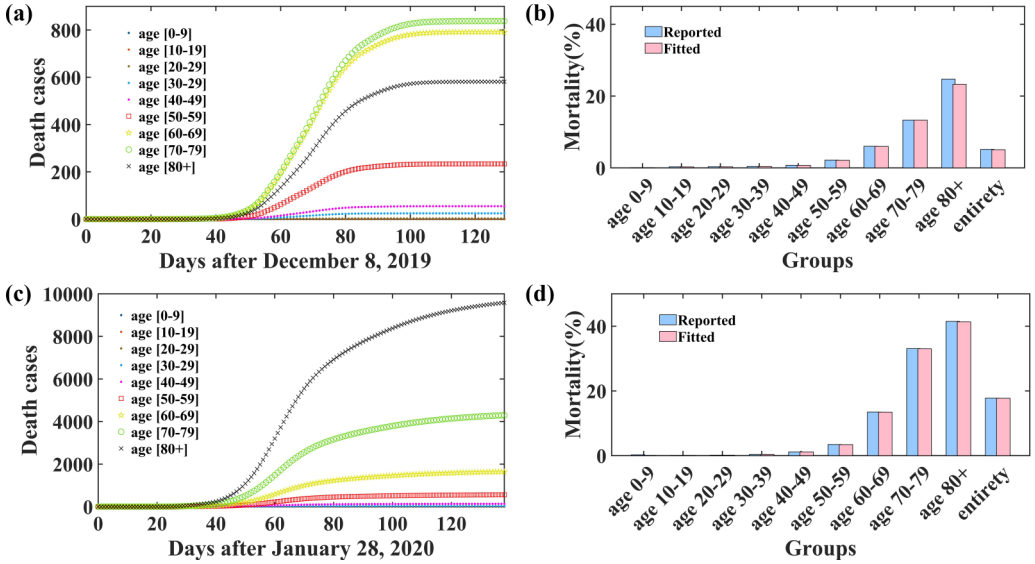

FIG. 13. Mortality rate of different age groups and evolution of the number of deaths. (a) Empirical data of mortality rate of different age groups in Wuhan versus model prediction. (b) Change in the number of deaths by age groups in Wuhan. (c) Empirical mortality data for all age groups in Lombardy vs model prediction. (d) Number of deaths in Lombardy in each age group. In (a) and (c), the legends from top to bottom are age [0-9] to [80]. In (b) and (d), the light blue and pink cylinders, respectively, represent the actual and simulated mortalities in each group, where the x axis marks the groups with growing age.

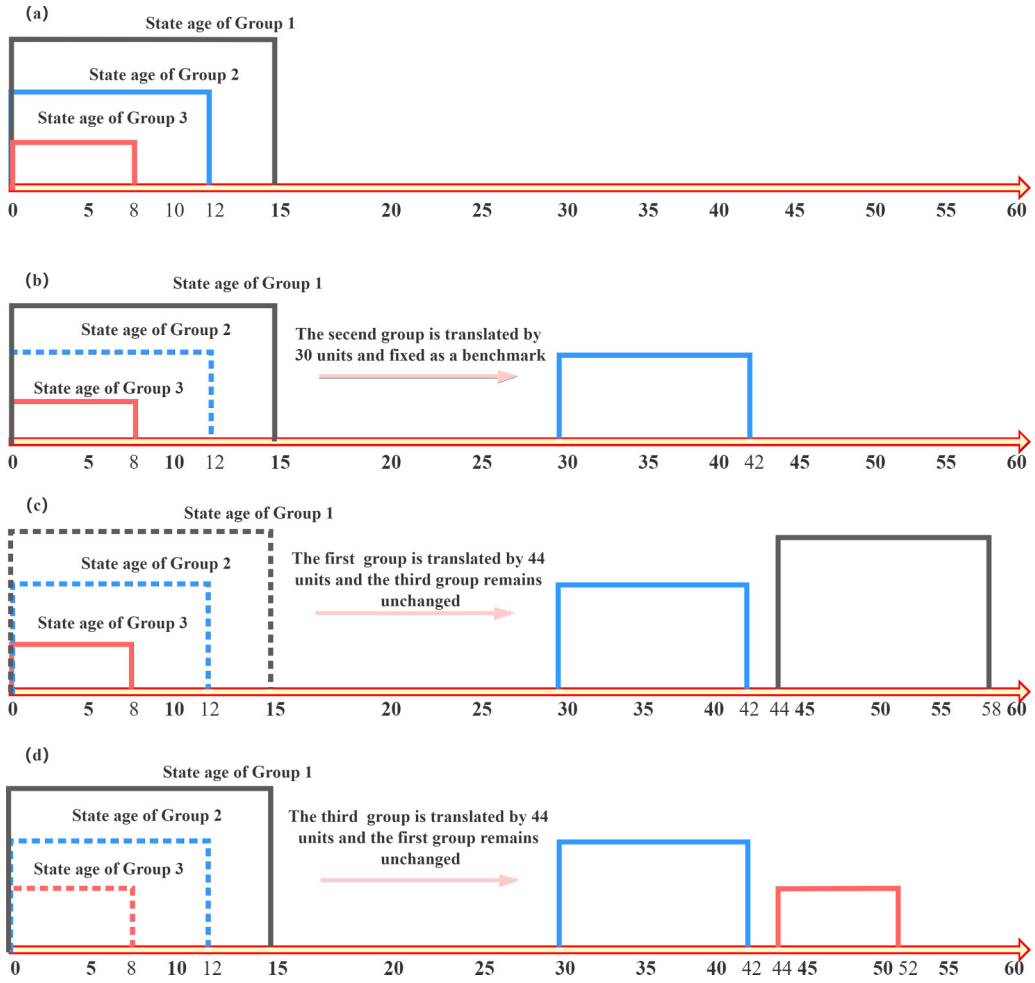

FIG. 14. Weighted state age of each age group. (a) Initial state age of the three groups of patients. (b) The state age of the second group is translated by 30 units and fixed as a benchmark. (c) The state age of the first group is translated by 44 units, and that of the third group remains unchanged. (d) The state age of the third group is translated by 44 units, and that of the first group remains unchanged. The width of each box represents the maximal state age of each age group.

transformation LT of different stages is given by

$$\mathrm{LT} = \frac{|P|_{M\to F}}{D_M \cdot R^T_{M\to F}}, \tag{D4}$$

Finally, the transition fractions from M state to F state for different age groups are

$$P_{M\to F} = LT \cdot \mathrm{ext}(1, y) \cdot R_{M\to F} = \begin{pmatrix} p^{M\to F}_{1,1} & \cdots & p^{M\to F}_{1,y} \\ \vdots & \ddots & \vdots \\ p^{M\to F}_{s,1} & \cdots & p^{M\to F}_{s,y} \end{pmatrix}, \tag{D5}$$

where the rows represent stages, the columns represent groups, and $p^{M\to F}_{i,j}$ denotes the transition fraction from M to F state in the jth age group at stage i. $\mathrm{ext}(n, m)$ is an auxiliary matrix of n rows and m columns whose elements are all one. The patient age distribution matrix D_F is

$$D_F = \begin{pmatrix} p^{M\to F}_{1,1} \cdot d^M_{1,1} & \cdots & p^{M\to F}_{1,y} \cdot d^M_{1,y} \\ \vdots & \ddots & \vdots \\ p^{M\to F}_{s,1} \cdot d^M_{s,1} & \cdots & p^{M\to F}_{s,y} \cdot d^M_{s,y} \end{pmatrix}. \tag{D6}$$

Where R and D_M and $|P|$ are known (see Tables III and VIII.). Similarly, we can obtain the estimation of the transition fractions among other states for different age groups at different stages. To simplify the model, we assume that the age distribution of M-state patients at each stage is identical ($d^M_{1,j} = d^M_{2,j} = \ldots = d^M_{s,j}$, $j = 1, 2, \ldots, y$), as shown in Table VIII. The reference transition fraction matrix among the states is also given in Table VIII. Finally, as shown in Fig. 13, we divide the patient population into nine groups.

APPENDIX E: A SCHEME TO REALIZE DIFFERENT ADMISSION STRATEGIES: TRANSLATION WEIGHTING

We have developed a translation weighting scheme to achieve different admission strategies. We set the priority weight of the ith age group as w_i and denote τ_i as the state age of the ith age group. After weighting, the new state age τ'_i becomes

$$\tau'_i = \tau_i + w_i. \tag{E1}$$

TABLE IX. Average wards per 100 000 people.

Country/Resource	General ward beds per 100 000 people	ICU beds per 100 000 people
China	400	3.6
Korea	530	10.6
Italy	333	12.5
Spain	269	9.7
Justification	[40]	

TABLE X. Proportion of confirmed cases in each age group in each country.

Age distribution of confirmed cases				
Groups/Country	China	Italy	Korea	Spain
0–9	0.009312	0.006245	0.012344	0.002639
10–19	0.01229	0.008122	0.052919	0.005433
20–29	0.081013	0.04061	0.272823	0.05018
30–39	0.170129	0.069164	0.106507	0.096366
40–49	0.191865	0.128128	0.133589	0.150725
50–59	0.224033	0.198045	0.183732	0.186374
60–69	0.192134	0.173838	0.126316	0.165787
70–79	0.087706	0.185173	0.066411	0.158476
80+	0.031519	0.188301	0.045359	0.18402
Justification		[41]		

The state age of patients in the age group i is increased by w_i units (days). The patients are then admitted according to the weighted state age under FCFS.

We divide the patients into three age groups, which are patients younger than 70 years old, patients between 70 and 80 years old, and patients older than 80 years. Figure 14 presents examples of several weight combinations. For example, in Fig. 14(b), the weights of the three age groups are $w_1 = 0$, $w_2 = 30$, and $w_3 = 0$, respectively. The state ages of the three age groups after weighting are also illustrated.

APPENDIX F: MEDICAL RESOURCES PER CAPITA AND THE AGE DISTRIBUTION OF CONFIRMED CASES IN FOUR COUNTRIES

We collect data on ward resources per 100 000 people per capita in China, South Korea, Italy, and Spain, as well as the fractions of COVID-19 confirmed cases in each age group, as shown in Tables IX and X.

[1] Weekly Operational Update: Coronavirus disease 2019 (COVID-19) World Health Organization 4 September 2020 (2020).

[2] M. U. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, L. Du Plessis, N. R. Faria, R. Li, W. P. Hanage *et al.*, The effect of human mobility and control measures on the COVID-19 epidemic in China, Science **368**, 493 (2020).

[3] S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, and M. Lipsitch, Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period, Science **368**, 860 (2020).

[4] X. Hao, S. Cheng, D. Wu, T. Wu, X. Lin, and C. Wang, Reconstruction of the full transmission dynamics of COVID-19 in Wuhan, Nature (London) **584**, 420 (2020).

[5] Y.-S. Long, Z.-M. Zhai, L.-L. Han, J. Kang, Y.-L. Li, Z.-H. Lin, L. Zeng, D.-Y. Wu, C.-Q. Hao, M. Tang *et al.*, Quantitative assessment of the role of undocumented infection in the 2019 novel coronavirus (COVID-19) pandemic, arXiv:2003.12028 (2020).

[6] Z.-M. Zhai, Y.-S. Long, J. Kang, Y.-L. Li, L. Zeng, L.-L. Han, Z.-H. Lin, Y.-Q. Zeng, D.-Y. Wu, M. Tang *et al.*, State-by-State prediction of likely COVID-19 scenarios in the United States and assessment of the role of testing and control measures, medRxiv 2020.04.24.20078774 (2020), doi: 10.1101/2020.04.24.20078774.

[7] Z.-M. Zhai, Y.-S. Long, M. Tang, Z. Liu, and Y.-C. Lai, Optimal inference of the start of COVID-19, Phys. Rev. Res. **3**, 013155 (2021).

[8] S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton *et al.*, Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe, Nature (London) **584**, 257 (2020).

[9] J. Dehning, J. Zierenberg, F. P. Spitzner, M. Wibral, J. P. Neto, M. Wilczek, and V. Priesemann, Inferring change points in the spread of covid-19 reveals the effectiveness of interventions, Science **369**, eabb9789 (2020).

[10] J. S. Jia, X. Lu, Y. Yuan, G. Xu, J. Jia, and N. A. Christakis, Population flow drives spatio-temporal distribution of COVID-19 in China, Nature (London) **582**, 389 (2020).

[11] S. T. Ali, L. Wang, E. H. Lau, X.-K. Xu, Z. Du, Y. Wu, G. M. Leung, and B. J. Cowling, Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions, Science **369**, 1106 (2020).

[12] N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg *et al.*, Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand, MRC Centre for Global Infectious Disease Analysis, Imperial College London (2020).

[13] N. G. Davies, A. J. Kucharski, R. M. Eggo, A. Gimma, W. J. Edmunds, T. Jombart, K. O'Reilly, A. Endo, J. Hellewell, E. S. Nightingale *et al.*, Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study, The Lancet Public Health **5**, e375 (2020).

[14] I. F. Miller, A. D. Becker, B. T. Grenfell, and C. J. E. Metcalf, Disease and healthcare burden of COVID-19 in the United States, Nat. Med. **26**, 1212 (2020).

[15] E. J. Emanuel, G. Persad, R. Upshur, B. Thome, M. Parker, A. Glickman, C. Zhang, C. Boyle, M. Smith, and J. P. Phillips, Fair allocation of scarce medical resources in the time of Covid-19, New Eng. J. Med. **382**, 2049 (2020).

[16] D. B. Whiteand B. Lo, A framework for rationing ventilators and critical care beds during the COVID-19 pandemic, J. Ame. Med. Asso. **323**, 1773 (2020).

[17] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong *et al.*, Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China, J. Ame. Med. Asso. **323**, 1061 (2020).

[18] X. Yang, Y. Yu, J. Xu, H. Shu, H. Liu, Y. Wu, L. Zhang, Z. Yu, M. Fang, T. Yu *et al.*, Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study, Lancet Respir. Med. **8**, 475 (2020).

[19] G. Grasselli, A. Zangrillo, A. Zanella, M. Antonelli, L. Cabrini, A. Castelli, D. Cereda, A. Coluccello, G. Foti, R. Fumagalli *et al.*, Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2 admitted to ICUs of the Lombardy Region, Italy, J. Ame. Med. Asso. **323**, 1574 (2020).

[20] S. Garg, Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019 COVID-NET, 14 States, March 1–30, 2020, MMWR Morb. Mortal. Wkly. Rep. **69**, 458 (2020).

[21] C. P. E. R. E. Novel *et al.*, The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China, Zhonghua Liu Xing Bing Xue Za Zhi **41**, 145 (2020).

[22] T. C. COVID and R. Team, Severe outcomes among patients with coronavirus disease 2019 (COVID-19)-united states, February 12-March 16, 2020., MMWR Morb. Mortal. Wkly. Rep. **69**, 343 (2020).

[23] https://www.epicentro.iss.it/coronavirus/sars-cov-2-dashboard (2020).

[24] http://mzj.wuhan.gov.cn/zwgk_918/fdzdgk/ggfw/shfl/201901/t 20190125_1000644.shtml (2020).

[25] https://www.citypopulation.de/en/italy/localities/lombardia/ (2020).

[26] M. Starnini, J. P. Gleeson, and M. Boguñá, Equivalence between Non-Markovian and Markovian Dynamics in Epidemic Spreading Processes, Phys. Rev. Lett. **118**, 128301 (2017).

[27] M. Feng, S.-M. Cai, M. Tang, and Y.-C. Lai, Equivalence and its invalidation between non-markovian and markovian spreading dynamics on complex networks, Nat. Commun. **10**, 4677 (2019).

[28] Z.-H. Lin, M. Feng, M. Tang, Z. Liu, C. Xu, P. M. Hui, and Y.-C. Lai, Non-markovian recovery makes complex networks more resilient against large-scale failures, Nat. Commun. **11**, 1 (2020).

[29] K. Gaythorpe, N. Imai, G. Cuomo-Dannenburg, M. Baguelin, S. Bhatia, A. Boonyasiri, A. Cori, Z. Cucunuba Perez, A. Dighe, I. Dorigatti *et al.*, Report 8: Symptom progression of COVID-19, Imperial College London (2020).

[30] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu *et al.*, Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study, The Lancet **395**, 1054 (2020).

[31] A. Pan, L. Liu, C. Wang, H. Guo, X. Hao, Q. Wang, J. Huang, N. He, H. Yu, X. Lin *et al.*, Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China, J. Ame. Med. Asso. **323**, 1915 (2020).

[32] D. Cereda, M. Tirani, F. Rovida, V. Demicheli, M. Ajelli, P. Poletti, F. Trentini, G. Guzzetta, V. Marziano, A. Barone *et al.*, The early phase of the COVID-19 outbreak in Lombardy, Italy, arXiv:2003.09320 (2020).

[33] https://en.wikipedia.org/wiki/covid-19_pandemic_in_italy (2020).

[34] http://wjw.wuhan.gov.cn/ (2020).

[35] R. J. Tibshirani and B. Efron, An introduction to the bootstrap, Mono. Stat. Appl. Prob. **57**, 1 (1993).

[36] G. Chowell, Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts, Infect. Dis. Model. **2**, 379 (2017).

[37] https://en.calameo.com/books/005243496f711f39d24b9 (2020).

[38] https://covid19.healthdata.org/italy/lombardia (2020).

[39] https://www.qianzhan.com/analyst/detail/220/200312-d2e29fce.html (2020).

[40] https://www.statista.com/chart/21105/number-of-critical-care-beds-per-100000-inhabitants/ (2020).

[41] https://www.insidermonitor.com/covid19/zh/ (2020).