Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.jfma-online.com

Original Article

Classification of community-acquired outbreaks for the global transmission of COVID-19: Machine learning and statistical model analysis

Wei-Chun Wang [a,1], Ting-Yu Lin [a,1], Sherry Yueh-Hsia Chiu [b], Chiung-Nien Chen [c,d], Pongdech Sarakarn [e], Mohd Ibrahim [f], Sam Li-Sheng Chen [g], Hsiu-Hsi Chen [a], Yen-Po Yeh [a,h,*]

[a] *Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan*
[b] *Department of Health Care Management and Healthy Aging Research Center, Chang Gung University, Taoyuan, Taiwan*
[c] *Center for Functional Image and Interventional Therapy, National Taiwan University, Taipei, Taiwan*
[d] *Department of Surgery, National Taiwan University Hospital, Taipei, Taiwan*
[e] *Epidemiology and Biostatistics Department, Faculty of Public Health, Khon Kaen University, Khon Kaen, Thailand*
[f] *Department of Community and Family Medicine, University Malaysia Sabah, Malaysia*
[g] *School of Oral Hygiene, College of Oral Medicine, Taipei Medical University, Taipei, Taiwan*
[h] *Changhua Public Health Bureau, Changhua, Taiwan*

Received 26 April 2021; received in revised form 6 May 2021; accepted 7 May 2021

KEYWORDS
COVID-19;
Machine learning;
Cluster analysis;
Community-acquired outbreak

Background: As Coronavirus disease 2019 (COVID-19) pandemic led to the unprecedent large-scale repeated surges of epidemics worldwide since the end of 2019, data-driven analysis to look into the duration and case load of each episode of outbreak worldwide has been motivated.
Methods: Using open data repository with daily infected, recovered and death cases in the period between March 2020 and April 2021, a descriptive analysis was performed. The susceptible-exposed-infected-recovery model was used to estimate the effective productive number (R_t). The duration taken from $R_t > 1$ to $R_t < 1$ and case load were first modelled by using the compound Poisson method. Machine learning analysis using the K-means clustering method was further adopted to classify patterns of community-acquired outbreaks worldwide.
Results: The global estimated R_t declined after the first surge of COVID-19 pandemic but there

* Corresponding author. Changhua Public Health Bureau No.162, Sec. 2, Jhongshan Rd., Changhua City, Changhua Country 500, Taiwan.
E-mail address: lgyeh@hotmail.com (Y.-P. Yeh).

[1] These authors contributed equally to this article.

https://doi.org/10.1016/j.jfma.2021.05.010

were still two major surges of epidemics occurring in September 2020 and March 2021, respectively, and numerous episodes due to various extents of Nonpharmaceutical Interventions (NPIs). Unsupervised machine learning identified five patterns as "controlled epidemic", "mutant propagated epidemic", "propagated epidemic", "persistent epidemic" and "long persistent epidemic" with the corresponding duration and the logarithm of case load from the lowest (18.6 ± 11.7; 3.4 ± 1.8)) to the highest (258.2 ± 31.9; 11.9 ± 2.4). Countries like Taiwan outside five clusters were classified as no community-acquired outbreak.

Conclusion: Data-driven models for the new classification of community-acquired outbreaks are useful for global surveillance of uninterrupted COVID-19 pandemic and provide a timely decision support for the distribution of vaccine and the optimal NPIs from global to local community.

Introduction

An emerging human coronavirus, severe acute respiratory syndrome, namely Coronavirus disease 2019 (COVID-19),[1] was first detected by the end of 2019[2] from the city of Wuhan, Hubei Province, China and then rapidly spread from hotspot to hotspot worldwide in the early phase before pandemic as indicated in the accompanying article of this special issue.[3] Accordingly, World Health Organization (WHO) declared COVID-19 as a public health emergency of international concern (PHEIC) on 30th January 2020.

As little is known about this new coronavirus that led to unawareness about the high transmission of COVID-19 through pre-symptomatic and asymptomatic COVID-19 cases during incubation period,[4,5] and slow reaction about public health system, this novel pathogen took only three months to cause a worldwide pandemic announced on 11th March 2020 and has caused long persistent epidemic since then. The fatal consequences of this long and persistent pandemic are three-fold. Infected elderly or people with underlying diseases are prone to serious outcomes, progress faster, and present higher case fatality.[6–8] This has also damaged the existing healthcare system especially on several aging countries.[9,10] The containment policies have been suggested including visa restrictions, denying travelers from areas with outbreaks, inbound quarantine measures, nationals/residents isolation, border control, flight/harbor suspensions, mandatory personal protective behavior and enlarged social distance in many countries attacked by COVID-19.[11–13] The policies apparently inflict heavy impacts on economic, civilization, and human life.

Attempts made to investigate the unprecedent epidemic trend in a systematic way would be beneficial to contain community-acquired outbreaks and provide a good guidance for the distribution of vaccine. In order to have well preparedness, a highly awareness is helpful to control the transmission scale and to reduce the severity of symptoms in the face of second epidemic waves.[14–17] In the early epidemic phases, rapidly identifying the clinical, epidemiological and pathogenologic characteristics of pathogens from infective cases and clarifying the travel history, occupation, contact history, cluster gathering (TOCC) for patients are helpful for health authorities to make contingency plan on implementing containment measures in community to prevent the community-acquired infection.[16–18] However, when time goes and COVID-19 pandemic still exists very few researches have been conducted to look into why there are uninterrupted and repeated episodes with high case load of community-acquired outbreaks during COVID-19 pandemic.

Traditionally, the global epidemic curve modelled by the susceptible-exposed-infected-recovered (SEIR) to yield effective reproductive number (R_t) may be sufficient to monitor the epidemic of emerging infectious disease. However, it seems better to develop a new surveillance system for monitoring the global and the local community-acquired outbreaks when facing long pandemic period worldwide. The aim of this study was to classify the community-acquired outbreak from different surges of COVID-19 pandemic using data-driven models taking into account the duration taken from $R_t > 1$ to $R_t < 1$ and case load given each duration.

Methods

Data sources

A daily basis data on the number of reported COVID-19 cases, recovery, and death were retrieved from the open data repository maintained by Johns Hopkins University Center for Systems Science and Engineering (CSSE).[19,20] To assure the quality of data, the frequencies reported by CSSE were cross-validated with that of the WHO situation reports, regional health authorities, and other web-based information sources.[19] In addition to the frequencies on global scale, the CSSE also reported the data on country and region level. Using such an open data repository in the period between March 1st 2020 and April 10th 2021, a big-data driven framework was facilitated to assess the epidemiological trends, time to lift social distancing, and classify the patterns of community-acquired outbreaks.

Data-driven framework

Estimating effective reproductive number

Following the method elaborated by Daley and Gani,[21] the reproductive number can be estimated by using information on the frequencies of susceptible, cases, recovered cases, and deaths. The corresponding effective reproductive number at time t (R_t) is then derived. Note that the total population of a country/region of interest is the summation of for quantities. By using the frequencies on global and countries/region level, the value of R_t of the scale of interest and the force of COVID-19 transmission can be evaluated. For an effective reproductive number larger than 1 ($R_t > 1$), the transmission of COVID-19 persists and the occurrence of outbreak is certain. On the other hand, a R_t less than one indicates the containment of transmission force and the end of outbreak.[22]

Classification of community-acquired outbreak based on case load and duration from $R_t > 1$ to $R_t < 1$

In conventional infectious epidemiology, whether the effective reproductive number is larger than 1 ($R_t > 1$) is often used to assess the spread of SARS-CoV-2 even given containment measures. In this scenario of community-acquired outbreaks of COVID-19 pandemic, the duration of outbreak retaining larger than one ($R_t > 1$) provides valuable information to define different types of community-acquired outbreaks. Here we define the duration as the time taken from $R_t > 1$ to $R_t < 1$. The longer the duration, the severe is the epidemic called persistent community-acquired outbreaks caused by COVID-19. Therefore, the duration from $R_t > 1$ to $R_t < 1$ might be one of useful criteria for classifying community-acquired outbreaks. In addition to duration, other information used for community-acquired outbreaks is pertaining to cumulative confirmed cases to reflect case load given the duration of persistent community-acquired outbreaks.

To model the contribution of duration and case load to the classification of community-acquired outbreaks, we applied a new compound Poisson regression model and unsupervised machine learning analysis with K-means clustering method as described in statistical analysis.

As the global resurgence was observed after June, 2020 while most of countries lifted social distancing, the classifications of community-acquired outbreaks were analyzed in two surge periods, first surge period (January ~ June, 2020) and second surge period (July ~ December, 2020), respectively. Additionally, although country or region might have several outbreaks during the surge period, the time to $R_t < 1$ with the maximum peak of R_t were selected to represent their outbreak for each country in the following compound Poisson regression and unsupervised machine learning analysis.

Statistical analysis

Descriptive analysis

Several descriptive analyses have been performed, including time trends of epidemiological profiles of cumulative COVID-19 cases and deaths, and the index for lifting social distancing (social distancing index, SDI) developed by Chen et al.[23] used to assess the balance between COVID-19 disease burden represented by the number of COVID-19 cases and the medical resource capacity captured by the number of recovery and case fatality rate of COVID-19. For a sustained outbreak with incased COVID-19 cases and unmet medical needs aroused by these cases, the index will be larger than 1. This scenario thus calls for a strict social distancing measure to reduce the spread of COVID-19 to the extent the medical care capacity can catchup. Following this rationale, when the SDI index is thus lower than 1 lifting social distancing can be considered. The degree of confidence in lifting social distancing by using the inverse of the decile of SDI from the lowest (0.1) to the highest (1) have also been proposed.[23]

Analysis with the SEIR model

The deterministic compartment model was applied to estimating the transmission coefficient, recovery rate and death rate based on time-series data including the number of reported infected cases, recovered cases, and deaths. The predicted and the observed were presented with each other.

The degree of community-acquired with compound Poisson regression analysis

In order to take into account the impact of the duration and case load as indicated above for classification of community-acquired outbreak, a compound Poisson regression model was applied. Let T_a denotes the numbers of days from $R_t > 1$ to $R_t < 1$ among 365 days or longer for a country or region, the mean number of days of outbreak for each country and region was assessed by

$$\mathrm{Ta} \sim Poisson(\lambda) \tag{4}$$

$X_1, X_2, \ldots, X_{Ta}$ are the daily reported cases on the 1st,2nd, …, Ta day representing the discrete time of duration and $\mathrm{Y} = \sum_{i=1}^{Ta} X_i$ is the random sum of the reported cases and follows the compound Poisson distribution.

$$\mathrm{Log}\left(\sum_{i=1}^{Ta} X_i\right) = \alpha + \beta_a(\,Country\,/\,region), \tag{5}$$

where the Ta is captured by a Poisson distribution and the cases occurred in each day is captured by a Gamma distribution. The vector of regression coefficient, β_a, thus represents the degrees of community-acquired COVID-19 outbreak for each country and region. We selected a country with moderate epidemic and sufficient large population for a stable figure, such as Sweden, as the reference group. The statistical analyses for the compound Poisson model were conducted using Tweedie's compound Poisson-Gamma mixture model by using the procedure of HPGENSELECT written by SAS program.

Machining learning approach with the K-means clustering analysis

We examined the patterns of disease outbreak by using K-means clustering technique. The K-means clustering is one of the popular machining learning algorithms. The K-means algorithm uses iterative refinement based on k clusters to ensure the minimized centroid for the sum of the squared distance between the data points of a cluster.[24,25] The best estimate for number of clusters k was based on the aligned box criterion statistics. After selection, the optimal number of clusters was set as 5 for unsupervised clustering. All statistical analyses were performed with SAS 9.4 and SAS Viya software.

Results

Global epidemics of COVID-19

Fig. 1 shows the epidemic curve of daily confirmed COVID-19 cases from March 1st 2020 to April 10th 2021. The declaration of COVID-19 as pandemic on March 11th 2020 by WHO was entirely based on a remarkably higher basic reproductive number as shown in Fig. 1 (a) (in orange). The epidemics in the globe had lasted from the day of declaration until May 2021 and showed a declining trend until September 2020 after the adoption of various extents of Nonpharmaceutical Interventions (NPIs) in global regions although the extent of execution varied from place to place. Since then, there was a second surge of epidemic until December 2020. Although the epidemics was tentatively contained between January and February 2021 after NPIs have been re-operated in various regions in the globe and the initial uptake of vaccine has started in certain regions, there was a third surge of epidemic since March 2021.

The corresponding trends of global SDI is also shown in Fig. 1 (a) (in green). As the magnitude of global SDI was larger than 1 until April 2021 it is very difficult to lift social distancing from the global perspective.

Continental epidemics of COVID-19

The similar trends as seen in global epidemics were also noted in other continental regions. During the COVID-19 pandemic from January 2020 to April 2021, the major contribution of COVID-19 cases around the world was mainly from Europe, followed by North America, South America, Asia and Oceania, and Africa, as shown in Fig. 1(b)—(f)). Notably, after initial outbreak, it took around two to three weeks for R_t close to 1 in Asia Pacific. The times to R_t below 1 were longer for other continents, including 57 days, 75 days, 157 days, and 170 days in North America, Europe, Africa, and South America, respectively. After R_t close to 1, the continent-specific time trends for R_t fluctuated between 1 and 2, which indicate the COVID-19 has become persistent and endemic. Regarding the size of SDI, all the values of SDI were above 1 during our study period. However, all time trends of SDI for each continent shows a tendency of accelerating recovery and improving case-fatality rate, namely the decline in SDI, after the first surge.

Global and continent-specific COVID-19 transmission and death rates

Fig. 2 shows the trends of global and continent-specific transmission coefficients and case-fatality rates using the SEIR model. Higher transmission coefficients were observed at initial outbreak phase globally and all other continents. The transmission coefficients then declined to a lower range between 0.1 and 0.3 after April 2020 but still alternating between different scales in the surge of epidemic and off-and-on with various extents of NPIs implemented in different continents. The similar time trends of case-fatality rates were also noted with higher case-fatality rates at the beginning of outbreaks and then a substantial decline from June 2020 onwards, indicating the improvement of quality of care for COVID-19 patients with time (Fig. 2 (g)—(l)). Compared with other continents, the higher case-fatality rates ranging between 0.02 and 0.04 were observed in South America (Fig. 2 (j)) and Africa (Fig. 2 (k)).

Classification of global community-acquired outbreak

Fig. 3 shows the ranges of regression coefficients estimated from the compound Poisson regression model (−10.4 to 5.5 in the first surge and −12.6 to 3.6 in the second surge), representing a wide ranges of various types of epidemic in all countries worldwide taking Sweden as the reference group (Supplementary Table 1). Table 1 also shows the results of the quintiles distribution of regression coefficients regarding the classification of five clusters.

Fig. 4 shows the results of the K-means clustering method with five patterns identified in both surge periods by using information on the duration of epidemic and case load. The manifested segmentation between clusters was observed in first surge period. The orders of clusters were ranked by duration and case load. Table 1 shows the average duration and case load of five clusters. Cluster 1 (in purple) had the lowest duration and the logarithm of case load (18.6 ± 11.7; 3.4 ± 1.8) and the cluster 5 had the highest figures (258.2 ± 31.9; 11.9 ± 2.4). Also, the regression coefficients were also ranked according to five clusters classified by K-means clustering method. Five patterns for the classification of community-acquired outbreaks are labelled as "controlled epidemic", "mutant propagated epidemic", "propagated epidemic", "persistent epidemic" and "long persistent epidemic".

In the second surge period, there are some overlapped segmentations within clusters. Compared with the first surge period, the risk of COVID-19 transmission was higher (relative risk: 2.20, 95% CI: 1.46—3.31) in the second surge period. Basically, the clusters can be ranked by duration and case load. Note that the duration taken to reach $R_t < 1$ in cluster 2 was shorter than cluster 3 but case load was larger in cluster 2 than cluster 3.

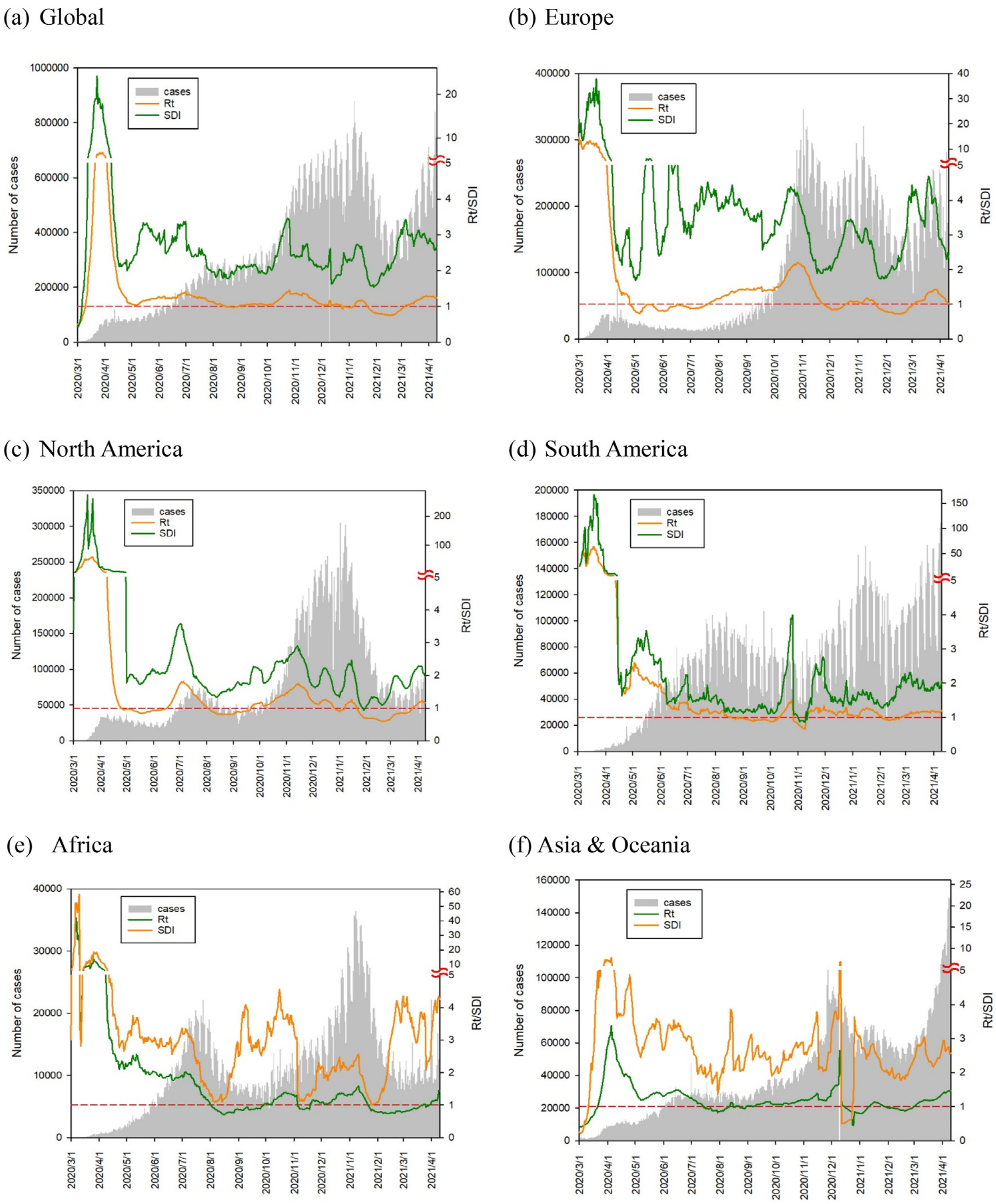
However, the regression coefficient was more likely to depend on case load rather than duration based on data in second surge period. The average duration and cumulative logarithm cases should be simultaneously used to ascertain the type of community-acquired outbreak.

Classification of community-acquired outbreak in Asia

In Asia, the classification of community-acquired outbreak was also categorized by five clusters (Fig. 5). The patterns were similar to the global clustering in both of two surge periods. In first surge period, country/region, such as

Vietnam and Macau in cluster 1 with lowest time to $R_t < 1$ and cumulative logarithm cases could be indicated as the outbreak with the well-controlled type. Country/region, such as South Korea and Hong Kong in the cluster 2 would make much effort but required the longer time to contain the outbreak. However, countries with outbreak without effective control in the cluster 4 or cluster 5 resulted in more cases and longer time before reaching to $R_t < 1$. Countries or regions including Laos, and Taiwan did not have any community-acquired outbreak in first surge period.

Time required to $R_t < 1$ was shorter in second surge period than the first surge period, particularly in cluster 4 and cluster 5. Compared with the clustering in first surge period, countries or regions might stay in the same clusters, or change to other clusters. The countries or regions with better containment measurements, the cluster were


(a) Global

(b) Europe

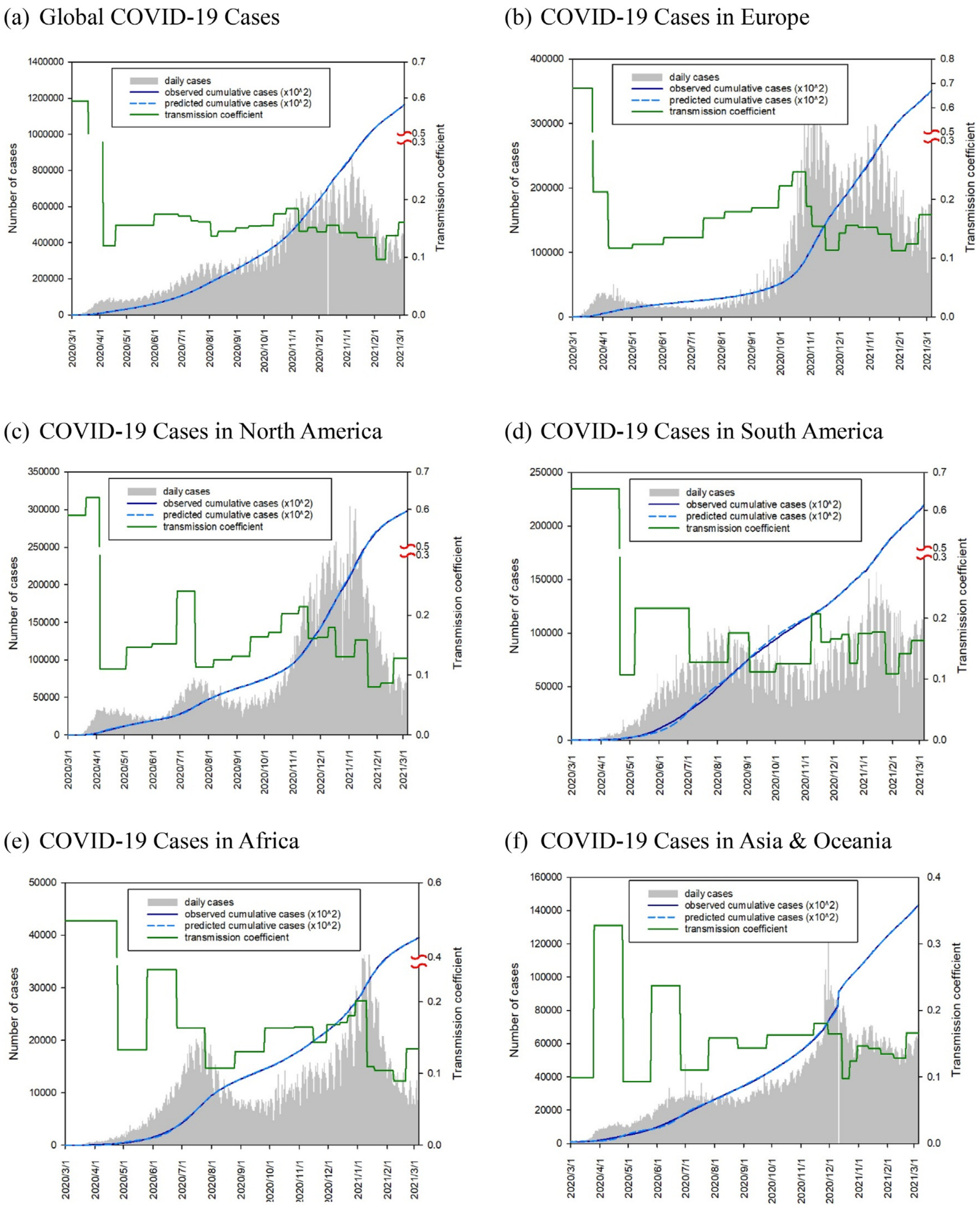
(c) North America

(d) South America

(e) Africa

(f) Asia & Oceania

Figure 1 Global and continent-specific epidemic curves, reproductive number, and index for lifting social distancing.


(a) Global COVID-19 Cases

(b) COVID-19 Cases in Europe

(c) COVID-19 Cases in North America

(d) COVID-19 Cases in South America

(e) COVID-19 Cases in Africa

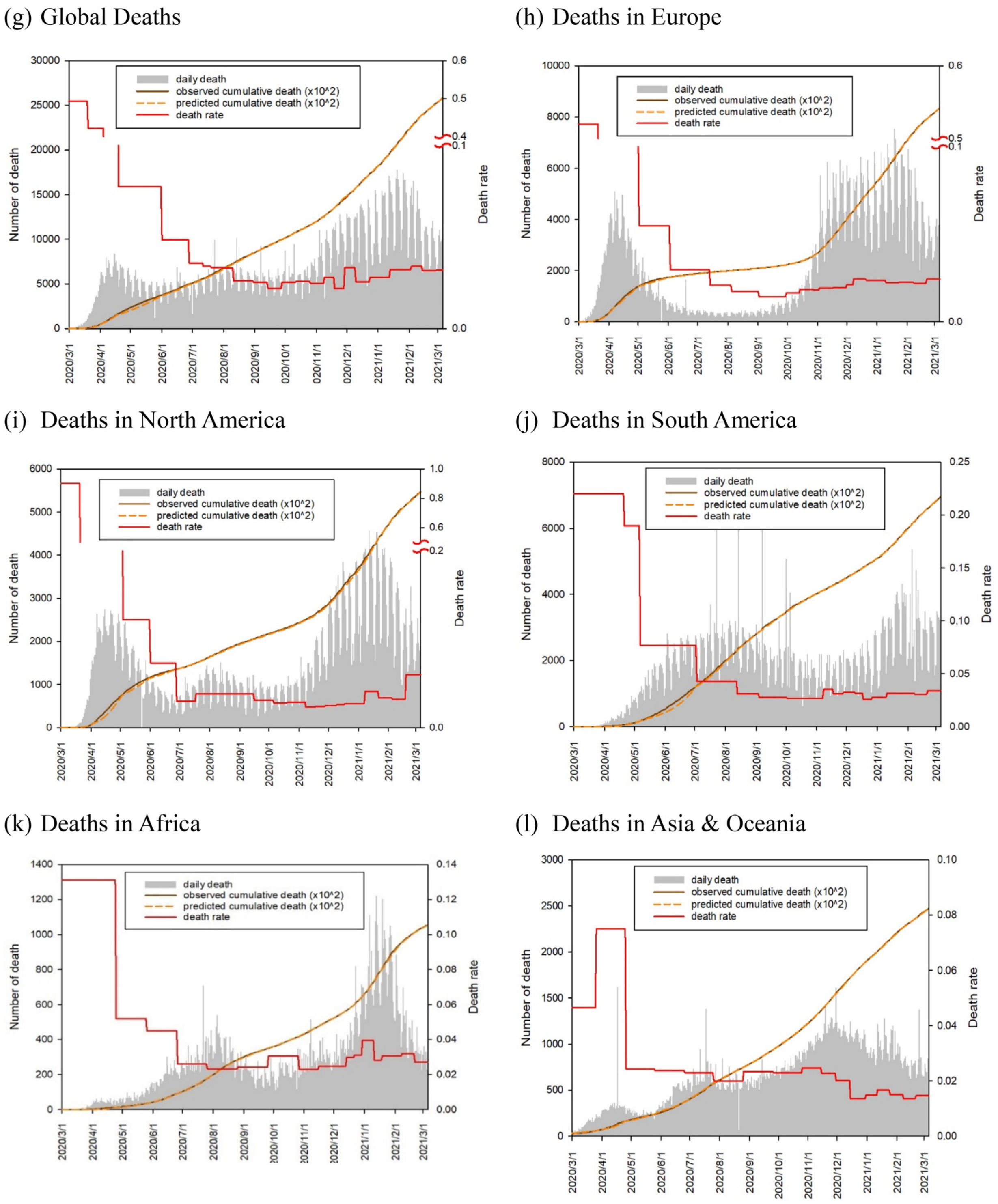
(f) COVID-19 Cases in Asia & Oceania

Figure 2 Global and continent-specific observed cases and deaths, estimated transmission rates, and death rates over time during the pandemic periods.

changed to the cluster with lower duration as well as the lower case load.

Discussion

In this article, we first reported the global and continental epidemics of COVID-19 in the light of daily reported confirmed cases, the estimated effective reproductive numbers, and index for lifting social distancing. We then compared the observed cumulative cases and deaths with the predicted ones based on the SEIR model, from which the estimated transmission coefficients and effective reproductive number (R_t) were derived. Consistent with the accompanying article in the same issue,[26] the first surge period had the highest transmission coefficients and also R_t after the first outbreak reported from China and the rapid spread from hotspot to hotspot worldwide in the early period of transmission caused by this novel pathogen until the declaration as pandemic because of delay of providing appropriate containment measures.



(g) Global Deaths

(h) Deaths in Europe

(i) Deaths in North America

(j) Deaths in South America

(k) Deaths in Africa

(l) Deaths in Asia & Oceania

Figure 2 (*continued*).

During the long persistent pandemic period, the numbers of resurgence were observed in countries with infection. The reasons behind the increased number of resurgence might be associated with the easing of NPI, community-acquired outbreak, or may be due to importation of cases.[27] It is therefore important to identify different types of community-acquired outbreaks. However, the use of conventional effective reproductive number may not be sufficient to capture these diversified types of community-acquired outbreaks for each county or region. We thus proposed a data-driven framework to improve the application of R_t with the SEIR model in order to have a better classification of community-acquired COVID-19 outbreaks worldwide.

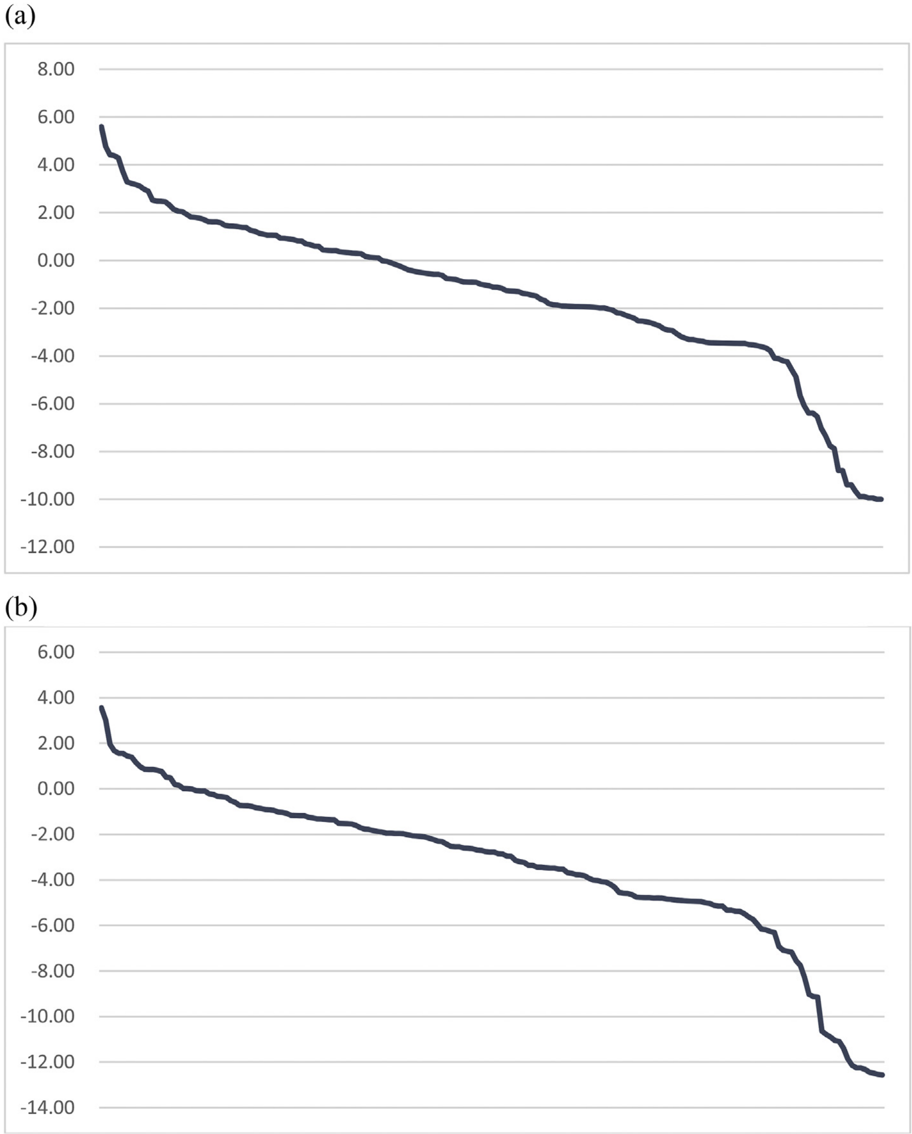


Figure3 Regression coefficient of compound Poisson model. (a) First Surge Period: January ∼ June (Reference group: Sweden). (b) Second Surge Period: July ∼ December (Reference group: Sweden).

Methodologically, we used two indicators, duration and case load of epidemic, to capture the heterogeneity of patterns associated with whether community-acquired outbreaks persisted with time. The longer the duration, the more likely to have persistent epidemic. The higher the case load the less likely the outbreak might be contained. Note that both duration and case load should be considered simultaneously. Note that the case load might be a dominated factor in the light of our data-driven approach. The higher regression coefficient in the cluster 2 had shorter duration but higher case load compared with the cluster 3 (see Table 1).

Accordingly, five patterns for the classification of community-acquired outbreaks are identified and labelled as ″controlled epidemic”, ″”, ″mutant propagated epidemic”, ″propagated epidemic”, ″persistent epidemic” and ″long persistent epidemic” from cluster 1 to cluster 5 identified by K-means clustering analysis. Interestingly, we found that the minimum time to $R_t < 1$ was around 3 weeks and case load was 30—70 cases for measuring the lowest odds of community-acquired outbreak. It should be also noted that the cluster 5 had the highest duration and case load, lasting for almost eight months and having colossal number of cases. The time window and case load are plausible as it might take one cycle of incubation period to contain outbreak if NPIs are effective. However, the outbreak period might be longer to move from cluster 1 to cluster 2 or cluster 3 if country or region's NPIs are less effective and may also be caused by the emerging viral variants. Moreover, countries with high attack rate would be classified as cluster 4 or cluster 5 due to the lift of NPIs or high transmissible mutant virus. The former reason is that nonpharmaceutical interventions to restrict human movement have proven to contain the spread of the virus. However, nonpharmaceutical interventions with limited physical movement by lockdown, adopted social distancing and personal protective mask were effective strategies but very challengeable for government. After COVID-19 pandemic,

Table 1 The average time to $R_t < 1$, average cumulative cases, and average estimated coefficients by resurgence patterns and risk comparison between first and second periods, compound Poisson regression model analysis.

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Relative Risk[a] (95% CI)
First Surge Period (January ∼ June, 2020)						
Time to $R_t < 1$[b]	18.6 ± 11.7	50.7 ± 20.4	62.3 ± 20.3	130.9 ± 22.9	258.2 ± 31.9	
Cumulative Logarithm Cases [b]	3.4 ± 1.8	7.3 ± 1.0	10.4 ± 1.2	11.0 ± 1.6	11.9 ± 2.4	
Beta Coefficient [b]	−7.29 ± 2.1	−2.52 ± 1.1	0.57 ± 1.4	0.98 ± 1.6	1.97 ± 2.3	
Second Surge Period (July ∼ December, 2020)						
Time to $R_t < 1$ [b]	22.0 ± 13.5	40.3 ± 14.0	64.5 ± 20.0	83.0 ± 15.5	158.4 ± 38.9	
Cumulative Logarithm Cases [b]	4.3 ± 2.0	10.7 ± 1.3	7.8 ± 0.9	11.7 ± 1.4	12.5 ± 1.6	
Beta Coefficient [b]	−9.61 ± 2.7	−2.02 ± 1.3	−4.91 ± 0.9	−1.02 ± 1.4	−0.20 ± 1.6	
Second Surge Period vs. First Surge Period						2.20 (1.46−3.31)

[a] Relative Risk = (Beta Coefficient in Second Surge Period/Beta Coefficient in First Surge Period).
[b] Presented as mean ± SD.

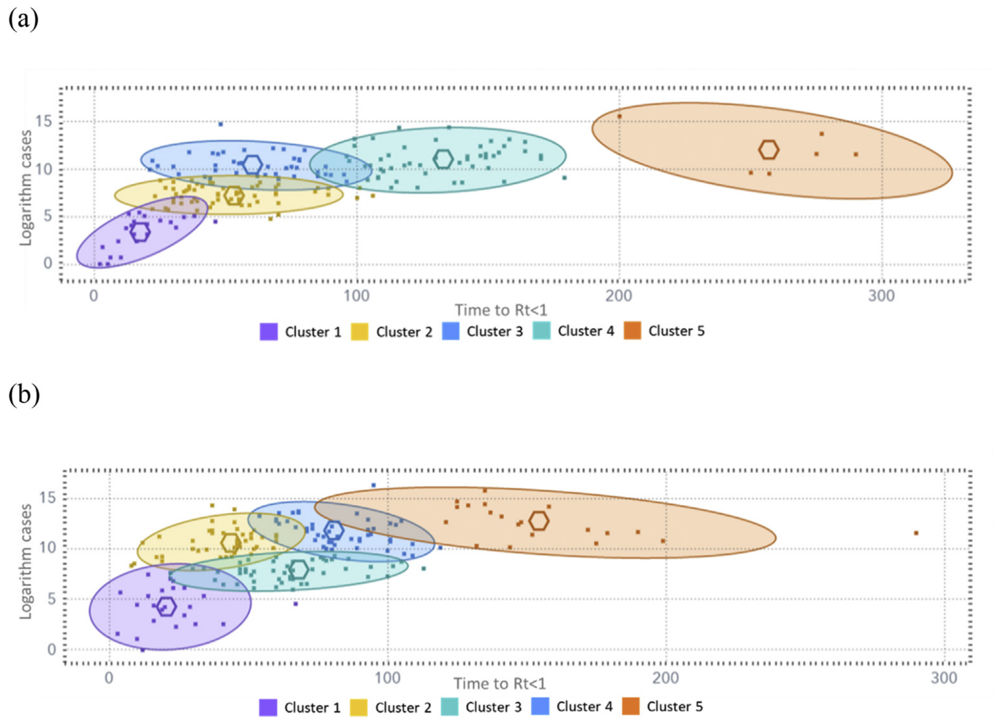
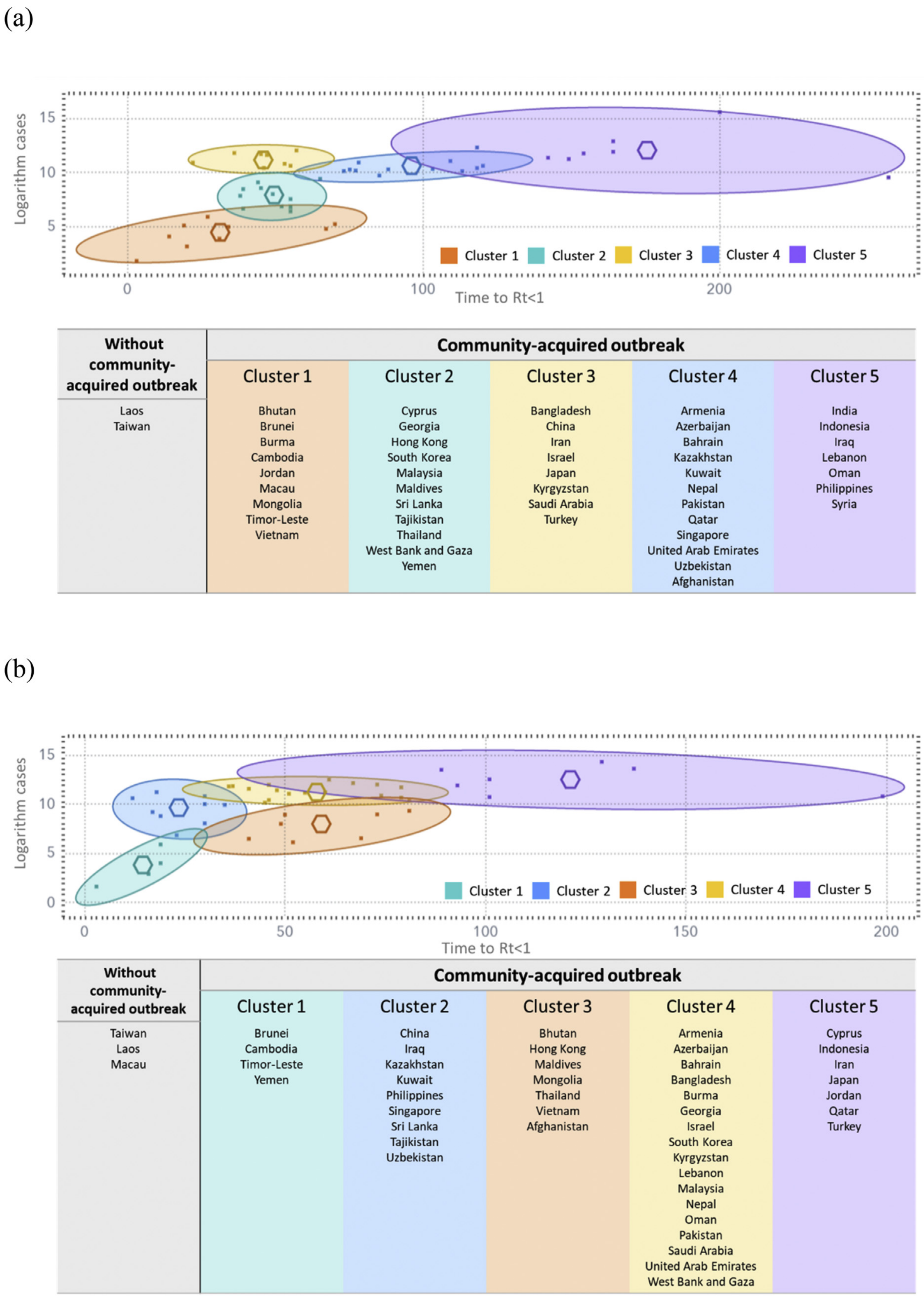


Figure 4 Global resurgence patterns by cluster K-means analysis. (a) First Surge Period (January ∼ June, 2020). (b) Second Surge Period (July ∼ December, 2020).

therefore, whether the spread of COVID-19 will be controllable in each continent highly depends on the effectiveness of NPIs. We demonstrated the effective reproductive number and SDI for global, continent, country, and region levels in Asia to elucidate the global dynamic change of two factors in relation to COVID-19 through the pandemic period (after March, 2020). The effective reproductive number was applied to monitoring the resurgence of the cluster infections in each continent. After the outbreak period, the transmission episode across continents were similar. The effective reproductive number was generally declined from high in the outbreak period (more than 3) approaching to 1 in the pandemic period. The effective reproductive number stably ranging between 0.5 and 2 during the pandemic period indicates COVID-19 has become endemic in community globally. Moreover, even though the effective reproductive numbers across the different periods were smaller than 1, the social distancing was not able to lift due to the high SDI (larger than 1) in each continent.

The chronological order of evolving from cluster 1 to cluster 5 is also consistent with biological plausibility from the experience of emerging infectious disease. A future outbreak will soon come after observing local first few X cases in the daily reports of past epidemic situation whenever the local authorities take containment measures too late.[28] Without the effective anti-viral therapies and vaccination, the observation on the natural course leading to the pandemic of COVID-19 was coherent with the



(a)

Without community-acquired outbreak	Community-acquired outbreak				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Laos Taiwan	Bhutan Brunei Burma Cambodia Jordan Macau Mongolia Timor-Leste Vietnam	Cyprus Georgia Hong Kong South Korea Malaysia Maldives Sri Lanka Tajikistan Thailand West Bank and Gaza Yemen	Bangladesh China Iran Israel Japan Kyrgyzstan Saudi Arabia Turkey	Armenia Azerbaijan Bahrain Kazakhstan Kuwait Nepal Pakistan Qatar Singapore United Arab Emirates Uzbekistan Afghanistan	India Indonesia Iraq Lebanon Oman Philippines Syria

(b)

Without community-acquired outbreak	Community-acquired outbreak				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Taiwan Laos Macau	Brunei Cambodia Timor-Leste Yemen	China Iraq Kazakhstan Kuwait Philippines Singapore Sri Lanka Tajikistan Uzbekistan	Bhutan Hong Kong Maldives Mongolia Thailand Vietnam Afghanistan	Armenia Azerbaijan Bahrain Bangladesh Burma Georgia Israel South Korea Kyrgyzstan Lebanon Malaysia Nepal Oman Pakistan Saudi Arabia United Arab Emirates West Bank and Gaza	Cyprus Indonesia Iran Japan Jordan Qatar Turkey

Figure 5 Resurgence patterns by cluster K-means analysis in Asia. (a) First Surge Period (January ∼ June, 2020). (b) Second Surge Period (July ∼ December, 2020).

pandemic pattern of the previous influenza pandemic pattern that came in waves, suggesting that implementing public health measures effectively before the first wave can only relax the damage on waves later but cannot be fully exempted from an initial outbreak.[14,15]

The study has some limitations. First, the time axis in the epidemic curves was the date of reported case from open-accessible repository rather than the date of laboratory report. The dates of testing report or symptom onset are more essential in understanding the spread of an emerging infectious disease. However, as our goal is to classify new pattern of persistent community-acquired outbreaks such influence may be trivial. An epidemic curve with the dates of testing report or symptom onset can give a more immediate insight on the virus transmissions. Second, complete and accurate information on community-acquired outbreak is required for the validation of the developed methods. Further study can employ an additional and validation process on more specific countries result using trusted public information. It would be interesting to incorporate more information, such as human mobility, in order to more comprehensively evaluate the community transmissions of the virus.

In conclusion, data-driven models for the new classification of community-acquired outbreaks are useful for global surveillance of uninterrupted COVID-19 pandemic and provide a timely decision support for the distribution of vaccine and the optimal NPIs from global to local community.

Disclosure statement

The authors have nothing to disclose.

Funding

This study was supported by Ministry of Science and Technology, Taiwan (MOST 109-2327-B-002-009).

Declaration of competing interest

The authors have no conflicts of interest relevant to this article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jfma.2021.05.010.

References

1. Guo YR, Cao QD, Hong ZS, Tan YY, Chen SD, Jin HJ, et al. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak - an update on the status. *Mil Med Res* 2020;**7**:11.
2. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China [published correction appears in Lancet. 2020 Jan 30]. *Lancet* 2020;**395**:497—506.
3. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* 2020;**368**:395—400.
4. Rothe C, Schunk M, Sothmann P, Bretzel G, Froeschl G, Wallrauch C, et al. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *N Engl J Med* 2020;**382**: 970—1.
5. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann Intern Med* 2020;**172**:577—82.
6. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese center for disease control and prevention. *J Am Med Assoc* 2020;**323**:1239—42.
7. Wang W, Tang J, Wei F. Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China. *J Med Virol* 2020;**92**:441—7.
8. Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study [published correction appears in Lancet Respir Med. *Lancet Respir Med* 2020;**8**(4):e26. 2020;8:475-481.
9. Grech V. Unknown unknowns - COVID-19 and potential global mortality. *Early Hum Dev* 2020;**144**:105026.
10. Grasselli G, Zangrillo A, Zanella A, Antonelli M, Cabrini L, Castelli A, et al. Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2 admitted to ICUs of the lombardy region. *Italy. JAMA*. 2020;**323**:1574—81.
11. World Health Organization. *Coronavirus disease 2019 (COVID-19): situation report*vol. 38; 2020.
12. World Health Organization. *Coronavirus disease 2019 (COVID-19): situation report*vol. 50; 2020.
13. World Health Organization. *Coronavirus disease 2019 (COVID-19): situation report*vol. 67; 2020.
14. Hatchett RJ, Mecher CE, Lipsitch M. Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proc Natl Acad Sci U S A* 2007;**104**:7582—7.
15. Morse SS. Pandemic influenza: studying the lessons of history. *Proc Natl Acad Sci U S A* 2007;**104**:7313—4.
16. McLean E, Pebody RG, Campbell C, Chamberland M, Hawkins C, Nguyen-Van-Tam JS, et al. Pandemic (H1N1) 2009 influenza in the UK: clinical and epidemiological findings from the first few hundred (FF100) cases. *Epidemiol Infect* 2010; **138**:1531—41.
17. Chowell G, Ammon CE, Hengartner NW, Hyman JM. Transmission dynamics of the great influenza pandemic of 1918 in Geneva, Switzerland: assessing the effects of hypothetical interventions. *J Theor Biol* 2006;**241**:193—204.
18. Bernard Stoecklin S, Rolland P, Silue Y, Mailles A, Campese C, Simondon A, et al. First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020. *Euro Surveill* 2020;**25**:2000094.
19. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time [published correction appears in Lancet Infect Dis. *Lancet Infect Dis* 2020;**20**(9):e215. 2020;20: 533-534.
20. Miller M. 2019 novel coronavirus COVID-19 (2019-nCoV) data repository. *Bull Assoc Can Map Libr Arch* 2020:47—51.
21. Daley DJ, Gani J. *Epidemic modelling: an introduction (No. 15)*. Cambridge University Press; 2020.
22. Nishiura H, Chowell G. The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In: *Mathematical and statistical estimation*

approaches in epidemiology. Dordrecht: Springer; 2009. p. 103—21.

23. Chen SL, Yen AM, Lai CC, Hsu CY, Chan CC, Chen TH. An index for lifting social distancing during the COVID-19 pandemic: algorithm recommendation for lifting social distancing. *J Med Internet Res* 2020;**22**. e22469. Published 2020 Sep. 17.
24. Pollard D. Strong consistency of K-means clustering. *Ann Stat* 1981;**9**:135—40.
25. Barber David. *Bayesian reasoning and machine learning*. Cambridge University Press; 2012.
26. Ku MS, Huang LM, Chiu SYH, Wang WC, Jeng YC, Yen MY, et al. Continental transmission of emerging COVID-19 on the 38o North latitude. *J Formos Med Assoc* 2021;**120**:S19—25.
27. Eurosurveillance editorial team. Rapid risk assessment from ECDC: resurgence of reported cases of COVID-19 in the EU/EEA, the UK and EU candidate and potential candidate countries. *Euro Surveill* 2020;**25**:2007021.
28. World Health Organization. *The first few X cases and contacts (FFX) investigation protocol for coronavirus disease 2019 (COVID-19) (No. WHO/2019-nCoV/FFXprotocol/2020.2)*. World Health Organization; 2020.