Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Annals of Epidemiology 75 (2022) 67–72

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: sciencedirect.com/journal/annals-of-epidemiology

Original article

What we can learn from the exported cases in detecting disease outbreaks – a case study of the COVID-19 epidemic

Le Bao*, Xiaoyue Niu, Ying Zhang

326 Thomas Building, Department of Statistics, The Pennsylvania State University, University Park, PA

ARTICLE INFO

Article history:
Received 25 October 2021
Revised 14 June 2022
Accepted 20 September 2022
Available online 24 September 2022

Keywords:
Disease indicators
Exported case data biases
Detection delay
Detection threshold
Disease surveillance system

ABSTRACT

Purpose: Early warning in the travel origins is crucial to prevent disease spreading. When travel origins have delays in reporting disease outbreaks, the exported cases could be used to estimate the epidemic.
Methods: We developed a Bayesian model to jointly estimate the epidemic prevalence and detection delay using the exported cases and their arrival and detection dates. We used simulation studies to discuss potential biases generated by the exported cases. We proposed a hypothesis testing framework to determine the epidemic severity.
Results: We applied the method to the early phase of the COVID-19 epidemic of Wuhan, United States, Italy, and Iran and found that the indicators estimated from the exported cases were consistent with the domestic data under certain scenarios. The exported cases could generate various biases if not modeled properly. We presented the required number of exported cases for determining different severity levels of the outbreak.
Conclusions: The exported case data is a good addition to the domestic data but also has its drawbacks. Utilizing the diagnosis resources from all countries, we advocate that countries work collaboratively to strengthen the global infectious disease surveillance system.

Introduction

Travel is a potent force in the emergence and spread of diseases [1]. Early warning and rapid response in the travel origins are crucial to prevent disease spreading. However, many travel origins have delays in reporting the disease outbreak. Under this circumstance, those exported cases, who traveled from the origin and were tested positive at the destination, became a valuable and informative data source for detecting the disease outbreaks in the travel origins.

The traveler data were mostly used as imported cases to model the epidemic at the travel destinations. For instance, [2] estimated indicators of COVID-19 outbreak in Nigeria using both the local cases and the imported cases. The traveler data were also used to assess the potential for the virus to spread across international borders after a local outbreak was confirmed by the domestic cases (see [3–5] as examples). A few studies focused on using the traveler data as exported cases to detect the outbreak at the travel origin [6–8,16–19]. These studies all used the number of exported cases and aggregated those numbers to the end of the study period and provided estimates for the time at the end of the study period.

Here we used the number of exported cases, the diagnostic dates, and the arrival dates (if available) to provide daily estimates of some key indicators of the epidemic during the early stage of the COVID-19 outbreak in four travel origins, Wuhan (China), Iran, Italy, and United States. The additional information in the diagnostic and arrival dates allowed us to approximate the detection delay among travelers and include some undiagnosed travelers in estimating the prevalence. They also lead to a better estimate of the growth rate compared to evenly distributing the total number of exported cases over the study period. We investigated the usefulness and limitations of using the exported cases in understanding the epidemic and provided a disease outbreak detection criterion based on the cumulative number of exported cases for stakeholders to make decisions.

#All authors contributed equally.
The authors have no conflicts of interest to disclose.

* Corresponding author. 326 Thomas Building, Department of Statistics, Penn State University, University Park, PA 16802.
E-mail address: lebao@psu.edu (L. Bao).

https://doi.org/10.1016/j.annepidem.2022.09.005
1047-2797/

Table 1
The start date of the local epidemic, the end date of the study period, the number of days in the study period

	Start date of the local epidemic	End date of the study period	Number of days in the study period	Number of exported cases
Wuhan	Dec 1, 2019	Jan 23, 2020	54	11
U.S.	Jan 20, 2020	Mar 13, 2020	54	25
Italy	Jan 31, 2020	Feb 28, 2020	29	57
Iran	Jan 9, 2020	Feb 23, 2020	46	4

the total number of exported traveler case reports within the study period and the total number of exported destinations within the study period.

Materials and methods

Data description

We chose four travel origins, Wuhan (China), the United States, Italy, and Iran, in which the COVID-19 outbreaks were detected relatively early compared to their neighbors in their region. We obtained the exported case reports within the period specified in Table 1 from [9] with reverification of their travel histories using government and media reports. For each travel origin, we also obtained the domestic case reports [10,11] to model the domestic growth curve as a comparison with the one estimated from the exported cases. Reasons for choosing the start and end dates of each travel origin were provided in Supplementary S1. The details of obtaining the outbound travel volume data, criteria used to exclude certain cases, and full data table were also in Supplementary S1.

For each travel origin, we also obtained the domestic case reports to model the domestic growth curve as a comparison with the one estimated from the exported cases. Domestic case reports for Wuhan were obtained from [10], and the data for the United States, Italy, and Iran were obtained from COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [11].

Statistical models for traveler cases

We assumed that in the initial period of the epidemic the disease prevalence in the general population increased exponentially: $\rho_t = exp(\beta_0 + \beta_1 t)$, and referred to β_1 as the exponential growth rate. Another important indicator, the basic reproduction number R_0, is the expected number of new infections caused by one infected individual during the infectious period. R_0 can be related to β_1 by $R_0 = \frac{1}{\int_0^\infty exp(-\beta_1 \tau)\omega(\tau)d\tau}$, where $\omega(\tau)$ is the density of the serial interval [12]. We assumed the exported cases follow a binomial distribution given the daily travel volume, disease prevalence in the general population, and a traveler bias correction factor. In addition, we assumed that a certain proportion of infected travelers were not detected by the end of the study (right censoring) and used the interval between arrival dates and detection dates to infer the censoring proportion. We used Bayesian inference to estimate the parameters. Detailed statistical models are provided in Supplementary S2.

Estimation based on domestic case reports

To investigate the potential biases generated by using only the exported cases, we compared the estimates from the exported case data with the ones from the domestic data. [10] carefully studied and reconstructed the full transmission dynamics in the early period of the COVID-19 epidemic in Wuhan by fitting a seven-compartment model. From their R_0 estimate of 3.54 with a 95% confidence interval (3.40,3.67), we could derive the exponential growth rate before January 23, 2020 (assumed to be a constant) being 0.192 with a 95% credible interval (0.185,0.199). Unfortunately, similar results were not readily available for the other travel origins due to limitations of the data and knowledge about their early stages. For the other three origins, we fitted the exponential growth rate by using domestic records in [11] and their corresponding R_0 to serve as a comparison with the one estimated from the exported case data. In this study, we set the serial interval T_c to follow a gamma distribution with mean (sd): 7.5 (3.4) in Wuhan and the United States [13], 6.6 (4.86) days in Italy [14], and 4.55 (3.33) days in Iran [15]. For Wuhan, the prevalence was taken from [10]. For the United States, Italy, and Iran, the prevalence was calculated from their reported numbers [11]. Note that the numbers in Wuhan included pre-symptomatic, ascertained, and unascertained cases, while the other three countries' numbers only included the ascertained cases.

Determining the severity of the outbreak based on the number of exported cases

International Health Regulations (2005) and the Global Outbreak Alert and Response Network (GOARN) considered the imported/exported human cases as a sign of potential public health risk. However, the guideline did not provide how the total number of exported cases could indicate the intensity of the epidemic. We developed a tool for policymakers to determine the intensity of a country's epidemic based on the number of exported cases. Statistically, we could form the question as a hypothesis testing problem. On each day, based on the cumulative number of exported cases up to that day, we could test whether the exponential growth rate is significantly above a certain threshold, for example, 0.1:

$$H_0 : \beta_1 = 0.1 v.s. H_1 : \beta_1 > 0.1$$

We could also test whether the prevalence rate is significantly above a certain threshold, for example

$$H_0 : \rho_t = 0.001 v.s. H_1 :_t > 0.001$$

The detailed test procedure was provided in Supplementary S3.

We developed a Shiny App for calculating the detection criteria in which users can specify their own significance level, initial prevalence, and detection threshold.

Results

Figure 1 visualized how the mean estimates and 95% uncertainty bounds of β_1, R_0, and the number of cases per 100,000 population changed over time. The points in Figure 1 were not from a single time series but represented the moving window of estimates from data up to each time point. The estimates became more stable with smaller uncertainties as more cases accumulated. The estimated exponential growth rate β_1 generally showed an increasing trend after the first exported case. The red lines in Figure 1 indicated the estimates from domestic data. Note that the numbers in Wuhan included pre-symptomatic, ascertained, and unascertained cases [10], while the other three countries only included ascertained cases. The unascertained cases here include asymptomatic and some mildly symptomatic cases that are not detected.

For Wuhan, the estimates from the exported cases were consistent with the ones estimated from domestic data. For the rest of

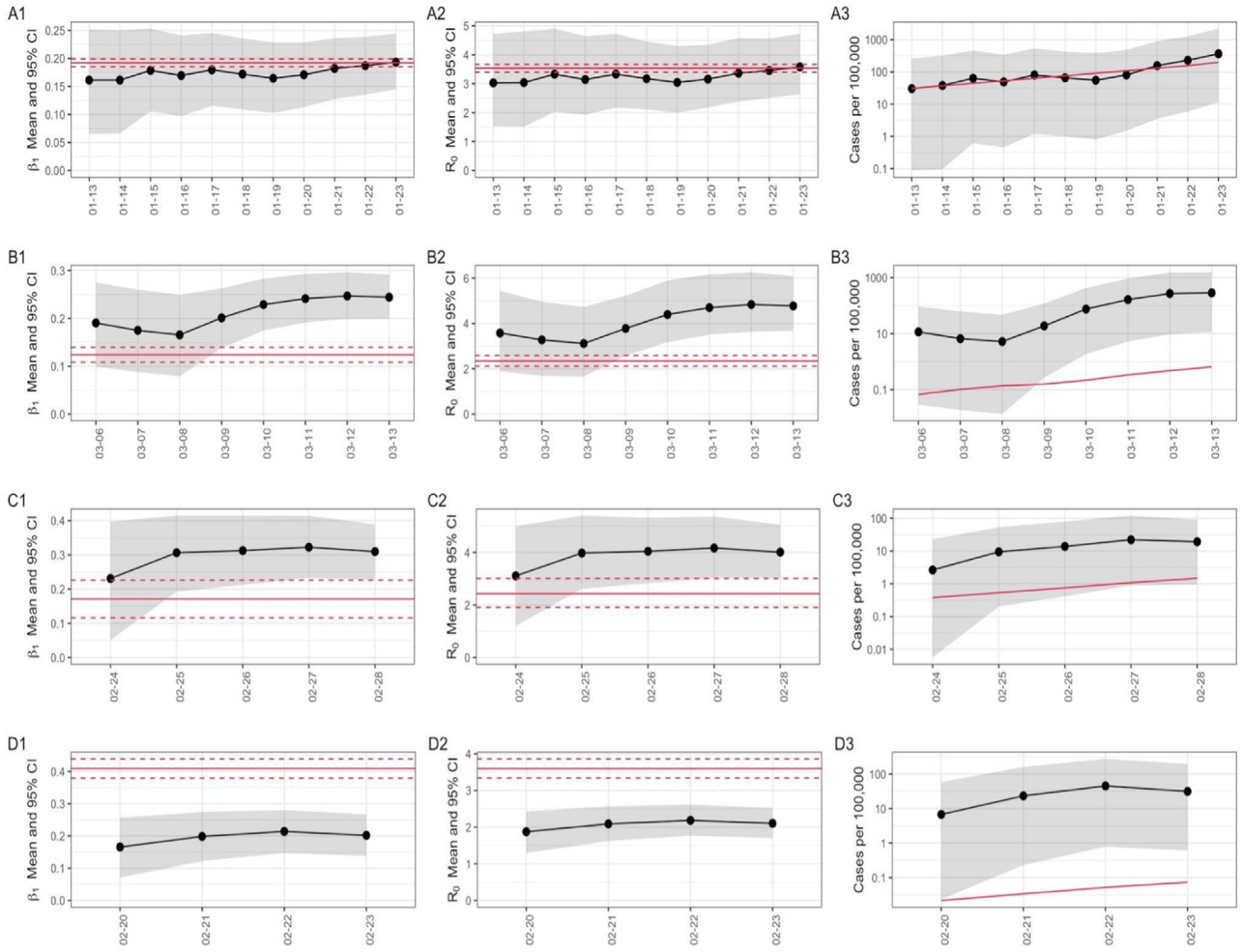

Fig. 1. Estimated key indicators using exported case data. Posterior mean and 95% credible interval of exponential growth rate β_1 (left), basic reproduction number R_0(middle), and cases per 100,000 population (right, on log scale with y-axis labels on the original scale) in the early stage of the COVID-19 outbreak in (*A*) Wuhan, China, (*B*) United States, (*C*) Italy, and (*D*) Iran. The black dots are the posterior mean and the gray bands are the 95% credible intervals estimated from the exported case data. The red horizontal lines indicate the mean (solid) and 95% confidence interval (dashed) estimated from domestic data for exponential growth rate β_1 and reproduction number R_0. The red lines in plots for cases per 100,000 indicate the estimated domestic prevalence (*10*) for Wuhan and the reported domestic prevalence (*11*) for the United States, Italy, and Iran. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the travel origins, prevalence estimated from exported data were all higher than the domestic reported ones. There are several possible reasons for the discrepancies. First, the domestic prevalence of those three countries only included ascertained cases while our estimated prevalence included some unascertained cases by allowing a proportion of infected travelers did not get diagnosed before the end of the study period. Second, there might be a certain proportion of unreported domestic cases in the JHU COVID-19 Data Repository during the early period which resulted in a lower domestic prevalence. Third, the initial number of cases might be larger than the one that appeared in the first government or the media report which would lead to an overestimate of the growth rate, β_1. Fourth, the United States, Italy, and Iran had 60.0%, 89.5%, and 75% missing arrival dates while Wuhan had only 9.1%. The large proportion of missing arrival dates might have affected the estimation of the detection delay. Finally, the travelers might not represent the general population thus could bear a higher or lower infection rate.

We also compared our prevalence estimates with existing literature using traveler data. For Wuhan, we compared our results with [6,7]. Wu et al. [6] used data up to January 25 and estimated the number of infected people as 75,815 (304–130,330). Our study end date was January 23 due to Wuhan lockdown and our estimated infection size on January 23 was 69,410 (2091–403,900). [7] used data up to January 22 to estimate the number of symptomatic infected individuals on January 18 as 4000 (1000, 9700), by assuming all individuals with symptoms were detected before January 22. Using data up to January 22, we estimated the number of infected individuals on January 18 as 17,881 (677–95,845). This number was closer to the one estimated by [10] using domestic data, which was 14,478. We also compared our estimate with Tuite et al. [17,18] for Italy and Iran. They modeled the traveler data using models by [16] and assumed that all traveler cases were detected at destination countries. For Italy, [17] estimated the domestic outbreak size to be 3971 (2907–5297) by Feb 29, 2020. Our estimates showed that only 35.0% (13.2%–59.5%) exported cases had been observed and the domestic outbreak size was 11,573 (571–54,232). For Iran, [18] estimated the domestic outbreak size to be 18,300 (3770–53,470) by Feb 23, 2020. Our estimates showed that 49.0% (15.8%–81.1%) of exported cases had been observed and the domestic outbreak size was 26,047 (487–158,411). For United States, our estimates showed that 26.6% (5.1%–49.1%) of exported cases had been observed by Mar 13, 2020. In addition to the prevalence, [19] developed a model for estimating the detection rate of exported cases and used Wuhan's traveler data as an illustrating example. Their estimated detection proportion for Wuhan was 38%

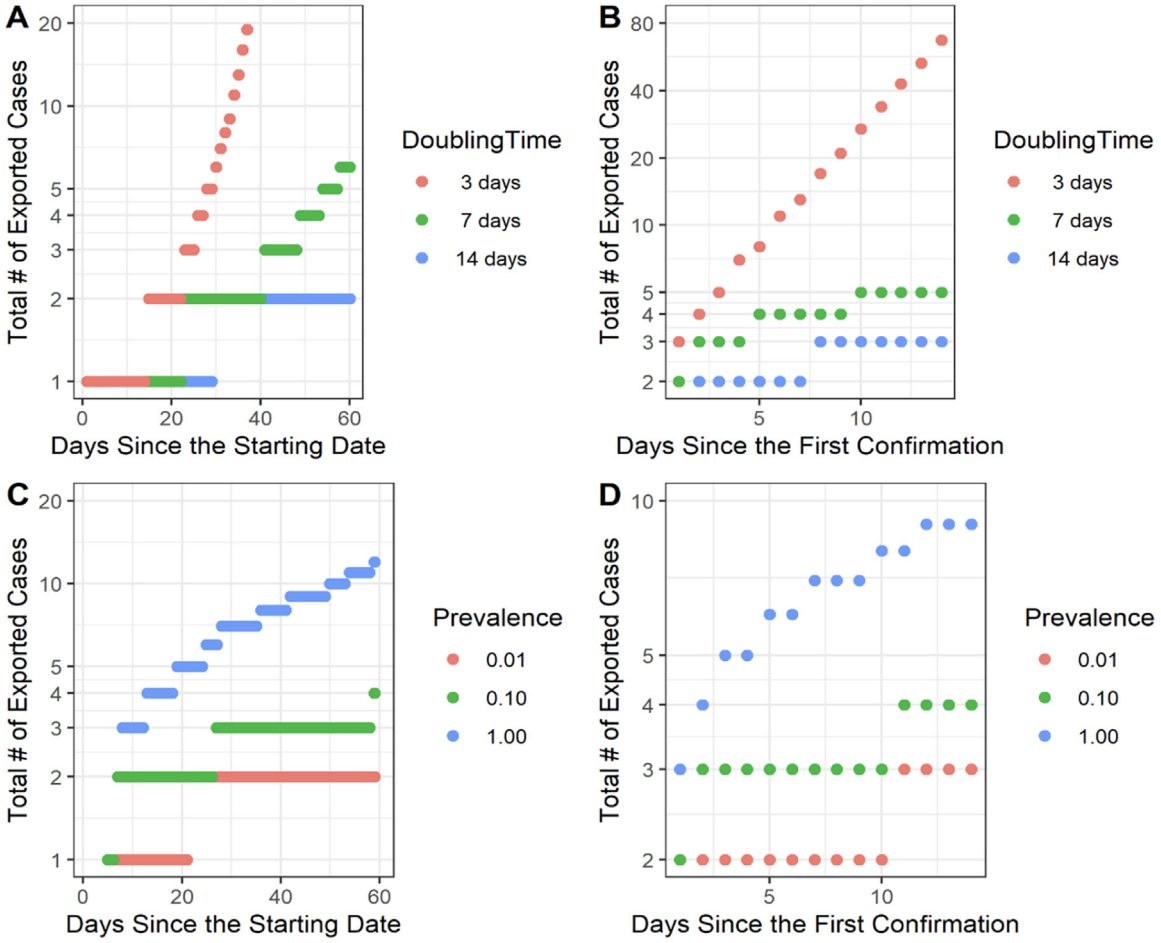

Fig. 2. Hypothesis tests of the doubling time and prevalence. The tests were conducted under three different lengths of doubling time: 3 days (red), 7 days (green), and 14 days (blue) in *A* and *B*; and under three different prevalence rates per thousand people: 0.01 (red), 0.1 (green), and 1 (blue) in *C* and *D*. The dotted lines show the required cumulative numbers of exported cases to reject the null hypothesis and favor a shorter doubling time or a higher prevalence rate under significance level 0.05. The x-axis indicates the number of days since the initial local infection in *A* and *C*, and the number of days since the first exported case in *B* and *D*. The y-axis indicates the cumulative total of exported cases at the day indicated by the x-axis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(22%–64%), similar to our estimated detection rate of 39.3% (12%–65.8%).

Figure 2 presented the epidemic severity detection results for the various thresholds. Sub-figures A and B illustrate the minimum required number of exported cases for detecting a doubling time shorter than the threshold; C and D illustrate the minimum required number of exported cases for detecting a prevalence higher than the threshold. When making decisions to detect a future disease outbreak on a certain day, if we have relatively accurate information about how long it has been since the initial infection date, we can use A and C to compare the observed total number of exported cases up to this day. At a pre-specified significance level, say 0.05, if the observed total number exceeded the number of the corresponding day in the figure, we would be able to tell with 95% confidence that the local epidemic has been doubling faster than the threshold or the local prevalence has reached to a level higher than the threshold. For instance, assuming 1000 outbound travelers per day, we may look at the total number of exported cases that had been detected within the first month since the initial local infection (day 30 on the x-axis of Figs. 2*A* and *C*). One exported case suggested that the local epidemic doubled in less than 14 days; two exported cases allowed us to conclude that the epidemic doubled in less than 7 days and the prevalence was above 0.01 per thousand; five cases indicated a doubling time less than 3 days; three and seven exported cases were required to conclude a prevalence greater than 0.1 per thousand and 1 per thousand, respectively. If we do not know the initial infection date, we would use B and D to reach a conclusion. The total number of exported cases and how quickly those cases accumulated together provided evidence of the intensity of the outbreak: a larger number of confirmed infections among outbound travelers was needed to detect a shorter doubling time or a higher prevalence; the required number of cases increased as the epidemic went into a later period.

A Shiny App for implementing the above procedure can be found here: https://lebao.shinyapps.io/growthratetest/, where users could specify the general population size, the detection threshold, the daily traveler sizes, the number of simulations, and the significance level, etc. With this tool, policymakers could adjust the parameters to fit their own country's situation, and quickly determine how severe their country's epidemic is based on the exported cases.

Discussion

We used the COVID-19 epidemic as an example to illustrate how the exported cases could be used to detect a disease outbreak. We proposed the moving window of the study end date to offer a new perspective of mimicking the real decision-making process. Our inclusion of the detection delay parameter relaxed the assumption that all exported cases had been detected by the end

of the study period and could correct the potential underestimation due to the right censoring. Detailed information on symptom onset date and natural history of disease would be more accurate than our approximation using the arrival and diagnostic dates. Both the bias correction term and the detection delay terms could potentially vary by destination, which was not considered here due to limited data availability.

In addition to the comparisons with the prevalence using domestic reports, we performed posterior predictive checks (Supplementary S4), sensitivity analysis (Supplementary S5), power analysis for the test (Supplementary S6), and a series of simulation studies (Supplementary S7) to validate our method and evaluate the robustness of the model. Our posterior predictive samples were more compatible with the observed traveler case reports over time (Supplementary Figure S1). Our simulation studies showed that, if the exponential growth phase of the true prevalence was relatively short while we assumed an exponential form throughout, with the help of the bias correction term, α, the fitted prevalence curve would stay close to the truth for a week or so before drifting away (Supplementary Figure S6); our model could still detect the domestic prevalence at 0.01% or lower thresholds, but it lost the power of detecting the domestic prevalence at threshold 0.1% and above (Supplementary Figure S7). If the traveler data had a bias, the estimated domestic prevalence and the exponential growth rate would be also biased in the same direction. The bias of the model with α went to zero as more case reports being observed, but the bias of the model without α remained constant over time (Supplementary Figure S8). When we misspecified the start date of the epidemic earlier than the true epidemic start date, the prevalence would be over-estimated, and the exponential growth rate would be under-estimated. As more traveler cases being observed, the growth rate estimate went to the true value (Supplementary Figure S9). Completely missing or misspecified arrival dates affected the accuracy of the estimates, and it demonstrated the importance of reporting arrival dates at the early stage of an outbreak. The bias term α can help alleviate the bias caused by misspecified or missing arrival dates (Supplementary Figure S10). Finally, and more fundamentally, how well one can estimate the domestic prevalence using the exported cases depends on how representative the traveler samples were. If the traveler data had biases, we would need a large amount of data to estimate the intercept difference between the observed time series of case reports and the fitted exponential trend in order to detect the bias and uncover the true prevalence. Due to the sparsity of the traveler data, this information was relatively weak and resulted in a large uncertainty of the bias estimates. Studies that can potentially help estimate the bias include a better understanding of the travelers' demographics, such as age, gender, sub-national region of residence, and whether the traveler is a foreign tourist or a resident of the travel origin. The bias estimate could be improved if we could (1) classify people into different risk groups based on those factors using external data, for example, case reports from other countries with good surveillance systems; and (2) know the distribution of risk groups among travelers and the general population, respectively. If so, a more informative prior distribution of the bias parameter can be derived to account for factors that lead to systematic biases between travelers and the general population.

Rapid detection of a disease outbreak is crucial for government intervention and raising public awareness. Utilizing the diagnosis resources from all countries, the exported case data is a good addition to the domestic data. The proposed method could help detect new disease outbreaks, as well as resurgence and new hot spot locations during the current COVID-19 epidemic. We advocate that countries should work in a collaborative way by sharing the traveler cases information in a timely manner, including arrival dates, detection dates, and travel history. Working together, we would strengthen the global infectious disease surveillance system, which is especially important in detecting disease outbreaks in countries where public health infrastructure is rudimentary or nonexistent.

Ethics approval

Ethics approval is not needed because all the data involved are available in the public domain.

Author contributions

LB, YZ, and XN designed the study. LB and YZ collected and verified the data. LB, YZ, and XN analyzed the data, interpreted the results, and wrote the article.

Acknowledgments

This research was supported by the National Institute of Health/ National Institute of Allergy and Infectious Diseases grant number R01AI136664 and the Huck Institutes of the Life Sciences, Penn State University.

We thank Xihong Lin and Haidong Wang for helpful discussions; Kaiyi Wu, Zhiyang Liang, Xinyi Yang and Qianyi Zhao for exported case report verification.

References

[1] Wilson ME. Travel and the emergence of infectious diseases. Emerg infect dis. 1995;1:39.
[2] Adegboye OA, Adekunle AI, Gayawan E. Early transmission dynamics of novel coronavirus (COVID-19) in Nigeria. Int J Environ Res Public Health 2020;17:3054.
[3] Bogoch II, Creatore MI, Cetron MS, Brownstein JS, Pesik N, Miniota J, et al. Assessment of the potential for international dissemination of Ebola virus via commercial air travel during the 2014 west African outbreak. Lancet 2015;385:29–35.
[4] Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis 2020;20:553–8.
[5] Boldog P, Tekeli T, Vizi Z, Dénes A, Bartha FA, R¨ost G. Risk assessment of novel coronavirus COVID-19 outbreaks outside China. J Clin Med 2020;9:571.
[6] Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet 2020;395:689–97.
[7] N. Imai, I. Dorigatti, A. Cori, C. Donnelly, S. Riley, NM Ferguson, "Estimating the potential total number of novel coronavirus cases in Wuhan City, China" (Technical Report, Imperial College London, 2020).
[8] Chinazzi M, Davis JT, Gioannini C, Litvinova M, Rossi L, Xiong X, et al. Preliminary assessment of the international spreading risk associated with the 2019 novel coronavirus (2019- nCoV) outbreak in wuhan city. Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University; 2020.
[9] Xu B, Gutierrez B, Mekaru S, Sewalk K, Goodwin L, Loskill A, et al. Epidemiological data from the COVID-19 outbreak, real-time case information. Sci Data 2020;7:1–6.
[10] Hao X, Cheng S, Wu D, Wu T, Lin X, Wang C. Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. Nature 2020;584:420–4.
[11] The Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, (COVID-19) data repository, https://github.com/CSSEGISandData/COVID-19, 2020. Accessed June 9, 2021.
[12] Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proceed Royal Society B: Biol Sci 2007;274:599–604.
[13] Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New Eng J Med 2020;382:1199–207.
[14] Cereda D, Manica M, Tirani M, Rovida F, Demicheli V, Ajelli M, et al. The early phase of the Covid-19 outbreak in Lombardy, Italy. Epidemics 2021;37. doi:10.1016/j.epidem.2021.100528.
[15] Aghaali M, Kolifarhood G, Nikbakht R, Mozafar Saadati H, Hashemi Nazari SS. Estimation of the serial interval and basic reproduction number of Covid-19 in Qom, Iran, and three other countries: a data-driven analysis in the early phase of the outbreak. Transbound Emerg Dis 2020;67:2860–8.
[16] Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, et al. Pandemic potential of a strain of influenza A (H1N1): early findings. Science 2009;324(5934):1557–61.
[17] Tuite AR, Ng V, Rees E, Fisman D. Estimation of COVID-19 outbreak size in Italy. Lancet Infect Dis 2020;20(5):537.

[18] Tuite AR, Bogoch II, Sherbo R, Watts A, Fisman D, Khan K. Estimation of Coronavirus Disease 2019 (COVID-19) Burden and Potential for International Dissemination of Infection from Iran. Ann Intern Med 2020;172(10):699–701.

[19] Niehus Rene, et al. Using observational data to quantify bias of traveller-derived COVID-19 prevalence estimates in Wuhan, China. Lancet Infect Dis 2020;20(7):803–8.