Monitoring real-time transmission heterogeneity from Incidence data

Yunjun Zhang[1*], Tom Britton[2], and Xiaohua Zhou [1*,3,4,5]

1 Department of Biostatistics, School of Public Health, Peking University, Beijing, China
2 Department of Mathematics, Stockholm University, Stockholm, Sweden
3 Beijing International Center for Mathematical Research, Peking University
4 School of Mathematical Sciences, Peking University
5 Center for Statistical Science, Peking University,Beijing, China

* yunjun.zhang@pku.edu.cn, azhou@math.pku.edu.cn

Abstract

The transmission heterogeneity of an epidemic is associated with a complex mixture of host, pathogen and environmental factors. And it may indicate superspreading events to reduce the efficiency of population-level control measures and to sustain the epidemic over a larger scale and a longer duration. Methods have been proposed to identify significant transmission heterogeneity in historic epidemics based on several data sources, such as contact history, viral genomes and spatial information, which is sophisticated and may not be available, and more importantly ignore the temporal trend of transmission heterogeneity. Here we attempted to establish a convenient method to estimate real-time heterogeneity over an epidemic. Within the branching process framework, we introduced an instant-individualheterogenous infectiousness model to jointly characterized the variation in infectiousness both between individuals and among different times. With this model, we could simultaneously estimate the transmission heterogeneity and the reproduction number from incidence time series. We validated the model with both simulated data and five historic epidemics. Our estimates of the overall and real-time heterogeneities of the five epidemics were consistent with those presented in the literature. Additionally, our model is robust to the ubiquitous bias of under-reporting and misspecification of serial interval. By analyzing the recent data from South Africa, we found evidences that the Omicron might be of more significant transmission heterogeneity than the Delta. Our model based on incidence data was proved to be reliable in estimating the real-time transmission heterogeneity.

Author summary

The transmission of many infectious diseases is usually heterogeneous in time and space. Such transmission heterogeneity may indicate superspreading events (where some infected individuals transmit to disproportionately more susceptible than others), reduce the efficiency of the population-level control measures, and sustain the epidemic over a larger scale and a longer duration. Classical methods of monitoring epidemic spread centered on the reproduction number which represent the average transmission potential of the epidemic at the population level, but failed to reflect the systematic variation in transmission. Several recent methods have been proposed to identify significant

transmission heterogeneity in the epidemics such as Ebola, MERS, COVID-19. However, these methods are developed based on some sophisticated information such as contact history, viral genome and spatial information, of the confirmed cases, which are typically field-specific and not easy to generalize. In this study, we proposed a simple and generic method of estimating transmission heterogeneity from incidence time series, which provided consistent estimation of heterogeneity with those records with sophisticated data. It also helps in exploring the transmission heterogeneity of the newly emerging variant of Omicron. Our model enhances current understanding of epidemic dynamics, and highlight the potential importance of targeted control measures.

Introduction

The transmission of infectious disease is typically uneven or heterogeneous in terms of time and space due to a complex mixture of host, pathogen and environmental factors [1–6]. High level of transmission heterogeneity may indicate superspreading events (SSEs) in which certain individuals infect a greater large number of secondary cases than average [1], invoking the so-called 20-80 rule. It has been documented that the SSEs considerably reduced the efficiency of population-level control measures [1] and played a key role in dramatically driving the spread of many pathogens in scale and duration, including severe acute respiratory syndrome (SARS) [7], Middle East Respiratory Syndrome (MERS) [8], Ebola [3, 4] and COVID-19 [6, 9, 10]. Therefore, monitoring the degree of transmission heterogeneity and its change could be vital for epidemic forecasting and efficient intervention in infectious disease epidemiology.

Mathematically, the transmission heterogeneity is represented by the variation in offspring distribution, namely, the distribution of secondary cases that may be generated by a given infectious case . Classical methods of estimating heterogeneity rely heavily on reconstructing the offspring distribution. As the epidemiological links among reported cases are complex, this reconstruction poses considerable challenges in both data collection and model building. According to different types of data used in the reconstruction, the existing methods of inferring heterogeneity can be grouped into three categories. The first category are methods based on contact-tracing-data. By interviewing patients to document their contacts with other infected patients, all or most of the cases could be positioned in the network of transmission, and the resulting empirical offspring distribution could be directly used to estimate the transmission heterogeneity [1, 4, 10, 11].

The second category is based on virus-sequence-data. For many pathogens, in particular RNA viruses, evolutionary processes occur on the same timescale as epidemiological processes, which makes it possible to extract epidemiological information from genetic analysis [12, 13]. Many studies showed that the virus phylogeny reconstructed from the virus sequence sampled from the infected individuals reflected the underlying transmission history of the epidemic, with the branching events in a phylogeny corresponding to transmission events in the past. By incorporating the level of heterogeneity into the likelihood function of the virus phylogeny, it is possible to estimate the heterogeneity as well as other epidemiological parameters from the sampled sequence data [2, 14, 15].

For the third category, individual-level spatial information has been integrated to reconstruct the transmission history in recent years. By developing a continuous-time spatiotemporal transmission model with a distance-based kernel to characterized the infectiousness between individuals as a function of the mutual distance, it is possible to infer explicitly the mean offspring distribution of each case and hence to infer the transmission heterogeneity and other epidemiological parameters [3, 9, 16].

Although considerable progress has been made for analyzing heterogeneity, these

methods also showed some theoretical and practical limitations. Firstly, all these methods required context-specific information which could be hard to obtain and/or could be erroneous. For example, the contact tracing in epidemiological investigation may be time-consuming and subjective [17] and has to be limited to a certain number of infected cases. In viral genetic analysis, the commonly used correspondence between the reconstructed viral phylogeny and the transmission history may be biased if there are within-host evolution and recombination in viral genomes [18]. When incorporating the spatial information, the model simply assumes that transmission occurred mostly within close residence because of the lack of detailed individual movement data, which is only appropriate under certain control measures [3, 9].

In addition, most of existing studies assumed a constant level of heterogeneity for an epidemic under study, which may in fact grow and/or decline through the epidemic. This simplification would bring some computational benefit but failed to characterize the temporal change of heterogeneity over the epidemic. Although Lau et al [3, 9] compared the degree of heterogeneity in different periods of an outbreak (i.e., before and after deploying the control measures), it could still be hard to reflect the real-time development of the epidemic and consequently lead to inadequacy in epidemic control to a certain extent. ,

Monitoring real-time transmission dynamics from incidence data has drawn a lot of research efforts. Several tools for the estimating of real-time reproduction number based on incidence data had been developed with successful applications [19–21], but the study on real-time transmission heterogeneity is so far rather limited. In some recent studies, researchers suggested the relationship between the transmission heterogeneity and the incidence over an epidemic [22–25], but none have attempted to accurately delineate the heterogeneity with incidence data and to compare with those records in literatures. In this study, we attempted to develop a simple method to estimate the transmission heterogeneity on the basis of incidence data. Specifically, we extended the homogeneous transmission model in [19, 20] to allow for the variation of infectiousness at different times and among different people, and consequently generated real-time estimates of transmission heterogeneity and reproduction number simultaneously. Moreover, we evaluated this model with both simulated data and historic epidemic data, which turned out to be consistent with that of those involving contact-tracing or spatial data. Our model performed robust even in the presence of measurement errors such as under-reporting or misspecification of serial interval. We further explored the transmission heterogeneity of the new SARS-CoV-2 variant Omicron based on the incidence time series from South Africa.

Materials and methods

Renewal process model of transmission

We considered an outbreak observed regularly (in days, weeks or months) over the time period $1 \leq t \leq T$. Let I_t be the incidence or number of newly infected cases at time t, and the epidemic curve till time t is denoted as $\bar{I}_1^t = \{I_1, I_2, \cdots, I_t\}$. For simplicity, we excluded the possibility of imported case during the study period. However, this restriction could be relaxed by discriminating the effect on newly infections of local/imported cases as in [20].

We adopted the renewal process to model the transmission of the infectious disease. Under the standard renewal process model [19], the newly infected at time t (i.e., I_t) is generated by all the infectious individuals who had been infected before time t according to a Poisson relation as:

$$I_t|\bar{I}_1^{t-1} \sim \text{Pois}(R_t\Lambda_t) \tag{1}$$

where "|" stands for conditions and Pois stands for *Poisson* distribution. The parameter of R_t is the instantaneous reproduction number, representing the average number of secondary cases that caused by a random case at time t if circumstances remained the same after that [19, 26]. The quantity $\Lambda_t = \sum_{s=1}^{t-1} I_s w_{t-s}$, known as the total infectiousness, characterizes how many past effective cases contribute to the newly observed case-count at time t. The weight w_s defines the impact of each past case on the newly infection, which could be approximated by the generation time distribution or the serial interval distribution.

Instant-individual reproduction number

In this study, we aimed to extend the standard model to allow for transmission heterogeneity during the transmission process. To characterize the effect of each infected individual on new infection at a particular time point, we introduce the "instant-individual reproduction number" (IIRN), denoted as $v_{s,t}^i$, representing the expected number of secondary cases generated at time t by the i-th individual infected at time s (where $s < t$). We also use the Poisson distribution to model the stochastic effect in transmission [1], so the number of secondary cases caused by a particular case (i.e., offspring distribution) in the given context is $\text{Pois}(v_{s,t}^i)$. In addition, we adopted the assumption that the offspring distributions of different cases were independent, so the incidence I_t is the sum of these *Poisson*-distributed variables. In other words, I_t is *Poisson*-distributed with the composite rate of $v_t = \sum_{s<t,i} v_{t,s}^i$.

The concept of IIRN provides a new tool to explore the variation of infectiousness between different individuals and among different times. Next we study how the standard renewal process model and two recently proposed heterogenous transmission models fit within this framework. The standard renewal process model is a homogeneous transmission model, which assumed a constant IIRN for all the infected cases who had been infected at the same time. In other words, the standard model is identical to assume $v_{s,t}^i = w_{t-s}R_t$. Hence the composite rate at time t is $v_t = \sum_{s\leq t,i} v_{t,s}^I = R_t\Lambda_t$. This model, while useful for monitoring the average transmission potential, fails to account for the variation in infectiousness particular found in the those superspreading events.

Another common method of allowing for transmission heterogeneity is an instant-level heterogeneity model [22, 25]. This model extended the standard model (1) by replacing the instantaneous reproduction number R_t with an instant-related random variable for all the infected cases, that is,

$$v_{t,s}^i = w_{t-s}\eta_t, \text{ where } \eta_t \sim \Gamma(k_t, \frac{k_t}{R_t}).$$

where $\Gamma(\cdot,\cdot)$ stands for *Gamma* distribution in the shape-rate parameterizations. Therefore, the composite rate under this model is $v_t = \sum_{s<t,i} v_{t,s}^i = \Lambda_t\eta_t \sim \Gamma(k_t, \frac{k_t}{\Lambda_t R_t})$. And the incidence I_t is Negative Binomial distribution as (NegB indicating *Negative Binomial distribution*):

$$I_t|\bar{I}_1^{t-1} \sim \text{NegB}(k_t, \frac{k_t}{\Lambda_t R_t + k_t})$$

This model accounted for the variation in infectiousness at different times, which could be useful in epidemic forecasting in the long term [22, 25]. But this model overlooked the variation in infectiousness of different infectious individuals, and hence failed to identify the exact degree of heterogeneity from incidence data (showed in Results).

Recently, Johnson et al [27] proposed an individual-level heterogeneity model to characterize transmission heterogeneity within the renewal process framework. The authors assumed random infectiousness for each infected individual at the time of being infected (e.g. at time s), so its infectiousness in later time steps could be calculated as

$$v_{t,s}^i = w_{t-s}\eta_s^i, \text{ where } \eta_s^i \sim \Gamma(k_s, \frac{k_s}{R_s}).$$

With this model, the composite rate of newly infection at time t is $v_t = \sum_{s\leq t,i} v_{t,s}^i = \sum_s w_{t-s}\Theta_s$, where $\Theta_s = \sum_i \eta_s^i \sim \Gamma(k_s * I_s, \frac{k_s}{R_s})$. Θ_s was referred to as the disease momentum [27], representing the total infectiousness of all the cases infected at time s. As the weighted summary of *Gamma* variables is not *Gamma* distributed, the incidence I_t can only be approximated by

$$I_t|\bar{I}_1^{t-1} \sim \text{Pois}(\sum_s w_{t-s}\Theta_s).$$

Although the individual level transmission heterogeneity has been characterized in this model, it not only overlooked the instant-level heterogeneity but also introduced a large number of nuisance parameters of disease momentums $\{\Theta_s\}$. These nuisance parameters destroyed the independence of incidence data among different times, and incurred considerable computational complexity in the analysis of incidence time series, which hinder the accuracy of estimating parameters of interest. Simulation study showed unstable estimation results of transmission dynamics [27].

Instant-individual heterogeneity model

For directly transmitted diseases such as SARS-CoV, MERS, Ebola, or COVD-19, the instant individual reproduction number is affected by a complex mixed factors of host, pathogen and environmental factors [1, 28]. Therefore the reproduction number is specific to time and individual. Here we assumed $v_{s,t}^i$ to be a random variable, and its values are drawn independently, for each individual i and each instant t, from a *Gamma* distribution with mean of $w_{t-s}R_t$ and the rate of $\frac{k_t}{R_t}$, that is,

$$v_{t,s}^i \sim \Gamma(w_{t-s}k_t, \frac{k_t}{R_t}) \tag{2}$$

Under this random IIRN assumption, heterogeneous transmission stems from the variation in reproduction numbers of different individuals and at different times. And superspreading events were likely triggered by those important realizations from the right-hand tail of the distribution of IIRN, which indicated a random mixture of host, pathogen and environmental factors of assisting the rapid transmission of disease [28].

The parameter k_t in (2), referred to as (instantaneous) dispersion number, was introduced to control the transmission heterogeneity. Similar to the explanation of instantaneous reproduction number R_t in [26], the instantaneous dispersion number k_t also controls the variation in the offspring distribution of a random infected case. Suppose the transmission dyanmics remains the same (i.e., the R_t and k_t keep constant) during the infectious time of the i-th case, its individual reproduction number over the whole infectious period is the sum of independent IIRNs over all infectious instants, that is $v_s^i = \sum_{t\geq s} v_{t,s}^i \sim \Gamma(k_t, \frac{k_t}{R_t})$. As a consequence of this *Gamma-Poisson* mixture, the offspring distribution of the particular case is Negative Binomial distribution as

$$I_s^i \sim \text{NegB}(k_t, \frac{k_t}{R_t + k_t})$$

with the mean of $\mu = E(I_s^i) = R_t$ and variance $\sigma^2 = R_t(1 + R_t/k_t)$. The offspring distribution was identical to the standard model of transmission heterogeneity in [1].

Obviously, the dispersion number k_t is an empirical measure of degree-of-transmission heterogeneity, with smaller k_t indicates higher variance in offspring distribution (i.e., higher level of heterogeneity). When k_t decreases both the likelihood of super- and that of sub-spreading events increase [22]. Traditionally, it is regarded as *significant* transmission heterogeneity when k_t gets smaller than 1 [1].

Based on the random IIRN assumption, the total effect of all the infected cases on the newly infection at time t was the sum of their independent IIRNs, that is, $v_t = \sum_{s\leq t,i} v_{t,s}^i \sim \Gamma(k_t\Lambda_t, \frac{k_t}{R_t})$. Furthermore, the incidence I_t is Negative-Binomial distributed as

$$I_t|\bar{I}_1^{t-1} \sim \text{NegB}(k_t\Lambda_t, \frac{k_t}{R_t+k_t}),$$

that is,

$$P(I_t|\bar{I}_1^{t-1}, w, R_t, k_t) = \binom{\Lambda_t k_t + I_t - 1}{\Lambda_t k_t - 1} (\frac{R_t}{R_t+k_t})^{I_t} (\frac{k_t}{R_t+k_t})^{\Lambda_t k_t}, \tag{3}$$

This incidence model is referred to as the Instant-individual heterogeneity model.

If assuming that the transmission dynamics (i.e., reproduction number R_t and dispersion number k_t) was constant, it is possible to obtained the overall estimate of both transmission heterogeneity and reproduction number simultaneously by fitting the observed incidence time series to this model. Additionally, in real epidemics, the transmission dynamics may vary with time because of changes in host and environmental factors. A common framework for monitoring the temporal trend of transmission dynamics is to assume constant transmissibility potential and heterogeneity over a time period $[t-\tau+1, t]$, measured by $R_{t,\tau}$ and $k_{t,\tau}$ [19]. With this assumption, the likelihood of the incidence $I_{t-\tau+1}, \cdots, I_t$ given the transmission dynamics ($\{R_{t,\tau}, k_{t,\tau}\}$) and conditioned on the previous incidences $I_1, \cdots, I_{t-\tau}$ is

$$P(I_{t-\tau+1}, \cdots, I_t|\bar{I}_1^{t-\tau}, R_{t,\tau}, k_{t,\tau}) = \prod_{s=t-\tau+1}^{t} \binom{\Lambda_s k_{t,\tau} + I_s - 1}{\Lambda_s k_{t,\tau} - 1} (\frac{R_{t,\tau}}{R_{t,\tau}+k_{t,\tau}})^{I_s} (\frac{k_{t,\tau}}{R_{t,\tau}+k_{t,\tau}})^{\Lambda_s k_{t,\tau}}, \tag{4}$$

On the basis of this joint likelihood function of both reproduction number and dispersion number, it is possible to infer the real-time transmission heterogeneity from the incidence data, which gives a more complete view of the characteristics of disease spreading. In particular, the maximum likelihood estimation of the reproduction number with this new likelihood function is given by $\hat{R}_{t,\tau} = \frac{\sum_{s=t-\tau+1}^{t} I_s}{\sum_{s=t-\tau+1}^{t} \Lambda_s}$, which coincides with that of the homogeneous model [19, 36]. This property guarantes that the estimation of reproduction number with our model is robust to the bias of constant under-reporting rate (shown in Results). It is also possible to derive the posterior distribution of R_t and k_t by using a Bayesian framework.

Simulation of Incidence time series

We applied the Instant-individual heterogeneity (IIH) model to simulated datasets to test its accuracy under various levels of transmission heterogeneity and reproduction number. Each simulation began with 10 infected index cases, and stopped after 24 days. We assumed constant reproduction number R and dispersion number k, and simulated the newly infection according to the likelihood of the incidence in (3).

Specifically, we set three levels of reproduction number R as 1.1, 1.3 and 1.5; and four levels of dispersion number k as 0.2, 0.5, 2, and 5. We also varied the window size used in the estimation as 7, 14 and 21 days. The serial interval distribution was set as a

gamma distribution with mean of 5.2 days and the standard deviation of 1.72 days as in the COVID-19 [29]

We chose the incidence data from the last time window to perform estimation. We assumed non-informative priors of uniform distribution over [10^{-6},100] and [0.1,10] for the reproduction number and the dispersion number respectively. Both the maximum a posteriori (MAP) estimation and the 95% highest posterior density (HPD) interval of reproduction number and dispersion number were generated.

The simulation was repeated 100 times under each condition. Three criteria were used to evaluate the accuracy of the estimation. Firstly, the *relative* root mean squared errors (RMSEs) were calculated for the estimation of R and k respectively. The relative RMSE was defined as:

$$\sqrt{\frac{\sum_i(\hat{\theta}_i/\theta-1)^2}{n-1}},$$

where θ is the true value of parameter, and $\hat{\theta}_i$ is the estimation of parameter based on the i-th simulation. n stands for the number of simulations

Secondly, the coverage of the 95% HPD of reproduction number R was calculated. Thirdly, the probability of correctly identifying heterogeneity, namely the proportion of simulations where both the true dispersion number k and its estimate were larger or smaller than 1, was calculated for the estimation of k.

Analyzing real epidemic data

We also applied the instant-individual heterogeneity model to disease incidence time series from five past outbreaks where the levels of heterogeneity were estimated on the basis of contact tracing data or individual level spatial information. The commonly used transmission heterogeneity model in [22] (referred to as the instant-level heterogeneity model) was also used to analyze these incidence time series under the same setting for comparison.

We retrieved the epidemic curves, as well as the mean and standard deviation of the serial intervals of these epidemics from the literature (Table 1). These epidemics were classified into two groups according to the way of estimating transmission heterogeneity in previous studies. The first group (static scenario) includes three epidemics, i.e., COVID-19 in Hongkong, China between 2020-01-24 and 2020-04-28 (referred to as COVID-19 in Hongkong), COVID-19 in Tianjing, China between 2020-01-21 and 2020-02-15 (referred to as COVID-19 in Tianjing), and MERS in several places in South Korea between 2015-05-11 and 2015-06-26 (referred to MERS in South Korea). For each of these outbreaks, previous study assumed constant transmission parameters over the study period and estimated the overall R and k on the basis of contact-tracing data [10, 38, 39]. Here we followed this assumption and applied the IIH model and instant-level model to the incidence data over the same period to get the overall estimation of k and R, which were compared with the corresponding records in literatures.

The second group (time-varying scenario) includes two outbreaks: one is the Ebola epidemic between Aug 04, 2014 (week 36), and March 29, 2015 (week 13), in the capital Freetown of Sierra Leone (referred to as Ebola, Sierra Leone); the other is the COVID-19 in five counties (i.e., Cobb, DeKalb, Gwinnett, Fulton and Dougherty) in Georgia, United State during the period between March 1, 2020 and May 3, 2020 (referred to as COVID-19, Georgia). For each of the epidemics, previous studies analyzed the transmission heterogeneity of several periods on the basis of individual-level spatial information as well as population density data [3, 9, 16]. To make a comparison with the recorded temporal trends of heterogeneity in literatures, we assumed constant transmission dynamics over a sliding time window to reveal the

Table 1. Description of 5 historic epidemic data analyzed.

Category	Disease	Location	Duration of Outbreak	Mean (SD) serial interval	Reference for Mean (SD)	Source of Incidence time series
Static scenario	COVID-19	Hongkong, China	from 2020-01-24 to 2020-04-28	5.2 (1.72)	[29]	[30]*
	COVID-19	Tianjing, China	from 2020-01-21 to 2020-02-15	5.2 (1.72)	[29]	[10]
	MERS	South Korea	from 2015-05-11 to 2015-06-26	12.6 (2.8)	[31]	[32]
Time Varying scenario	Ebola	Freetown, Sierra Leone	from 2014-08-04 to 2015-03-29	15.3 (9.3)	[33]	[33]
	COVID-19	Georgia, United States	from 2020-03-01 to 2020-05-03	5.2 (1.72)	[29]	[34]*
	COVID-19	South Africa	from 2021-05-01 to 2022-01-09	5.2 (1.72)	[29]	[30]*

*Dataset were accessed on 2022-02-01

real-time estimation of k_t and R_t on the basis of the incidence data. We set the window length as 7 time-steps (i.e., days or weeks depending on the frequency of incidence data collection) for all these analyses, which was recommended in [20] when monitoring the temporal trend of reproduction number.

In addition, we also explored the transmission heterogeneity of the variant of Omicron by applying the IIH model to the incidence time series from the South Africa between 2021-05-01 and 2022-01-07. The real-time estimation of transmission dynamics was also generated as the same procedure in the time-varying scenario.

During these studies, the discrete distribution of the serial interval was then obtained by assuming a gamma distribution truncated by Mean+3*SD of serial interval. We used Bayesian Monte Carlo Markov Chain algorithm to calculated the posterior distribution from the likelihood functions of (3) and (4) by assuming non-informative priors of uniform distribution over [10^{-6},100] and [0.1,10] for the reproduction number and the dispersion number respectively. The resulting 95% high probability domain (95% HPD) could be directly compared with those 95% confidence intervals in literature. The inference algorithm was implemented via the open-sourced python package of pymc3 [35]. All the codes for this study are available online: https://github.com/yunPKU/infer_heterogeneity_from_incidence

Sensitivity Analysis

Underreporting and misspecification of serial interval are ubiquitous biases for the analysis of epidemiological data [9, 36]. To explore the effect of these biases on the estimation of real-time dispersion number and reproduction number, we performed sensitivity analysis on the basis of the epidemic data of Ebola, Sierra Leone [33]. Firstly, we explore the effect of underreporting on our analysis by testing 4 reporting rates (i.e., ρ =0.8, 0.6, 0.4, 0.2). With each rate, we generated synthetic incidence time series in the Ebola epidemic by increasing the recorded incidence data proportionally.

Secondly, we tested the errors in the serial interval by analyzing the Ebola epidemic data with biased serial interval distribution. We performed estimation with three values of bias for the mean (i.e., -7 days, 7 days, and 14 days) and three biases for std (i.e., -3.5 days, 3.5 days and 7 days) respectively.

Results

Evaluation on simulated data

With the simulated data, our model could accurately estimate the overall dispersion number and the reproduction number providing sufficient data (Figure 1) . As the window length increased, the relative RMSEs of these two estimates k and R showed a decreasing trend under all simulation settings. Also, the probability of identification of k and the coverage of 95% HPD of R increased with the window length.

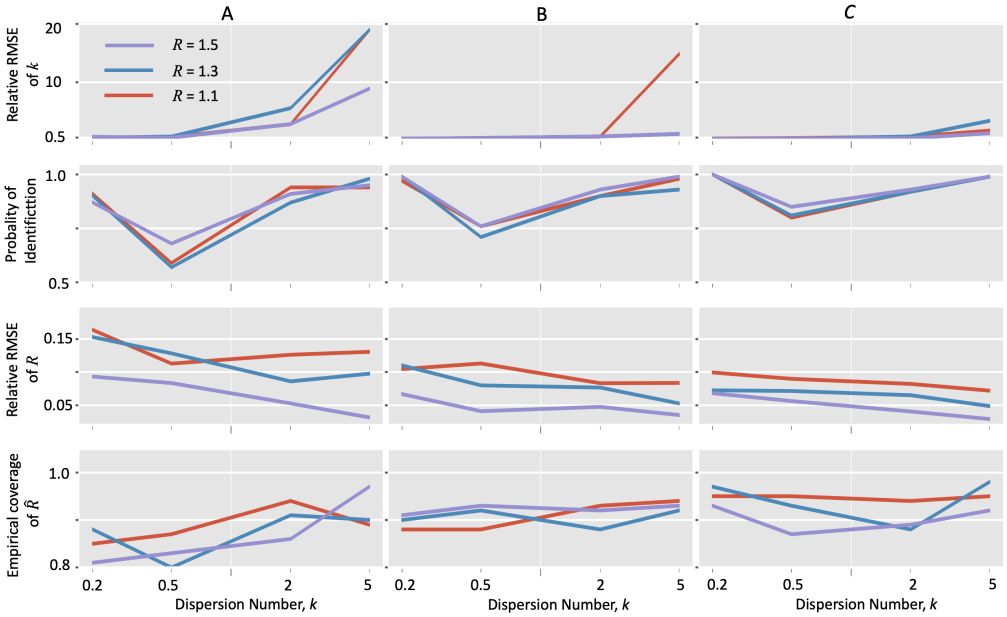

Fig 1. Accuracy of the instant-individual heterogeneity model in estimating transmission dynamics with simulated data. Incidence data were generated with the instant-individual heterogeneity model (3) with given reproduction number and dispersion number. Each simulation began with 10 cases and stopped at 24 days. The relative root mean squared error (RMSE) and the coverage of 95% high probability density interval were calculated for the estimation of reproduction number R. The relative RMSE and the probability of identification (defined in the section of methods) were calculated for the estimation of dispersion number k. A. Estimation under window size = 7 days; B. Estimation under window size = 14 days; C. Estimation under window size = 21 days;

It should be noted that the true values of k affected the performance of these estimates in different ways. On the one hand, the relative RMSEs of estimating k under the homogenous conditions (i.e., true $k > 1$) were larger than those under the heterogeneous conditions (i.e., true $k < 1$). This observation is consistent with the simulation study of dispersion number based on the methods of moment [1, 37], which found that the dispersion number is likely to be overestimated for small sample size. However, the probabilities of identifying of k under the homogeneous conditions were closer to 0.9, suggesting that our model could correctly identify this homogeneous condition. In addition, as to the estimation of R, the relative RMSE decreases and the coverage of 95% HPD increases when the true k increaseed, suggesting that the estimate of R is more accurate for the homogeneous situation.

Validation with Real epidemics

Static scenario

When analyzing the incidence data of three epidemics, our estimates of the dispersion number k were 0.30 (95% HPD: 0.20 0.43), 0.54 (95% HPD: 0.16 1.33), and 0.30 (95% HPD: 0.12 0.57) for the epidemics of COVID-19 in Hongkong (2020) and the COVID-19 in Tianjin (2020), and the MERS in South Korea (2015), respectively. These suggested significant transmission heterogeneity in these outbreaks. Our estimates were consistent with those revealed by previous studies based on contact tracing data of the three epidemics (Figure 2) [10, 38, 39]

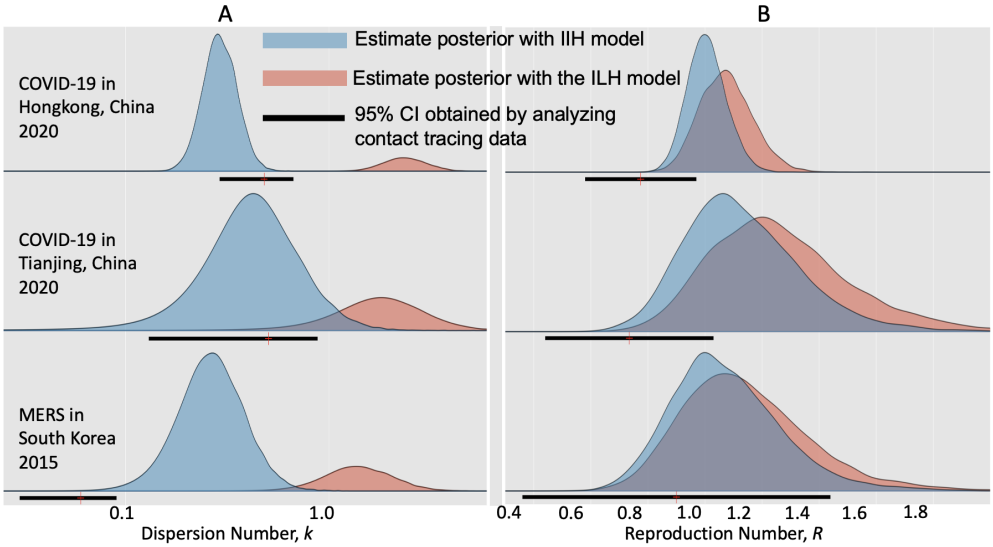

Fig 2. Comparison of estimating transmission dynamics of three epidemics with the instant-individual heterogeneity (IIH) model and the instant-level heterogeneity (ILH) model in [22]. During each epidemic, transmission dynamics (i.e., reproduction number R and dispersion number k) were assumed constant. Colored areas showed the posteriors of the estimates by analyzing incidence times series. Black solid lines represented the estimates in literatures obtained by analyzing the contact tracing data of these epidemics [10, 38, 39]. A: Estimation of reproduction number (R); B: Estimation of dispersion number (k).

When analyzing the incidence data with the instant-level heterogeneity model [22], the estimates of k were 2.50 (95% HPD: 1.48 3.30), 2.23 (95% HPD: 0.69 5.39), and 1.60 (95% HPD: 0.68 3.14) for these three epidemics respectively, which exceeded the threshold value of 1 and hence failed to recognize significant transmission heterogeneity in these outbreaks.

As to the estimation of reproduction number R, both the IIH model and the instant-level model gave consistent estimates with previous studies (Figure 2 B), while the estimates of the IIH model were closer to those estimates from the contact tracing data.

Time-varying scenario

By assuming that the transmission parameters remanin constant over a time window 7 steps (i.e., days or weeks depending on the frequency of incidence data collection), we obtained the real-time estimation of the dispersion number (k_t) as well as the

reproduction number (R_t) over an epidemic. Firstly, we analyzed the weekly incidence of probable and confirmed cases of Ebola between August 4th, 2014, and March 29th, 2015, in the capital Freetown of Sierra Leone. By setting the reference time of 2014-11-01 as in [16], the whole duration was divided into 5 periods (P1 P5, Figure 3).

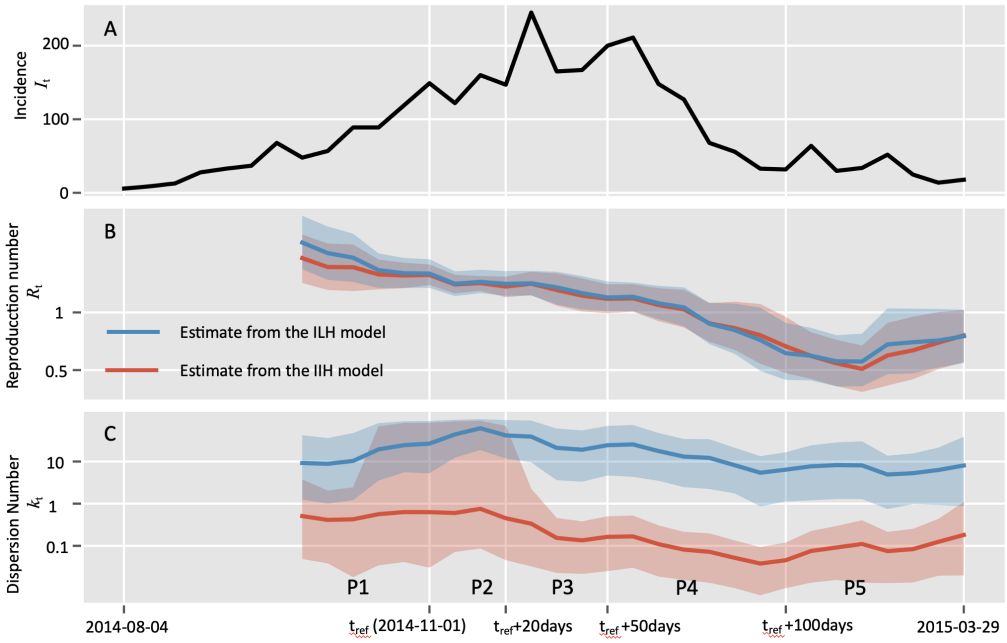

Fig 3. Comparison of estimating real-time transmission dynamics of the Ebola epidemic between Aug 04, 2014 (week 36), and March 29, 2015 (week 13), in the capital Freetown of Sierra Leone. Transmission dynamics (i.e., reproduction number R and dispersion number k) were assumed constant over a window of 7 weeks, and the estimates were obtained by analyzing the incidence data of the time window. Solid lines show the mean estimates from two methods, i.e., red curves and blue curves represent the estimation from the instant-individual heterogeneity model (IIH) and the instant-level heterogeneity (ILH) model respectively. The shaded areas show the 95% high probability density (HPD) intervals. As in [16], the reference time t_{ref} was set as 2014-11-01, and the whole time period was divided into five periods as: from 2014-10-20 to t_{ref} (period 1), t_{ref} to t_{ref} +20 days (period 2), t_{ref} +20 days to t_{ref} +50 days (period 3), t_{ref} + 50 days to t_{ref} + 100 days (period 4), and thereafter (period 5).A: Incidence data of the confirmed and probable cases; B: Estimation of reproduction number (R_t); C: Estimation of dispersion number (k_t).

The estimated dispersion number (k_t) from the instant-individual heterogeneity model remained stable during the first two periods (P1 and P2) and decreased since the third period and then reached the lowest level around 0.1 in the forth period. At last, the k_t bounced up to around 0.2 in the last period (Figure 3 C). This temporal trend of k_t was consistent with previous study based on individual level spatial information, suggesting the transmission heterogeneity were becoming more significant as the epidemic went on and might be crucial to driving the spreading of Ebola disease in the study area [16]. In contrast, the instant-level model generated much higher estimate of dispersion number k_t which remained above 1, suggesting it failed to reveal the significant transmission heterogeneity during this outbreak (Figure 3 C).

We also noted that both the IIH model and the instant-level model gave similar estimation of the real-time reproduction number, which showed a declining trend in most part of the period, and was below 1 since the middle of the fourth period (Figure 3

B).

Secondly, we validated the IIH model with the COVID-19 incidence data, between March 1, 2020 and May 3, 2020, in five counties of Georgia state, USA (Figure 4). The estimated real-time dispersion number (k_t) in all the five counties declined from the level of above or closer to 1 during period 1 (before Apr 03) to the level of closer to 0.1 in period 3 after Apr 17 (Figure 4C), suggesting significant transmission heterogeneity of COVID-19 in all these counties [9]. Notably, the transmission heterogeneity became mostly significant in the rural area (Dougherty) with the estimated k_t reached the lowest level of around 0.01 in the second period, which was consistent with the documented superspreading event in this county [40]. In contrast, the instant-level model, generated the real-time estimation of k_t being above 1, which failed to identify the significant transmission heterogeneity in all these counties.

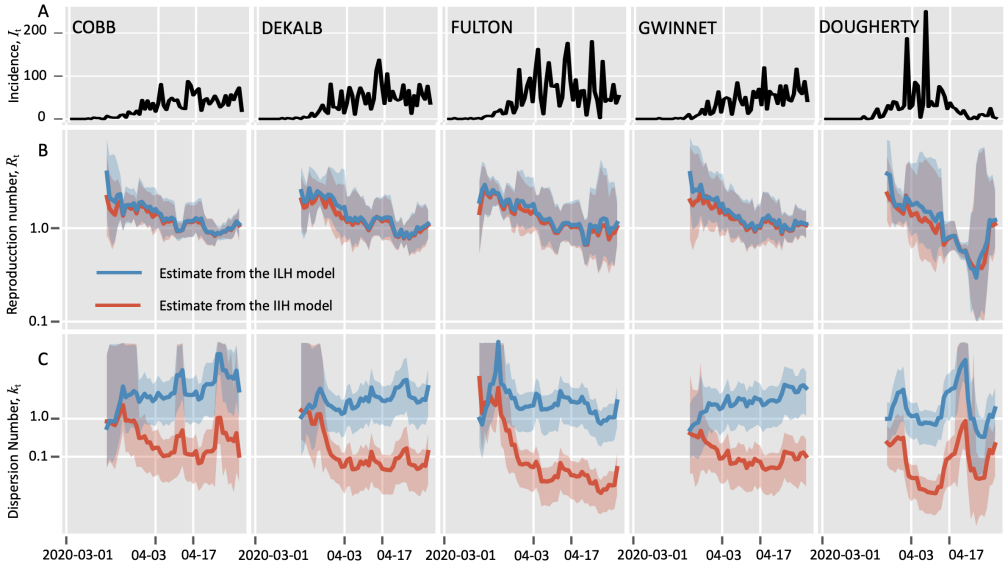

Fig 4. Comparison of estimating real-time transmission dynamics of the COVID-19 epidemic between March 1, 2020 and May 3, 2020, in five counties of Georgia state, USA. Transmission dynamics (i.e., reproduction number R and dispersion number k) were assumed constant over a window of 7 days, and the estimates were obtained by analyzing the incidence data of the time window. Solid lines show the mean estimates from two methods, i.e., red curves and blue curves represent the estimation from the instant-individual heterogeneity model (IIH) and the instant-level heterogeneity (ILH) model respectively. The shaded areas show the 95% high probability density (HPD) intervals. As in [9], the reference time was set as April 3rd, 2021 when the shelter-in-place order was announced. The whole study period was divided into three periods, i.e., before April 3rd, between April 3rd and April 17th, after April 17th. A: Incidence data of the confirmed and probable cases; B: Estimation of reproduction number (R_t); C: Estimation of dispersion number (k_t).

The IIH model and the instant-level model gave similar estimation of reproduction number R_t (Figure 4B). We found that the reproduction numbers in four countries (i.e., except for Gwinnet) declined below 1 short after Apr-17 (i.e., 2 weeks after the shelter-in-place order), suggesting the order was effective to reduce the transmission of COVID-19. Similar to the findings in [9], our IIH model also indicated that the urban area of Dougherty was the first country where R_t declined below 1.

Sensitivity analysis

By analyzing the synthetic data with the IIH model, we found that as the real-time dispersion number (k_t) decreased as the reporting rate decreased, suggesting that the estimation of heterogeneity was conservative if there were a lot of missing cases. This finding is consistent with [16]. Fortunately, this effect of reporting rate was not considerable even when the reporting rate decreased to 0.4 (i.e., 60% cases were missing), where the estimation of k_t was still covered by the 95% HPD obtained under the 100% reporting rate (Figure 5B). Also, the temporal trends of k_t estimated under different reporting rates were similar, suggesting the surveillance of the temporal trend of the heterogeneity with the IIH model was robust to the bias of underreporting.

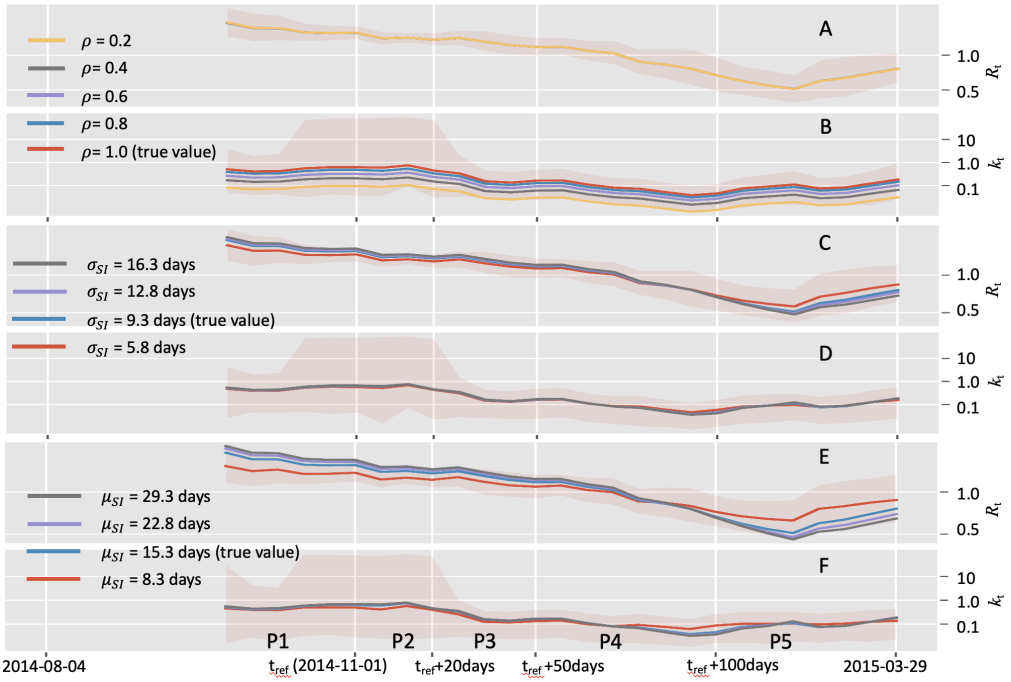

Fig 5. Effects of constant underreporting rates and misspecification of the serial interval on estimating transmission dynamics with the instant-individual heterogeneity model. Synthetic data incorporating missing cases were generated on the basis of the incidence data from the Ebola epidemic between Aug 04, 2014 (week 36), and March 29, 2015 (week 13), in the capital Freetown of Sierra Leone. Colored lines show the mean estimates and the shaded areas show the 95% high probability density intervals under the true values. A and B: Estimation under different reporting rates; C and D: Estimation from different specification of the serial interval mean; E and F: Estimation from different specification of the serial interval standard deviation.

In addition, we found that the estimation of R_t with the IIH model was unaffected by the reporting rate (Figure 5A). The underlying reason is that the maximum likelihood of R_t under our model is identical to that of the homogeneous transmission model [19], the estimation of R_t was robust to missing cases providing the fraction of cases observed is time-independent through the epidemic.

It has been reported that the misspecification of the serial interval (or generation interval) is a large potential source of bias when estimating reproduction number from observed incidence data [36]. However, we found that estimation of the dispersion number k_t was robust to the biases either in the mean or in the std of the serial interval

(Figure 5 D and F). The effects were small and were covered by the 95% HPDs under the true values.

As in [36], the estimation of R_t showed more visible changes than k_t because of the biases in serial interval (Figure 5 C and E). Generally, shorter serial interval (either because of change in mean or of change in std) may lead to lower estimate R_t when the true value is high and higher estimate R_t when the true value low.

Estimating real-time transmission heterogeneity of Omicron

To get a timely estimate of the transmission heterogeneity of Omicron, we applied the IIH model to the incidence data in South Africa between 2021-05-01 and 2022-01-07 [30] (acessed on 2022-02-01). This duration includes the third wave of COVID-19 caused by the Delta variant from May 2021 to September 2021, and the early stage of the potential wave caused by Omicron. With this incidence data, we could not only reveal the transmission heterogeneity of Omicron, but also we made a comparison with that of Delta.

During the period of 2021-12-01 to 2022-01-07 (referred to as Omicron wave), we estimated the overall estimation of reproduction number and the dispersion parameter were 0.99 (95% HPD: 0.85, 1.15) and $3.43 * 10^{-4}$ (95% HPD: $2.05 * 10^{-4}$,$5.14 * 10^{-4}$) respectively. To make a comparison, we focused on the epidemic wave caused by the Delta variant between 2021-06-01 to 2020-08-01 (referred to as Delta wave) during which the epidemic also experienced growth and declining. The overall estimation of reproduction number and the dispersion parameter were 1.04 (95% HPD: 0.97, 1.13) and $9.27 * 10^{-4}$ (95% HPD: $6.03 * 10^{-4}$,$1.27 * 10^{-3}$) respectively. Notably that the overall dispersion number in the Omicron wave was lower than that in the Delta wave.

By setting the window size of 7 days, we got the real-time estimation of transmission dynamics during these two periods. During the Omicron wave, the estimated reproduction number R_t reached the peak value of 2.15 on 2021-12-03 and then declined to the level around 0.9 after 2021-12-15. The underlying reason for this decrease in R_t was the deploying of control measures by the South Africa government as indicated by the government stringency index [41]. We also noted that the estimated dispersion number k_t declined since 2021-12-01 and reached a stable level about $3 * 10^{-4}$ in the middle of Dec 2021.

During the Delta wave, however, we estimated reproduction number R_t remained around 1 during this period which was smaller than the amount in the early of Dec 2021. In addition, the estimated dispersion number k_t remained close to 10^{-3}, which was higher than the stable level in the end of Dec 2021. Therefore, the overall and real-time estimation of transmission dynamics of these two period hint us that Omicron might not only have higher transmissibility but also a greater potential for superspreading.

Discussion

In this study, we proposed a reliable, flexible and generic model to estimate real-time heterogeneity using incidence time series. When it was applied to the epidemic of Ebola in Sierra Leone and the epidemic of COVID-19 in the state of Georgia, USA, the series of daily/weekly heterogeneities, according to its estimation, paralleled with the trends reported by previous studies based on individual spatial data [3, 9].

Besides this model successfully estimated the overall heterogeneity of three epidemics. Specifically, the overall heterogeneity (in terms of dispersion number k) were estimated to be 0.30 (95% HPD: 0.20 0.43), 0.54 (95% HPD: 0.16 1.33), and 0.30 (95% HPD: 0.12 0.57) in COVID-19 epidemic in Hongkong [38], COVID-19 epidemic in Tianjing [10], and MERS epidemic in South Korea [39], respectively, which were all

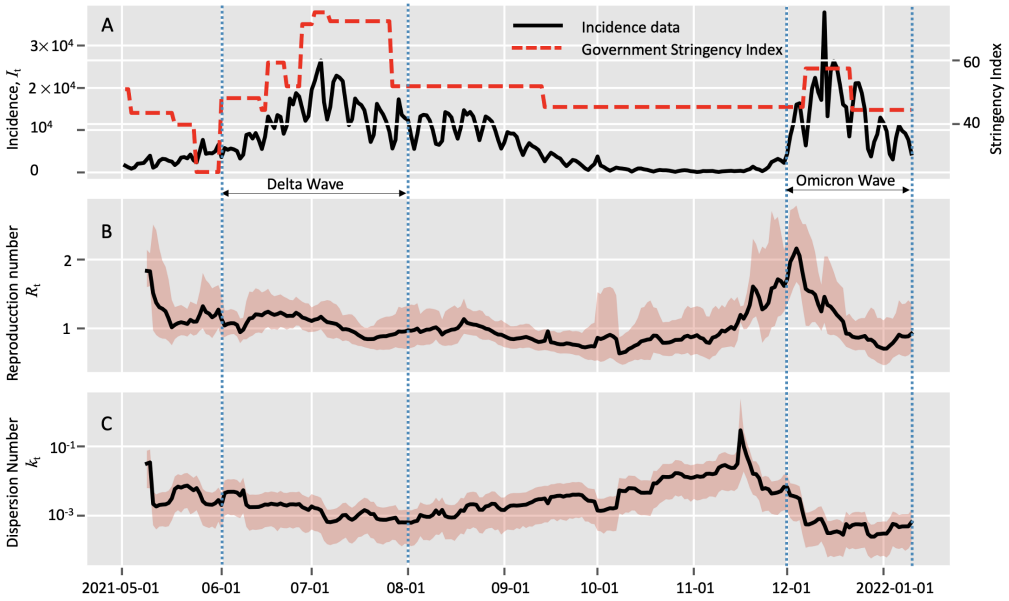

Fig 6. Estimation of real-time transmission dynamics of the COVID-19 epidemic between 2021-05-01 and 2022-01-09 in South Africa. Transmission dynamics (i.e., reproduction number R and dispersion number k) were assumed constant over a window of 7 days, and the estimates were obtained by analyzing the incidence data of the time window. Solid lines show the mean estimates and the shaded areas show the 95% high probability density (HPD) intervals. A: Incidence data of the confirmed cases and government stringency data in South Africa; B: Estimation of reproduction number (R_t); C: Estimation of dispersion number (k_t).

consistent with the heterogeneities revealed by previous studies based on contact-tracing data.

Transmission heterogeneity is a ubiquitous feature in the spread of infectious disease due to a mixture of factors involving host, pathogen and environment. Accurate estimating real-time heterogeneity is vital for prediction of future epidemics and exploring targeted interventions. Existing methods of inferring transmission heterogeneity rely heavily on sophisticated data to reconstruct the offspring distribution and largely ignore the temporal change in heterogeneity. One existing model, which involves instant-level heterogeneity [22, 25], could only allow for part of the variation and hence failed to reveal accurate real-time heterogeneity. As evidenced in our analysis of the instant-level heterogeneity model, its estimation of transmission heterogeneity (in terms of dispersion number k) of all the real epidemics remained above the threshold of 1, indicating no significant heterogeneity in these epidemics, which completely deviated from the records in literature. Our model, however, addressed the heterogeneity with a flexible and generic way to estimate the real-time heterogeneity on the basis of incidence data, which is easy to implement and was proved reliable.

The benefits of our model stem from the two theoretical advantages. Firstly, we introduced the assumption of random instant-individual reproduction number to characterize the variation of infectiousness between different people and at different times. Both these variations were important source of the heterogeneity in transmission and therefore should be characterized in the model. This assumption is applicable for directly transmitted disease such as SARS-CoV, MERS, Ebola, and COVD-19, where the infectiousness of a particular individual at a particular instant was determined by the properties of the host and pathogen and environmental circumstances [1, 28].

Secondly, our model is easy to implement as it employs only incidence data. We deduced the joint likelihood function of incidence data on both the reproduction number (R_t) and transmission heterogeneity (k_t), which enabled us easily to monitor these epidemiological parameters simultaneously.

When comparing the precision of different methods, we found that our estimation was less precise with broader credible intervals than the results based on contact-tracing data for the two outbreaks (i.e., MERS in South Korea 2015 and COVID-19 in Tianjin China, 2020) with smaller size (i.e., 100 200 cases). For the outbreak of COVID-19 in HongKong with more than 1,000 cases, our estimation had better precision than the result from contact-tracing data in terms of narrower credible interval. This might be related with the sample size of the outbreak, and our model might be more applicable to larger size epidemics.

This merit of our model could allow for fast and timely epidemiological surveillance, possibly even for the new SARS-CoV-2 variant of Omicron, which has been spreading wildly across the world since its first detection in November 2021 in Gauteng Province, South Africa. We estimated the heterogeneity (in terms of dispersion number k) was $k \approx 3.43 * 10^{-4}$ in December 2021 in South Africa, which was more significant than that of the Delta wave (i.e., $k \approx 10^{-3}$) [42]. The more significant heterogeneity of Omicron, together with its higher reproduction number, might be able to explain its unprecedentedly fast spreading. So far, little is known about the transmission heterogeneity of Omicron, and the traditionally used data for heterogeneity analysis including contact tracing data, viral sequence data and individual spatial-information have not been fully available for the analysis of its transmission heterogeneity. Our results also highlighted the need of taking more efficient measure of to reduce people gathering and the possible superspreading events [28, 43].

During the implementation of our model, the serial interval distribution is required to approximate the infectiousness profile w_s. This distribution information may not be correctly obtained at the early stage of newly emerging infectious disease or may be biased for some pathogens where infectiousness occurs before symptoms. Fortunately, our model performed robust to the misspecification of serial interval (showed in results). Additionally, we could also relieve this dependence by integrating detailed epidemiological linkage data to estimate the serial interval separately [20] or extending the inference framework to incorporating estimation of serial interval distribution and transmission dynamics simultaneously as in [44].

When interpreting the results, we regarded the transmission heterogeneity estimated based on the incidence of confirmed cases accumulating over a time window till time t as the result at that time. Since the confirmation of a case occur after the time of its infection, together with the delay due to the accumulation of data, our estimation of transmission heterogeneity definitely fell behind the reality. This delay might make our estimation misleading if the underlying transmission dynamics change rapidly during the period. We could reduce the delay by applying our model to the transformed infection data which was generated by accounting for the possible delay between infection and diagnosis [21, 45]. In addition, we could also optimize the time length of data accumulation size based on certain performance constrain such as short-term predictive accuracy [46] to get a timely and accurate estimation of transmission dynamics.

The analysis with our model could be biased by the fact that we assumed all cases be detected when analyzing the incidence data. We also showed with synthetic data that our model performed robust as long as the reporting rate (e.g., being 40%) was constant through the epidemic. However, the reporting rate could vary with time in reality because of improved case ascertainment or case definition, or testing capacity.

In this study, we utilize the *Gamma* distribution to characterize the transmission heterogeneity, which has been widely used in other studies. It also should be noted that

the *Gamma* distribution is not suitable for all types of heterogeneity in the transmission. For example, the ongoing vaccination could incur heterogeneity as some people are vaccinated and others are not. This type of heterogeneity should play an important role especially when modelling the transmission heterogeneity in the pandemic of COVID-19, which should probably be Bimodal-distributed instead of Gamma distributed.

In summary, we proposed a simple and generic model to estimate the real-time transmission heterogeneity based on incidence data. This model could help epidemiologists better understand the complex mechanism in disease spreading, especially for those that are lack of more detailed data.

Acknowledgments

The authors are grateful to Professor Jantien Backer for helpful discussion about the dataset of Ebola epidemic, and Ms. Jian Zhang for his careful examination of the manuscript. Y.J.Z and X.H.Z acknowledge support from National Natural Science Foundation of China (Grant number: 82041023), the Bill & Melinda Gates Foundation (Grant number: INV-016826) . T.B. is supported by The Swedish Research Council (grant 2020-04744). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005;438(7066):355–359.

2. Li LM, Grassly NC, Fraser C. Quantifying transmission heterogeneity using both pathogen phylogenies and incidence time series. Molecular biology and evolution. 2017;34(11):2982–2995.

3. Lau MS, Dalziel BD, Funk S, McClelland A, Tiffany A, Riley S, et al. Spatial and temporal dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic. Proceedings of the National Academy of Sciences. 2017;114(9):2337–2342.

4. Faye O, Boëlle PY, Heleze E, Faye O, Loucoubar C, Magassouba N, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. The Lancet Infectious Diseases. 2015;15(3):320–326.

5. Lakdawala SS, Menachery VD. Catch me if you can: superspreading of COVID-19. Trends in Microbiology. 2021;29(10):919–929.

6. Lewis D. Superspreading drives the COVID pandemic–and could help to tame it. Nature. 2021;590(7847):544–547.

7. Lee N, Hui D, Wu A, Chan P, Cameron P, Joynt GM, et al. A major outbreak of severe acute respiratory syndrome in Hong Kong. New England Journal of Medicine. 2003;348(20):1986–1994.

8. Stein RA. Super-spreaders in infectious diseases. International Journal of Infectious Diseases. 2011;15(8):e510–e513.

9. Lau MS, Grenfell B, Thomas M, Bryan M, Nelson K, Lopman B. Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2

transmission in Georgia, USA. Proceedings of the National Academy of Sciences. 2020;117(36):22430–22435.

10. Zhang Y, Li Y, Wang L, Li M, Zhou X. Evaluating transmission heterogeneity and super-spreading event of COVID-19 in a metropolis of China. International journal of environmental research and public health. 2020;17(10):3705.

11. Althaus CL. Ebola superspreading. The Lancet Infectious Diseases. 2015;15(5):507–508.

12. Ypma RJ, van Ballegooijen WM, Wallinga J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics. 2013;195(3):1055–1062.

13. Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS computational biology. 2013;9(3):e1002947.

14. Barido-Sottani J, Vaughan TG, Stadler T. A multitype birth–death model for Bayesian inference of lineage-specific birth and death rates. Systematic biology. 2020;69(5):973–986.

15. Zhang Y, Leitner T, Albert J, Britton T. Inferring transmission heterogeneity using virus genealogies: Estimation and targeted prevention. PLoS computational biology. 2020;16(9):e1008122.

16. Lau MS, Gibson GJ, Adrakey H, McClelland A, Riley S, Zelner J, et al. A mechanistic spatio-temporal framework for modelling individual-to-individual transmission—With an application to the 2014-2015 West Africa Ebola outbreak. PLoS computational biology. 2017;13(10):e1005798.

17. Malmberg H, Britton T. Inflow restrictions can prevent epidemics when contact tracing efforts are effective but have limited capacity. Journal of The Royal Society Interface. 2020;17(170):20200351.

18. Volz EM, Romero-Severson E, Leitner T. Phylodynamic inference across epidemic scales. Molecular Biology and Evolution. 2017;34(5):1276–1288.

19. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. American journal of epidemiology. 2013;178(9):1505–1512.

20. Thompson R, Stockwin J, van Gaalen RD, Polonsky J, Kamvar Z, Demarsh P, et al. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. Epidemics. 2019;29:100356.

21. Huisman JS, Scire J, Angst DC, Neher RA, Bonhoeffer S, Stadler T. Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2. medrxiv. 2021; p. 2020–11.

22. Parag KV. Sub-spreading events limit the reliable elimination of heterogeneous epidemics. Journal of the Royal Society Interface. 2021;18(181):20210444.

23. Lee H, Nishiura H. Sexual transmission and the probability of an end of the Ebola virus disease epidemic. Journal of theoretical biology. 2019;471:1–12.

24. Churcher TS, Cohen JM, Novotny J, Ntshalintshali N, Kunene S, Cauchemez S. Measuring the path toward malaria elimination. Science. 2014;344(6189):1230–1232.

25. Schneckenreither G, Herrmann L, Reisenhofer R, Popper N, Grohs P. Assessing the heterogeneity in the transmission of infectious diseases from time series of epidemiological data. medRxiv. 2022;.

26. Fraser C. Estimating individual and household reproduction numbers in an emerging epidemic. PloS one. 2007;2(8):e758.

27. Johnson KD, Beiglböck M, Eder M, Grass A, Hermisson J, Pammer G, et al. Disease momentum: estimating the reproduction number in the presence of superspreading. Infectious Disease Modelling. 2021;6:706–728.

28. Prentiss M, Chu A, Berggren KK. Superspreading events without superspreaders: using high attack rate events to estimate Nº for airborne transmission of COVID-19. MedRxiv. 2020;.

29. Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, et al. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. Eurosurveillance. 2020;25(17):2000257.

30. Hannah Ritchie LRGCACGEOOJHBMDB Edouard Mathieu, Roser M. Coronavirus Pandemic (COVID-19). Our World in Data. 2020;.

31. Cowling BJ, Park M, Fang VJ, Wu P, Leung GM, Wu JT. Preliminary epidemiological assessment of MERS-CoV outbreak in South Korea, May to June 2015. Eurosurveillance. 2015;20(25):21163.

32. Shin SY, Seo DW, An J, Kwak H, Kim SH, Gwack J, et al. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. Scientific reports. 2016;6(1):1–7.

33. Backer JA, Wallinga J. Spatiotemporal analysis of the 2014 Ebola epidemic in West Africa. PLoS computational biology. 2016;12(12):e1005210.

34. Georgia coronavirus cases and deaths. Data provided by USAFacts.;. Available from: `https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/georgia`.

35. Salvatier J, Wiecki TV, Fonnesbeck C. Probabilistic programming in Python using PyMC3. PeerJ Computer Science. 2016;2:e55.

36. Gostic KM, McGough L, Baskerville EB, Abbott S, Joshi K, Tedijanto C, et al. Practical considerations for measuring the effective reproductive number, R t. PLoS computational biology. 2020;16(12):e1008409.

37. Gregory R, Woolhouse M. Quantification of parasite aggregation: a simulation study. Acta tropica. 1993;54(2):131–139.

38. Adam DC, Wu P, Wong JY, Lau EH, Tsang TK, Cauchemez S, et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. Nature Medicine. 2020;26(11):1714–1719.

39. Chowell G, Abdirizak F, Lee S, Lee J, Jung E, Nishiura H, et al. Transmission characteristics of MERS and SARS in the healthcare setting: a comparative study. BMC medicine. 2015;13(1):1–12.

40. Days After a Funeral in a Georgia Town, Coronavirus 'Hit Like a Bomb';. Available from: `https://www.nytimes.com/2020/03/30/us/coronavirus-funeral-albany-georgia.html`.

41. Hale T, Angrist N, Goldszmidt R, Kira B, Petherick A, Phillips T, et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nature Human Behaviour. 2021;5(4):529–538.

42. Wang J, Chen X, Guo Z, Zhao S, Huang Z, Zhuang Z, et al. Superspreading and heterogeneity in transmission of SARS, MERS, and COVID-19: A systematic review. Computational and Structural Biotechnology Journal. 2021;19:5039–5046.

43. Wong F, Collins JJ. Evidence that coronavirus superspreading is fat-tailed. Proceedings of the National Academy of Sciences. 2020;117(47):29416–29418.

44. White LF, Wallinga J, Finelli L, Reed C, Riley S, Lipsitch M, et al. Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. Influenza and other respiratory viruses. 2009;3(6):267–276.

45. Goldstein E, Dushoff J, Ma J, Plotkin JB, Earn DJ, Lipsitch M. Reconstructing influenza incidence by deconvolution of daily mortality time series. Proceedings of the National Academy of Sciences. 2009;106(51):21825–21829.

46. Parag KV, Donnelly CA. Using information theory to optimise epidemic models for real-time prediction and estimation. PLoS computational biology. 2020;16(7):e1007990.