Time-series modeling of epidemics in complex populations: detecting changes in incidence volatility over time

Rachael Aber[1], Yanming Di[2], and Benjamin D. Dalziel[1,3]

[1]Department of Integrative Biology, Oregon State University, Corvallis, Oregon, USA

[2]Department of Statistics, Oregon, Oregon State University, Corvallis, Oregon, USA

[3]Department of Mathematics, Oregon State University, Corvallis, Oregon, USA

Abstract

Trends in infectious disease incidence provide important information about epidemic dynamics and prospects for control. Higher-frequency variation around incidence trends can shed light on the processes driving epidemics in complex populations, as transmission heterogeneity, shifting landscapes of susceptibility, and fluctuations in reporting can impact the volatility of observed case counts. However, measures of temporal volatility in incidence, and how volatility changes over time, are often overlooked in population-level analyses of incidence data, which typically focus on moving averages. Here we present a statistical framework to quantify temporal changes in incidence dispersion and detect rapid shifts in the dispersion parameter, which may signal new epidemic phases. We apply the method to COVID-19 incidence data in 144 United States (US) counties from the January 1st, 2020 to March 23rd, 2023. Theory predicts that dispersion should be inversely proportional to incidence, however our method reveals pronounced temporal trends in dispersion that are not explained by incidence alone, but which are replicated across counties. In particular, dispersion increased around the major surge in cases in 2022, and highly overdispersed patterns became more frequent later in the time series. These findings suggest that heterogeneity in transmission, susceptibility, and reporting could play important roles in driving large surges and extending epidemic duration. The dispersion of incidence time series can contain structured information which enhances predictive understanding of the underlying drivers of transmission, with potential applications as leading indicators for public health response.

Author summary

Understanding patterns in infectious disease incidence is crucial for understanding epidemic dynamics and for developing effective public Traditional metrics used to quantify incidence patterns often overlook variability as an important characteristic of incidence time series. Quantifying variability around incidence trends can elucidate important underlying processes, including transmission heterogeneity. We developed a statistical framework to quantify temporal changes in case count dispersion within a single time series and applied the method to COVID-19 case count data. We found that conspicuous shifts in dispersion occurred across counties concurrently, and that these shifts were not explained by incidence alone. Dispersion increased around peaks in incidence such as the major surge in cases in 2022, and dispersion also increased as the pandemic progressed. These increases potentially indicate transmission heterogeneity, changes in the susceptibility landscape, or that there were changes in reporting. Shifts in dispersion can also indicate shifts in epidemic phase, so our method provides a way for public health officials to anticipate and manage changes in epidemic regime and the drivers of transmission.

Introduction

Time series of infectious disease incidence appear, to varying degrees, "noisy", showing higher frequency fluctuations (e.g., day-to-day or week-to-week fluctations) around trends at the broader temporal ranges typical for epidemic curves (e.g., months or years). Short-term fluctuations in incidence time series are caused in part by variable reporting, but may also reflect the population-level impacts of transmission heterogeneity, and changes in the landscape of susceptiblility (1; 2; 3; 4; 5; 6; 7; 8). Metrics of variability in incidence time series may therefore carry information regarding underlying drivers of transmission, and offer a relatively unexplored avenue for understanding epidemic dynamics.

Contact tracing data has revealed temporal changes in the variability of individual reproductive numbers, quantified by shifts in the dispersion parameter of the offspring distribution in branching process models (7; 8). Similar evidence has been recovered through statistical reconstruction of transmission networks, indicating temporal trends in the level of dispersion at different phases of an epidemic (3). However, the scaling from individual-level transmission heterogeneity to population-level epidemic dynamics is not fully understood. In addition, traditional contact tracing is very resource intensive, and although new approaches using digital technologies may improve its speed and scalability (9), it would be helpful to have complementary population-level analyses that can estimate heterogeneity using incidence data, which is more widely available. The importance of considering population-level variability and its relationship to individual-level variability is further highlighted by the finding that a combination of individual-based and population-based strategies were required for SARS-CoV-2 control during the early phases of the pandemic in China (6). An important challenge therefore is to develop methods that can detect changes in population-level variability in incidence time series, and to interpret these changes in terms of underlying transmission processes.

Emerging statistical techniques are leveraging variability in epidemic time series to enhance understanding of disease dynamics at the population level. For example, a recently-developed method uses population-level incidence data to the dispersion parameter of the offspring distribution, which quantifies heterogeneity in secondary cases generated by an infected individual (5). It is also possible to estimate the dispersion parameter from the distribution of the final size of a series of localized outbreaks (10). Clustering of cases has also been estimated directly from incidence data (11). Another important application links variability in incidence to epidemic phases; for example, changes in the mean and interannual coefficient of variation of measles incidence have been used to identify a country's position on the path to elimination, providing insights into vaccination strategies and epidemiological dynamics (12). Analysis of the shape of epidemic curves for influenza in cities may identify contexts where incidence is focussed more intensely (proportionally more infections in a smaller span of time) with implications on the sensitivity of cities to climate forcing and for surge capacity in the health system (4; 13).

What drives indicence dispersion and how does it relate to the underlying branching process of transmission, and to observations of cases? Under a wide range of configurations for a branching process model of contagion spread, the number of infected individuals I_t at time t will have a negative binomial distribution (14; 15), $I_t \sim NB\left(\mu_t, \theta_t\right)$, where μ_t is the expectation for I_t and θ_t is the dispersion parameter. The variance is related to the mean and dispersion parameters by $\mathrm{Var}[I_t] = \mu_t + \mu_t^2/\theta_t$, so smaller values of the dispersion parameter θ_t correspond to increasing amounts of dispersion, which increase the amounts by which the variance in realized number infected I_t exceeds the expected value, μ_t. Conversely, the distribution of I_t tends to a Poisson distribution (where the variance equals mean) as θ_t becomes large. The negative binomial distribution may also accurately model a time series if there is a changing process mean within a time step: for example, if the mean of a Poisson distribution itself follows a gamma distribution, the resulting distribution is negative binomial. Negative binomial regression (in contrast to Poisson regression) can account for unobserved heterogeneity, time dependence in the rate of a process and contagion within a time step that all lead to overdispersion (16).

An interpretation of the dispersion parameter for a time series model of counts is that events are $1 + \theta^{-1}$ times as "crowded" in time relative to a Poisson process with the same mean (17) (see Supplemental Information). For example, $\theta = 1$ corresponds to a situation where the average number of infections in the same time step as a randomly selected case will exceed the Poisson expectation by a factor of two. In a simple example relevant to surge capacity in healthcare systems, $\theta = 1$ implies that a random infectious individual visiting the emergency department at a hospital would find it on average to be twice as crowded with other infectious individuals (infected by the same pathogen) than expected for a Poisson process with the same incidence rate.

In a sufficiently large host population, and when the infectious pathogen can be assumed to spread in nonoverlapping generations, the number of infections each generation is often modeled as

$$I_{t+1} \sim NB(\mu_t = R_t I_t, \theta_t = I_t) \tag{1}$$

where time-varying reproductive number R_t gives the expected number of secondary infections aqcuired from an infected indivual at time t, and the generation time is set to 1 without loss of generality (14; 18). Setting $\theta_t = I_t$ arises from the assumption that individuals who acquire the infection at time t form independent lineages with identically distributed local rate parameters. In applications this model for $theta$ beomes $\theta_t = C_t/\rho_t$ where C_t represents reported cases and ρ the reporting rate, which relates reported cases to the true number of infections as $C_t = \rho_t I_t$. However, this requires that susceptible depletion in one lineage does not affect another, that transmission rates are equal across lineages, and that reporting rates do not vary across lineages.

In practice, these assumption will not often hold, and our aim in this paper is to develop, test and apply an alternative approach, which makes data-driven estimates of θ_t, including identifying timepoints when θ is changing rapidly, which may help to reveal the impacts of heterogeneity in transmission, susceptibility, and reporting.

Methods

By definition incidence volatility is fast relative to broadscale epidemic dynamics. Consequently, in order to estimate incidence volatility we first model incidence at broad spatiotemporal scales using natural splines (19). To allow for diverse shapes in the broadscale epidemic dynamics, these are fitted within a moving window

$$\log\left(\frac{\mu_t}{N}\right) = \sum_{j=1}^{J} \beta_j^{(t)} h_j(t) \tag{2}$$

where N represents population size, $h_j(t)$ are basis functions, J is the degrees of freedom for the splines, and $\beta_j^{(t)}$ are fitted parameters for a symmetrical window of half-width Δ, centered at t, i.e., extending from $t - \Delta$ to $t + \Delta$. The degrees of freedom to be used for the splines, and the width of the moving window will depend on the application. Explaination of the specific choices we used J and Δ for our application to COVID-19 cases in US counties is described below.

Modeling the underlying epidemic dynamics based on log-transformed incidence allows us to address the statistical effects of population size on the relationship between the mean and variance in count data, which would otherwise confound our analysis. Specifically, since population size influences the mean and variance of case count data, it impacts dispersion in different-sized populations that are otherwise identical. Accordingly, population size appears as an offset in our model of broad-scale incidence changes. That is,

$$\log(\mu_t) = \sum_{j=1}^{J} \beta_j^{(t)} h_j(t) + \log(N) \tag{3}$$

The form of the probability mass function (PMF) for infections at a time step is:

$$f_t(I) = \binom{I+\theta-1}{I} \left(\frac{\mu}{\mu+\theta}\right)^I \left(\frac{\theta}{\mu+\theta}\right)^\theta \tag{4}$$

where μ is estimated via the linear predictor outlined above.

We estimate θ_t from observed incidence data using an iteratively reweighted least-squares (IRLS) procedure for mean estimation, combined with the optimize function in R, which uses a combination of golden section search and successive parabolic interpolation, to compute θ_t. Specifically, within each time window, the spline model with an offset term was used to estimate a series of μ_s values for $s = t - \Delta$ to $s = t + \Delta$ via IRLS, as implemented in the `NBPSeq` R package (20). A single value of θ_t was then calculated for the entire time window by maximizing the likelihood function, which is based on the negative binomial probability mass function defined above.

In addition to fitting the model at each time step, we developed a likelihood-ratio test (LRT) to test the hypothesis that θ has changed at each time step. This test involves fitting and comparing two models: a null model (no θ change) and a two-part model (with a θ change). For the null model, a single θ value was fitted for the entire time window. For the θ-change model, separate θ values were fitted for the left (from $t - \Delta$ to t) and right (from t to $t + \Delta$) halves of the time window.

Very large θ values correspond to processes that are operationally identical to a Poisson process. Accordingly, the test does not produce a p-value if any of the three θ estimates exceed a user-specified threshold. In the application below, we set this threshold at 10^3, meaning that θ estimates with temporal crowding within 0.1% of that expected for a Poisson process were considered effectively Poisson.

Similarly, values of θ very close to 0 focus all of the mass of the PMF on 0, representing a scenario where the probability of observing any infections approaches zero. As with the Poisson-like tolerance described in the previous paragraph, our algorithm does not produce a p-value if any of the three θ estimates are below a user-specified threshold. This threshold will depend on the presence of contiguous sections of the time series being analyzed during which no cases are observed. In the application below, we set this threshold to 10^{-3}, because θ values below this level correspond to 0 frequencies that greatly exceed those in the data.

With both upper and lower θ thresholds—–corresponding to Poisson-like and zero tolerances, respectively—–maximum likelihood estimates (MLEs) of θ beyond these thresholds exhibited unbounded behavior. When θ exceeded the upper threshold, corresponding to processes operationally identical to a Poisson process, the MLE tended to grow arbitrarily large, with the likelihood surface reaching its maximum at the upper boundary of the calculated domain. Conversely, when θ fell below the lower threshold, representing extreme overdispersion with probability mass concentrated near zero, the MLE approached zero, and the likelihood surface peaked at the lower boundary of the domain. This behavior reflects the inability of the model to reliably estimate θ when it lies outside the specified thresholds (Fig. 1a)

Application to simulated data

We evaluated the robustness of our framework to a range of population sizes, magnitudes of dispersion changes, and shapes of underlying incidence trends by generating 2,000 simulated epidemic curves with known parameters. Epidemic trends were modeled as smoothed incidence series derived from 16-week sections randomly selected from US COVID-19 data (described below), scaled to reflect different population sizes ranging from 10^3 to 10^7. For each simulated trajectory, dispersion parameters (θ_1 and θ_2) were assigned to the two halves of the selected 16-week window, and case counts were simulated using a negative binomial distribution, where the mean (μ) was based on the smoothed incidence trend scaled by the population size. The values of θ_1 and θ_2 were drawn from a uniform distribution spanning 10^{-2} to 10^2, with 10% of simulations set to have no change in dispersion ($\theta_1 = \theta_2$). Extremely large differences in dispersion (absolute log-ratio > 3) were capped by setting $\theta_2 = \theta_1$.

Application to empirical data

We applied our framework to COVID-19 case data for the United States at the administrative level of counties, compiled by The New York Times, based on reports from state and local health agencies between Jan 4, 2020, and March 18, 2023 (21), and using county population sizes estimated for 2021 from the United States Census Bureau (22). Cumulative cases for the largest three counties in each state were converted to weekly counts by keeping the last observation from each week and differencing to compute new cases. Occasionally, reported cumulative case counts were not monotonically increasing due to corrections posted by local agencies as they resolved incoming data. As a result, approximately 0.24% of estimated new cases across all counties in the dataset were negative and these were set to zero. For each county, we analyzed overlapping 16-week windows, shifting one week (i.e., one timestep) at a time. Within each window, the framework estimated the dispersion parameter (θ) using a natural spline with three degrees of freedom to model the broad-scale trend in incidence. Outputs included estimated dispersion parameters (θ_1, θ_2, for the left and right halves of each window, and θ for the entire window), likelihood ratio test statistics, p-values for changes in dispersion at the midpoint of the window, and flags for boundary conditions such as failure to reject Poisson-like dispersion or collapse to extreme overdispersion.

Results and Discussion

The simulation studies indicate our LRT framework accurately detects changes in dispersion, with p-values averaging 0.5 when the effect size is zero and shrinking toward zero as the effect size increases (Fig. 1). The framework is also robust to the range of population sizes we addressed in the empirical data—county population sizes ranged from approximately 48 thousand to 9.9 million, and we tested the framework on simulated populations between 10 thousand and 10 million, yielding accurate estimates within the range of θ values, $10^{-}2 \leq \theta \leq 10^2$.

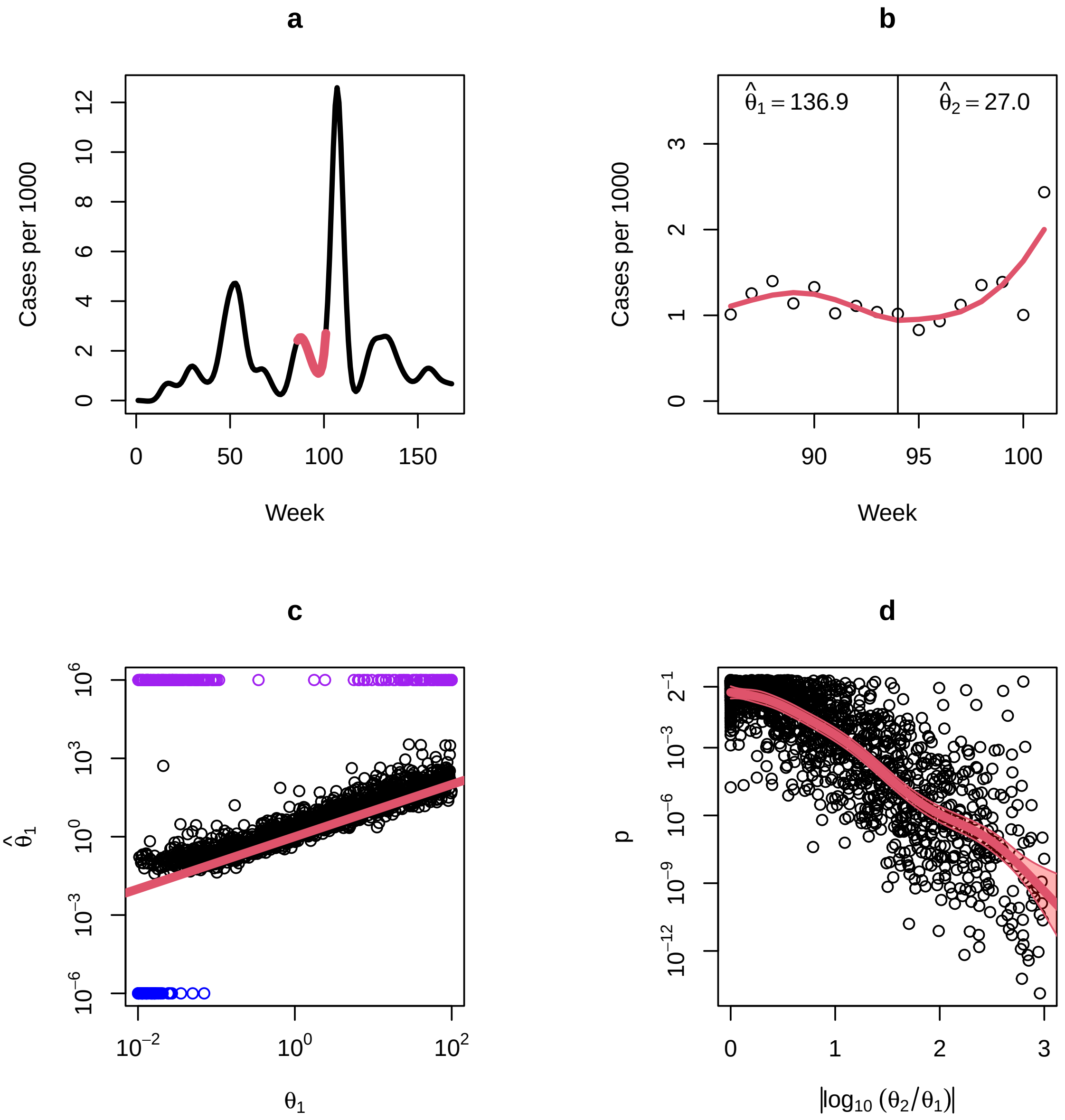

Figure 1: Detecting dispersion changes in case count time series. a: Weekly incidence of COVID-19 in the United States, with time measured in weeks since January 4, 2020, showing an example of a randomly-selected 16 week period used as an incidence trend when in simulation-based validation of the LRT test (red). b: Cases in in one county (Douglas County, Nebraska) over the sample time period with estimated incidence trend (red) and estimated dispersion values on either side of the midpoint. c: Estimated θ versus true θ in simulation studies combining a randomly-selected section of the national incidence curve with a random population size and set of dispersion values. Estimated values outside of tolerance plotted in purple (close to Poisson) and blue (close to collapsing to zero). d: Statistical power of the LRT test.

Paragraph on results shown in figure 2.

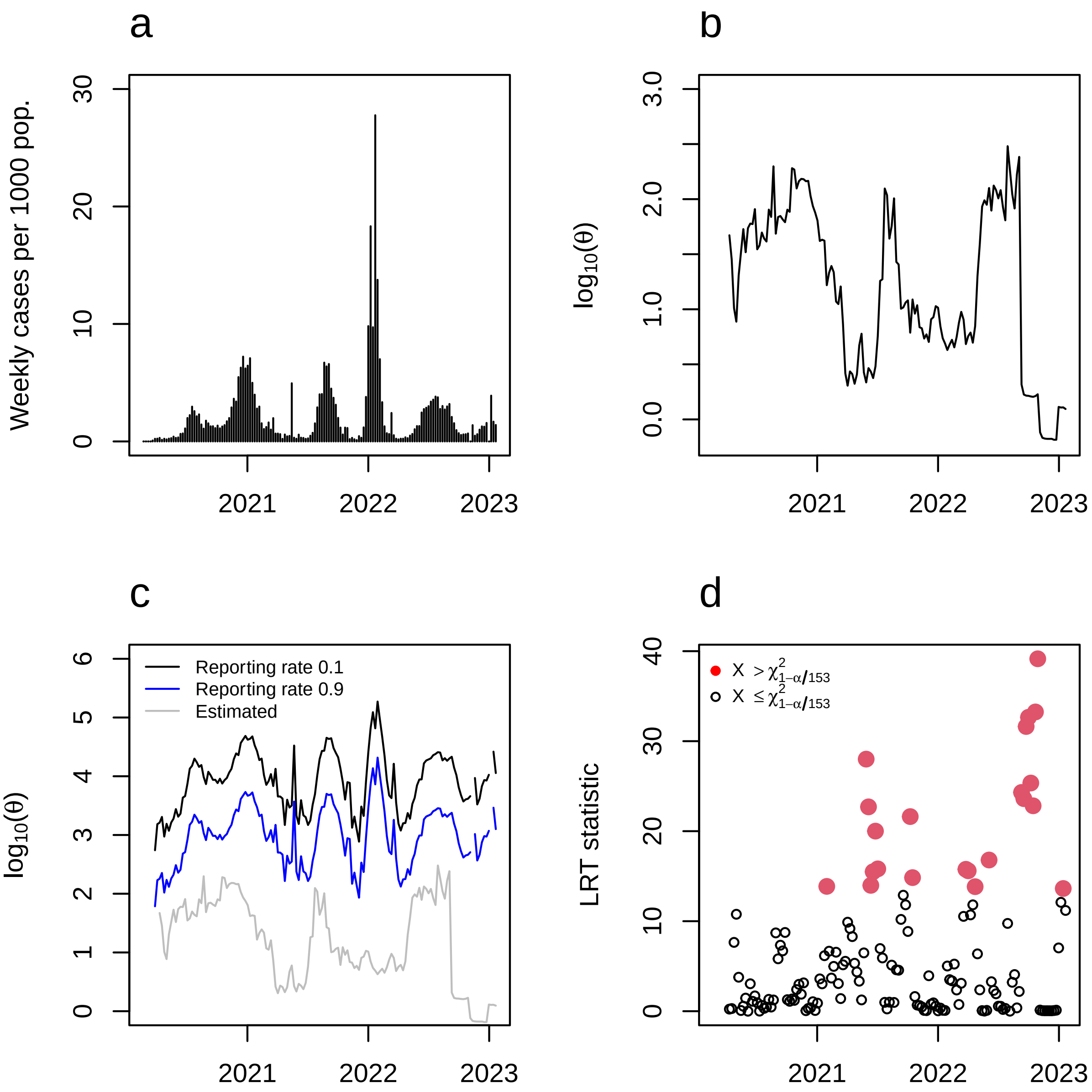

Figure 2: Dispersion analysis of weekly COVID-19 case data for one county. a: Weekly incidence in Jefferson County, Alabama. b: Estimated dispersion parameter over time. c: Comparison of estimated dispersion (grey) with predicted values under the standard model of $\theta_{t+1} = C_t/\rho_t$ where C_t represents reported cases and ρ_t the reporting rate, for constant values of $\rho_t = 0.1$ (black) and $\rho_t = 0.9$ (blue). This wide range of constant values is intended to enclose the *theta* values that would be predicted by the standard model under under variable ρ. d: Likelihood ratio test (LRT) statistic over time, highlighting statistically large changes in dispersion (red), defined as p-values less than the Bonferroni-corrected 5% quantile of a chi-square random variable with one degree of freedom.

Acknowledgments

RA work on this project was supported by X and Y. BDD work on this project was supported by the National Science Foundation (NSF) grants (X, Y) and by the David and Lucile Packard Foundation.

References

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] Lloyd-Smith JO. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PloS one. 2007;2(2):e180.

[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] Dalziel BD, Kissler S, Gog JR, Viboud C, Bjørnstad ON, Metcalf CJE, et al. Urbanization and humidity shape the intensity of influenza epidemics in US cities. Science. 2018;362(6410):75–79.

[5] Kirkegaard JB, Sneppen K. Superspreading quantified from bursty epidemic trajectories. Scientific Reports. 2021;11(1):24124.

[6] Sun K, Wang W, Gao L, Wang Y, Luo K, Ren L, et al. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. Science. 2021;371(6526):eabe2424.

[7] Guo Z, Zhao S, Lee SS, Hung CT, Wong NS, Chow TY, et al. A statistical framework for tracking the time-varying superspreading potential of COVID-19 epidemic. Epidemics. 2023;42:100670.

[8] Ko YK, Furuse Y, Otani K, Yamauchi M, Ninomiya K, Saito M, et al. Time-varying overdispersion of SARS-CoV-2 transmission during the periods when different variants of concern were circulating in Japan. Scientific Reports. 2023;13(1):13230.

[9] Kretzschmar ME, Rozhnova G, Bootsma MC, van Boven M, van de Wijgert JH, Bonten MJ. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. The Lancet Public Health. 2020;5(8):e452–e459.

[10] Blumberg S, Lloyd-Smith JO. Inference of R 0 and transmission heterogeneity from the size distribution of stuttering chains. PLoS computational biology. 2013;9(5):e1002993.

[11] 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. Plos one. 2023;18(5):e0286012.

[12] Graham M, Winter AK, Ferrari M, Grenfell B, Moss WJ, Azman AS, et al. Measles and the canonical path to elimination. Science. 2019;364(6440):584–587.

[13] Wallinga J. Metropolitan versus small-town influenza. Science. 2018;362(6410):29–30.

[14] Kendall DG. Stochastic processes and population growth. Journal of the Royal Statistical Society Series B (Methodological). 1949;11(2):230–282.

[15] Grenfell BT, Bjørnstad ON, Finkenstädt BF. Dynamics of measles epidemics: scaling noise, determinism, and predictability with the TSIR model. Ecological monographs. 2002;72(2):185–202.

[16] Barron DN. The analysis of count data: Overdispersion and autocorrelation. Sociological methodology. 1992:179–220.

[17] Lloyd M. Mean crowding'. The Journal of Animal Ecology. 1967:1–30.

[18] Bjørnstad ON, Finkenstädt BF, Grenfell BT. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. Ecological monographs. 2002;72(2):169–184.

[19] Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. BMC Medical Research Methodology. 2019;19:1–16.

[20] Di Y, Schafer D, Cumbie J, Chang J. NBPSeq: Negative Binomial Models for RNA-Sequencing Data. R package version 03 0, URL http://CRAN R-project. 2015.

[21] Times TNY. Coronavirus (Covid-19) Data in the United States. GitHub Repository. 2021. Accessed: July 11, 2021. Available from: https://github.com/nytimes/covid-19-data.

[22] U S Census Bureau. Annual County Resident Population Estimates: 2020-2021. US Census Bureau Datasets. 2021. Accessed: July 11, 2021. Available from: https://www2.census.gov/programs-surveys/popest/datasets/2020-2021/counties/totals/.