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

Rachael Aber[1,2], Yanming Di[2], Benjamin 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

* aberr@oregonstate.edu

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 transmission in complex populations, as transmission heterogeneity, shifting landscapes of susceptibility, and fluctuations in reporting can directly impact the volatility of observed case counts. However, measures of incidence variability–and how they change over time–are often overlooked in population-level analyses of incidence data. Here we present a statistical framework to quantify temporal changes in incidence dispersion and detect discrete shifts in the dispersion parameter, which may signal new epidemic phases. We apply the method to COVID-19 incidence data in 144 US counties from the January 1st, 2020 to March 23rd, 2023. While standard theory predicts that dispersion should be inversely proportional to incidence (i.e., that the dispersion parameter should be equal to the incidence at the previous time step), our method reveals pronounced temporal trends in dispersion that are not explained by incidence alone, but which are replicated across counties. 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 may play causal 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 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 health responses. However, traditional metrics used to quantify incidence patterns often overlook variability as an important characteristic of incidence time series. Quantifying higher-frequency variation around incidence trends can elucidate important underlying processes, including transmission heterogeneity or changes in the susceptibility landscape. We developed a statistical framework to quantify temporal changes in incidence dispersion within a single time series. We used negative binomial regression to account for factors that might confound the estimation of dispersion, allowing us to use a single parameter to infer the degree of incidence dispersion. We applied the method to

COVID-19 incidence data from 144 US counties from January 1st, 2020 to March 23rd, 2023, and 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 that some individuals spread SARS-CoV-2 (the virus that causes COVID-19) to many more people than others, that the susceptibility landscape changed, or that there were changes in reporting. Shifts in dispersion can indicate shifts in epidemic phase as well as provide information about important underlying biological processes. Therefore, our method to estimate dispersion 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 fluctuations) around trends at the broader temporal ranges typical for epidemic curves (e.g., months or years). Short-term fluctuations in incidence time series are often caused in part by variable reporting, but also reflect the population-level impacts of transmission heterogeneity, and/or changes in susceptiblility [1–5]. Metrics of variability in incidence time series could therefore carry information regarding underlying drivers of transmission, and offer a relatively unexplored avenue for understanding epidemic dynamics.

"Metrics of variability" refers to estimates of the extent to which incidence is dispersed relative to its expected value. The expectation could be provided by a moving average of past incidence, a process-based model, or another type of population-level model. While the expectation itself will vary over time, there may also be important changes in the level of dispersion around the deterministic skeleton of the epidemic. Individual-level heterogeneity in transmission scales up to affect population-level dynamics [1], so variability in epidemic trajectories at the population level may provide information about individual-level variability in the transmission process (transmission heterogeneity).

Contact tracing data has revealed temporal variation in transmission heterogeneity, which is quantified using the dispersion parameter of the offspring distribution. Working from contact tracing data, researchers found that the dispersion parameter of the offspring distribution varied during different phases of the COVID-19 epidemic in Hong Kong [4], which is important for understanding and managing the population-level dynamics of disease transmission and outbreak control. Similarly, the dispersion parameter of the offspring distribution varied over time as new variants emerged [5]. However, traditional contact tracing is very resource intensive, and although new approaches using digital technologies may improve its speed and availability [6], there is a need for 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 was required for SARS-CoV-2 control [3].

Statistical techniques that make inference based on the variability of epidemic time series at the population level are emerging. Working from population-level incidence data, the offspring distribution dispersion parameter was recovered [2]. Another area of interest is how incidence variability is related to different phases of an epidemic. For example, the mean and interannual coefficient of variation of measles incidence were

used to construct a metric indicative of where a location may be on the path to elimination of a pathogen [7]. Analyzing variability in terms of bursts of incidence is also important for planning surge capacity in public health systems [8]. Dispersion is a way to specifically capture clustering of cases in epidemic time series. An index of effective aggregate dispersion (EffDI) was proposed to elucidate clusters of infection directly from incidence data [9]. In the current work, our interest is in the dispersion of epidemic time series.

Incidence dispersion dynamics are distinct from, but related to, the dispersion parameter of the offspring distribution. Infectious disease transmission follows a branching process, as new infections result from exposure to individuals who are currently infectious. Under a wide range of configurations for a branching process model, including the vast majority of parsimonious ones, the number of infected individuals I_t at time t will have a negative binomial distribution [10]: $I_t \sim NB(\mu_t, \theta_t)$, where μ_t is the expected value for I_t and θ_t is the dispersion parameter. The dispersion parameter is related to the variance of incidence by $\mathrm{Var}[I_t] = \mu_t + \mu_t^2/\theta_t$. Importantly, 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. This model is common in population biology (e.g., in the study of measles [11]).

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 [12]. 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 the healthcare system, $\theta = 1$ 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.

Modeling μ_t and θ_t is important to the prediction of epidemic trajectories. A time-series approximation to compartment epidemic models uses $\mu_t = \lambda I_{t-1}$ where λ is the current local rate at which infections at time $t-1$ produces new infections at time t and $\theta_t = I_{t-1}$.

$$I_t = NB(\mu = \lambda I_{t-1}, \theta_t = I_{t-1}) \tag{1}$$

The assertion that $\theta_t = I_{t-1}$ comes from the assumption that each of the individuals who acquired the infection at $t-1$ form "independent lineages" with identically distributed "local" rate parameter. Under these conditions, $\theta_t = I_{t-1}$ is a good approximation for sufficiently large populations ([10], [13]). However, this model assumes that subpopulations of infectious individuals emanating from one infected individuals are independent and statistically identical to those from another. That is, the model assumes susceptible depletion in one lineage does not affect another, transmission rates are equal across lineages, and reporting rates do not vary across lineages.

In practice, the above assumptions may not hold, and thus we instead estimate θ_t to understand how heterogeneity in transmission, susceptibility and reporting (and other factors) are driving epidemic dynamics. Instead of assuming $\theta_t = I_{t-1}$, drivers of θ_t can be investigated, first by quantifying how it changes over time and to what degree it is governed by I_t. We adopt a more general/phenomenological model for μ_t. Instead of $\mu_t = \lambda I_{t-1}$, which is a special kind of moving window, we use a natural spline in time to model the mean, which is able to approximate a wider range of autoregressive and other models.

Furthermore, the rate of the case-count-generating process is often changing within a time step, which cannot be accommodated by a Poisson model. The negative binomial distribution may 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 [14]. 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 [15]. This is an additional reason that the dispersion parameter of a negative binomial distribution should be used to measure changes in meaningful case count clustering in settings with differing base population/incidence.

We develop a method that quantifies the evolution of dispersion along incidence time series, allowing for the detection of changes in clustering that are not due to changes in population size or overall burden of incidence. We apply the method to COVID-19 *weekly* new cases (to avoid irregularities in daily data reporting, such as when all weekend cases are reported on Monday) in large US counties to investigate the relationships between incidence, dispersion and epidemic dynamic regimes over a portion of the COVID-19 pandemic.

Materials and methods

Introduction to the method

Classical approaches [11] model incidence at a time step using a negative binomial variable with expectation equal to the epidemic intensity and dispersion parameter equal to previous incidence. However, other processes besides the current number infected might affect dispersion, so we instead investigate changes in dispersion over time to understand processes that may leave a signal in dispersion.

Our general framework is that incidence at a time step is drawn from a negative binomial distribution with time-varying mean and dispersion parameter (the dispersion parameter varies more slowly than the mean parameter). The model is formulated with a linear predictor that includes a natural spline in time with three degrees of freedom to account for autocorrelation in case counts. Natural splines are cubic splines which are linear outside of the boundary knots [16]. A recently proposed negative binomial regression model for time series of counts also accommodates serial dependence [17].

A persistent challenge in investigating changes in variability has been "spurious correlation" with population size. Since population size influences the mean and variance of case count data and thus could have an impact on estimates of dispersion, we adjust for population size using an offset in the model. There is an offset term in order to directly model counts (here, COVID-19 cases) per unit of observation (here, per individual).

$$\log(E[Y_i]/n_i) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i) \tag{2}$$

$$\log(E[Y_i]) - \log(n_i) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i) \tag{3}$$

$$\log(E[Y_i]) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i) + log(n_i) \tag{4}$$

Thus, the form of the probability mass function for incidence at a time step is:

$$f_t(I) = \binom{I + \theta - 1}{I} \left(\frac{\mu}{\mu + \theta}\right)^I \left(\frac{\theta}{\mu + \theta}\right)^\theta \tag{5}$$

where μ for the time step is estimated via the linear predictor outlined above. The form of the expectation and variance from this model is below.

$$E(I) = \mu \tag{6}$$

$$Var(I) = \mu + \frac{\mu^2}{\theta} \tag{7}$$

Note that high variability/dispersion corresponds to low values of the dispersion parameter, θ. In addition to fitting the model at each time step, we developed a straightforward likelihood-ratio test (LRT) that can be applied at each time step. This involves fitting both a null model (no θ change) and a two-part (θ-change) model. In sum, our method identifies shifts in population-level dispersion in incidence while accounting for population size.

Application to simulated data

To test the validity and power of the LRT, we simulated both Gaussian and uniform epidemic curves with an attack rate (proportion of the population who become infected) of 0.1. Epidemic curves spanning 60 time steps each were produced, and the likelihood-ratio test (LRT) procedure was applied to each at its known change point in θ. Varying the magnitude of the θ change, location of the change in the curve, population size underlying the curve, and curve shape (as mentioned above) allowed us to test the validity and power of our approach across a range of situations.

Application to empirical data

We estimated μ_t and θ_t using iteratively reweighted least-squares (procedure implemented via the NBPSeq R package [18]) using a window around each time step. For each window, a series of μ_t was estimated using a spline function in time, and a single value of θ was estimated for that window. By moving the window one time step at a time, a time series of θ_t was produced. We investigated large counties (the three counties with the largest populations in each state), due to power constraints and convergence issues.

Results

We found that the LRT method is robust across population sizes (for population sizes included in the empirical data) (Fig. 1 e, f). The criteria for adequate test performance are that the average p-value is 0.5 when the effect size is zero, and lower average p-values are observed with increasing effect sizes.

The negative binomial framework for estimating θ is also robust across the population sizes examined, and a large portion of inaccurate estimates correspond to counties with small population sizes outside of the range considered (S1 Fig). In row one and two of Fig. 1, we illustrate that an increase in θ is associated with decreased dispersion in simulated case count time series. This pattern is observed in the empirical data, and is independent of whether incidence is high or low, as seen in Fig. 2, where an example time series of $log_{10}\theta$ estimated from empirical data in Jefferson County, AL shows decreasing values around periods in which cases increase in volatility. For example, consider the period of time around the beginning of 2022, in which high case count variability coincides with low values of $log_{10}\theta$. Additionally, the dispersion parameter changed as new variability regimes began ($log_{10}\theta$ values colored red in Fig. 2 b). There were departures from the dispersion parameter under the
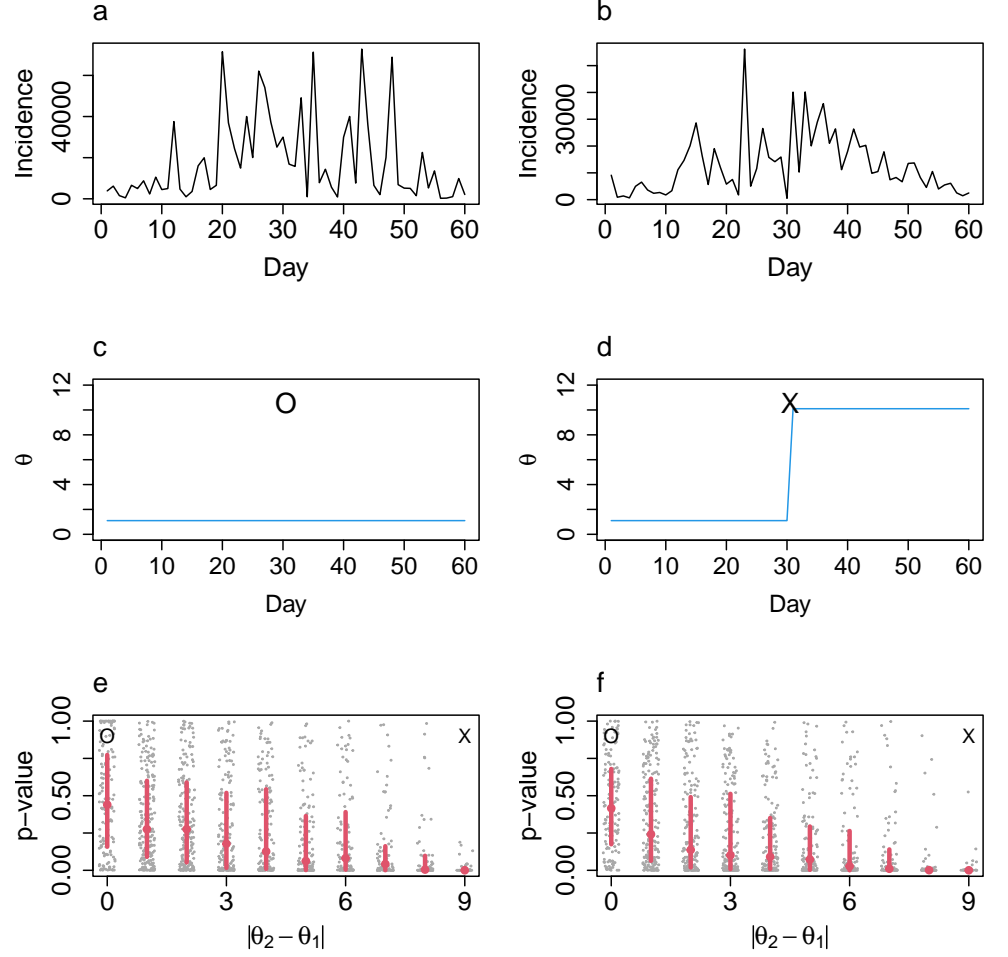
Fig 1. Detecting dispersion changes in incidence time series in populations of different sizes. a: Simulated case counts when dispersion is constant. b: Simulated case counts when dispersion changes. c: Constant dispersion parameter used in generation of the above. d: Changing dispersion parameter used in generation of the above. e, f: Performance of the LRT with simulated data that have different absolute differences in θ (horizontal axis of each panel) illustrates p-value distribution in population size of 50,000 (e) and population size of 10,000,000 (f), which represent the range in the empirical data. O and X mark the null and alternative hypotheses indicated in panels c and d. Red vertical lines represent the interquartile range.

model in Equation 1, especially around the beginning of 2022 and the end of the time frame examined (Fig. 2 b). The dispersion parameter under the model in Equation 1 can be plotted by taking a series of previous incidence and adding an offset to account for the reporting rate ($log10(\theta_t)$ in the model = $log10(I_{t-1})$ = $log10(cases_{t-1}/reportingrate) = log10(cases_{t-1}) - log10(reportingrate)$).

Highly overdispersed incidence patterns were observed for many of the examined counties more frequently later in time series, consistent with more heterogeneity in transmission, susceptibility and reporting. The most dispersed category in Fig. 3 (e) reaches its highest proportion near the end of the time frame examined. In addition, there are increases in dispersion around the peaks in incidence in the dataset (Fig. 3 c,
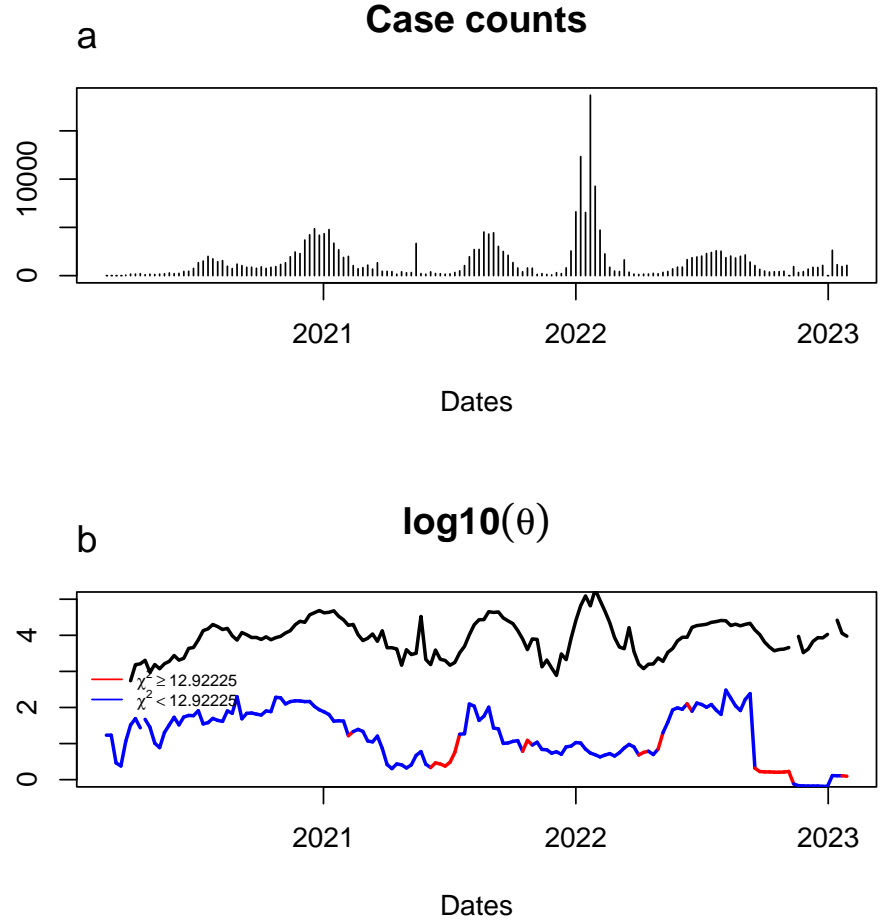
Fig 2. Negative binomial regression/LRT method applied to case counts between 2020-01-04 and 2023-03-18 from Jefferson County, AL. a: Case counts. b: Time series of $log_{10}\theta$ colored by whether the associated LRT statistic is greater than the Bonferroni-corrected level 0.05 quantile of a chi-square random variable with one degree of freedom (red indicates areas of putative change in θ). The black line represents the dispersion parameter under the model from Equation 1, assuming a reporting rate of 0.10.

d). Evidence for a change in θ was observed across many counties (evidenced by a concentration of low p-values around peak incidence) (Fig. 3 f). Note that these p-values should be corrected for multiple testing when used for inference instead of visualization.

Raising variance relative to mean (higher dispersion) implies spatiotemporal "crowding" of cases (i.e. localized surges) which may necessitate more surge capacity in hospitals and testing centers. Therefore, it may be the case that there were more surges on the way up to or on the way down from peak incidence. It appears that there are more surges on the way up to, and on the way down from, peak incidence. This indicates less diffuse epidemic dynamics that are potentially more subject to climate forcing [19], and increased locally experienced mean density [12].
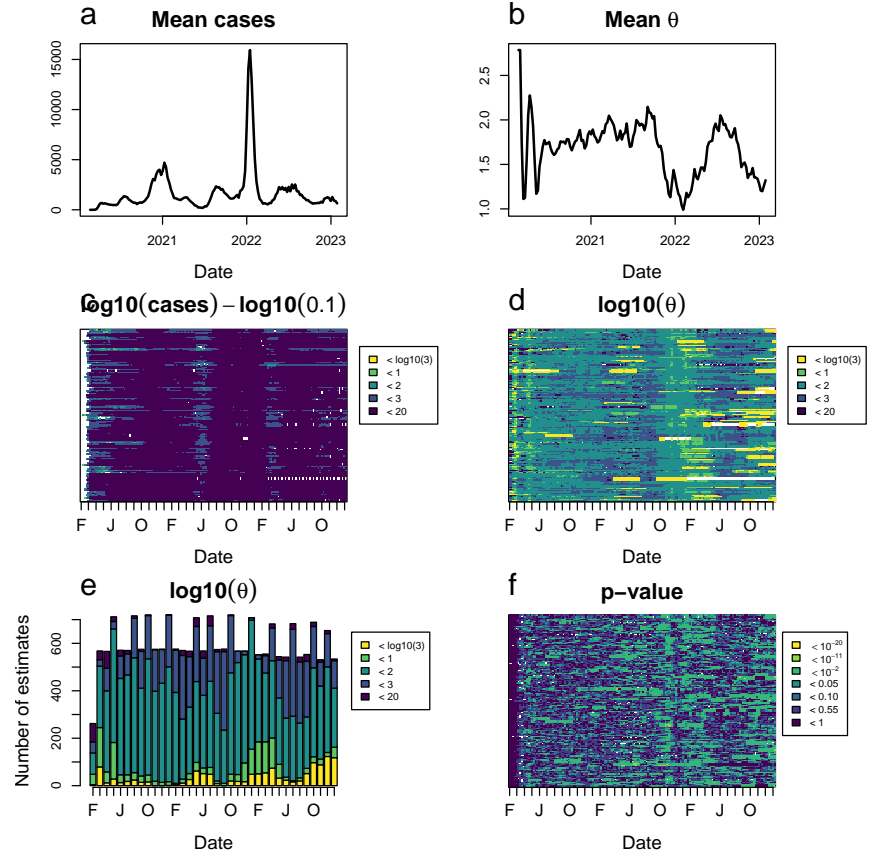
Fig 3. Incidence and dispersion between 2020-01-04 and 2023-03-18 in large counties in the US. a: Mean COVID-19 cases of the 144 US counties over time. b: Mean θ of the 144 US counties over time. c: log10(case counts) - log10(reporting rate) over time for each of the large counties (y-axis). This represents $log_{10}\theta$ under the model in Equation 1. d: $log_{10}\theta$ over time for each of the large counties (y-axis). e: Binned $log_{10}\theta$ estimates for all counties over time. f: LRT p-values over time for each of the large counties (y-axis).

Discussion

We developed an approach to quantify clustering of cases in epidemic time series that is less susceptible to detecting artifacts based on population size and incidence. Our method forms part of a larger initiative to investigate variability in incidence as an important attribute of epidemic time series using novel metrics. For instance, burst-tree decomposition of time series has also facilitated computation of a burst-size distribution for a series given a specified time window [20], allowing comparison of variability within one location over time. Spatial variation in superspreading potential has been investigated through risk maps of superspreading environments [21], so future work could investigate the correspondence between our dispersion metric and indicators of a high risk of superspreading.

Methods that use incidence time series are crucial due to the ease of obtaining this type of data, so the timing and geographical allocation of public health resources can be achieved with limited resources. Additionally, population-wide disease control approaches are often less effective than those which are targeted to individuals in high-transmission contexts [1], so identifying candidate time periods when transmission

heterogeneity is high may catalyze the development of more efficient control strategies. 192

Application of the methods we developed to empirical case count time series 193
spanning from the beginning of 2020 to early 2023 revealed distinct increases in 194
dispersion both near the end of the time frame considered and near peaks in 195
COVID-19 incidence. These results imply that we can revise our understanding of 196
case count dispersion: dispersion is high *near* peak incidence, suggesting that 197
dispersion is not simply governed by incidence. These departures from the level of 198
dispersion expected under the model represented in Equation 1 represent time periods 199
in which the assumptions may be violated and cases may be highly clustered in time. 200

Though large cities may be subject to more "smooth" epidemic dynamics, our 201
contribution highlights the circumstances under which dynamics are less smooth in 202
large populations. 203

Previous research to evaluate bursty dynamics based on Influenza-like Illness (ILI) 204
times series showed that epidemics in smaller communities were concentrated on 205
narrower windows of the influenza season; the proportion of disease incidence that 206
occurred in a given week was a metric of interest [19]. So, additional research is 207
needed to understand the relationship between burstiness in small communities and 208
temporally changing dispersion in these areas. 209

Since there are regimes of the COVID-19 epidemic that are subject to increased 210
dispersion across large counties, these may be candidate time periods for rolling out 211
extra surge capacity in large counties. But, as mentioned above, further research is 212
needed to understand whether this phenomenon occurs in small and mid-sized 213
counties. 214

Conclusion 215

We presented an approach to quantify clustering of cases in epidemic time series that 216
does not detect artifacts based on population size and incidence. While our estimation 217
framework facilitates comparison of dispersion between counties and also over time in 218
a given county, our LRT framework additionally allows for detection of changing 219
dispersion. Application of these methods to empirical incidence time series spanning 220
from the beginning of 2020 to early 2023 revealed distinct increases in dispersion both 221
near the end of the time window and near peaks in COVID-19 incidence. These 222
findings will assist in the allocation of public health resources, especially in the 223
planning of surge capacity. Since there are regimes of the COVID-19 epidemic that are 224
subject to increased dispersion across large counties, these may be candidate time 225
periods for rolling out extra capacity. However, further research is needed to 226
understand this phenomenon in small counties, as well as for other pathogens. 227

Supporting information 228

S1 Fig.   Estimates of the dispersion parameter from simulated data. Points 229
colored red for underlying population sizes less than 50,000. 230

Acknowledgments 231

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. doi:10.1038/nature04153.
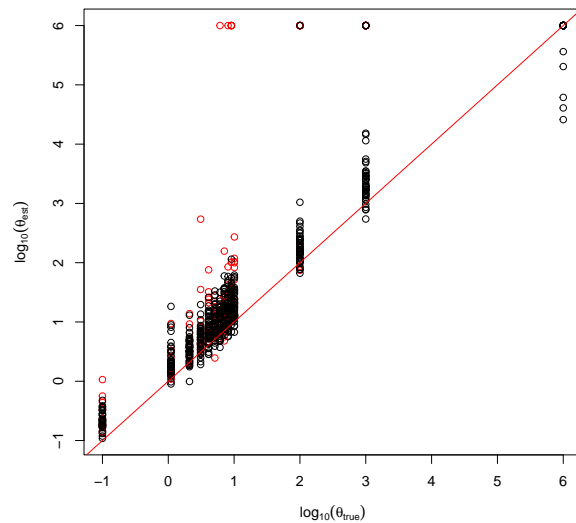
2. Kirkegaard JB, Sneppen K. Superspreading quantified from bursty epidemic trajectories. Scientific Reports. 2021;11(1):24124. doi:10.1038/s41598-021-03126-w.

3. 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. doi:10.1126/science.abe2424.

4. 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.

5. 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.

6. Kretzschmar ME, Rozhnova G, Bootsma MCJ, Van Boven M, Van De Wijgert JHHM, Bonten MJM. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. The Lancet Public Health. 2020;5(8):e452–e459. doi:10.1016/S2468-2667(20)30157-2.

7. 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. doi:10.1126/science.aau6299.

8. Wallinga J. Metropolitan versus small-town influenza. Science. 2018;doi:10.1126/science.aav1003.

9. 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. doi:10.1371/journal.pone.0286012.

10. Kendall DG. Stochastic Processes and Population Growth. Journal of the Royal Statistical Society: Series B (Methodological). 1949;11(2):230–264. doi:10.1111/j.2517-6161.1949.tb00032.x.

11. 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. doi:10.2307/3100024.

12. Lloyd M. 'Mean Crowding'. The Journal of Animal Ecology. 1967;36(1):1. doi:10.2307/3012.

13. BJøRNSTAD ON, Finkensta RF, Grenfell BT. DYNAMICS OF MEASLES EPIDEMICS: ESTIMATING SCALING OF TRANSMISSION RATES USING A TIME SERIES SIR MODEL;72(2).

14. Cook JD. Notes on the Negative Binomial Distribution. NoJournal; p. 5.

15. Barron DN. The Analysis of Count Data: Overdispersion and Autocorrelation. Sociological Methodology. 1992;22:179–220. doi:10.2307/270996.

16. Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. BMC Medical Research Methodology. 2019;19(1):46. doi:10.1186/s12874-019-0666-3.

17. Davis RA, Wu R. A negative binomial model for time series of counts. Biometrika. 2009;96(3):735–749. doi:10.1093/biomet/asp029.

18. Yanming Di DWS. NBPSeq: Negative Binomial Models for RNA-Sequencing Data; 2022. Available from: `https://CRAN.R-project.org/package=NBPSeq`.

19. Dalziel BD, Kissler S, Gog JR, Viboud C, Bjørnstad ON, Metcalf CJE, et al. Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. Science. 2018;362(6410):75–79. doi:10.1126/science.aat6030.

20. Jo HH, Hiraoka T, Kivelä M. Burst-tree decomposition of time series reveals the structure of temporal correlations. Scientific Reports. 2020;10(1):12202. doi:10.1038/s41598-020-68157-1.

21. Loo BPY, Tsoi KH, Wong PPY, Lai PC. Identification of superspreading environment under COVID-19 through human mobility data. Scientific Reports. 2021;11(1):4699. doi:10.1038/s41598-021-84089-w.