

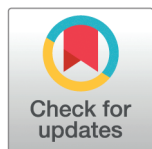
RESEARCH ARTICLE

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

Rachael Aber^{1,2,3*}, Yanming Di², Benjamin D. Dalziel^{1,4}

1 Department of Integrative Biology, Oregon State University, Corvallis, Oregon, United States of America, **2** Department of Statistics, Oregon State University, Corvallis, Oregon, United States of America, **3** Data Sciences, Exponent, Inc., Bellevue, Washington, United States of America, **4** Department of Mathematics, Oregon State University, Corvallis, Oregon, United States of America

* raber@exponent.com



OPEN ACCESS

Citation: Aber R, Di Y, Dalziel BD (2025)

Time-series modeling of epidemics in complex populations: Detecting changes in incidence volatility over time. PLoS Comput Biol 21(7): e1012882.

<https://doi.org/10.1371/journal.pcbi.1012882>

Editor: Edward M. Hill, University of Liverpool, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: February 16, 2025

Accepted: May 29, 2025

Published: July 11, 2025

Copyright: © 2025 Aber et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The code used for the analysis reported in this paper can be accessed from

<https://github.com/rachaelaber/dispersion>. The data analyzed here are originally from The New York Times and The United States Census Bureau, accessed via

<https://github.com/nytimes/covid-19-data/> and <https://www2.census.gov/programs-surveys/popest/datasets/2020-2021/counties/totals/>, respectively.

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 to 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 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 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.

Author summary

Quantifying patterns in infectious disease incidence is crucial for understanding epidemic dynamics and for developing effective public health policy. Traditional

Funding: RA was supported in part by the Achievement Rewards for College Scientists (ARCS) Foundation Oregon through an ARCS Scholar Program award and in part by the Association for Computing Machinery (ACM) SIGHPC Computational and Data Science Fellowship. BDD was supported in part by a grant from the National Science Foundation under Award Number CBET-2200338, and by grants from the David and Lucile Packard Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

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 dispersion in time series of case counts and applied the method to COVID-19 case count data from U.S. counties. We found large shifts in incidence volatility (week-to-week variability in the numbers of new cases) that were synchronized across counties, and were not explained by broader-scale changes in mean incidence over time. Incidence 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 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 caused in part by variable reporting, but may also reflect the population-level impacts of transmission heterogeneity and changes in the landscape of susceptibility [1–8]. 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 estimate 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 of the offspring distribution 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 focused more intensely (proportionally more infections in a smaller span of time) with implications for the sensitivity of cities to climate forcing and for surge capacity in the health system [4,13].

What drives incidence 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], $I_t \sim NB(\mu_t, \theta_t)$, 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 $\text{Var}[I_t] = \mu_t + \mu_t^2/\theta_t$, 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 derivation in S1 Text). 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) as 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

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

where time-varying reproductive number R_t gives the expected number of secondary infections acquired from an infected individual 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 θ becomes $\theta_t = C_t/\rho_t$ where C_t represents reported cases and ρ_t 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 assumptions will not often hold, and our aim in this paper is to develop, test and apply an alternative approach which produces 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 modeled incidence at broad spatiotemporal scales using natural splines [19]. To allow for diverse shapes in the broadscale epidemic dynamics, spline modeling was conducted within a moving window such that for each half of the window

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

where μ_t is the mean of the negative binomial process at time t , N represents population size, $h_j(t)$ are basis functions, the degrees of freedom is equal to the number of knots k for the natural spline, $J = k + d + 1$, where d is the degree of the polynomial, and $\beta_j^{(t)}$ are fitted parameters. The window has half-width Δ , centered at t , i.e., extending from $t - \Delta$ to $t + \Delta$. The degrees of freedom (number of knots) to be used for the splines, and the width of the moving window will depend on the application. Explanation of the specific choices we used for our application to COVID-19 cases in US counties is provided 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,

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

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

$$f_t(I_t) = \binom{I_t + \theta_t - 1}{I_t} \left(\frac{\mu_t}{\mu_t + \theta_t}\right)^{I_t} \left(\frac{\theta_t}{\mu_t + \theta_t}\right)^{\theta_t} \quad (4)$$

where μ_t 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 function 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 function 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 1C).

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_t) was based on the smoothed incidence trend across the 144 counties 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 of θ_2 to $\theta_1 > 3$) 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 January 1, 2020, and March 23, 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 (the 144 counties used in the analysis) 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 case counts 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 for each half

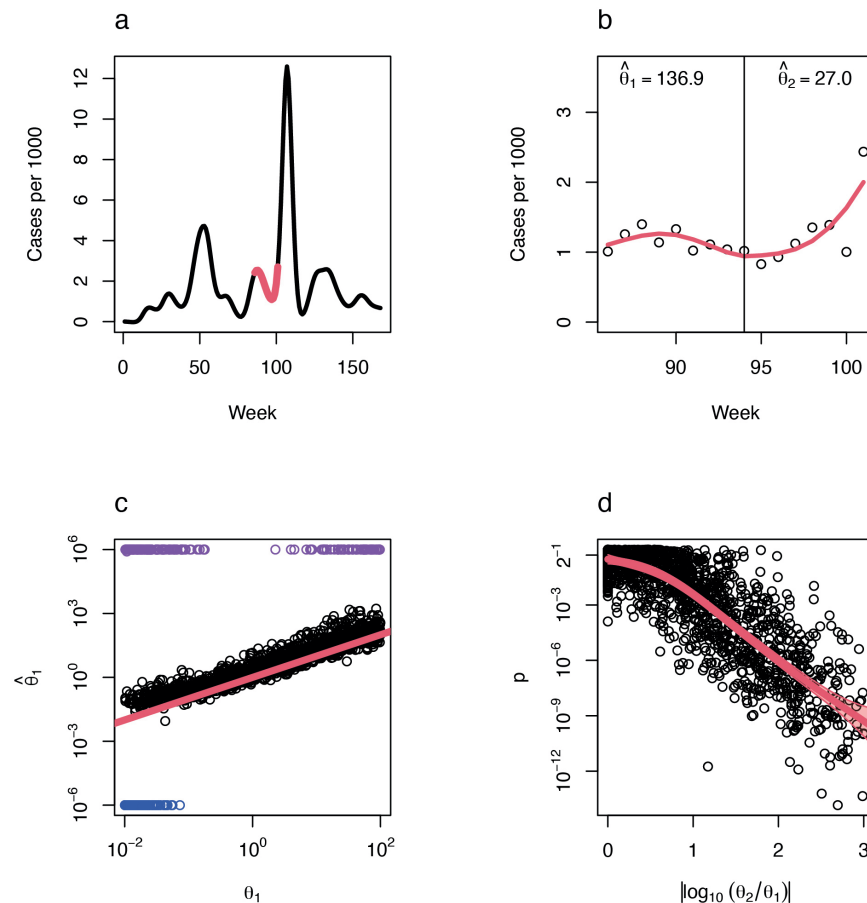


Fig 1. Detecting dispersion changes in case count time series. a: Weekly incidence of COVID-19 in the United States, with time measured in weeks since January 4, 2020, showing an example of a randomly-selected 16-week period used as an incidence trend in simulation-based validation of the LRT test (red). b: Cases 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_1 versus true θ_1 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), and a line with an intercept of zero and a slope of one plotted in red. d: Statistical power of the LRT test with smooth function (red line) and a 99.7% confidence interval for predicted p (red shading).

<https://doi.org/10.1371/journal.pcbi.1012882.g001>

of the window 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.

Selection of window width and spline degrees of freedom

Choices of window width and degrees of freedom for the natural splines were made by comparing the accuracy of θ estimates using simulated data over various values of both window width and degrees of freedom. Using 16-week windows and three degrees of freedom for the splines, our method did not systematically overestimate or underestimate true dispersion in this application to COVID-19 weekly case count data.

Results

Simulations indicate that the LRT framework accurately detects changes in dispersion, with p -values converging to 0.5 as the effect size approaches 0, reflecting the uniform distribution of p -values under the null hypothesis, and decreasing toward zero as the effect size increases (Fig 1D). The framework is also robust to the range of population sizes present in the empirical data—county populations ranged from approximately 48 thousand to 9.9 million, and we tested the framework on simulated populations between 10 thousand and 10 million. Across this range, the method produced accurate estimates for θ within $10^{-2} \leq \theta \leq 10^2$ (Fig 1C), encompassing all operationally relevant values for COVID-19 incidence data and many other infectious diseases. Lower values would concentrate the probability mass function (PMF) for cases almost entirely on 0, while higher values effectively correspond to a Poisson distribution.

Applying the method to COVID-19 cases in US counties enabled investigation of changes in dispersion (Fig 1B) around the overall epidemic trajectory (Fig 1A). Periods of increased case count variability (for example, around the start of 2023 in Fig 2A) corresponded with decreases in θ (see the corresponding time period in Fig 2B), indicating that dispersion was dynamic. Changes in dispersion exhibited both expected and unexpected patterns of variation relative to standard theory (S1 Fig). In some instances, θ_t varied inversely with incidence I_t , consistent with standard epidemic theory, while in other periods, deviations from this expectation occurred, potentially signaling shifts in underlying transmission dynamics.

Notably, significant changes in the dispersion parameter were observed during major epidemic transitions. For example, during the beginning of 2022, and at the end of the time series, when the pandemic was transitioning toward endemicity as the landscape of susceptibility was evolving [23]. The landscape of susceptibility was evolving as a larger proportion of cases involved reinfections. These findings underscore the complex behavior of the dispersion parameter, which not only varied with changes in case count regimes but also revealed departures from the model expectations described by Eq (1), which are consistent with changes in the underlying drivers of transmission.

As mentioned in the introduction, reporting rate and case count can be used to arrive at θ_t used in the model of incidence in Eq 1. Fig 2C displays two θ_t time series with differing assumed reporting rates, alongside the θ_t time series estimated directly from the case counts using our method. When the reporting rate is assumed to be higher at a time point, a lower θ_t would be used in the Eq 1 model.

Dispersion increased markedly around the peak in incidence during the major 2022 wave, from late December 2021 to early February 2022 (Fig 3A and 3B display averages over counties and Fig 3C and 3D display results across counties). This is in strong contrast to standard epidemic theory which predicts that dispersion should decrease as incidence rises, and suggests that the assumptions underlying the model in Eq 1 do not hold during these time periods. One potential explanation for this is that transmission rates may not be independent across lineages. For example, one high-transmission lineage may spur another as the pathogen spreads through more complex landscapes of susceptibility and transmission risk during later pandemic waves. A high concentration of low p -values around peak incidence (Fig 3F) corroborates widespread changes in θ across counties, reinforcing the statistical significance of this pattern. While these p -values should be corrected for multiple testing if used for inference rather than visualization, the overall trend suggests a systematic departure from theoretical expectations. Highly overdispersed patterns were also observed more frequently later in the time series (Fig 3D), pointing to increasing heterogeneity in transmission,

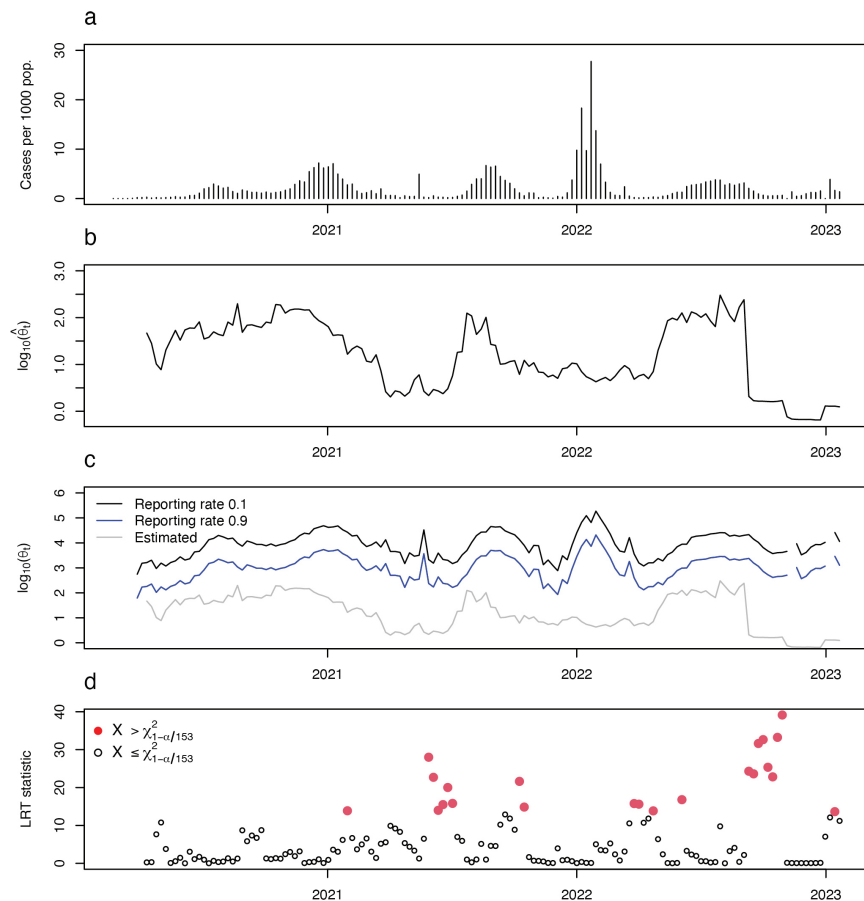


Fig 2. Dispersion analysis of weekly COVID-19 case data for Jefferson County, Alabama. Results for all counties are shown in Fig 3. a: Weekly reported COVID-19 incidence. b: Estimated dispersion parameter ($\hat{\theta}_t$) over time. c: Comparison of estimated dispersion (gray) with predicted values from the standard model $\hat{\theta}_{t+1} = C_t/\rho_t$, where C_t is reported cases and ρ_t the reporting rate at time t . Predictions are shown for fixed $\rho_t = 0.1$ (black) and $\rho_t = 0.9$ (blue), chosen to encompass the range of θ expected under variable ρ . d: Likelihood ratio test (LRT) statistic over time. Statistically significant changes in dispersion (red) correspond to p-values below the Bonferroni-corrected 5% threshold of a chi-square distribution with one degree of freedom.

<https://doi.org/10.1371/journal.pcbi.1012882.g002>

susceptibility, and reporting during the later phases of the pandemic. In both the 2022 wave and later in the pandemic, localized surges indicated by higher dispersion may have played a larger role in pandemic dynamics than expected, including potentially placing increased demand for surge capacity on hospitals and testing centers.

Discussion

Our method forms part of a larger interest in investigating variability in infections 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 [24], allowing comparison of variability within one location over time.

Spatial variation in superspreading potential has been investigated through risk maps of superspreading environments [25], and future work could investigate the correspondence

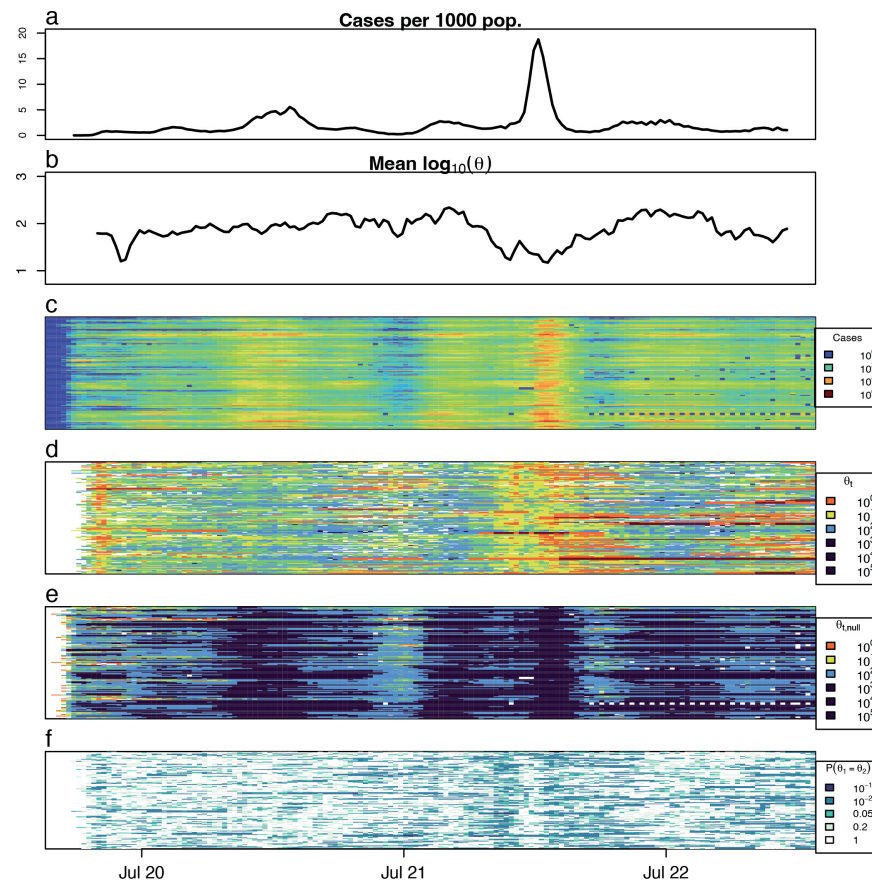


Fig 3. Incidence and dispersion between Jan 4, 2020 and March 18, 2023, in large counties in the US. a: Mean COVID-19 cases of the 144 US counties over time (total cases over the counties divided by total population over the counties multiplied by 1,000). b: Mean $\log_{10} \theta_t$ of the 144 US counties over time. NAs produced by the method (see text) were removed from the average. c: $\log_{10}(\text{cases}_t)$ over time for each of the 144 counties, where county is the y-axis. d: $\log_{10} \theta_t$ over time for each of the 144 counties. e: Expected value of $\log_{10} \theta_t$ under the null model, assuming a reporting rate of 0.5 for each county. f: LRT p-values over time for each county.

<https://doi.org/10.1371/journal.pcbi.1012882.g003>

between dispersion in case count time series, as quantified here, and indicators of a high risk of superspreading, with the potential to further elucidate drivers of transmission risk across scales, and more finely resolve landscapes of susceptibility. Additionally, as population-wide disease control efforts may be less effective than those which are focused to individuals in high-transmission contexts [1], identifying candidate time periods when transmission heterogeneity is high may catalyze the development of more effective control strategies, particularly those that connect vulnerable populations with resources at critical times.

The finding that dispersion increased rather than decreased during the 2022 surge challenges theoretical expectations and suggests that fundamental assumptions about the scaling of transmission dynamics may require reevaluation. One hypothesis is that transmission heterogeneity could play a role in driving large surges, amplifying incidence beyond what homogeneous models predict. In particular, our finding of departures from theoretical expectations of case count dispersion indicates violations of the assumptions underlying the commonly used time-series epidemic model shown in Eq 1: lineages are likely dependent—that is, one high-transmission lineage may spur another. It was found that

stratification of contacts across multiple dimensions prevents underestimation of R_0 [26] and evolving contact structures have also been discussed in the literature [6], both of which could result in the observed departures. Stratification of the transmission network and its evolution is synonymous with individual-level heterogeneity in transmission and its evolution, which 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). More specifically, it was also recently demonstrated that increasing the number of strata over which populations are organized increases R_0 , or doesn't affect R_0 if the additional stratification is "random" (random mixing hypothesis) [26].

Future work could investigate how bursts of highly clustered transmission events could generate feedback that accelerates epidemic spread, which, if true, could refine predictive models of contagion dynamics in complex populations. For instance, contact-tracing effort could be directed towards the candidate time periods, both to perform confirmatory analyses of transmission heterogeneity, and to increase the availability of an informative public health resource around surges.

A primary limitation of the method presented here is that careful consideration must be given to the choice of parameter values based on the user's specific application. Specifically, the choice of window half-width must be informed by simulation of data and estimation of known θ . Similarly, the choice of spline degrees of freedom must optimize the accuracy of θ estimates. In our case, we are additionally limited by course (weekly) data.

Our results highlight some of the limitations of theoretical assumptions about the role of incidence dispersion in epidemic dynamics, such as during peak incidence periods of the COVID-19 pandemic. Models that incorporate time-varying incidence dispersion can be used to quantify the role of transmission heterogeneity in epidemic dynamics in complex populations, and can help practitioners to identify candidate time periods and locations for confirmatory analysis of superspreading, as well as for public health intervention. Our findings suggest the importance of considering nonindependent transmission rates across lineages when modeling epidemics in complex populations.

Supporting information

S1 Text. Derivation of the relationship between the dispersion parameter and the mean crowding parameter.

(PDF)

S1 Fig. Spearman's correlation between the dispersion parameter and COVID-19 cases. A 32-week sliding window is used to compute each correlation, and 1,000 bootstrap replicates in which county labels are permuted are used to compute the 95% confidence interval.

(PDF)

Author contributions

Conceptualization: Rachael Aber, Yanming Di, Benjamin D. Dalziel.

Data curation: Rachael Aber, Benjamin D. Dalziel.

Formal analysis: Rachael Aber, Benjamin D. Dalziel.

Funding acquisition: Rachael Aber, Benjamin D. Dalziel.

Methodology: Rachael Aber, Yanming Di, Benjamin D. Dalziel.

Project administration: Rachael Aber.

Software: Rachael Aber, Yanming Di, Benjamin D. Dalziel.

Supervision: Rachael Aber.

Validation: Rachael Aber, Benjamin D. Dalziel.

Visualization: Rachael Aber, Benjamin D. Dalziel.

Writing – original draft: Rachael Aber, Benjamin D. Dalziel.

Writing – review & editing: Rachael Aber, Yanming Di, Benjamin D. Dalziel.

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):355–9. <https://doi.org/10.1038/nature04153> PMID: 16292310
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. <https://doi.org/10.1371/journal.pone.0000180> PMID: 17299582
3. Lau MSY, 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. *Proc Natl Acad Sci U S A*. 2017;114(9):2337–42. <https://doi.org/10.1073/pnas.1614595114> PMID: 28193880
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 U.S. cities. *Science*. 2018;362(6410):75–9. <https://doi.org/10.1126/science.aat6030> PMID: 30287659
5. Kirkegaard JB, Sneppen K. Superspreading quantified from bursty epidemic trajectories. *Sci Rep*. 2021;11(1):24124. <https://doi.org/10.1038/s41598-021-03126-w> PMID: 34916534
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. <https://doi.org/10.1126/science.abe2424> PMID: 33234698
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. <https://doi.org/10.1016/j.epidem.2023.100670> PMID: 36709540
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. *Sci Rep*. 2023;13(1):13230. <https://doi.org/10.1038/s41598-023-38007-x> PMID: 37580339
9. 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. *Lancet Public Health*. 2020;5(8):e452–9. [https://doi.org/10.1016/S2468-2667\(20\)30157-2](https://doi.org/10.1016/S2468-2667(20)30157-2) PMID: 32682487
10. Blumberg S, Lloyd-Smith JO. Inference of $R(0)$ and transmission heterogeneity from the size distribution of stuttering chains. *PLoS Comput Biol*. 2013;9(5):e1002993. <https://doi.org/10.1371/journal.pcbi.1002993> PMID: 23658504
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. <https://doi.org/10.1371/journal.pone.0286012> PMID: 37253038
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):584–7. <https://doi.org/10.1126/science.aau6299> PMID: 31073065
13. Wallinga J. Metropolitan versus small-town influenza. *Science*. 2018;362(6410):29–30. <https://doi.org/10.1126/science.aav1003> PMID: 30287649
14. Kendall DG. Stochastic processes and population growth. *J Roy Statist Soc Ser B: Statist Methodol*. 1949;11(2):230–64. <https://doi.org/10.1111/j.2517-6161.1949.tb00032.x>
15. Grenfell BT, Bjørnstad ON, Finkenstädt BF. Dynamics of measles epidemics: scaling noise, determinism, and predictability with the TSIR model. *Ecol Monogr*. 2002;72(2):185–202. [https://doi.org/10.1890/0012-9615\(2002\)072\[0185\]:domesn2.0.co;2](https://doi.org/10.1890/0012-9615(2002)072[0185]:domesn2.0.co;2)
16. Barron DN. The analysis of count data: overdispersion and autocorrelation. *Sociol Methodol*. 1992;22:179. <https://doi.org/10.2307/270996>

17. Lloyd M. Mean crowding. *J Anim Ecol*. 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. *Ecol Monogr*. 2002;72(2):169–84. [https://doi.org/10.1890/0012-9615\(2002\)072\[0169\]:domees\]2.0.co;2](https://doi.org/10.1890/0012-9615(2002)072[0169]:domees]2.0.co;2)
19. Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. *BMC Med Res Methodol*. 2019;19:1–16.
20. Di Y, Schafer D, Cumbie J, Chang J. NBPSeg: negative binomial models for RNA-sequencing data; 2015
21. Times T. Coronavirus (Covid-19) data in the United States. 2021. [cited 2021 July 11]. <https://github.com/nytimes/covid-19-data>
22. U S Census Bureau. Annual County Resident Population Estimates: 2020-2021. US Census Bureau Datasets. 2021. [cited 2021 July 11]. <https://www2.census.gov/programs-surveys/popest/datasets/2020-2021/counties/totals/>
23. Lavine JS, Bjornstad ON, Antia R. Immunological characteristics govern the transition of COVID-19 to endemicity. *Science*. 2021;371(6530):741–5. <https://doi.org/10.1126/science.abe6522> PMID: 33436525
24. Jo H-H, Hiraoka T, Kivelä M. Burst-tree decomposition of time series reveals the structure of temporal correlations. *Sci Rep*. 2020;10(1):12202. <https://doi.org/10.1038/s41598-020-68157-1> PMID: 32699282
25. Loo BPY, Tsoi KH, Wong PPY, Lai PC. Identification of superspreading environment under COVID-19 through human mobility data. *Sci Rep*. 2021;11(1):4699. <https://doi.org/10.1038/s41598-021-84089-w> PMID: 33633273
26. Manna A, Dall'Amico L, Tizzoni M, Karsai M, Perra N. Generalized contact matrices allow integrating socioeconomic variables into epidemic models. *Sci Adv*. 2024;10(41):eadk4606. <https://doi.org/10.1126/sciadv.adk4606> PMID: 39392883