Detecting changes in dispersion in COVID-19 incidence time series using a negative binomial model

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

* aberr@oregonstate.edu

Abstract

Metrics of variability are often overlooked and useful ways to understand epidemic dynamics. For instance, superspreading of SARS-CoV-2 (the virus that causes COVID-19) can be elucidated by utilizing such metrics. Our method identifies shifts in population-level dispersion (a measure of case clustering) in COVID-19 cases, allowing a more complete and predictive understanding at both the individual and population level, and allowing practitioners to prepare surge capacity at certain points in an epidemic. Although classical theory predicts that there will be less dispersion (less clustering) when incidence is higher, we considered a more general negative binomial regression framework to investigate processes that may also affect the spread of cases. We investigated changes in dispersion and found that there are increases in dispersion around peaks in incidence in many US counties. In addition, highly overdispersed patterns occur more frequently later in time series, consistent with more heterogeneity in transmission, susceptibility, and reporting. Our method is robust to differences in population size and incidence, allowing for quantification of dispersion-potentially indicative of superspreading dynamics-without artifactual contributions from other features.

Author summary

Understanding disease spread is crucial for managing epidemics, but traditional metrics often overlook the variability in how viruses like SARS-CoV-2 (the virus that causes COVID-19) spread. We developed a method to identify shifts in the degree of dispersion (case clustering) patterns of SARS-CoV-2 within a single time series. By examining spatiotemporal differences in how the virus spreads, we can better predict and prepare for surges in cases. We used negative binomial regression to account for factors that might confound the estimation of case count dispersion, allowing us to use a single parameter to infer the degree of dispersion (clustering) of cases. We found that around incidence peaks, the spread of SARS-CoV-2 became more erratic, with some individuals potentially spreading the virus to many more people than others. Additionally, as the pandemic progressed, the spread of the virus also became more erratic, suggesting increased differences in factors such as how people transmit the virus or how susceptible they are to the virus. Our method accurately measures disperison regardless of population size and incidence, providing a way to understand

case count clustering across a range of locations. This can help public health officials better anticipate and manage outbreaks, especially during times when the virus spreads unpredictably.

Introduction

Time series of observed infectious disease incidence are, to varying degrees, "noisy", showing higher frequency oscillations around trends at broader temporal scales. Highly variable incidence that characterizes noisy data arise from imperfect and variable reporting (i.e, measurement error), but also suggests transmission heterogeneity (superspreading), demographic/environmental heterogeneity, or changes in population effective reproduction number (R). Therefore, variability can contain information important to understanding epidemic dynamics and societal responses. Yet metrics of variability are often overlooked ways to understand these dynamics, and techniques based on variability in epidemic time series are still emerging. One area of interest is how variability is related to different phases of an epidemic. For instance, the mean and interannual coefficient of variation of measles incidence was used to construct a metric indicative of where a location may be on the path to elimination of a pathogen [1]. It may be possible to use other variability metrics to determine what dynamic regime of an epidemic is taking place. Additionally, it was recently found that the time-varying transmission heterogeneity for COVID-19 decreased over time and was significantly associated with interventions to slow spread in Hong Kong [2]. Individual-level variability in transmission is most often studied using contact-tracing data. However, contact tracing data requires intense investment of resources [3], so analysis of incidence data may often be more feasible. Therefore, it is desirable to detect signatures of a change in transmission heterogeneity using case count time series. Specifically, individual-level heterogeneity in transmission scales up to affect population-level dynamics [4], so variability in epidemic trajectories at the population level may provide information about individual-level variability in the transmission process. In sum, variability in population-level incidence time series may therefore provide information about what phase or dynamic regime an epidemic is in, as well as potentially indicating the level of heterogeneity at finer spatial and temporal scales, in transmission, susceptibility, reporting and/or that resulting from environmental/demographic stochasticity. To that end, an index of effective aggregate dispersion (EffDI) was proposed to elucidate clusters of infection directly from incidence data [5]. Analyzing variability in terms of bursts of incidence is also important for planning surge capacity [6]. Sun et al. [7] found a combination of individual-based and population-based strategies was required for SARS-CoV-2 control, further highlighting the importance of considering population-level variability and its relationship to individual-level variability. One potentially useful variability metric in the context of case count time series is dispersion. The dispersion of a case count time series forms part of a negative binomial model used to model case counts at a given time step. The negative binomial model provides a way to more flexibly model the variance as a function of the mean, as opposed to the equality of mean and variance implied by the Poisson distribution. Importantly, dispersion is related to a previously proposed "mean crowding" parameter, which is the mean number per individual of other individuals in the same quadrat [8]. If we instead think of mean crowding from the perspective of an infectious individual in an epidemic system, incidence experienced in their local spatial-temporal neighborhood will be higher than the global mean when dispersion is high. It is useful to consider dispersion in case count time series, as (absent other processes that influence dispersion such as variation in reporting) one can think of degree of dispersion as degree of clustering/crowding of

cases, which is related to concepts like transmission heterogeneity. One of the reasons why studying variability in incidence time series is not more widely done is because it is difficult to disentangle the effects of population size/incidence on variance. Since mean case counts is directly related to population size/incidence, we would infer that variance in case counts is large in large population/incidence settings by default. Furthermore, the rate of the case-count-generating process is often changing within a timestep, 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 timestep: for example, if the mean of a Poisson distribution itself follows a gamma distribution, the resulting distribution is negative binomial [9]. The result is that we should use the dispersion parameter of a negative binomial distribution to measure changes in meaningful case count clustering in settings with differing base population/incidence, not the variance. 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 timestep that all lead to overdispersion [10]. We developed 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 applied the method to COVID-19 incidence data in 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 theory put forward by Grenfell et al. [11] proposed that incidence at a time step can be modeled by a negative binomial variable with expectation equal to the epidemic intensity and dispersion parameter equal to previous incidence:

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

However, other processes besides the current number infected might affect dispersion, so we instead investigated changes over time to understand important processes that may leave a signal in dispersion. As mentioned above, a persistent challenge in investigating changes in variability has been "spurious correlation" with population size. Since population size influences mean and variance in count data and thus could have an impact on estimates of dispersion, we robustly adjusted for population size using an offset in the model. In sum, our method identifies shifts in population-level dispersion in incidence while accounting for population size. The general framework is that incidence at a time step is drawn from a negative binomial distribution with time-varying mean and dispersion parameters (that vary more slowly than the mean). 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 [12]. A recently proposed negative binomial regression model for time series of counts also accommodates serial dependence [13]. There is an offset term in order to directly model counts (here, COVID-19 cases) per unit of observation (here, per individual):

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

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

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

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

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

Here, μ for the time step is estimated via the linear predictor outlined above. This form has expectation and variance as follows:

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

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

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 could be applied at each time step. This involves fitting both a null model (no θ change) and a two-part (θ change) model.

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 of 0.1 – epidemic curves over 60 timesteps each were produced, and a likelihood-ratio test (LRT) procedure was applied to each. 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 one θ using iterative reweighted least-squares (procedure implemented via the NBPSeq R package [14] and from Di et al. [15]) using a window around each time step. For each window, μ_t was estimated using a spline function in time, and the single value of θ was estimated for the window. By moving the window one time step at a time, a time series for θ_t was produced. We investigated large counties (largest three counties in each state), due to power constraints.

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 low average p-values are observed with increasing effect size. The negative binomial framework for estimating θ is also robust across the population sizes examined (see S1 Fig). In row one and two of Fig. 1, we illustrated that an increase in θ is associated with decreased variability in simulated incidence time series. This pattern is observed in the empirical data, and is independent of whether incidence increases or decreases, as seen in Fig. 2(a, c).

Highly overdispersed incidence patterns were observed for many counties more frequently later in time series, consistent with more heterogeneity in transmission, susceptibility and reporting. The most dispersed category in Fig.3 (a) reaches its highest proportion near the end of the timeframe examined. In addition, there are increases in dispersion around the peaks in incidence in the dataset (Fig. 3 b, c). The

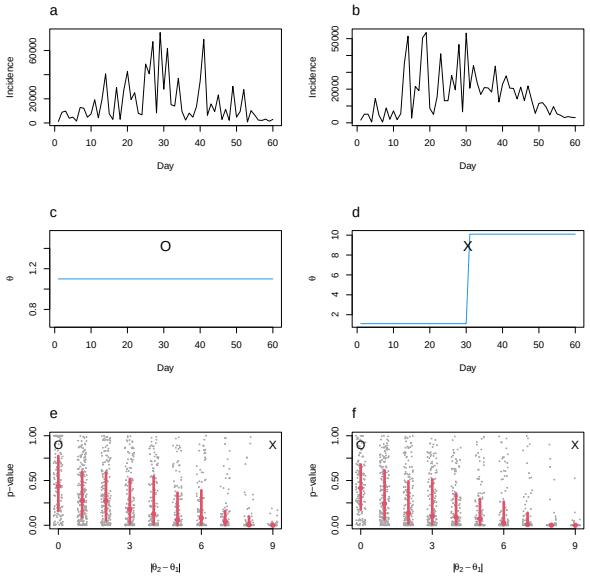


Fig 1. Detecting dispersion changes in incidence time series in populations of different sizes. A: Simulated incidence when dispersion is constant. B: When dispersion changes during the epidemic. C: Constant dispersion used in generation of above. D: Changing dispersion used in generation of above. E: Performance of the method with simulated data that has different absolute differences in theta (horizontal axis of each pane) illustrates p-value distribution across different population sizes (each pane is one population size). O and X mark the null and alternative hypotheses indicated in panels C and D.

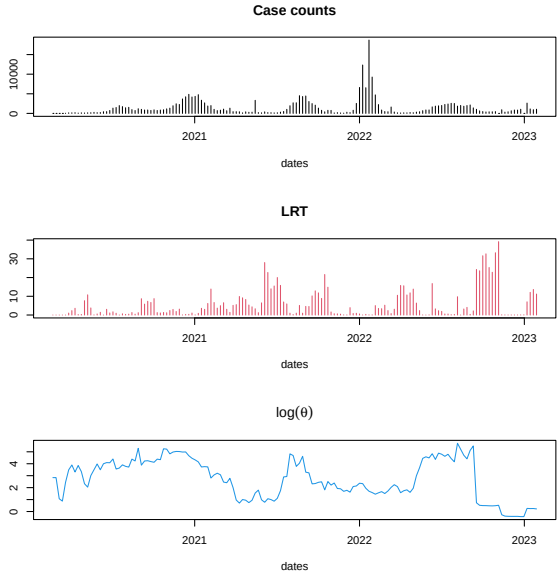


Fig 2. Method applied to case counts between 2020-01-04 and 2023-03-18 for Jefferson County, AL. A: Case counts . B: LRT statistic C: Log dispersion parameter.

evidence for a change in θ was observed across many counties (evidenced by a concentration of low p-values around peak incidence) (Fig. 3 d).

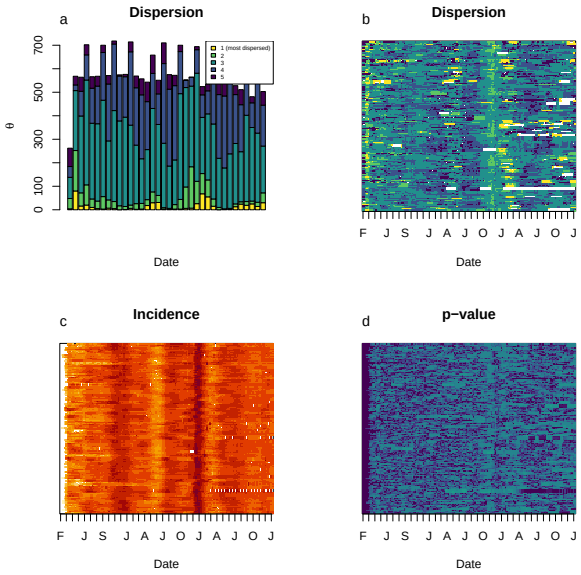


Fig 3. Incidence and dispersion between 2020-01-04 and 2023-03-18 in large counties in the US. A: Binned log of the dispersion parameter over time. B: Log of the dispersion parameter over time as well as for each of the large counties (y-axis). C: Log incidence (new cases per individual) over time as well as for each of the large counties (y-axis). D: LRT p-values over time as well as for each of the large counties (y-axis).

Raising variance relative to mean 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 are more surges on the way up to or on the way down from peak incidence. Additionally, it may indicate less diffuse epidemics that are potentially more subject to climate forcing [16], or increased locally experienced mean density [8].

Discussion

We presented an approach to quantify clustering of cases in epidemic time series that does not detect artifacts based on population size and incidence. Our method forms part of a larger push to investigate variability in incidence time series 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 [17], allowing comparison of variability within one location over time. Spatial variation in superspreading potential has been investigated through e.g., risk maps of superspreading environments [18], so future work could investigate the correspondence between our variability metric and indicators of high risk of superspreading. Methods that use incidence time series are crucial part due to the ease of obtaining incidence data, so the timing/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 [4], so models that incorporate transmission heterogeneity may catalyze the development of more efficient control strategies. Our results imply that we can revise our understanding of case count dispersion: dispersion/crowding is high near peak incidence, suggesting

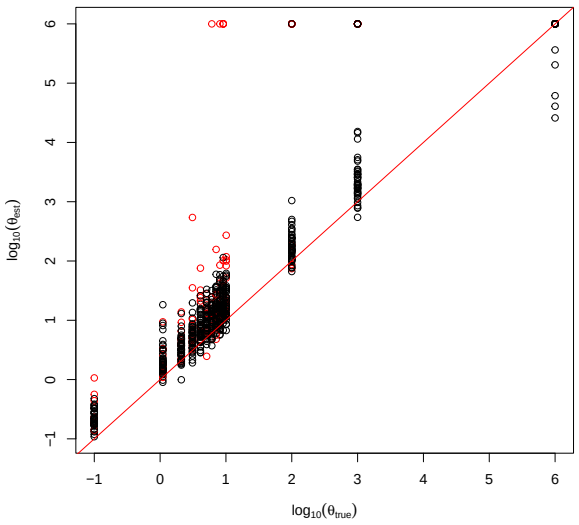
that low dispersion is not simply governed by high incidence. Though large cities may be subject to more "smooth" epidemic dynamics, our contribution highlights the circumstances under which dynamics are less smooth in large counties. Previous research to evaluate bursty dynamics based on Influenza-like Illness (ILI) times series showed that epidemics in smaller communities are concentrated on narrower windows of the influenza season - the proportion of disease incidence that occurred in a given week was a metric of interest [16]. So, additional research is needed to understand the correspondence between burstiness of small communities and the potential impact of temporally changing dispersion in these areas.

Conclusion

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

Supporting information

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



Acknowledgments

177

References

1. 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.

2. Adam D, Gostic K, Tsang T, Wu P, Lim WW, Yeung A, et al. Time-varying transmission heterogeneity of SARS and COVID-19 in Hong Kong; 2022. Available from: `https://www.researchsquare.com/article/rs-1407962/v1`.

3. 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.

4. 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.

5. 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.

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

7. 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.

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

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

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

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. 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.

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

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

15. Yanming D, W SD, S CJ, H CJ. The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. Statistical Applications in Genetics and Molecular Biology. 2011;10(1):1–28.

16. 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.

17. 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.

18. 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.