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 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 in certain months. Although classical theory predicts that there will be less dispersion when incidence is higher, we considered a more general negative binomial regression framework to account for processes that may also affect the spread of cases. We investigated changes in dispersion and found that there are increases in dispersion around holiday periods in many US counties, concurrent with incidence increases. 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-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 diseases like COVID-19 spread. We developed a method to identify shifts in the spread patterns of SARS-CoV-2 (the virus that causes COVID-19) 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 influence the spread of cases, allowing us to use a single parameter to infer the degree of clustering of cases. We found that during holiday periods, the spread of SARS-CoV-2 becomes more erratic, with some individuals spreading the virus to many more people than others. As the pandemic progressed, the spread of the virus became erratic, suggesting increased differences in factors such as how people transmit the virus or susceptibility. Our method accurately measures spreading heterogeneity regardless of population size and incidence, providing a way to understand superspreading 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 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]. Similarly, 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]. 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 [3]. Analyzing variability in terms of bursts of incidence is also important for planning surge capacity [4]. Sun et al. [5] 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.

Dispersion around a rolling mean (moving window) of an incidence time series may contain information about the size and frequency of local outbreaks. Dispersion may be a more useful metric than variance because the same value of the dispersion parameter will result in a larger variance/mean ratio when the mean is larger, thus capturing clustering in the true sense, and not just reflecting the statistical artifact that large mean results in large variance. From the perspective of an infectious individual, incidence experienced in their local spatial-temporal neighborhood will be higher than the global mean when dispersion is high. A 'mean crowding' parameter was proposed, which is the mean number per individual of other individuals in the same quadrat [6]. This is a useful way to think about 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. Specifically, individual-level heterogeneity in transmission scales up to affect population-level dynamics [7], so dispersion in epidemic trajectories at the population level may provide information about individual-level variability in the transmission process. Individual-level variability in transmission is most often studied using contact-tracing data. However, contact tracing data requires intense investment of resources [8], so analysis of incidence data may often be more feasible.

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. For instance, if we model a process as Poisson, we assume that if the mean of the process is large, the variance will be large as well. 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 in an

epidemic, 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: 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 variability in settings with differing base population/incidence, not the variance of a Poisson model. In distributions that form part of the exponential family, the dispersion parameter determines the mean-variance relationship.

Negative binomial regression (in contrast to Poisson regression) can account for unobserved heterogeneity, time dependence in the rate of a process and contagion 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 variability that are not due to changes in population size or overall burden of incidence. We apply 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. From a modern theoretical ecology perspective, investigating beyond the first moment of a process has also been identified as important: ecological experiments are typically geared towards assessing the impacts of the mean strength of causal processes, however the variance about mean effects have been mostly ignored as a driver in biological assemblages, but may be as important as the mean [11].

Materials and methods

Introduction to the method

Classical theory put forward by Grenfell et al. [12] proposed that incidence can be modeled by a negative binomial variable with expectation equal to the epidemic intensity and dispersion 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 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 [13]. A recently proposed negative binomial regression model for time series of counts also accommodates serial dependence [14]. 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 is:

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

This form has expectation and variance as follows:

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

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

Application to simulated data

For validity/power simulations, we used both Gaussian and uniform epidemic curves with an attack rate of 0.1 in the simulated set of time series–3,000 epidemic curves over 60 days each were produced, and a likelihood-ratio test (LRT) procedure was applied to each. Varying the effect size, location of the breakpoint, population size, and curve shape allowed us to test the validity and power of our approach S1 Appendix.

Application to empirical data

To evaluate whether the negative binomial conditional distribution is needed (opposed to a Poisson/quasi-Poisson conditional distribution with the same model of process mean), inspection of mean-variance relationships using a diagnostic plot is possible [15], along with evaluation of dispersion statistics. We estimated μ_t and θ using iterative reweighted least-squares (procedure implemented via the NBPSeq R package [16] and from Di et al. [17]) with a moving window approach. For each window, μ_t was estimated using a spline function in time, and a 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 (top 4% in each state), due to power constraints.

Results

We found that the negative binomial/LRT method is robust to differences in population size (for population sizes of at least 10,000) (Fig. 1). For an analytical derivation, see the Supplemental Materials S1 Appendix. 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. In row one and two of Fig. 1, we illustrated that a drop in θ is associated with increased variability in simulated incidence time series, and that the same relationship is observable in the empirical time series, with an increase in θ corresponding to a decrease in variability around the trend in incidence.

Highly overdispersed incidence patterns were observed more frequently later in time series, consistent with more heterogeneity in transmission, susceptibility and reporting. Interestingly, the most dispersed category in Fig.2 (a) reaches its highest proportion near the end of the timeframe examined. In addition, there are increases in dispersion around the holiday periods in the dataset (Fig. 2.(b)), concurrent with increases in incidence (Fig. 2(c)). The evidence for a change in θ was observed across many counties (evidenced by concentration of low p-values concurrent with peak incidence) (Fig. 2 (d)).

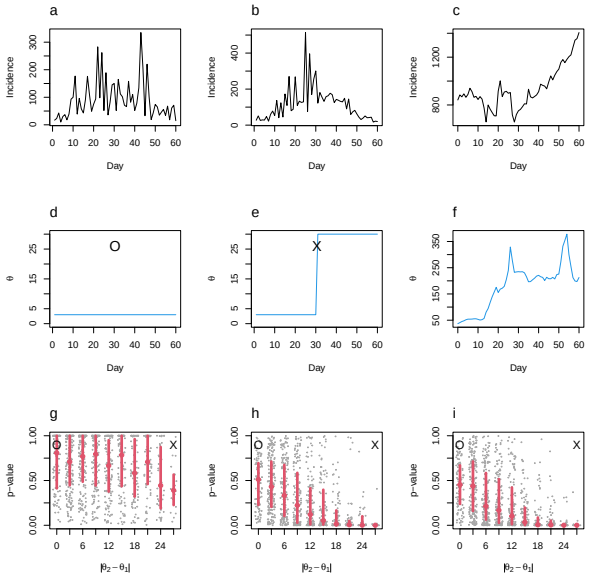

Fig 1. Detecting dispersion changes in incidence time series in populations of different sizes. A: Simulated incidence when dispersion is constant. B: When dispersion changes during the epidemic. C: Daily COVID-19 cases in Jefferson County, AL between 2020-09-10 and 2020-11-09 (X-axis is days since beginning of inclusion period for the model fit). D: Constant dispersion used in generation of above. E: Changing dispersion used in generation of above. F: Dispersion estimates from model fit to the above series. G: Performance of the method with simulated data that has different absolute differences in theta (horizontal axis of each pane) illustrates similar 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 D and E.

The occurrence of high dispersion at times of peak incidence is of interest because it has more impact on variance/mean relationships than when incidence is lower. In other words, what makes a change in dispersion meaningful is how it affects variance-mean relationships. For instance, if dispersion is high in a high incidence setting, the variance-mean ratio would be larger than for the same dispersion in a smaller incidence setting. A change from $\theta = 1000$ to $\theta = 100$ is operationally significant for large populations during times of peak incidence due to this variance-mean scaling:

$$var = \mu + \mu^2/\theta \tag{8}$$

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. Additionally, it may indicate less diffuse epidemics that are potentially more subject to climate forcing [18], or increased locally experienced mean density [6]. Additionally, θ seems to behave like a state variable. That is, its rate of change is dependent on its own value and the values of other state variables (like incidence), which determine the "state" of the system (Fig. 3). State variables depend only on the current state and not on the history or the process, so they allow for the formulation of predictable relationships, such as those found in the laws of thermodynamics. This predictability and simplicity are crucial for both theoretical studies and practical applications, and creating a reliable description of this (epidemic) system is overdue.

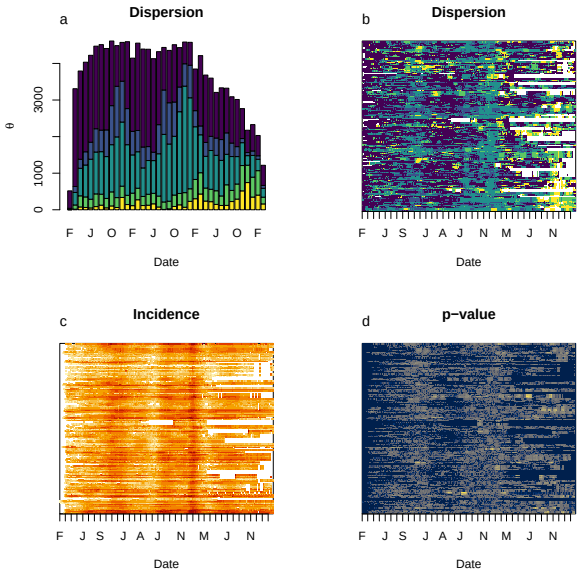

Fig 2. Incidence and dispersion between 2020-02-20 and 2023-03-19 in large counties in the US. A: Binned log of the dispersion parameter over time. B: Log of the dispersion parameter over time as well as for each of the large counties (y-axis). C: Log incidence (new cases per individual) over time as 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).

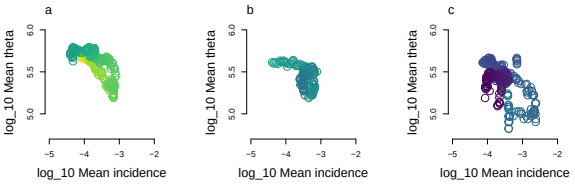

Fig 3. Log of the mean (over a single time series) dispersion parameter v. log of the mean incidence (over a single time series; normalized by population size).

Discussion

We presented an approach to compare dispersion 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 as an important attribute of epidemic time series.

Burst-tree decomposition of time series' has also facilitated computation of a burst-size distribution for a series given a specified time window [19], allowing comparison of variability within one location over time. Similarly, spatial variation in superspreading potential has been investigated through e.g., risk maps of superspreading environments [20]. Methods that use incidence time series are a crucial part of this research area 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 [7], 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 is high at unexpected times (peak incidence) and corresponds to significant increases in variance when incidence is high. 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. In addition, a big city with more (smaller) hospitals could effectively be analogous to a collection of small towns, and therefore experience a benefit of reduced variance-mean relationships, especially during periods of peak incidence. This implies that there may be merit to revising current public health strategies. 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 [18]. So, additional research is needed to understand the balance between burstiness of small communities and the potentially less severe effect of increased dispersion.

Conclusion

CO_2 Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit.

Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat. Ut neque ipsum, luctus id lacus ut, laoreet scelerisque urna. Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed. Nam condimentum sem eget mollis euismod. Nullam dui urna, gravida venenatis dui et, tincidunt sodales ex. Nunc est dui, sodales sed mauris nec, auctor sagittis leo. Aliquam tincidunt, ex in facilisis elementum, libero lectus luctus est, non vulputate nisl augue at dolor. For more information, see S1 Appendix.

Supporting information

S1 Fig. Bold the title sentence. Add descriptive text after the title of the item (optional).

S2 Fig. Lorem ipsum. Analytical approach: robust to population size changes.

S1 File. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod

ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Video. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Appendix. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Table. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

Acknowledgments

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

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. 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.
4. Wallinga J. Metropolitan versus small-town influenza. Science. 2018;doi:10.1126/science.aav1003.
5. 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.
6. Lloyd M. 'Mean Crowding'. The Journal of Animal Ecology. 1967;36(1):1. doi:10.2307/3012.
7. 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.
8. 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.

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. Benedetti-Cecchi L. The Importance of the Variance Around the Mean Effect Size of Ecological Processes. Ecology. 2003;84(9):2335–2346. doi:10.1890/02-8011.

12. 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.

13. 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.

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

15. Ver Hoef JM, Boveng PL. QUASI-POISSON VS. NEGATIVE BINOMIAL REGRESSION: HOW SHOULD WE MODEL OVERDISPERSED COUNT DATA? Ecology. 2007;88(11):2766–2772. doi:10.1890/07-0043.1.

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

17. 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.

18. 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.

19. 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.

20. 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.