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 can be elucidated by utilizing such metrics. Our method identifies shifts in population-level incidence dispersion, 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 consider a more general negative binomial regression framework to account for processes that may also affect the spread of cases. We investigate changes in dispersion and find 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 changes in incidence and to population size, allowing for quantification of dispersion-indicative of superspreading dynamics-without artifactual contributions from these features.

Author summary

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

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, [1] use the mean and interannual coefficient of variation of measles incidence to construct a metric indicative of where a location may be on the path to elimination of the pathogen. 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. Recently, an index of effective aggregate dispersion (EffDI) was proposed to elucidate clusters of infection from incidence data (Schneckenreither et al. 2023). Analyzing variability in terms of bursts of incidence is also important for planning surge capacity (Wallinga 2018). Sun et al. [3] 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. 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 (Lloyd 1967). 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 (Lloyd-Smith 2005), 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 (Kretzschmar et al. 2020), 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. Furthermore, the rate of the case-count-generating process is often changing in an epidemic, which cannot be accomodated 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 (Cook 2009). 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.

$$f(y_i|\theta_i, \phi) = \exp\left(\frac{\theta_i y_i - b(\theta_i)}{a_i(\phi)} - c(y_i, \phi)\right) \quad (1)$$

$$Var(y_i|\phi) = b''(\theta_i)a_i(\phi) \quad (2)$$

Where θ_i is the natural (canonical) parameter and ϕ is the dispersion parameter.

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 (Barron 1992).

We develop 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 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 (Benedetti-Cecchi 2003).

Materials and methods

Introduction to the method

Classical theory put forward by Grenfell et al. (2002) proposed that incidence can be modeled by a negative binomial variable with expectation equal to the epidemic intensity and dispersion equal to previous incidence:

$$\mathrm{I_t} = NB(\mu = \lambda_t, \theta_t = I_{t-1}) \quad (3)$$

However, other processes besides the current number infected might affect dispersion, so we instead investigated changes in to understand important processes that may leave a signal in dispersion. In other words, we estimated over time. 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 as follows:

The linear predictor 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 (Perperoglou et al. 2019). A recently proposed negative binomial regression model for time series of counts also accommodates serial dependence (Davis and Wu 2009). b) There is an offset term in order to directly model counts (here, COVID-19 cases) per unit of observation (here, per individual):

$$log(E[Y_i]/n_i) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i) \quad (4)$$
$$\log(E[Y_i]) - log(n_i) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i) \quad (5)$$
$$log(E[Y_i]) = \beta_1 (h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i) + log(n_i) \quad (6)$$

So, the form of the probability mass function is:

$$xf_t(I) = \binom{I+\theta-1}{I} \frac{\mu}{\mu+\theta}^{I} \frac{\theta}{\mu+\theta}^{\theta} \quad (7)$$

This form has expectation and variance as follows:

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

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 the LRT test 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.

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 (Ver Hoef and Boveng 2007), along with evaluation of dispersion statistics. We then estimated μ_t and θ using iterative reweighted least-squares (procedure implemented via the NBPSeq R package (https://CRAN.R-project.org/package=NBPSeq) and from Di et al. (2011)) 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

Results

We found that the negative binomial method is robust to changes in population size (for population sizes of at least 10,000). For an analytical derivation, see Supplement 1. The criteria 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 Figure 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 reaches its highest proportion near the end of the timeframe (Figure 1.(a)) In addition, there are increases in dispersion around the holiday periods in the dataset (Figure 1.(b)), concurrent with

increases in incidence (Figure 1.(c)). The evidence for a change in θ was observed across many counties (evidenced by concentration of low p-values concurrent with peak incidence) (Figure 1.(d)).

Fig 1. 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.

Fig 2. 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).

The occurrence of high dispersion at times of peak incidence is of interest because it has more impact on variability than when incidence is lower. In other words, what makes a change in dispersion meaningful is how it affects variance and 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:

$$var = \mu + \mu^2/\theta$$

So, a small increase in dispersion could have large impacts on the variance in large populations at times of peak incidence. 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 (Dalziel et al. 2018), or increased locally experienced mean density (Lloyd 1967). To illustrate this using the empirical data, we observed an increase from $\theta = 2.707735e + 16$ to $\theta = 21.8711$, which, conditional on the observed case count of 3, would lead to an increase in variance from around 3 to around 3.4. In contrast, if the observed case count were 300, variance would go from around 300 to around 4415.019.

Additionally, θ seems to behave like a state variable. That is, its rate of change is dependent on its value and the values of other state variables (like incidence), which determine the "state" of the system. 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 3. Fig 3. caption

Table 1. Table caption Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam.

Heading1				Heading2			
*cell*1*row*1	cell2 row 1	cell3 row 1	cell4 row 1	cell5 row 1	cell6 row 1	cell7 row 1	cell8 row 1
*cell*1*row*2	cell2 row 2	cell3 row 2	cell4 row 2	cell5 row 2	cell6 row 2	cell7 row 2	cell8 row 2
*cell*1*row*3	cell2 row 3	cell3 row 3	cell4 row 3	cell5 row 3	cell6 row 3	cell7 row 3	cell8 row 3

Table notes 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.

LOREM and IPSUM nunc blandit a tortor

3rd level heading

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.

1. react
2. diffuse free particles
3. increment time by dt and go to 1

Sed ac quam id nisi malesuada congue

Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

- First bulleted item.
- Second bulleted item.
- Third bulleted item.

Discussion

We presented an approach to compare variability 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 (Jo et al. 2020), 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 (Loo et al. 2021). 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 goal can be achieved with limited resrouces. Additionally, population-wide disease control approaches are often less effective than those which are targeted to individuals in high-transmission contexts

(Lloyd-Smith et al. 2005), 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 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 (Dalziel et al. 2018). So, additional research is needed to understand the balance between burstiness of small communities and the potentially less severe effect of increased dispersion.

Though some kinds of time dependence in the rate can cause autocorrelation, as can contagion (if it occurs outside of set periods), and heterogeneity (if an omitted variable is correlated in time) (Barron 1992), our focus on dispersion makes sense because we are concerned with the clustering of cases from the point of view of an individual case. Also, demographic structure (e.g., age structure) has the potential to affect temporal autocorrelation in transmission rate - the effects of age structure can be captured by a model that includes an infection rate that varies over time (Earn et al. 1998).

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.
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. Research Square. 2022;doi:https://doi.org/10.21203/rs.3.rs-1407962/v1.
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.