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

Rachael Aber[1], Yanming Di[2], and Benjamin D. Dalziel[1,3]

[1]Department of Integrative Biology, Oregon State University, Corvallis, Oregon, USA
[2]Department of Statistics, Oregon, Oregon State University, Corvallis, Oregon, USA
[3]Department of Mathematics, Oregon State University, Corvallis, Oregon, USA

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 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 the 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 findings suggest that heterogeneity in transmission, susceptibility, and reporting could play important roles in driving large surges and extending epidemic duration. The dispersion of incidence time series can contain structured information which enhances predictive understanding of the underlying drivers of transmission, with potential applications as leading indicators for public health response.

Author summary

Understanding patterns in infectious disease incidence is crucial for understanding epidemic dynamics and for developing effective public Traditional 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 case count dispersion within a single time series and applied the method to COVID-19 case count data. We found that conspicuous shifts in dispersion occurred across counties concurrently, and that these shifts were not explained by incidence alone. Dispersion increased around peaks in incidence such as the major surge in cases in 2022, and dispersion also increased as the pandemic progressed. These increases potentially indicate 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 fluctations) 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 susceptiblility (1; 2; 3; 4; 5; 6; 7; 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 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 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 countrys 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 focussed more intensely (proportionally more infections in a smaller span of time) with implications on the sensitivity of cities to climate forcing and for surge capacity in the health system (4; 13).

What drives indicence 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), $I_t \sim NB\left(\mu_t, \theta_t\right)$, 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 $\mathrm{Var}[I_t] = \mu_t + \mu_t^2/\theta_t$, 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 Supplemental Information). 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) than 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

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

where time-varying reproductive number R_t gives the expected number of secondary infections aqcuired from an infected indivual 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. 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 assumption will not often hold, and our aim in this paper is to develop, test and apply an alternative approach, which makes 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 model incidence at broad spatiotemporal scales using natural splines (19). To allow for diverse shapes in the broadscale epidemic dynamics, these are fitted within a moving window

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

where N represents population size, $h_j(t)$ are basis functions, J is the degrees of freedom for the splines, and $\beta_j^{(t)}$ are fitted parameters for a symmetrical window of half-width Δ, centered at t, i.e., extending from $t - \Delta$ to $t + \Delta$. The degrees of freedom to be used for the splines, and the width of the moving window will depend on the application. Explaination of the specific choices we used J and Δ for our application to COVID-19 cases in US counties is described 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,

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

The form of the probability mass function for infections at a time step is:

$$f_t(I) = \binom{I+\theta-1}{I}\left(\frac{\mu}{\mu+\theta}\right)^I\left(\frac{\theta}{\mu+\theta}\right)^\theta \tag{4}$$

where μ is estimated via the linear predictor outlined above.

We estimated θ_t given observed incidence using an iteratively reweighted least-squares (IRLS) procedure for mean estimation in conjunction with an optimization procedure to compute θ_t. That is, for each time window, a series of μ_s from $s = t - \Delta$ to $s = t + \Delta$ was estimated using the spline and offset term via IRLS in the NBPSeq R package (20). Then, a single value of θ_t for the time window was computed via an optimization procedure.

In addition to fitting the model at each time step, we developed a likelihood-ratio test (LRT) that could be applied at each time step to test the hypothesis that θ has changed. This involves fitting and comparing both a null model (no θ change) and a two-part (θ-change) model. For the null model, an optimization for a single θ value for the time window was performed, and for the θ-change model, an optimization for a θ value during the first half of the window was performed, as well as an optimization for the second half.

Since very large θ correspond to processes that are operationally nearly identical to a Poisson process, the test will not produce a p-value if any of the three values are greater than a user-specified threshold. In the applicatio below we set this threshold at 10^3, meaning that theta estimates with temporal crowding that was within 0.1% of that expected for a Poisson process was considered effectively Poisson.

Very large θ estimates are often produced in small populations when the process can't be distinguished from a Poisson process. Therefore, we have very small $\phi = 1/\theta$ values be considered zero, at which point we fail to reject the Poisson hypothesis.

In addition, we imposed restrictions on θ estimates such that $\theta = 0$ (point mass) results in no p-value being produced in the test result. So, large ϕ are considered values where we fail to reject the point mass hypothesis.

Additionally, we included an option to retrieve the likelihood of a range of theta values in addition to these restrictions.

Application to simulated data

Our simulation samples start points and theta values, and incidence is modeled according to overall incidence trends.

To test the validity and power of the methods, we epidemic curves with known values of μ and θ. We varied the magnitude of the θ change, location of the change in the curve, population size underlying the curve, and curve shape to test the validity and power of our approach across a range of scenarios. Epidemic curves were created from a deterministic skeleton (template) that either assumed a flat distribution of incidence over time or a Gaussian-shaped distribution, with parameters controlling the total epidemic size, peak time, peak width, and duration. Stochastic noise was then added to these templates by sampling from a negative binomial distribution with mean given by the template values.Each curve was split into two equal pieces and separate dispersion parameters θ_1 and θ_2 were used in each respective piece, before and after the breakpoint. This allowed us to simulate epidemic curves with variation in dispersion, as well as the special case when $\theta_1 = \theta_2$.

Data availability and processing

We applied our method to cumulative COVID-19 case counts in United States (US) counties (21). Population sizes were from taken the US Census Bureau (22). Cumulative cases for the largest three counties in each state were converted to weekly counts by keeping the last observation from each week and differenced to compute new cases. Missing values for new cases were all at the beginning of the pandemic and were imputed as zero. Approximately 0.24% of new cases were negative due to corrections in the cumulative data. These were also imputed as zero. We used a width paramter of $\delta = 8$ week and set the degrees of freedom parameter for the natural splines to $J = 3$. [This seems like a wide window and a low-df spline. Why did we choose these and what happens if we use different values? Some sort of sensitivity analysis is probably needed, even if it is just to run the analysis using a smaller window and larger df]

Results and Discussion

Our estimates of θ were robust across the population sizes included in the empirical analysis of U.S. counties. However, inaccuracies were more common in counties with small populations outside the examined range.We found that the LRT method of testing for sharp changes in θ is also robust across population sizes represented in the empirical data (Fig 1 e, f). Adequate test performance for the LRT is characterized by an average p-value of 0.5 when the effect size is zero, with decreasing average p-values as the effect size increases.

Acknowledgments

RA work on this project was supported by X and Y. BDD work on this project was supported by the National Science Foundation (NSF) grants (X, Y) and by the David and Lucile Packard Foundation.

References

References

[1] Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005;438(7066):355–359.

[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.

[3] Lau MS, 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. Proceedings of the National Academy of Sciences. 2017;114(9):2337–2342.

[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 US cities. Science. 2018;362(6410):75–79.

[5] Kirkegaard JB, Sneppen K. Superspreading quantified from bursty epidemic trajectories. Scientific Reports. 2021;11(1):24124.

[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.

[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.

[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. Scientific Reports. 2023;13(1):13230.

[9] Kretzschmar ME, Rozhnova G, Bootsma MC, van Boven M, van de Wijgert JH, Bonten MJ. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. The Lancet Public Health. 2020;5(8):e452–e459.

[10] Blumberg S, Lloyd-Smith JO. Inference of R 0 and transmission heterogeneity from the size distribution of stuttering chains. PLoS computational biology. 2013;9(5):e1002993.

[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.

[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):584–587.

[13] Wallinga J. Metropolitan versus small-town influenza. Science. 2018;362(6410):29–30.

[14] Kendall DG. Stochastic processes and population growth. Journal of the Royal Statistical Society Series B (Methodological). 1949;11(2):230–282.

[15] 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.

[16] Barron DN. The analysis of count data: Overdispersion and autocorrelation. Sociological methodology. 1992;p. 179–220.

[17] Lloyd M. Mean crowding'. The Journal of Animal Ecology. 1967;p. 1–30.

[18] Bjørnstad ON, Finkenstädt BF, Grenfell BT. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. Ecological monographs. 2002;72(2):169–184.

[19] Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. BMC Medical Research Methodology. 2019;19:1–16.

[20] Di Y, Schafer D, Cumbie J, Chang J. NBPSeq: Negative Binomial Models for RNA-Sequencing Data. R package version 03 0, URL http://CRAN R-project. 2015;.

[21] The New York Times. Coronavirus (COVID-19) Data in the United States. GitHub. 2024;Available from: github.com/nytimes/covid-19-data.

[22] United States Census Bureau. Annual County Resident Population Estimates: April 1, 2020 to July 1, 2021. US Census Bureau Population Estimates Program. 2021;Available from: www2.census.gov/programs-surveys/popest/datasets/2020-2021/counties/totals.