

RESEARCH ARTICLE

Superspreading of SARS-CoV-2 in the USA

Calvin Pozderac *, Brian Skinner

Department of Physics, Ohio State University, Columbus, Ohio, United States of America

* pozderac.16@osu.edu

Abstract

A number of epidemics, including the SARS-CoV-1 epidemic of 2002-2004, have been known to exhibit superspreading, in which a small fraction of infected individuals is responsible for the majority of new infections. The existence of superspreading implies a fat-tailed distribution of infectiousness (new secondary infections caused per day) among different individuals. Here, we present a simple method to estimate the variation in infectiousness by examining the variation in early-time growth rates of new cases among different subpopulations. We use this method to estimate the mean and variance in the infectiousness, β , for SARS-CoV-2 transmission during the early stages of the pandemic within the United States. We find that $\sigma_{\beta}/\mu_{\beta} \gtrsim 3.2$, where μ_{β} is the mean infectiousness and σ_{β} its standard deviation, which implies pervasive superspreading. This result allows us to estimate that in the early stages of the pandemic in the USA, over 81% of new cases were a result of the top 10% of most infectious individuals.

 OPEN ACCESS

Citation: Pozderac C, Skinner B (2021) Superspreading of SARS-CoV-2 in the USA. PLoS ONE 16(3): e0248808. <https://doi.org/10.1371/journal.pone.0248808>

Editor: Yury E Khudyakov, Centers for Disease Control and Prevention, UNITED STATES

Received: September 28, 2020

Accepted: March 7, 2021

Published: March 25, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pone.0248808>

Copyright: © 2021 Pozderac, Skinner. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: We use publicly available data taken from the data set provided by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The paper describing the data can be found at: [https://doi.org/10.1016/S1473-3099\(20\)30120-1](https://doi.org/10.1016/S1473-3099(20)30120-1). The data can be

Introduction

The temporal growth of an epidemic is often characterized by either a time scale (such as the doubling time) [1, 2] or by the reproduction rate R_0 , which indicates the average number of new infections produced by each infected individual [3]. Estimates of R_0 for the current pandemic of SARS-CoV-2 range from 1.4 to 3.8 [4–7]. Neither of these numbers, however, gives any information about the distribution of infectiousness among individuals—i.e., whether new infections arise relatively uniformly from all infected individuals, or whether new infections are driven primarily by a small number of highly infectious individuals. The latter case is commonly referred to as “superspreading”, and different epidemics exhibit superspreading to different degrees. For example, during the outbreak of SARS CoV-1 in 2002-2004, over 80% of cases were observed to result from the top 20% most infectious individuals [8, 9]. Understanding the degree of superspreading in the current pandemic of SARS-CoV-2 is crucial for developing strategies to mitigate continued spread and informing an educated reopening procedure [10–13].

Here we present a simple and direct method to understand how the infectiousness (also called the “reproduction rate” of the disease) varies among infected individuals. At late times after the onset of an epidemic, the number of infected individuals is large, and consequently any statistical fluctuations in the growth rate are relatively small, so that the growth rate is well characterized by the mean infectiousness, μ_{β} . However, at early times, when there are relatively few cases, the growth rate is stochastic and the degree of randomness depends on the variance

found at https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series under the filename `time_series_covid19_confirmed_US.csv`.

Funding: The author(s) received no specific funding for this work.

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

in infectiousness, σ_β^2 , between individuals (Fig 1a). By examining the variance in growth rate across subpopulations at these early times (Fig 1b), we are able to infer the variation in the distribution of infectiousness. In our analysis we divide the US cases into counties and observe how the variance in growth rate across them evolves as the number of cases increases.

Formalizing this idea, we present a derivation of the variance in the exponential growth rate, or number of new cases per infected individual per day, $\Delta I/I$, using an SIR framework that incorporates a probability distribution for the infectiousness of a given individual. Our result implies a simple method for estimating the mean, μ_β , and variance, σ_β^2 , of the infectiousness β . We apply this method to data for COVID-19 cases in the USA, and find a mean infection rate of $\mu_\beta = 0.18$ cases/day and standard deviation of $\sigma_\beta \approx 0.59$ cases/day. Since the standard deviation is considerably larger than the mean, with $\sigma_\beta/\mu_\beta \approx 3.2$, we conclude that superspreading is prevalent. By our estimate, these results imply that at least 81% of new cases are caused by the top 10% of most infectious individuals. Our method, which uses only a direct measurement of variance in detected case data in the USA, is consistent with estimates of superspreading using surveillance data [14], secondary-case data [15], and more complicated estimates of cluster size distribution using Markov Chain Monte Carlo [16].

Results

Variance in growth rate in the SIR model

We derive a relation between the variance in the case growth rate and the variance in individual infectiousness between individuals in the population. We start with a standard discrete-time SIR model [17], which is governed by the following difference equations:

$$\begin{aligned} \Delta S &= -\beta I \frac{S}{N} \\ \Delta I &= \beta I \frac{S}{N} - rI \\ \Delta R &= rI \end{aligned} \tag{1}$$

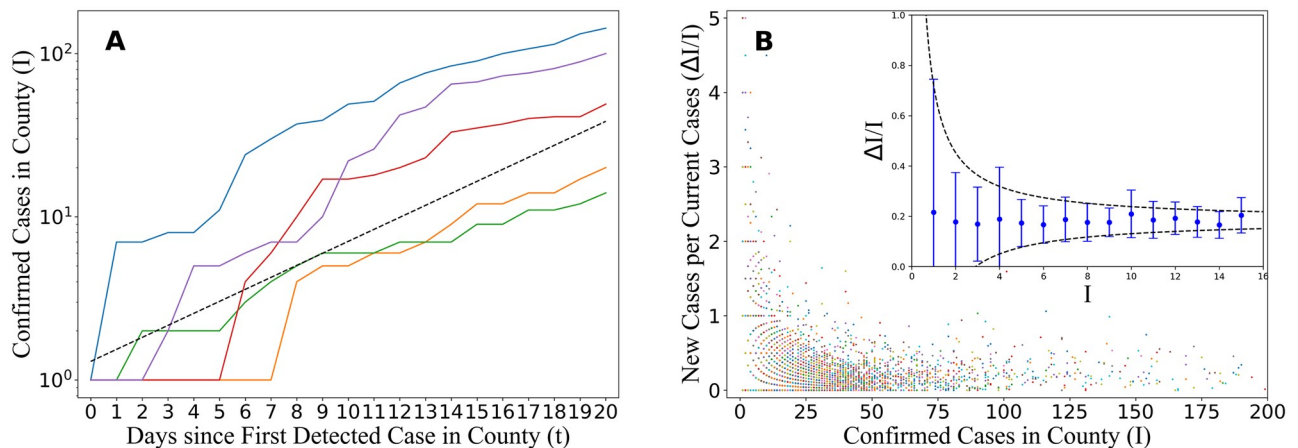


Fig 1. (a) Illustration of the variance in early-time growth rate of new cases. At early times, there is noticeable variance in the growth rate between counties. As the number of cases grows, all counties stabilize towards the average growth rate $I \sim (1 + \mu_\beta)^t$, (dashed black line) where t is the number of days since the first case in a county. The counties shown are Boulder, CO (blue), St. Mary, LA (purple), Vanderburgh, IN (red), Mesa, CO (orange), and Jones, GA (green). (b) The number of daily infections per infected individual as a function of total infections. In the main figure, each point corresponds to a given county (across all US counties that never report $\Delta I < 0$) at a given time point (within the first 14 days after the first infection reported in that county). As the number of cases increase, all counties converge to the mean infection rate. The mean (points) and variance (bars) of $\Delta I/I$ at a given I are shown in the inset. The variance decreases like $(\mu_\beta + \sigma_\beta^2)/I$ (black lines).

<https://doi.org/10.1371/journal.pone.0248808.g001>

Here, N is the total population and S , I , and R are the time-dependent numbers of susceptible, infected, and recovered individuals, respectively. The parameters β and r encode the infectiousness and recovery rate of a disease within a population. The time is effectively discretized into days by the available data, so we use ΔI rather than the usual time derivative, dI/dt . The SIR description typically assumes fixed values for β and r across the population. However, in superspreading contexts there is a substantial variance in the infectiousness within a population [8, 9, 18, 19]. We account for this variation by introducing a probability distribution of infectiousness, $p(\beta)$, so that the probability for a randomly-selected individual to have infectiousness in the range $(\beta, \beta + d\beta)$ is given by $p(\beta)d\beta$.

For an individual with a given infectiousness, β , the probability of infecting exactly n others in a day follows the Poisson distribution, $\text{Pois}(n;\beta)$. The probability that a randomly selected individual will infect n others is given by combining the Poisson distribution with the distribution $p(\beta)$, giving

$$P(n) = \int_0^\infty d\beta \frac{e^{-\beta} \beta^n}{n!} p(\beta). \tag{2}$$

The first two moments of $P(n)$, μ_n and σ_n^2 , can be calculated independent of the form of $p(\beta)$:

$$\mu_n = \sum_{n=0}^\infty nP(n) = \mu_\beta \tag{3}$$

$$\sigma_n^2 = \sum_{n=0}^\infty (n - \mu_n)^2 P(n) = \mu_\beta + \sigma_\beta^2 \tag{4}$$

Eq (4) represents the variance, among all infected individuals, of the number of new infections caused by a single person in a given day. When there are I active cases, the mean number of new cases per infected person, $\Delta(I + R)/I$, is given by the average of I random variables drawn from the distribution $P(n)$. By the central limit theorem, it follows that $\text{Var}(\Delta(I + R)/I) = \sigma_n^2/I$. Additionally, in the SIR model with a finite total population N , $\Delta(I + R)/I = \beta S/N = \beta(1 - (I + R)/N)$ decreases as the susceptible population continually shrinks. Effectively, $p(\beta)$ is scaled by the factor $(1 - (I + R)/N)$, which represents the fraction of the population that remains susceptible. Consequently, $\mu_\beta \rightarrow \mu_\beta(1 - (I + R)/N)$ and $\sigma_\beta^2 \rightarrow \sigma_\beta^2(1 - (I + R)/N)^2$. Therefore the total variance in $\Delta(I + R)/I$ follows:

$$\text{Var}\left(\frac{\Delta(I + R)}{I}\right) = \frac{\mu_\beta(1 - \frac{I+R}{N}) + \sigma_\beta^2(1 - \frac{I+R}{N})^2}{I} \tag{5}$$

This result becomes simpler in the limiting case where there is no significant change in the susceptible population ($N \rightarrow \infty$) and no recovery ($R \rightarrow 0$). In this limit, we retrieve the case of simple exponential growth, for which [20]

$$\text{Var}\left(\frac{\Delta I}{I}\right) = \frac{\mu_\beta + \sigma_\beta^2}{I}. \tag{6}$$

In the limit $\sigma_\beta \rightarrow 0$, where every infected individual has the same infectiousness μ_β , the variance in the average infection rate is simply μ_β/I , which corresponds to the variance in a Poisson process with rate μ_β .

In the case of SARS-CoV-2, it is well established that there are asymptomatic carriers [21–23] who transmit the virus without being detected, as well as other infections that are

undetected or unreported. Current estimates typically predict that only 10 – 25% [24–26] of cases are detected. One can attempt to address this effect by assuming that there is a fixed detection probability, p_{det} , and that the entire infected population, regardless of symptoms, follows the same infectiousness distribution $p(\beta)$. In this case, there are many more infected individuals, $I \sim I_{\text{det}}/p_{\text{det}}$, than those detected, which reduces the statistical fluctuations in the growth rate and makes our calculation of σ_{β}^2 a lower bound. The effect of undetected cases is considered in more detail in the [S3 Appendix](#). In order to be conservative (especially given the possibility that asymptomatic cases have a lower rate of infection than symptomatic ones [27, 28]), the results we present here use $p_{\text{det}} = 1$.

Data for COVID-19 in the USA

We now turn our attention to data for total detected cases of COVID-19 in the USA, taken from the publicly available data set at Ref. [29]. In the following analysis we limit our consideration to only a short timescale (~ 14 days) after the first infection is detected in a given county. This limitation in time scale serves three main purposes; first, it is likely that through changes in policy, lockdown, social distancing, mask usage, etc., the average infectiousness within the population is time-dependent. By restricting ourselves to a relatively small window of early times, we may assume that there is a constant average infectiousness. Second, considering only beginning stages allows us to neglect the possible saturation of the susceptible population, effectively allowing us to take the $N \rightarrow \infty$ limit. Finally, the recovery period for COVID-19 ranges from 7-14 days [30, 31] and so by considering this two week period, we can treat our system as if there is limited recovery and $R \rightarrow 0$. These restrictions allow us to treat the USA data using the exponential case, [Eq \(6\)](#).

In our analysis, the population is divided into geographic regions and the variance is calculated across different trajectories $I(t)$. The US cases are divided by county. For each county, we calculate the average number of new cases per current case per day, $\Delta I/I$, for the first 14 days after the first infection is detected in that county. The variance in $\Delta I/I$ is then calculated among all counties that have a given fixed value of I (we present data only for values of I that have at least 250 corresponding counties). As shown in [Fig 2](#), the US data generally follows the predicted $\sim 1/I$ trend. An unbiased fit of the data gives $\text{Var}(\Delta I/I) \propto I^{-0.74}$. From [Eq \(6\)](#), we calculate $\mu_{\beta} + \sigma_{\beta}^2$ by averaging $\text{Var}(\Delta I/I) \times I$, weighted by the number of instances at each I value. One might worry that the main source of variation comes from differing average growth rates, μ_{β} , in various counties (e.g. rural vs. urban). However, we show in the [S2 Appendix](#) that variance in μ_{β} across counties is too small to explain the large observed variance in $\Delta I/I$.

We calculate μ_{β} from the entire USA population by averaging all values of $\Delta I/I$ weighted by the current number of infections. Equivalently, we sum the number of cases caused each day and then divide by the sum of the number of cases across those days. This procedure gives the mean infectiousness, μ_{β} , and thus from [Eq \(6\)](#) and the fitted slope in [Fig 2](#), we can infer σ_{β}^2 .

This calculation yields $\mu_{\beta} = 0.18$ cases/day and $\sigma_{\beta} = 0.59$ cases/day. The small value of $\mu_{\beta}^2/\sigma_{\beta}^2 = 0.096$, equivalent to the dispersion parameter [16, 32, 33], provides clear evidence for superspreading during early stages of the COVID-19 pandemic in the United States. (See [S7 Appendix](#) for discussion about defining the dispersion parameter in terms of the daily infection rate).

These results for μ_{β} and σ_{β} can be used to further quantify the extent of superspreading under the assumption that $p(\beta)$ follows a gamma distribution (as in Ref. [18]). In the Methods section we present a derivation of the cumulative share of infections, Y , caused by the top X portion of most infectious cases. The corresponding ‘‘Lorenz curve’’ $Y(X)$ is plotted in [Fig 3](#). This result implies (using our relatively conservative estimate of σ_{β}) that 81% of new infections

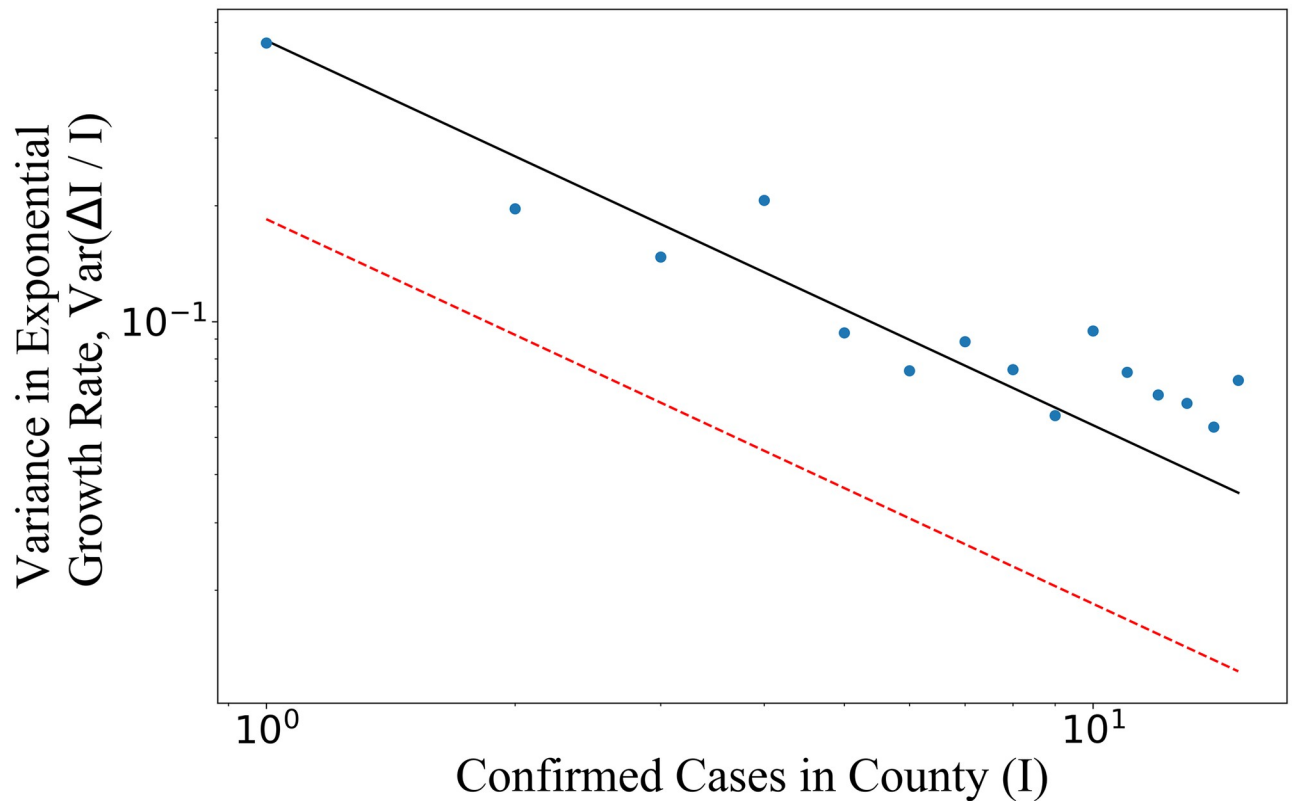


Fig 2. As the number of infections I in a given county increases, the variance in exponential growth rate, $\text{Var}(\Delta I/I)$, decreases as $(\mu_\beta + \sigma_\beta^2)/I$. Each data point at a given I is calculated by taking the sample variance in $\Delta I/I$ across all counties when they have I cases. We observe that the USA data (blue) is inconsistent with a model of uniform infectiousness, or $\sigma_\beta = 0$ (dashed red line). A fit to the data (solid black line) implies a large variance in infectiousness, such that $\sigma_\beta/\mu_\beta \approx 3.2$.

<https://doi.org/10.1371/journal.pone.0248808.g002>

are produced by the top 10% of most infectious individuals, while only about 4.5% of cases arise from the 80% of infected individuals with the lowest infection rates.

Discussion

As we have shown, a wide distribution $p(\beta)$ in infectiousness β leads to large statistical variation in the early-time growth rate of a disease. By calculating the variance in growth rate among different subpopulations one can infer the variance in $p(\beta)$. Our result for COVID-19 cases in the USA suggests that $\sigma_\beta/\mu_\beta \approx 3.2$, implying a relatively severe superspreading. If we further assume that $p(\beta)$ follows a gamma distribution (as in Ref. [18]), then we can produce a more direct estimate of the extent of superspreading (Fig 3). Our relatively simple and direct method, based on a calculation of variance in reported case data, can be contrasted with more complicated methods for inferring the dispersion parameter that are based on maximum likelihood estimation (e.g., Ref. [33] develops such a method using simulated data), cluster size distributions [16, 34], and surveillance or tracing data [14, 15]. These methods also tend to yield a lower-bound estimate for σ_β/μ_β . While studies based on testing and contact tracing (e.g., Refs. [18, 35–37]) remain the definitive method for assessing superspreading, the method we present here may provide a much simpler way of estimating its prevalence across a much larger population.

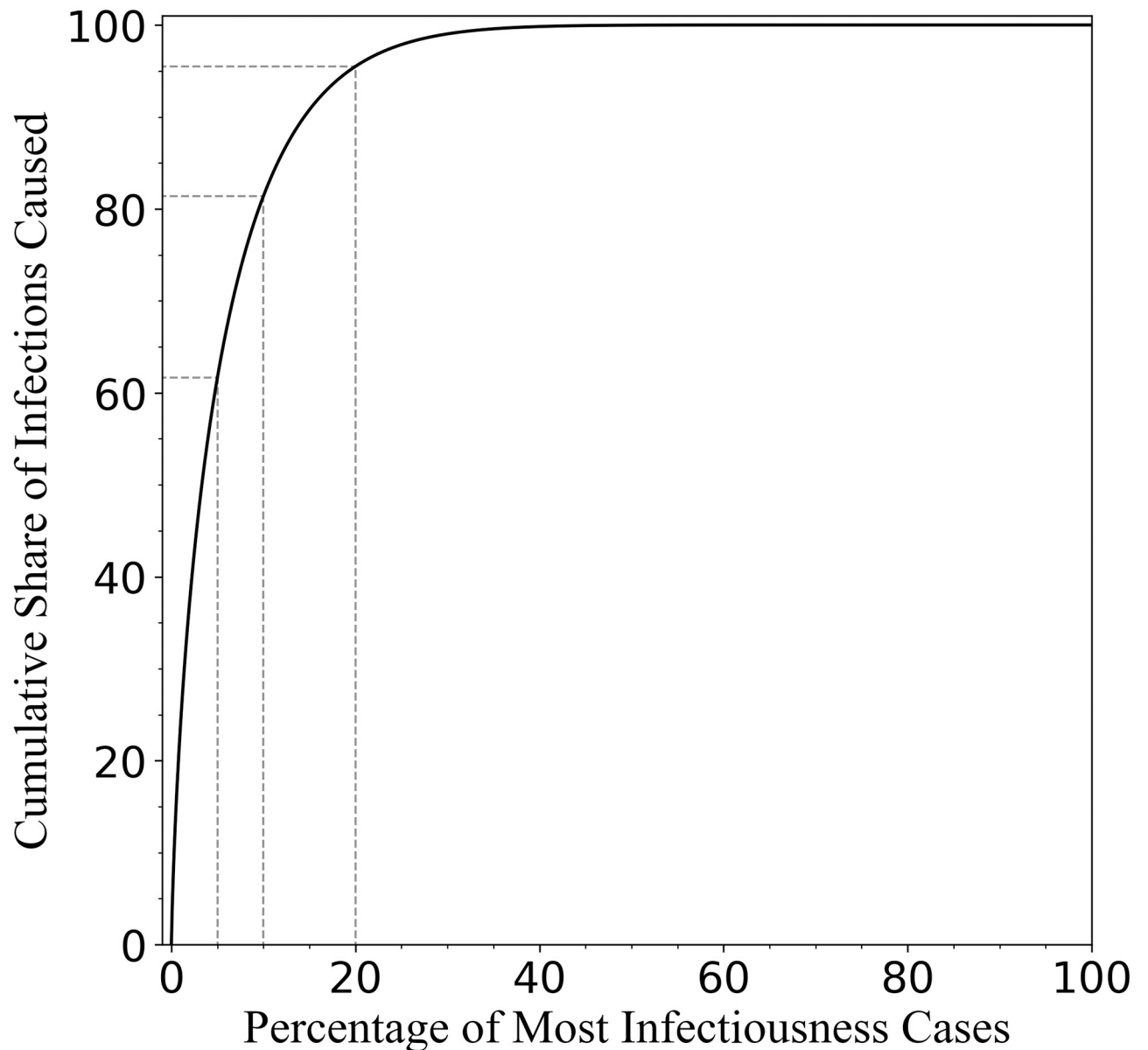


Fig 3. An estimated Lorenz curve for SARS-CoV-2 infections in the USA, which displays the percentage of new cases that are caused by a given cumulative percentage of most infectious individuals (solid black). A few points in the curve are highlighted (dashed grey lines): 61.7%, 81.4%, and 95.5% of new cases are caused by the top 5%, 10%, and 20% infectious cases, respectively. Accounting for undetected and asymptomatic cases would apparently make this curve steeper, corresponding to more severe superspreading.

<https://doi.org/10.1371/journal.pone.0248808.g003>

We emphasize that our analysis is unable to determine whether this large variance is a result of differing biological symptoms, social behavior, or other possible explanations. Additionally, this estimation is carried out for early times to minimize effects from a time varying $p(\beta)$ and therefore predominantly speaks to the infectiousness prior to widespread lockdown measures.

We close by commenting on a number of complicating factors that we did not include in our analysis and which, one might suspect, could alter our primary finding of a large value of $\sigma_{\beta}/\mu_{\beta}$. For example, we have assumed a uniform value of μ_{β} across different geographic locations; we have neglected undetected cases; we have ignored the possible variation in detection

rate p_{det} among different counties; we have effectively treated each county as an isolated population and have neglected cross-county interactions; and we have ignored the effects of the incubation period as well as the potential variation in incubation periods between individuals. In the Supplemental Information, we consider each of these mechanisms in turn and show that none of them can explain our result, so that our conclusion of prevalent superspreading of SARS-CoV-2 in the USA remains robust. In brief: the variation in μ_β among different geographic locations is too small to explain the observed variance in growth rate [S2 Appendix]; neglecting undetected cases leads to an *underestimate* of the variance σ_β^2 , so that our result is effectively a lower bound for the prevalence of superspreading [S3 Appendix]; variation in p_{det} between counties does not directly affect the variance in the growth rate $(\Delta I_{\text{det}})/I_{\text{det}}$, other than to provide an average of $p_{\text{det}} < 1$, which results in a lower-bound estimate of σ_β^2 [S4 Appendix]; cross county interactions tend to reduce the variance, so our result cannot be explained as a consequence of such interactions [S5 Appendix]; and variations in incubation period can only reduce the apparent variance in growth rate [S6 Appendix].

Methods

Data source

We use publicly available data taken from the data set provided by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [29] to estimate μ_β . Knowing μ_β enables us to determine σ_β by taking a best fit to Eq (6). Counties that recorded $\Delta I < 0$ at any point are discarded from the analysis due to the potential for recording error; such counties comprise ~20% of all counties.

Numerical simulation

We corroborate Eqs (5) and (6) using a numerical simulation of the trajectories of infection growth, $I(t)$, for a given distribution $p(\beta)$. Reference 18 has suggested that infectiousness follows a gamma distribution, and consequently, $P(n) = \text{NB}(n; \mu_\beta^2/\sigma_\beta^2, \mu_\beta/(\mu_\beta + \sigma_\beta^2))$ where NB is the negative binomial distribution [10, 16]. Using this assumption, we simulate the growth of the epidemic by assuming that a given individual i , with infectiousness β_i that is drawn randomly from $p(\beta)$, generates a number n_i of new cases each subsequent day that is drawn from $\text{Pois}(n_i; \beta_i)$. The simulation results confirm Eqs (5) and (6), as shown in S1 Appendix. Numerical simulations were performed using Python; the primary analysis is publicly available [38] and the simulations are available upon request to the corresponding author.

Derivation of the curve $Y(X)$

Following Ref. [18], we assume that the distribution of infectiousness, $p(\beta)$, follows a gamma distribution. This assumption also allows us to further quantify the degree of superspreading by deriving a mathematical relation for the curve $Y(X)$, where Y represents the proportion of infections produced by the top X fraction of most infectious individuals. In particular, one can calculate the fraction of individuals X_{β_0} with infectiousness larger than a given value β_0 , as well as the fraction of secondary infections Y_{β_0} that these individuals are expected to cause:

$$X_{\beta_0} = \int_{\beta_0}^{\infty} d\beta p(\beta) = Q\left(\frac{\mu_\beta^2}{\sigma_\beta^2}, \beta_0 \frac{\mu_\beta}{\sigma_\beta^2}\right) \tag{7}$$

$$Y_{\beta_0} = \int_{\beta_0}^{\infty} d\beta p(\beta) \frac{\beta}{\mu_\beta} = Q\left(1 + \frac{\mu_\beta^2}{\sigma_\beta^2}, \beta_0 \frac{\mu_\beta}{\sigma_\beta^2}\right), \tag{8}$$

where Q is the Regularized Gamma function. By eliminating β_0 we find

$$Y = Q\left(1 + \frac{\mu_\beta^2}{\sigma_\beta^2}, Q^{-1}\left(\frac{\mu_\beta^2}{\sigma_\beta^2}, X\right)\right). \quad (9)$$

Fig 3 displays the cumulative share of infections, Y , caused by the top X portion of most infectious cases.

Supporting information

S1 Appendix. Simulations.

(PDF)

S2 Appendix. Variance in μ_β .

(PDF)

S3 Appendix. Undetected cases.

(PDF)

S4 Appendix. Variance in testing.

(PDF)

S5 Appendix. Cross-county interactions.

(PDF)

S6 Appendix. Variance in incubation period.

(PDF)

S7 Appendix. Dispersion parameter comparison.

(PDF)

Acknowledgments

The authors are grateful to N. E. Skinner for helpful conversations.

Author Contributions

Conceptualization: Brian Skinner.

Data curation: Calvin Pozderac.

Formal analysis: Calvin Pozderac.

Investigation: Calvin Pozderac, Brian Skinner.

Methodology: Calvin Pozderac, Brian Skinner.

Project administration: Brian Skinner.

Software: Calvin Pozderac.

Supervision: Brian Skinner.

Visualization: Calvin Pozderac.

Writing – original draft: Calvin Pozderac.

Writing – review & editing: Calvin Pozderac, Brian Skinner.

References

1. Muniz-Rodriguez K, Chowell G, Cheung CH, Jia D, Lai PY, Lee Y, et al. Doubling Time of the COVID-19 Epidemic by Province, China. *Emerging Infectious Diseases*. 2020; 26(8):1912–1914. <https://doi.org/10.3201/eid2608.200219> PMID: 32330410
2. Zhou L, Liu JM, Dong XP, McGoogan JM, Wu ZY. COVID-19 seeding time and doubling time model: an early epidemic risk assessment tool. *Infectious Diseases of Poverty*. 2020; 9(76). <https://doi.org/10.1186/s40249-020-00685-4> PMID: 32576256
3. Murray JD. *Mathematical Biology*. Springer-Verlag; 2003.
4. Li Q, Guan X, Wu P, Wang X, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *New England Journal of Medicine*. 2020; 382(13):1199–1207. <https://doi.org/10.1056/NEJMoa2001316> PMID: 31995857
5. Riou J, Althaus CL. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance*. 2020; 25(4). <https://doi.org/10.2807/1560-7917.ES.2020.25.4.2000058> PMID: 32019669
6. Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R. High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2. 2020; 26(7).
7. Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*. 2020; 27(2). <https://doi.org/10.1093/jtm/taaa021> PMID: 32052846
8. Galvani AP, May RM. Dimensions of superspreading. *Nature*. 2005; 438:293–295. <https://doi.org/10.1038/438293a>
9. Stein RA. Super-spreaders in infectious diseases. *International Journal of Infectious Diseases*. 2011; 15(8):e510–e513. <https://doi.org/10.1016/j.ijid.2010.06.020>
10. Althouse BM, Wenger EA, Miller JC, Scarpino SV, Allard A, Hébert-Dufresne L, et al. Stochasticity and heterogeneity in the transmission dynamics of SARS-CoV-2; 2020.
11. O'Donoghue AL, Dechen T, Pavlova W, Boals M, Moussa G, Madan M, et al. Super-Spreader Businesses and Risk of COVID-19 Transmission. *medRxiv*. 2020; <https://doi.org/10.1101/2020.05.24.20112110>
12. Vespignani A, Tian H, Dye C, Lloyd-Smith JO, Eggo RM, Shrestha M, et al. Modelling COVID-19. *Nature Reviews Physics*. 2020; p. 1–3.
13. Weiner Z, Wong G, Elbanna A, Tkachenko A, Maslov S, Goldenfeld N. Projections and early-warning signals of a second wave of the COVID-19 epidemic in Illinois. *medRxiv*. 2020.
14. Lau MS, Grenfell B, Nelson K, Lopman B. Characterizing super-spreading events and age-specific infectivity of COVID-19 transmission in Georgia, USA. *medRxiv*. 2020; <https://doi.org/10.1101/2020.06.20.20130476>
15. Hasan A, Susanto H, Kasim M, Nuraini N, Triany D, Lestari B. Superspreading in Early Transmissions of COVID-19 in Indonesia. *medRxiv*. 2020. <https://doi.org/10.1101/2020.06.30.20143560> PMID: 32637975
16. Endo A, Abbott S, Kucharski AJ, Funk S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Research*. 2020; 5:67. <https://doi.org/10.12688/wellcomeopenres.15842.1>
17. Kermack WO, McKendrick AG, Walker GT. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character*. 1927; 115(772):700–721. <https://doi.org/10.1098/rspa.1927.0118>
18. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. *Nature*. 2005; 438:355–359. <https://doi.org/10.1038/nature04153>
19. Sneppen K, Taylor RJ, Simonsen L. Impact of Superspreaders on dissemination and mitigation of COVID-19. *medRxiv*. 2020; <https://doi.org/10.1101/2020.05.17.20104745>
20. Bliss CI, Fisher RA. Fitting the Negative Binomial Distribution to Biological Data. *Biometrics*. 1953; 9(2):182. <https://doi.org/10.2307/3001850>
21. Mahajan A, Solanki R, Sivadas N. Estimation of Undetected Symptomatic and Asymptomatic cases of COVID-19 Infection and prediction of its spread in USA. *medRxiv*. 2020; <https://doi.org/10.1101/2020.06.21.20136580>
22. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Euro Surveillance*. 2020; 25(10). <https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000180> PMID: 32183930

23. Nishiura H, Kobayashi T, et al. Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). *Int J Infect Dis*. 2020; 94:154–155. <https://doi.org/10.1016/j.ijid.2020.03.020> PMID: 32179137
24. Pedersen M, Meneghini M. Quantifying undetected COVID-19 cases and effects of containment measures in Italy: Predicting phase 2 dynamics. 2020; <https://doi.org/10.13140/RG.2.2.11753.85600>
25. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*. 2020; 368(6490):489–493. <https://doi.org/10.1126/science.abb3221> PMID: 32179701
26. Lu FS, Nguyen AT, Link NB, Lipsitch M, Santillana M. Estimating the Early Outbreak Cumulative Incidence of COVID-19 in the United States: Three Complementary Approaches. *medRxiv*. 2020; <https://doi.org/10.1101/2020.04.18.20070821> PMID: 32587997
27. Wang Y, Tong J, Qin Y, Xie T, Li J, vi J, et al. Characterization of an asymptomatic cohort of SARS-CoV-2 infected individuals outside of Wuhan, China. *Clinical Infectious Diseases*; <https://doi.org/10.1093/cid/ciaa629> PMID: 32442265
28. Chu DK, Akl EA, Duda S, Solo K, Yaacoub S, Schünemann HJ, et al. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. *The Lancet*. 2020; 395:1950–1951. [https://doi.org/10.1016/S0140-6736\(20\)31142-9](https://doi.org/10.1016/S0140-6736(20)31142-9) PMID: 32497510
29. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious diseases*. 2020; 20(5):533–534. [https://doi.org/10.1016/S1473-3099\(20\)30120-1](https://doi.org/10.1016/S1473-3099(20)30120-1)
30. Roman Wölfel VMC, Guggemos W, Seilmaier M, Zange S, Müller MA, Niemeyer D, et al. Virological assessment of hospitalized patients with COVID-2019. *Nature*. 2020; 581:465–469. <https://doi.org/10.1038/s41586-020-2196-x> PMID: 32235945
31. Bar-On YM, Flamholz A, Phillips R, Milo R. Science Forum: SARS-CoV-2 (COVID-19) by the numbers. *Elife*. 2020; 9:e57309. <https://doi.org/10.7554/eLife.57309>
32. Hébert-Dufresne L, Althouse BM, Scarpino SV, Allard A. Beyond R0: Heterogeneity in secondary infections and probabilistic epidemic forecasting. *medRxiv*. 2020; <https://doi.org/10.1101/2020.02.10.20021725>
33. 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):1–8. <https://doi.org/10.1371/journal.pone.0000180>
34. Kucharski AJ, Althaus CL. The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. *Eurosurveillance*. 2015; 20(25). <https://doi.org/10.2807/1560-7917.ES2015.20.25.21167> PMID: 26132768
35. Althaus CL. Ebola superspreading. *The Lancet Infectious Diseases*. 2015; 15(5):507–508. [https://doi.org/10.1016/S1473-3099\(15\)70135-0](https://doi.org/10.1016/S1473-3099(15)70135-0)
36. Melsew YA, Gambhir M, Cheng AC, McBryde ES, Denholm JT, Tay EL, et al. The role of super-spreading events in Mycobacterium tuberculosis transmission: evidence from contact tracing. *BMC Infectious Diseases*. 2019; 19(1):244. <https://doi.org/10.1186/s12879-019-3870-1> PMID: 30866840
37. Adegboye OA, Elfaki F. Network analysis of mers coronavirus within households, communities, and hospitals to identify most centralized and super-spreading in the arabian peninsula, 2012 to 2016. *Canadian Journal of Infectious Diseases and Medical Microbiology*. 2018; 2018. <https://doi.org/10.1155/2018/6725284>
38. Pozderac C. Python code used for analysis and figures; <https://github.com/calvinpozderac/COVID-19-Superspreading>