Preferential observation of large infectious disease outbreaks leads to consistent overestimation of intervention efficacy.

Jon Zelner^{1,2}, Nina Masters¹, Kelly Broen^{1,2}, and Eric Lofgren^{3,4}

¹Dept. of Epidemiology, University of Michigan School of Public Health

²Center for Social Epidemiology and Population Health (CSEPH), University of Michigan School of Public Health

³Paul G. Allen School for Global Animal Health, Washington State University

⁴Dept. of Math and Statistics, Washington State University

Abstract

Data from infectious disease outbreaks in congregate settings are often used to elicit clues about which types of interventions may be useful in other facilities. This is commonly done using before-and-after comparisons in which the infectiousness of pre-intervention cases is compared to that of post-intervention cases and the difference is attributed to intervention impact. In this manuscript, we show how a tendency to preferentially observe large outbreaks can lead to consistent overconfidence in how effective these interventions actually are. We show, in particular, that these inferences are highly susceptible to bias when the pathogen under consideration exhibits moderate-to-high amounts of heterogeneity in infectiousness. This includes important pathogens such as SARS-CoV-2, influenza, Noroviruses, HIV, Tuberculosis, and many others

Introduction

The SARS-CoV-2 pandemic has brought the implications of heterogeneous infectiousness and susceptibility for outbreak control to the forefront of public and scientific conversations around control of the COVID-19 pandemic (1,2). During the pandemic, outbreaks in congregate settings including cruise ships, jails, prisons (3), nursing homes, and long term care facilities have become key foci of SARS-CoV-2 infection and mortality. Superspreading, in which a small fraction of cases drives the bulk of transmission is common in such outbreaks. This is often the result of a confluence of factors in congregate settings including high rates of contact between individuals within the facility, concentration of highly susceptible individuals in a single location, and high viral load in a subset of infectious individuals. Data from such contexts are often used to elicit clues about which types of interventions may be useful in other facilities, using before-and-after comparisons in which the infectiousness of pre-intervention cases is compared to that of post-intervention cases and the difference is attributed to intervention impact. For example, this approach was used in an early and widely-cited paper suggesting that public health countermeasures during the SARS-CoV-2 outbreak on the *Diamond Princess* cruise ship was able to drive the on-board R_0 down from a value of nearly 15 to under 2, potentially preventing more than 2,500 cases (4). In this analysis, we explore the sensitivity of such inferences to the presence of heterogeneity in infectiousness, ranging from the moderate, exponential variation associated with classic SIR models to the significant overdispersion characteristic of superspreading.

Heterogeneity in SARS-CoV-2 transmission. Analyses dating to the beginning of the COVID-19 pandemic suggest that a large proportion of SARS-CoV-2 transmissions are likely attributable to a relatively small number of infectious individuals, with one study from Hong Kong suggesting that 20% of cases were responsible for 80% of observed transmissions(5), and similar dynamics observed in transmission in the U.S. state of Georgia (6). In a seminal analysis, Lloyd-Smith et al. (7)demonstrated how such heterogeneity in infectiousness was responsible for less-frequent but more explosive outbreaks characterized by high-yield super-spreading events (SSEs). This and other similar observations have led to suggestions that SARS-CoV-2 infection control should focus primarily on identifying the most infectious individuals and screening, isolating, and treating their exposed contacts. Despite the clear importance of understanding which interventions are effective at preventing and slowing SARS-CoV-2 transmission in congregate settings, little attention has been paid to the practical challenges of evaluating the efficacy of interventions meant to shorten or cut the ‘long tail’ of overdispersed transmissibility characteristic of the transmission of SARS-CoV-2 and other acute infectious diseases in congregate settings.

Figure 1: Flow diagram illustrating causation of observation by intervention.

In addition to well-documented heterogeneity in SARS-CoV-2 transmission, extensive variation has also been documented in the transmission of a wide variety of pathogens, including SARS-CoV-1 (8), Ebola (9), noroviruses (10), tuberculosis (11,12), and measles (13,14). Under the best-case scenario, superspreading is known to pose significant challenges for both the efficacy of intervention and the measurement of key transmission parameters (15,16). It has been noted before that selection bias is likely to compound these challenges, but this question has not been formally addressed.

Outbreaks cause interventions. Infectiousness heterogeneity poses numerous problems for infectious disease surveillance and control. In this analysis we focus in on the specific problem of observation bias. Simply put, outbreaks resulting in the application of interventions are likely to be significantly larger than those that do not. If we observe and analyze only the set of outbreaks requiring intervention, we are likely to overgeneralize the characteristics of this selected set of large outbreaks to the population as a whole. As the average value of the effective reproduction number, R , for a given pathogen at the population level declines, due to changing behavior, or the accumulation of naturally or vaccine-derived immunity in the population, analyzed outbreaks will necessarily become progressively less representative of the population as a whole, particularly for those pathogens exhibiting significant variation in infectiousness.

Pre-post comparisons may lead to erroneous conclusions. Interventions meant to arrest an ongoing outbreak are necessarily *reactive* in nature: An anomalous transmission pattern is detected, and interventions are applied in an effort to stop the outbreak ‘in its tracks’. Intervention effects are often measured using some kind of pre-post comparison, measuring change in average infectiousness prior to and after the institution of interventions. Previous work has shown that even with limited heterogeneity in infectiousness, such before-and-after comparisons routinely lead to overly optimistic estimates of intervention impact (17). This is because analyses of large outbreaks characterized by overt morbidity and mortality are far more likely than analyses of small, non-fatal outbreaks. Consequently, there is typically no pool of similar, uncontrolled outbreaks that can be used as controls to isolate intervention effects. Figure 1 illustrates how this bias towards analyzing the most severe outbreaks induces a causal linkage between the infectiousness of early cases and the application of a treatment. In other words, because interventions are not randomly assigned to outbreaks, there is a significant risk of *confounding by indication* (18) due to systematic differences between outbreaks receiving intervention and those that do not.

Nevertheless, the pre/post approach continues to be applied to the measurement of intervention effectiveness for a range of pathogens, including noroviruses (19). SARS (20,21), Hand, Foot, and Mouth Disease (22), pandemic influenza (23), Ebola (9), plague (24), and others. It is important to note that the results of many of these analyses are presented with the caveat that the effects of intervention may be difficult to disentangle from inherent variation in individual infectiousness, contact, etc. Nevertheless, for emerging infections such as SARS-CoV-2, there is often little choice but to rely on such pre/post comparisons. Our goal in the

present analysis is to use a transmission modeling approach to identify the situations in which such comparisons are most prone to bias, and to quantify the direction and magnitude of these biases.

Methods

For clarity, we will make a distinction between the population-level basic reproduction number, R_0 , which represents transmissibility throughout the entire population in the absence of immunity or intervention, and the setting-specific transmissibility, denoted R_s . This indicates the average transmissibility in a given setting, s , such as a jail, school, or nursing home, and may include the impact of measures such as social distancing that may reduce infectiousness, as well as accumulated immunity from natural infection and a potential vaccine.

Before-and-after effect estimates are typically made by estimating R_s^{pre} before, and R_s^{post} after the intervention. Implicit in this approach is an assumption that: 1) outbreaks selected for intervention and analysis are representative of the population of potential outbreaks, i.e. that $R_s^{\text{pre}} \approx R_s$, and that 2) $R_s^{\text{post}} \approx \rho R_s$, where $0 < \rho < 1$ is the reduction in transmissibility conferred by the intervention. Put another way, this framework implicitly assumes that the time-evolution of $R_s(t)$ is primarily a function of intervention, with values of an individual's infectiousness in setting s , $r_{i,s}$ representing simple random samples from the setting-specific distribution of infectiousness. In our analysis, we adapt the analytic approach in (17) to understand the implications of overdispersed, heterogeneous infectiousness for the evaluation of reactive interventions following the appearance of an anomalous cluster of cases.

Transmission model We use the approach developed by Zelner et al. to model heterogeneous infectiousness in a large norovirus outbreak (10), which builds on the framework of Lloyd-Smith et al. (7). As described above, each case has an *individual* reproduction number represented by a random variable, $r_{i,s}$ which follows a gamma distribution with dispersion parameter k_s and mean R_s , i.e. $r_{i,s} \sim \text{Gamma}(k_s, R_s/k_s)$ where R_s/k_s is equivalent to the shape parameters θ_s . This allows us to hold the expected value of R_s across all infectious individuals in a given setting constant while manipulating the variance: As $k \rightarrow 0$, the variance of individual infectiousness increases towards infinity, and as $k \rightarrow \infty$, the variance of $r_i \rightarrow 0$, i.e. all cases will have exactly the same infectiousness.

To allow for susceptible depletion, we consider a finite-population stochastic SIR model. To account for between-individual variation in infectiousness in the context of a mass-action model, we take advantage of the fact that the distribution of a sum of N gamma random variables with shape k and scale θ is Gamma distributed with shape kN . So, we draw the total force of infection on each day, $\lambda_s(t) = \sum_i \mathcal{J}(\tau_i = t) r_{i,s}$ from a Gamma distribution with shape parameter $k_s I_s(t)$ and scale parameter R_s/k_s , where τ_i is the onset time for individual i . So, the risk of infection to each susceptible individual on day t is equal $1 - \exp(-\lambda(t)/N)$. We also use this value to estimate the value of $R_s(t) = \lambda(t) / \sum_i \mathcal{J}(\tau_i = t)$.

Modeled outbreaks are started at $t = 1$ with the introduction of a single infectious individual into a population of 1000 people. Contact is assumed to be density dependent, i.e. the daily per-capita infectiousness of each case $\beta_i = \frac{r_i}{N}$. For simplicity, we assume that individuals are infectious only on the day they develop symptoms. Individuals infected on day t become infectious on day $t + 1$. We do this because models that estimate $R_s(t)$ typically report the mean infectiousness of cases with onset at day t . We utilize an SIR model rather than an SEIR model because of the potential for pre-symptomatic infectiousness among SARS-CoV-2 cases, and because the inclusion of a latent phase does not impact the overall outbreak dynamics, which are the focus of the current analysis.

Distinguishing regression toward the mean from intervention effects We hypothesize that *observed* outbreaks with true setting-specific reproduction numbers, R_s close to one will be disproportionately likely to exhibit larger-than-expected values of $R_s(t)$ in their early phases, followed by regression toward the mean value of R_s as the outbreak proceeds. When this coincides with the time an outbreak crosses a size threshold, c , that makes the outbreak severe enough to trigger intervention, the estimated protective effects attributed to the intervention may be large even when the intervention is completely ineffective.

Concretely, if the value of $R_s < 1$ in the absence of intervention, on average, the introduction of an infectious individual would be expected *a priori* to not result in an outbreak. When outbreaks of size $z > c$ occur, the sample mean of daily reproduction numbers prior to crossing the intervention threshold is will necessarily be greater than the mean at the beginning of the outbreak, outbreak size thresholds $c > 1$. This opens up the risk of what Gelman and Carlin term Type M errors of *magnitude* in analyses in which the estimator of a protective effect of an intervention is the change in the expected value of $R_s(t)$ before and after intervention. By conditioning on observing only outbreaks with final size $z \geq c$, we increase the likelihood of observing values of ΔR_0 that can be explained by regression toward the mean rather than the causal impact of intervention.

Simulation strategy To measure the probability that ΔR_0 will cross unity in the absence of intervention, i.e. the risk of misattribution of outbreak effects to regression toward the mean (RTM), we constructed a design analysis (25) with the following steps:

1. Select parameters $\theta = c, R_0, k$.
2. Introduce 1 infectious individual with r_i drawn uniformly at random from the true distribution of $r_{i,s}$ into a closed population of 1000 susceptible individuals.
3. Simulate the transmission process over a 14-day period 5000 times for each input parameter set.

Then, for the subset of simulated outbreaks with final size above the outbreak threshold, i.e. $z > c$:

1. Determine the day, ζ when the intervention threshold is crossed.
2. Calculate $R_s^{pre} = \sum_{i < \zeta} r_{i,s} / c$ and $R_s^{post} = \sum_{c \leq i < z} r_{i,s} / (z - c)$.
3. Estimate $\Delta R_s = R_s^{post} - R_s^{pre}$.
4. Calculate the proportion of outbreaks in which $R_s^{post} < 1$ and $R_s^{post} + \Delta R_s > 1$. This is the fraction of outbreaks in which a post-intervention drop in R_s which would be attributed to intervention is likely the result of RTM. This value is denoted $Pr(R_s^{pre} > 1, R_s^{post} < 1 | \theta)$.

In all simulations, we explore a range of R_s values from 2.5, representing the rates of uncontrolled SARS-CoV-2 transmission in a completely susceptible population, to values < 1 , suggestive of largely controlled transmission characterized by sporadic outbreaks.

Results

For misattribution of the impact of RTM to intervention effects to occur, the average value of $r_{i,s}$ prior to the triggering of interventions necessarily must be significantly greater than the mean value of R_s across all outbreaks. Figure 2 illustrates that this is the case across all levels of R_s and k included in our simulations, with the ratio of the observed $\hat{R}_s$ to the true R_s greatest for small values of k , with these differences becoming even more pronounced for higher values of R_s .

Figure 3 illustrates how these over-estimates may translate into large drops in R_s indicative of effective intervention may occur even if the true value of R_s is above 1, as in the analysis of the *Diamond Princess* SARS-CoV-2 outbreak in (4). Somewhat unintuitively, these effects are most pronounced for higher outbreak thresholds, greater heterogeneity, and larger values of R_s . These large drops and higher average pre-intervention infectiousness estimates can be explained by the fact that smaller outbreaks, i.e. those just above the threshold, are more common than very large ones and are more likely to be driven by a small number of early cases with disproportionately high infectiousness.

Finally, Figure 4 shows the proportion of outbreaks in which $\Delta R_s + R_s^{post} > 1$, i.e. when a drop in $R_s(t)$ suggestive of interventions contributing to the end of the outbreak is likely to be misattributed to regression toward the mean. This shows that for small values of the dispersion parameter - indicative of wide heterogeneity in infectiousness - the expected value of the setting-specific reproduction number can exhibit drops suggestive of effective intervention even when no intervention is applied $> 50\%$ of the time, and this is true even when $R_s > 1$. When the value of R_s is near the critical value of 1, the probability of misattribution approaches 100% of outbreaks regardless of the outbreak threshold used. For values of $0.5 < R_s < 1$, the probability of misattribution increases with R_s . This is a result of the very small probability of drawing a sequence of $r_{i,s}$ values that sum to greater than 1, with any outbreaks with values of $R_s < 1$ attributable to stochasticity in the transmission process rather than in the distribution of individual infectiousness. Notably, the outbreak size threshold has little effect on any of these results.

Discussion

Although our results are schematic in nature, they illustrate key challenges to estimating the impact of interventions on the course of outbreaks of acute pathogens such as SARS-CoV-2 in congregate settings. The finding that increasing amounts of heterogeneity are associated with misattribution of efficacy to ineffective interventions shows that the results of analyses by Cooper et al. (17) are only strengthened in the context of superspreading.

Figure 2: **Ratio of pre-intervention to true value of R_s .** Results are presented as a function of R_s and k_s for outbreak thresholds of 5, 10, 20 and 30 cases

Figure 3: **Average decline in R_s after crossing a pre-specified intervention threshold.** Results are presented as a function of R_s and k for outbreak thresholds of 5, 10, 20 and 30 cases.

Figure 4: **Probability of misattributing the end of an outbreak to interventions.** Results are presented as a function of R_s and k for outbreak thresholds of 5, 10, 20, and 30 cases

Another implication of these results is that cross-outbreak inferences must be made with significant care: The erroneous conclusion that an intervention was effective in an outbreak in which the true value of R_s was < 1 does not have direct implications for that outbreak, which was in truth self-limiting in nature. However, if this is taken as meaning that the intervention approach applied in that case was associated with a protective effect $\rho < 1$, this could have negative implications if this is used as a pretext for applying the same approach during an outbreak in which $R_s \gg 1$. For example, data from the Diamond Princess outbreak have been used to inform interventions in other congregate settings.

Similarly, just as the generalizability of inferences about infection control are limited by the highly-selected nature of such analyses, these results also show the potential peril associated with extrapolating backwards from single outbreaks to characterize the transmission characteristics at a population level. Although it is possible that the average infectiousness among pre-intervention cases on the *Diamond Princess* cruise ship was nearly 15, as suggested by Rocklöv et al. (4), our results show how even moderate amounts of heterogeneity can result in large overestimates of pre-intervention infectiousness, and can be characterized in large drops in the value of $R_s(t)$, even when the value of R_s is near the estimated population average value of 2-3 for SARS-CoV-2.

It is important that these findings not be interpreted as an argument against the use of before-and-after comparisons as a tool for characterizing intervention effects. Often, in rapidly emerging and changing situations, they are the only tool at hand. Instead, we argue for a more statistically rigorous approach to accounting for the selection bias and heterogeneity which may give rise to these overly-optimistic inferences. For example, conditioning on the probability of observing an outbreak of size $> c$ in the likelihood when fitting model parameters R_s and k_s and ρ_s would allow estimates to be adjusted for the observation bias at the center of these faulty inferences. One difficulty associated with this, however, and a weakness of the current analysis, is the exact threshold is typically unknown and is instead a function of the perceived severity of the outbreak among other factors that may be difficult or impossible to observe directly. Future analyses should examine whether the threshold can plausibly be treated as a nuisance parameter that is either estimated or integrated out during the process of model fitting.

The COVID-19 pandemic has shown clearly that transmission models are critical tools for addressing and interrupting infectious disease transmission in real time. At the same time, the pandemic has also revealed the serious public health consequences of model misspecification and the critical need to have a suite of validated tools at the ready to guide intervention during future events. Future work should focus on directly addressing the out-of-sample predictive utility of such models, which is the acid test of their utility as tools for guiding real-world infectious disease surveillance and intervention.

References

1. Liu Y, Eggo RM, Kucharski AJ. Secondary attack rate and superspreading events for SARS-CoV-2. *The Lancet*. 2020;395(10227):e47.
2. Lemieux J, Siddle KJ, Shaw BM, et al. Phylogenetic analysis of SARS-CoV-2 in the Boston area highlights the role of recurrent importation and superspreading events. *medRxiv*. 2020;2020.08.23.20178236.
3. Lofgren E, Lum K, Horowitz A, et al. The Epidemiological Implications of Incarceration Dynamics in Jails for Community, Corrections Officer, and Incarcerated Population Risks from COVID-19. *medRxiv*. 2020;2020.04.08.20058842.
4. Rocklöv J, Sjödin H, Wilder-Smith A. COVID-19 outbreak on the Diamond Princess cruise ship: Estimating the epidemic potential and effectiveness of public health countermeasures. *J Travel Med*. 2020;27(3).
5. Adam DC, Wu P, Wong JY, et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nature Medicine*. 2020;1–6.
6. Lau MSY, Grenfell B, Thomas M, et al. Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA. *PNAS*. 2020;117(36):22430–22435.
7. Lloyd-Smith JO, Schreiber SJ, Kopp PE, et al. Superspreading and the effect of individual variation on disease emergence. *Nature*. 2005;438(7066):355–359.
8. Shen Z, Ning F, Zhou W, et al. Superspreading SARS Events, Beijing, 2003. *Emerg Infect Dis*. 2004;10(2):256–260.
9. Althaus CL, Low N, Musa EO, et al. Ebola virus disease outbreak in Nigeria: Transmission dynamics and rapid control. *Epidemics*. 2015;11:80–84.
10. Zelner J, Adams C, Havumaki J, et al. Understanding the Importance of Contact Heterogeneity and Variable Infectiousness in the Dynamics of a Large Norovirus Outbreak. *Clinical Infectious Diseases*. 2019;ciz220.
11. Ypma RJF, Altes HK, van Soolingen D, et al. A Sign of Superspreading in Tuberculosis: Highly Skewed Distribution of Genotypic Cluster Sizes. *Epidemiology*. 2013;24(3):395–400.
12. Melsew YA, Gambhir M, Cheng AC, et al. The role of super-spreading events in Mycobacterium tuberculosis transmission: Evidence from contact tracing. *BMC Infect Dis*. 2019;19(1):244.
13. De Serres G, Markowski F, Toth E, et al. Largest Measles Epidemic in North America in a Decade Quebec, Canada, 2011: Contribution of Susceptibility, Serendipity, and Superspreading Events. *The Journal of Infectious Diseases*. 2013;207(6):990–998.
14. Seto J, Ikeda T, Tanaka S, et al. Detection of modified measles and super-spreader using a real-time reverse transcription PCR in the largest measles outbreak, Yamagata, Japan, 2017 in its elimination era. *Epidemiol. Infect*. 2018;146(13):1707–1713.
15. Garske T, Rhodes CJ. The effect of superspreading on epidemic outbreak size distributions. *Journal of Theoretical Biology*. 2008;253(2):228–237.
16. Blumberg S, Lloyd-Smith JO. Inference of R_0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. *PLoS Comput Biol*. 2013;9(5):e1002993.
17. Cooper B, Stone S, Kibbler C, et al. Systematic review of isolation policies in the hospital management of methicillin-resistant Staphylococcus aureus: A review of the literature with epidemiological and economic modelling. *Health Technol Assess*. 2003;7(39).
18. Salas M, Hotman A, Stricker BH. Confounding by Indication: An Example of Variation in the Use of Epidemiologic Terminology. *Am J Epidemiol*. 1999;149(11):981–983.
19. Heijne JCM, Teunis P, Morroy G, et al. Enhanced Hygiene Measures and Norovirus Transmission during an Outbreak. *Emerg. Infect. Dis*. 2009;15(1):24–30.
20. Wallinga J. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. *American Journal of Epidemiology*. 2004;160(6):509–516.
21. Krumkamp R, Duerr H-P, Reintjes R, et al. Impact of public health interventions in controlling the spread of SARS: Modelling of intervention scenarios. *International Journal of Hygiene and Environmental Health*. 2009;212(1):67–75.
22. Haydon DT, Chase-Topping M, Shaw DJ, et al. The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. *Proceedings of the Royal Society B: Biological Sciences*. 2003;270(1511):121–127.
23. Hens N, Calatayud L, Kurkela S, et al. Robust Reconstruction and Analysis of Outbreak Data: Influenza A(H1N1)v Transmission in a School-based Population. *American Journal of Epidemiology*. 2012;176(3):196–203.

24. Dean KR, Krauer F, Schmid BV. Epidemiology of a bubonic plague outbreak in Glasgow, Scotland in 1900. *Epidemiology*; 2018.
25. Gelman A, Carlin J. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspect Psychol Sci.* 2014;9(6):641–651.