Empirical Model of Spring 2020 Decrease in Daily Confirmed COVID-19 Cases in King County, Washington

Jared C Roach

Submitted to: JMIR Public Health and Surveillance
on: June 29, 2020

Table of Contents

Empirical Model of Spring 2020 Decrease in Daily Confirmed COVID-19 Cases in King County, Washington

Jared C RoachMD, PhD,

Corresponding Author:
Jared C RoachMD, PhD,
Phone: +12063245271
Email: jared.roach@isbscience.org

Abstract

Background: Projections of the near future of daily case incidence of COVID-19 are valuable for informing public policy. Near-future estimates are also useful for outbreaks of other diseases and other outcome measures. Short-term predictions are unlikely to be affected by changes in herd immunity. Both phenomenological and mechanistic models are valid approaches for epidemiology depending on the amount of available data and how much of the mechanism is known. Phenomenological models have fallen out of favor in recent decades.

Objective: To produce an empirical exponential model of the COVID-19 outbreak in King County, Washington during Spring 2020.

Methods: Confirmed case-count data from Public Health – Seattle & King County were fit to an exponential model with non-linear regression.

Results: Using an exponential model, incidence-count half-life (t1/2) is a more precise statistic than R because it does not require propagating uncertainty estimates from unmeasured variables. The parameters and predictions of this model remained stable for 77 days from March 26 to June 10, 2020 during a period of consistent public health policy.

Conclusions: In the absence of major net changes in factors that affect reproduction number (R), the two-parameter exponential model should be a standard model. This model had been standard for epidemiological analysis of pandemics since the early twentieth century but in recent decades has lost popularity to more complex compartmental models. An exponential model should be routinely included in reports describing epidemiological models as a reference, or null hypothesis. Exponential models should be fitted separately for time intervals that differ by any major changes in factors that affect R.

(JMIR Preprints 29/06/2020:21920)
DOI: https://doi.org/10.2196/preprints.21920

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Empirical Model of Spring 2020 Decrease in Daily Confirmed COVID-19 Cases in King County, Washington

Abstract

Background: Projections of the near future of daily case incidence of COVID-19 are valuable for informing public policy. Near-future estimates are also useful for outbreaks of other diseases and other outcome measures. Short-term predictions are unlikely to be affected by changes in herd immunity. Both phenomenological and mechanistic models are valid approaches for epidemiology depending on the amount of available data and how much of the mechanism is known. Phenomenological models have fallen out of favor in recent decades.
Objective: To produce an empirical exponential model of the COVID-19 outbreak in King County, Washington during Spring 2020.
Methods: Confirmed case-count data from Public Health – Seattle & King County was fit to an exponential model with non-linear regression.
Results: Using an exponential model, incidence-count half-life ($t_{1/2}$) is a more precise statistic than R because it does not require propagating uncertainty estimates from unmeasured variables. The parameters and predictions of this model remained stable for 77 days from March 26 to June 10, 2020 during a period of consistent public health policy.
Conclusions: In the absence of major net changes in factors that affect reproduction number (R), the two-parameter exponential model should be a standard model. This model had been standard for epidemiological analysis of pandemics since the early twentieth century but in recent decades has lost popularity to more complex compartmental models. An exponential model should be routinely included in reports describing epidemiological models as a reference, or null hypothesis. Exponential models should be fitted separately for time intervals that differ by any major changes in factors that affect R.

Keywords: reproduction number; reproduction ratio; reproductive number; reproductive ratio; COVID-19; SARS-CoV-2; half-life; SIR; compartmental model; epidemic.

Introduction

Projections of the near future of daily case incidence during epidemics are valuable for informing public policy as well as assuaging uncertainty in the minds of residents. Simple epidemiologic theory predicts an exponential change in viral incidence in a largely susceptible population, in which relatively few individuals are immune [1]. In an analysis of the 1918 influenza pandemic, Merritte Ireland reported, "If this was really the beginning of the great epidemic wave one should expect that if these series of data were plotted out on a logarithmic scale the increase from week to week would plot out as a straight line following the usual logarithmic rise of an epidemic curve" [2]. This basic insight that outbreaks of infectious disease should fit an exponential model remains valid a century later. An exponential model is simultaneously simple and mechanistic. There are a mere two parameters, yet these parameters enable modeling a biologically and epidemiological plausible mechanism. This mechanism is described by an initial case count and a fixed rate of transmission between individuals, and results in an exponential change in case counts. Furthermore, the parameters of the model are interpretable. In particular, the incidence-count half-life ($t_{1/2}$) has a specific, useful, and understandable meaning. Together, these five factors – historical acceptance, simplicity, plausibility, interpretability, and utility – make the exponential model best suited as the default model by which all other models should be compared.

Methods

Code

Code is provided as a supplemental file. All analyses were performed with R version 4.0.1 (R Foundation for Statistical Computing, Vienna, Austria; RRID:SCR_001905).

Data

Data is provided as a supplemental file. Data was acquired by transcription of the daily confirmed case counts daily from February 28 through June 17, 2020 [3]. Breakdown of daily data into subcategories including sex and ethnicity was not available in part due to privacy concerns related to potentially small aggregate bin sizes (personal communication from Public Health – Seattle & King County, May 28, 2020).

Starting Date

The model uses data starting with March 26, 2020. This model is robust to this choice. For example, for an analysis performed on May 11, choices of start date from March 26 to March 31 all result in a half-life between 24 and 25 days. Choices of later start dates produce longer half-lives, suggesting that the rate of decline may have been slowing, however, uncertainty in these estimates rises as they are derived from fewer data points. The choice of starting date (and end date, if specified) for data fit can be considered a parameter, so it is reasonable to consider the exponential model a three-parameter model – but if that choice is made for comparative purposes then other models should also have data-selection parameters enumerated. Extreme changes in initial date can change model predictions (e.g., Table 1, bottom two rows), and suggest a change in R across different time intervals.

Model Fit

The absolute residuals of the model are fairly constant and show little structure. This suggests that data deviating from the model fit is well described as “noise,” rather than some systematic effect, such as a change in public policy or public response to that policy (Figures 1-3). The residual on May 30 was one of the four highest residuals; it occurred seven days after Memorial Day and could represent a response to social gatherings on Memorial Day. However, the two flanking days do not have similarly high residuals, which would have been expected due to variations in viral and human behavior. There is also a mild weekly periodicity in the data, presumably corresponding to testing cycles. Therefore, smoothing over at least a week’s time is recommended; use of a two-parameter model achieves that goal.

Results

A model for King County during Spring 2020 is presented here as an example of exponential model fitting. The fit interval is bounded a start date representing widespread adoption of physical distancing public policy on March 26, 2020, with a data-freeze end date of June 17, 2020; spanning a time interval of 82 days. The March 26 bound was selected for two reasons. First, the reproduction number (R) of SARS-CoV-2 will have changed dramatically after physical distancing (aka, “social distancing”) policies were implemented in King County earlier in March. For example, Public Health – Seattle & King County (PHSKC) began encouraging physical distancing on March 10 and Washington State issued a “stay at home” directive on March 23. The inflection point of daily case counts of March 26 chosen as a bound was not caused by a rapid change in herd immunity. It was caused by a dramatic change in R secondary to intentional changes in human behavior. This cause is markedly different in character than the cause of the inflection point in most classical models, which is due to saturation of susceptible individuals. A May 2020 estimate of cumulative infections and therefore likely fraction of immune individuals in King County is only 2.1% [4], and this number has been slow to change. Therefore, an unmodified Susceptible, Infected and Recovered (SIR) model is not appropriate for this situation. Second, inspection of the raw data suggests an inflection point at or around this date. The highest confirmed case count to date occurred on April 1, but there is sufficient variability in the daily reported case counts that one cannot rule out a true peak of the curve occurring either several days earlier or later.

Nonlinear regression was performed on the daily confirmed case counts (Figure 1), with excellent fit ($R^2 = 0.87$). The equation for the fitted curve is, where N = predicted daily confirmed cases and t = days after March 26:

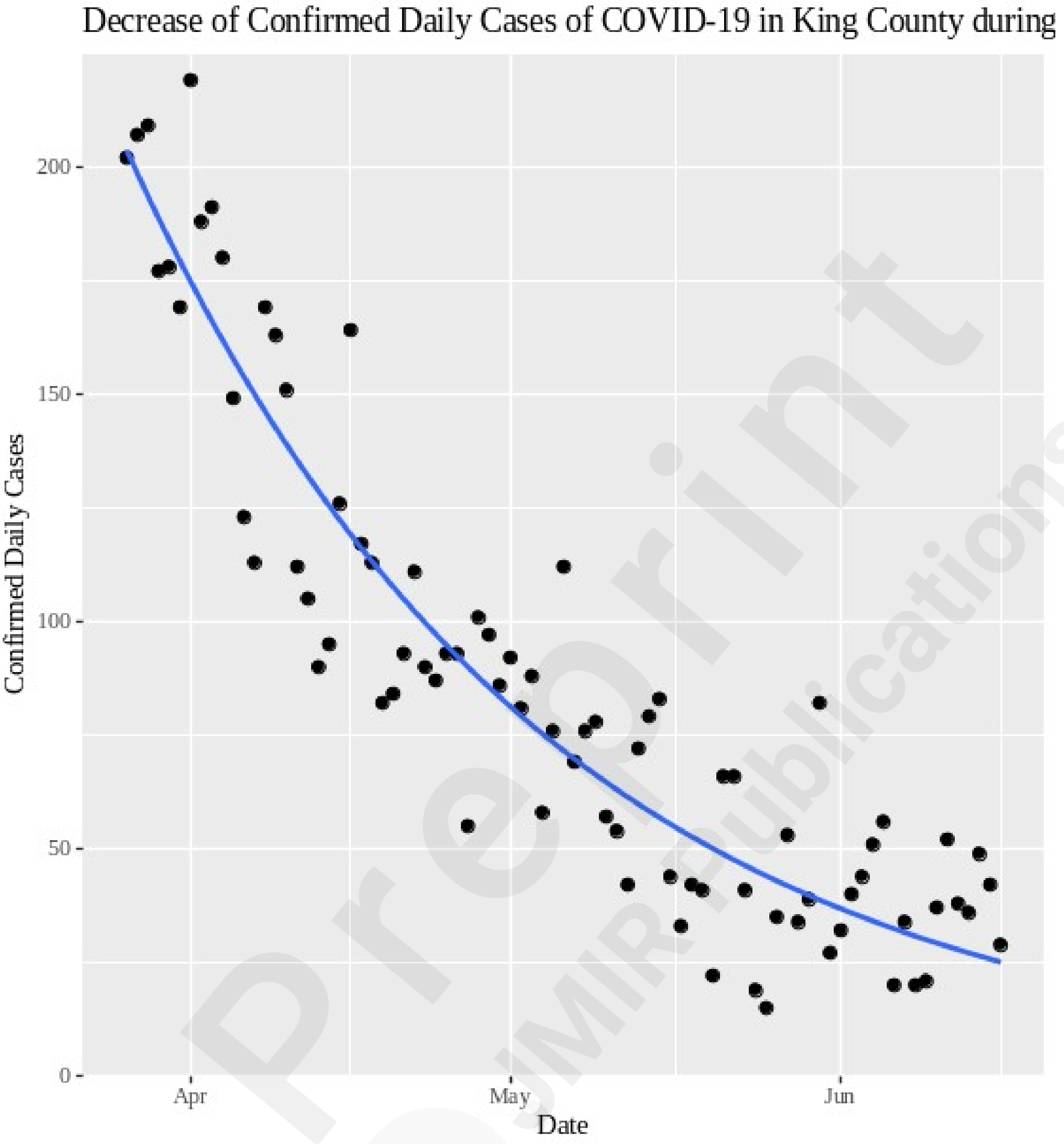
$$N=204\,e^{-0.025t} \qquad [1]$$

The two parameters of the model are the initial case count ($N_0 = 204$) and a rate constant ($\lambda = 0.025$). Consequently, the half-life ($t_{1/2}$) is

$$t_{1/2}=\frac{\ln 2}{\lambda}=\text{¿}\ 27.2\text{ days} \qquad [2]$$

Note that if the case count was increasing on a fitted interval, the half-life would be described instead as a doubling time. The King County model fit has a half-life ($t_{1/2}$) of 27.2 days (95% CI: 24.6—30.1). That is, the number of daily confirmed cases is expected to drop by 50% every 27 days. With approximately 80 cases observed on May 2, it predicted about 40 cases per day on May 29, 20 cases per day on June 25, and 10 cases per day on July 22, 2020. The Institute for Disease Modeling (IDM) model for COVID-19 assumes 4-day latency after infection followed by a period of 8-days of uniform infectiousness [5]. With these parameters, the model predicts R = 0.82 (95% CI: 0.80—0.83). The combination of simple model, large number of data points, and low residuals lead narrow confidence intervals for these analyses.

Figure 1. Exponential Fit to Model Decrease of Confirmed Daily Cases of COVID-19 in King County during Spring 2020. Each point is the number of confirmed cases of COVID-19 for each day as reported on June 17, 2020 by PHSKC. The date range is March 26 to June 16, 2020.



One test of a model’s utility is its ability to predict the future. Predictions from our model were first made on May 3 and ensconced in MedRxiv on May 11 [6]. The model parameters have remained stable and the predictions have largely been borne out. However, sub-analyses suggest that R has been increasing since March 26, slowly at first and more rapidly in recent weeks (Table 1). Alternatively, or in conjunction, increases in testing have enabled a higher ratio of confirmed to true case counts. It is difficult to establish from these data alone that these changes in R are significant (see Figure 3); they could result from a combination of progressive improvement in PHSKC’s estimates of previous days’ case counts, increased number of training data points, and noise.

Table 1. The model’s prediction of R is relatively stable when trained over all data starting with March 26. The model was first constructed on May 2, so predictions made on dates in April are hypothetical historical predictions; they were not prospective. If the model is trained only more recent data, starting with dates in May, longer half-lives are predicted, with R approaching 1.

Last Date Used to Train Model	Start Date Used to Train Model	Prospective Prediction	Number of Training Data Points	λ	Confirmed Cases Predicted on June 30	Half-life (days)	R
6/16/20	3/26/20	Yes	83	0.0255	17.6	27.2	0.815
6/10/20	3/26/20	Yes	77	0.0268	15.7	25.9	0.807
6/1/20	3/26/20	Yes	68	0.0271	15.3	25.6	0.805
5/20/20	3/26/20	Yes	56	0.0280	14.1	24.7	0.799
5/10/20	3/26/20	Yes	46	0.0283	13.7	24.5	0.798
5/2/20	3/26/20	Yes	38	0.0306	11.0	22.6	0.783
4/20/20	3/26/20	No	25	0.0349	7.4	19.9	0.756
4/10/20	3/26/20	No	15	0.0312	10.1	22.2	0.779
6/16/20	5/10/20	Yes	38	0.0146	26.6	47.4	0.890
6/16/20	5/25/20	Yes	23	0.0022	36.5	317.1	0.983

Discussion

Principal Results

Occam’s razor philosophizes that simple models are to be favored over more complex models unless there is compelling reason. Use of a simple historically accepted model limits temptation to shop for a model that fits a preconceived notion or policy. Simple models avoid overfitting to noise. Universal use of simple models increases comparability of results across studies; more complex models such as the SIR model are not as readily comparable as model specifications are not universal. For example, SIR models may differ in their inclusion of vital dynamics. In Spring 2020 in King County, there are no significant differences between the results of more complex models and a simple exponential model; the hypothesis that complex models are superior to the exponential model should be rejected.

This empirical model is sufficiently valuable that it should be used to inform public policy. Public policy should be crafted to keep R less than 1. If R is close to or greater than 1, more aggressive public policy measures are warranted. This model provides guidance on how close R is currently to 1. When R is less than 1, the length of the half-life allows appropriate expectations and allocation of resources. For example, if the number of cases declines by half only every month, it may make sense to increase public health measures and nonpharmaceutical interventions to speed the decline.

Limitations

A philosophy of this approach – the use of an exponential model to describe a disease outbreak – is to keep the model as simple as possible to avoid overfitting and to avoid assuming too much knowledge about the underlying processes and parameters of the outbreak. Major changes in factors that alter the biological plausibility of the model will invalidate its predictions. These will include rapid changes in herd immunity, such as occur near the peak incidence of an epidemic or in the wake of a change in public policy, such as “shelter in place”, that is widely adopted. Therefore, the exponential model is best used piecewise, with separate parameter fits to intervals punctuated by such events, as recommended by Faranda et al. [7], “Dynamical and statistical modeling should focus on limited stages of the epidemics and restrict the analysis to specific regions, thus accounting for large uncertainties” The model will work equally well in piecewise intervals with rising or falling daily case counts.

As this model intends to capture early indicators of outbreak progression, it focuses on reported cases counts, not deaths or hospitalizations. Deaths and hospitalizations, although often more reliably reported and so are used in models such as that of the IHME [8], are lagging indicators compared to incident case reports, but are excellent for predicting peak usage of resources.

Use of raw reported confirmed case counts to model the real incidence of COVID19 is subject to many caveats. They undercount true case incidence. However, if this factor is relatively constant, then the estimate of half-life and R is invariant. Confirmed case counts may not uniformly sample the population. The dynamics of the outbreak(s) in King County may be substantially different in different subpopulations within the county (e.g., herd immunity may be reached in a subset of long-term care facilities). Although they currently produce a good fit to an exponential model, the concurrence of parameters that sums to fit an exponential model may not persist. The characteristics of the population that receives tests vary over time. The number of tests performed may vary over time. The model does not include testing rate as a parameter; over the March to June time frame of the model, testing rate per day has not substantially altered in King County [3, 9, 10].

Comparison with Prior Work

The model presented here has consistencies with other data and models. In particular, reports by the Seattle Flu Study [11] and the Seattle Coronavirus Assessment Network [12] also show data that is consistent with an exponential decay in case rate in King County over the same period.

An excellent model for the COVID-19 outbreak in King County has been developed by IDM [4]. The IDM model is relied upon by both King County and Washington for public policy. Ideally an outbreak model should account for many hidden variables. These variables change over time. R is changing over time due to differences in population behaviors such as physical distancing. Likewise, the population is heterogenous, and includes subpopulations such as those in dense living situations. Each of these subpopulations may contribute uniquely to a real-world model; there is no guarantee that an encompassing systems model should fit an exponential. The IDM model attempts to capture many of these variables. Furthermore, it is parameterized with the premise that these variables can change rapidly, and produce rapid and large changes in the point estimate for R. Such rapid and large changes are not typically possible to model with a two-parameter model. An additional advantage of the IDM model is that it does not require piecewise fitting. Therefore, one would expect the IDM model to better capture very short-term fluctuations in R, but to be susceptible to overfitting. Conversely, a two-parameter model limits the possibility of overfitting and should be more robust to noisy data. Consistent with this expectation, the confidence intervals for the IDM and the exponential overlap across the entire trajectory, but the point estimates for R differ considerably. The IDM reported R=0.73 (95% CI: 0.3 to 1.2) for March 25; R=0.94 (95% CI: 0.55 to 1.33) for April 4; R=0.64 (95% CI: 0.28 to 1.0) for April 15; and R=0.89 (95% CI: 0.47 to 1.31) for April 22 [4,5,13]. The IDM models more volatility in their point estimate for R, varying from 0.94 to 0.64 and back to 0.89 over the course of three weeks. One can replicate this volatility by adding more parameters to the empirical modeling approach, such as fitting with a local smoothing algorithm (e.g., as shown in Figure 2). One possible interpretation is that the IDM model has some overfitting. Since the exponential model falls within the confidence interval of the IDM model, this inconsistency may not be relevant for informing policy, as long as confidence intervals are used for reporting data, and not point estimates.

The data and curve fit in Figure 1 best reflect a relatively smooth and continuous decline in cases coupled with a slowly increasing R. This is consistent with a society that is gradually decreasing effective social distancing, gradually increasing testing throughput, and gradually achieving pockets of herd immunity. These data and analyses are slightly inconsistent with an interpretation of rapid and substantial changes in R.

The average length (e.g., about 8 days) of the infectious period of COVID-19 is relatively short compared to the duration of modeled intervals (several months); the daily case rate should be roughly proportional to the prevalence. Therefore, estimation of the half-life using the method presented here is equally valid when applied to either prevalence or incidence. In some semantic contexts, it is more intuitive to attribute “half-life” to prevalence, which has a more tangible nature (“a *number* of people currently infectious”) than a daily case count which can be thought of as a *rate* with time as denominator, and therefore slightly less tangible.

Figure 2. Local curve fitting may either capture daily dynamics but more likely reveals overfitting to noise. Data points are identical to those in Figure 1. Curve is fit with the `geom_smooth` function (span=0.25), which produces a fit similar to some multiparameter models such as that of IDM. The increasing fit line at the end of June is consistent with a recently reported R=1.2 [16].

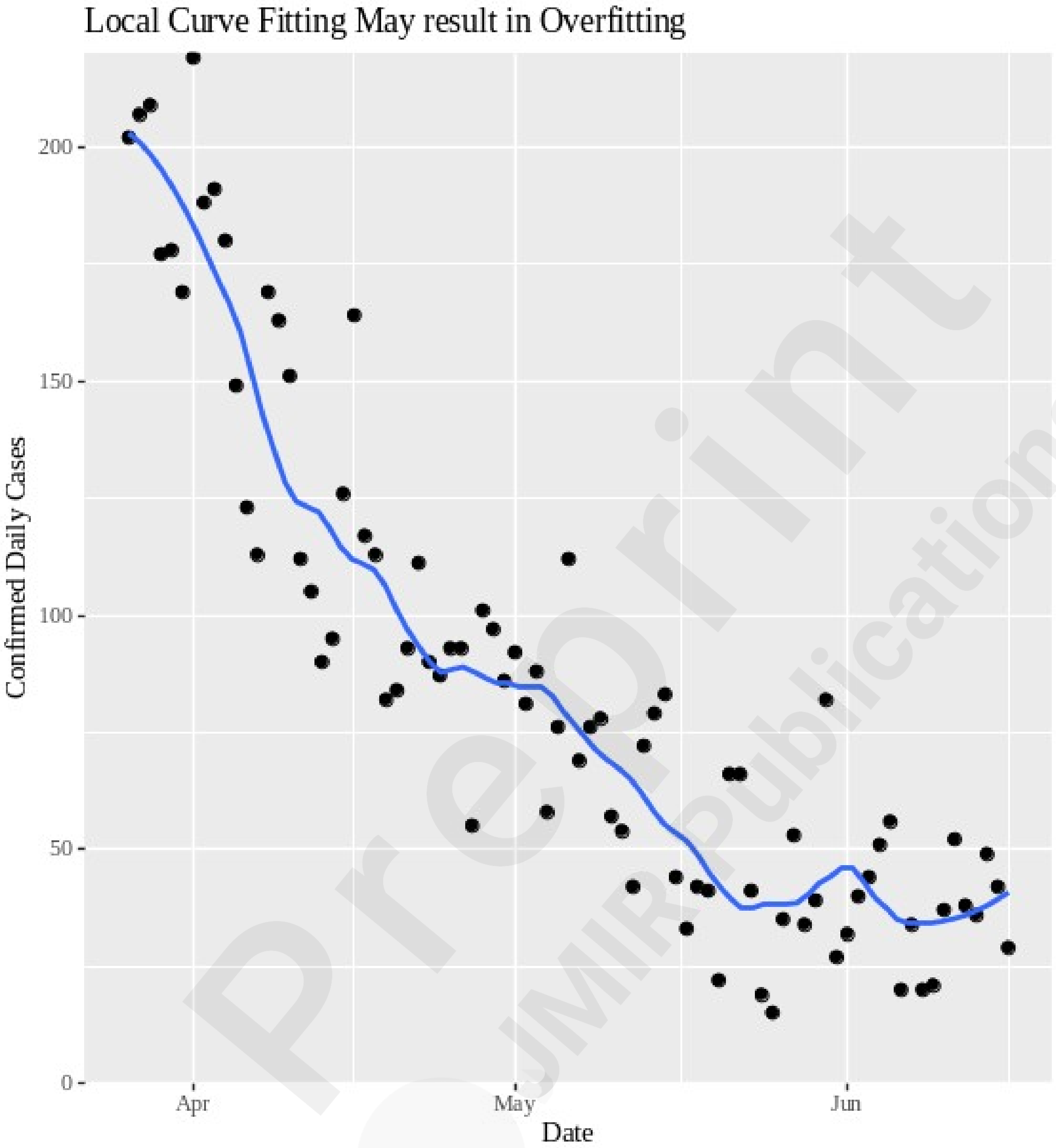
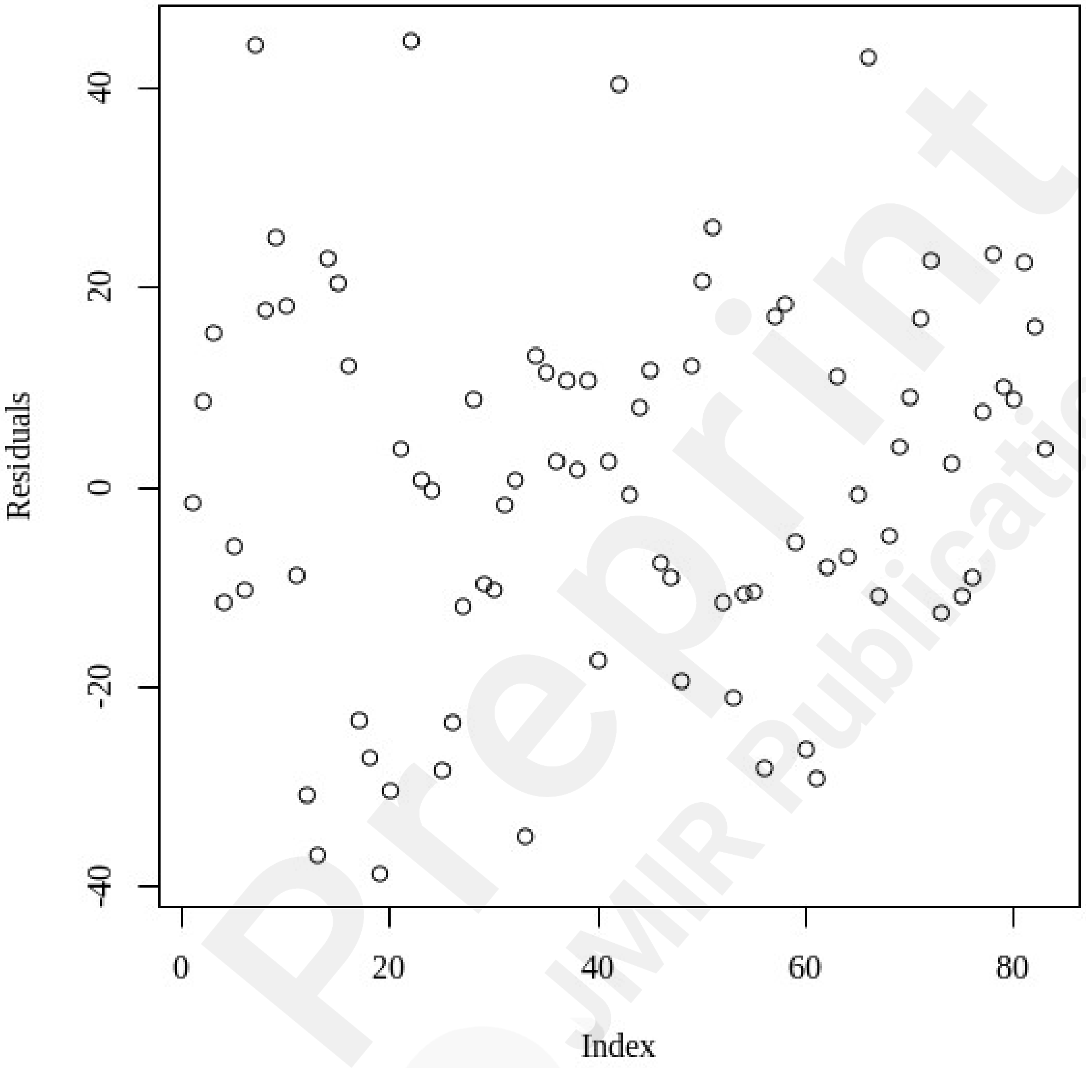


Figure 3. Model residuals show that there is little structure to the data, and no significant structure. Therefore, deviations of observed real values from model predictions are well described as noise, rather than attributed to failure of modeling assumptions.



Conclusions

The “incidence count half-life ($t_{1/2}$)” metric for outbreak modeling may be a more useful metric for communicating with the public than R_e or R_0. To date, R has largely been an academic statistic, and may be more useful for inferring details about the biology and mechanisms of transmission of a pathogen than for formulating public health policy. It is particularly awkward and counterintuitive that a parameter with a subscript of 0 (naught) is not a constant but rather is dependent on many observed and latent parameters. R is a mathematical paralog to the statistic ‘average lifetime’ (τ) used by physicists to describe radioactive decay. In that field τ is relatively deprecated compared to $t_{1/2}$; $t_{1/2}$ has many times greater usage in general and public communication than τ. Even if the relationship between R and half-life is modeled using a single parameter, the number of infectious days *(d)*,

$$R=e^{\frac{-d\ln 2}{t_{1/2}}}, \qquad [3]$$

A high uncertainty in R remains unavoidable because there is considerable uncertainty in *d*. This parameter *d* may vary between individuals and may vary considerably over time for each individual. The length of time an individual is infectious and how individual infectiousness varies over time and circumstances is not known. Therefore, there is more uncertainty for estimates of R than for half-life. Half-life depends only on the observed data; therefore, a good estimate of the uncertainty in half-life can be determined from the observed case-count data. This is not true for R. Furthermore, reproduction number parameters (e.g., R_e and R_0) depend on many other factors that they are typically incomparable between publications without expert interpretation and are neither well observed nor well known for SARS-CoV-2 [14, 15]. Since these factors must be estimated, a large uncertainty for R must be reported. This uncertainty in R overstates the amount of uncertainty relevant to at least some public health policy decisions. Also, due to the nonlinearity of the models, the confidence intervals for R can be asymmetric around the point estimate, increasing the likelihood of misinterpretations across a chain of public communication.

Note added in Proof

Following the data freeze analyzed in this report and the ensconcement of these predictions in MedRxiv [6], an inflection occurred in the trajectory of daily confirmed case counts in King County. The local minimum occurred between June 6 and June 16. This minimum occurred approximately 5-8 days (the expected lag time between changes in R and changes in confirmed case counts) after King County made a major public health policy change from “Phase 1” to “Phase 1.5” on June 5 [16]. One concludes that this several-month period from March 26 to June 10 between major public policy changes was well modeled by a simple exponential fit, and therefore that piecewise modeling of epidemics between major policy changes with separate exponential fits for each interval is a viable – and perhaps even the best – strategy, as long as the prevalence of herd immunity during these intervals is relatively constant.

Acknowledgements

Sui Huang and Noa Rappaport provided useful comments and criticisms.

Conflicts of Interest

None declared.

Abbreviations

COVID-19: Coronavirus disease 2019
IDM: Institute for Disease Modeling
PHSKC: Public Health — Seattle & King County
R: reproduction number
R^2: coefficient of determination
SIR: Susceptible, Infected and Recovered

References

1. Ross. An application of the theory of probabilities to the study of a priori pathometry. Proceedings of the Royal Society A. 1915.

2. Ireland MW, editor. The Medical Department of the United States Army in the World War, vol. IX: communicable and other diseases. Washington: U.S. Government Printing Office; 1928. p. 116-7.

3. Public Health – Seattle & King County. PHSKC website. www.kingcounty.gov/ depts/health/covid-19/data/daily-summary.aspx. 2020.

4. Thakkar, Burstein, Klein, and Famulare. Sustained reductions in transmission have led to declining COVID19 prevalence in King County, WA. Institute for Disease Modeling, Bellevue Washington. April 29, 2020.

5. Thakkar, Burstein, Klein, Schripsema, and Famulare. Physical distancing is working and still needed to prevent COVID-19 resurgence in King, Snohomish, and Pierce counties. Institute for Disease Modeling, Bellevue, Washington. April 10, 2020.

6. Roach JC. Empirical Model of Spring 2020 Decrease in Daily Confirmed COVID-19 Cases in King County, Washington. MedRxiv. May 20, 2020. doi=10.1101/2020.05.11.20098798 www.medrxiv.org/content/10.1101/2020.05. 11.20098798v1-3.

7. Faranda D, Castillo IP, Hulme O, Jezequel A, Lamb JSW, Sato Y, Thompson EL. Asymptotic estimates of SARS-CoV-2 infection counts and their sensitivity to stochastic perturbation. Chaos. 2020 May;30(5):051107. doi: 10.1063/5.0008834. PMID: 32491888; PMCID: PMC7241685.

8. IHME COVID-19 Health Service Utilization Forecasting Team, Christopher JL Murray. Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European Economic Area countries. medRχiv. doi: https://doi.org/10.1101/2020.04.21.20074732. 2020.

9. UW Virology COVID-19 Dashboard. depts.washington.edu/labmed/covid19. 2020.

10. Washington State Department of Health website. www.doh.wa.gov/ emergencies/coronavirus. 2020.

11. Chu HY, Englund JA, Starita LM, Famulare M, Brandstetter E, Nickerson DA, Rieder MJ, Adler A, Lacombe K, Kim AE, Graham C, Logue J, Wolf CR, Heimonen J, McCulloch DJ, Han PD, Sibley TR, Lee J, Ilcisin M, Fay K, Burstein R, Martin B, Lockwood CM, Thompson M, Lutz B, Jackson M, Hughes JP, Boeckh M, Shendure J, Bedford T; Seattle Flu Study Investigators. Early Detection of Covid-19 through a Citywide Pandemic Surveillance Platform. N Engl J Med. 2020 May 1. doi:10.1056/NEJMc2008646. PubMed PMID: 32356944; PubMed Central PMCID: PMC7206929.

12. Greater Seattle Coronavirus Assessment Network. SCAN COVID-19 Situation Report. publichealthinsider.com/wp-content/uploads/2020/05/SCAN_PUBLIC_ SITREP-30APR-5PM.pdf. Friday May 1, 2020.

13. Thakkar and Famulare. COVID-19 transmission was likely rising through April 22 across Washington State Institute for Disease Modeling, Bellevue, Washington. May 5, 2020.

14. Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. Complexity of the Basic Reproduction Number (R_0). Emerg Infect Dis. 2019;25(1):1-4. doi:10.3201/eid2501.171901.

15. Viceconte G, Petrosillo N. COVID-19 R_0: Magic number or conundrum? Infect Dis Rep. 2020 Feb 24;12(1):8516. doi: 10.4081/idr.2020.8516. PMID: 32201554; PMCID: PMC7073717.

16. King County. King County Safe Start Application: Moving from Modified Phase 1 to Phase 2. 2020. p.13.

Supplementary Files