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

Jared C. Roach
Institute for Systems Biology, Seattle, WA

ABSTRACT

Projections of the near future of daily case incidence of COVID-19 in King County are valuable for informing public policy. Projections made with the latest data may be better for some purposes than projections based on older data freezes of more reliable data. Incidence-count half-life ($t_{1/2}$) is a better statistic for many purposes than reproductive rate (R). Currently, an empirical fit to King County data predicts a half-life of 25.2 days (95% CI: 22.1—29.0), corresponding to R ≈ 0.8. The data and this curve best reflect a relatively smooth and continuous decline in cases coupled with a slowly increasing reproductive rate. With this model, the number of cases per day in King County is extrapolated to be 40 cases per day on May 24 and 20 cases per day on June 18, 2020.

INTRODUCTION

Projections of the near future of daily case incidence of COVID-19 in King County are valuable for informing public policy as well as assuaging uncertainty in the minds of residents. Any such projection, to be accurate, must account for many hidden variables. These variables change over time. One approach to best capturing the information in these variables is to use the most up-to-date, latest data. Although such data may be more incomplete and less reliable than an earlier data freeze, it may lead to a more useful model by better capturing the latest effects of dynamic changes in hidden variables. Over time, Public Health — Seattle & King County (PHSKC) improves previously reported confirmed case counts by removing duplicates, correcting residency information, adding newly received counts from previous dates, improving cause of death information, and other data cleaning. This creates a trade-off between using data that includes the last few reported days and using only older more reliable data.

An excellent model developed by the Institute for Disease Modeling (IDM) (Thakkar, Burstein, Klein, and Famulare, 2020) correctly recognizes that recent data reported by PHSKC may be updated in subsequent days by adding additional cases as they reach the database, and may be revised wholesale on occasion; therefore the IDM model does not include the most recent 10 days of PHSKC data. The model presented here is intended to complement the IDM model by incorporating these latest ten days of data. Therefore, one expects this model to better capture recent changes in underlying parameters, such as changes in population behavior or resolving flares in relatively isolated subpopulations, but at the risk of being more sensitive to issues associated with delayed case reporting and database maintenance. 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 more reliably assayed and reported and so are used in models such as that of the IHME (IHME COVID-19 Health Service Utilization Forecasting Team, 2020), are lagging indicators compared to incident case reports, but are excellent for predicting peak usage of resources.

To date, revisions to the PHSKC daily confirmed case reports fall into three categories: (1) additions to the last several days as new cases arrive in the database, (2) minor revisions to cases counts up to several weeks old, and (3) major revisions to the latest ten days of data. Revisions of the first type tend to have only a minor impact on estimates because they typically only have a noticeable magnitude of change for the last datapoint. They can also be adjusted statistically to add the expected proportion of delayed cases based on historically reported second-day adjustments. Revisions of the second type have negligible impact on estimates, as they seldom alter the counts for a given day by more than one or two counts. The third type of revision has happened only once, over the weekend, between May 2 and May 3, 2020; on May 3 a total of 105 cases counts were subtracted from the previous ten days, biased towards more recent dates, with over 20 counts being subtracted from previously reported counts for April 28, April 29, and April 30. This revision would have substantially altered the conclusions of the approach presented here and represent a major caveat to the interpretation and use of

these results; these results will be substantially misleading if major anomalies occur again. It is possible that this revision could have affected models with data freezes produced on May 2 (Famulare and Thakkar, 2020).

RESULTS

The model presented here is designed to predict the decrease in reported confirmed case counts using a simple mathematical function in order to avoid overfitting. It is intended for only short-term future extrapolations, on the order of weeks to a few months. Data before March 26, 2020 is not used in this model. The cutoff was selected for two reasons. First, the reproductive rate (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, PHSKC began encouraging physical distancing on March 10 and Governor Jay Inslee issued a "stay at home" directive on March 23. The cause of this inflection point – a dramatic change in reproductive rate to intentional changes in human behavior – is markedly different in character than the cause of the inflection point in most classical models (e.g., SIR), which is due to saturation of susceptible individuals; an unmodified SIR model is not appropriate for this situation. Second, inspection of the raw data suggests an inflection point at or around this date. Selecting a mathematical model to capture this inflection point would invite overfitting. The highest confirmed case count to date occurred on April 1, currently recorded as 215 case counts, 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. The analysis is robust to slightly different choices of threshold for data, including any date from March 25 to April 3, representing the height of the Spring 2020 COVID-19 outbreak in King County.

Simple epidemiologic theory predicts an exponential change in viral incidence in a largely susceptible population, in which relatively few individuals are immune (e.g., Ross, 1915). A current estimate of cumulative infections and therefore likely fraction of immune individuals in King County is 2.1% (Thakkar, Burstein, Klein, and Famulare, 2020). Therefore, it would make sense to estimate an exponential model case incidence. However, there are many factors in King County that would complicate such a simple model. The reproductive rate is changing over time due to differences in population behaviors such as physical distancing. Likewise, the population is unlikely to be homogenous, and may include 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 model should fit an exponential. Nevertheless, visual inspection of the raw data after March 26 suggests that an exponential fit would be excellent. Therefore, nonlinear regression was performed on the daily confirmed case counts (Figure 1), with excellent fit ($R^2 = 0.84$). The equation for the fitted curve is

$$\text{Daily confirmed cases} = 205\, e^{(-0.027 * \text{days after March 26})} \quad [1]$$

The model has a half-life ($t_{1/2}$) of 25.2 days (95% CI: 22.1—29.0). That is, the number of daily confirmed cases is expected to drop by 50% every 25 days. With approximately 80 cases observed on April 29, it predicts about 40 cases per day on May 24 and 20 cases per day on June 18, 2020. The IDM model (Thakkar, Burstein, Klein, Schripsema, and Famulare, 2020b) assumes 4-day latency after infection followed by a period of 8-days of uniform infectiousness. With these parameters, the effective reproduction rate (R) predicted by the model is 0.8 (95% CI: 0.77—0.82). This confidence interval for R falls within all reported confidence intervals from IDM. The IDM April 10 analysis reported R=0.73 (95% CI: 0.3 to 1.2) through March 25; the April 21 analysis reported R=0.94 (95% CI: 0.55 to 1.33) through April 4; the April 29 analysis reported R=0.64 (95% CI: 0.28 to 1.0) through April 15; the May 5 analysis reported R=0.89 (95% CI: 0.47 to 1.31) for April 22 (Thakkar, Burstein, Klein, Schripsema, and Famulare, 2020a; Thakkar, Burstein, Klein, Schripsema, and Famulare, 2020b; Thakkar, Burstein, Kleinand Famulare, 2020; Thakkar and Famulare, 2020). For comparing the dates between Figure 1 and the dates of IDM estimates of R, an offset of approximately one week should be used to account for the lag between initial infection and case report. One possible explanation for IDM's lower 0.64 estimate of R through

April 15 compared to the estimate of 0.80 in this report is the difference between incorporation of the most recent ten days of data, which at that time were consistent with a lessening of decline.

The philosophy of the approach here 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. However, a common concern in interpreting reported case data for public policy is that the fraction of reported cases compared to the number of true cases changes over time. Notably, as testing rate increases and encompasses more individuals with less severe symptoms, that fraction may increase. Testing rate has increased approximately 20% in King County between March 26 and May 17, 2020. The exact number of tests in King County not readily available. However, statewide, there has been about a thirty percent increase in testing over this period. Assuming that a relatively larger portion of this increase came in other counties with delayed outbreak onset, one can attribute a 20% increase in testing to King County. Figure 2 shows the results of adjusting the raw case counts for this change in testing rate, as well as adjusting the last two data points for delays in reporting. results in relatively little difference in parameter estimation. The expected half-life becomes 23 days (which was within the confidence interval previously). R^2 improves to 0.87 (from 0.84). This increased fit suggests that some or even a lot of the apparent "slowing" of physical distancing is explained by increased testing.

DISCUSSION

This empirical model is sufficiently valuable that it should be used to inform public policy. It is important for public health that R be less than 1. 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. By using fresh data, this model provides guidance on how close R is currently to 1. Current guidelines for modifying Washington State public policy (Inslee, 2020) include (1) sustained reduction in case counts for two weeks ("COVID-19 Disease Activity), and (2) readiness for contact tracing ("Case and Contact Investigations"). If indeed the number of cases declines by half only every 25 days or possibly longer – much longer if R is very close to 1 or becomes so as behavior changes, it may make sense to put less emphasis on waiting for the case count to reach a much lower level and more emphasis on increasing contact tracing and other public health measures to meet to the pandemic at a higher level of daily cases counts than may have been previously envisioned.

Use of raw reported confirmed case counts to model the real incidence of COVID19 is subject to many caveats. They undercount true case incidence, perhaps by a factor of ten. 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. If tests increase, then the true incidence may decline without a corresponding decline in confirmed incidence. The model does not currently include testing rate as a parameter. Indeed, over the March to May time frame of the model, testing rate per day has not substantially altered in King County (UW Virology COVID-19 Dashboard, 2020; PHSKC website, 2020; Washington State Department of Health website, 2020). Positive tests may be delayed by 1-2 weeks after infection. If an additional ten days embargo on the data is added, such as for the IDM model, there could be close to a month lag between a policy implementation and a statistically observable effect on the model. With extreme psychological, social, medical, and economic consequences of policy decisions, there is likely to be value in policy informed on the latest data, even if those data are subject to revision. However, policy based on just-in-time data such as the model presented here must be nimble enough to alter in the event of major model errors (like would have occurred during the week prior to 5/3/20). The data and curve fit in Figure 1 best reflect a relatively smooth and continuous decline in cases coupled with a slowly increasing reproductive rate. 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 reproductive rate that might drive conflicting back-to-back reports from King County Public Health such as those of May 4, "COVID-19 transmission has slowed," and May 8, "COVID-19 transmission rate could be rising in King County [...] after previous indications the transmission rate had fallen below a critical threshold" (King County, 2020).

The model presented here has consistencies with other data and models. In particular, recent reports by the Seattle Flu Study (Chu et al., 2020) and the Seattle Coronavirus Assessment Network (Greater Seattle Coronavirus Assessment Network, 2020) also show data that is consistent with an exponential decay in case rate. By design, this exponential model is smoother than other models. This design can create inconsistency with other models. The IDM models more volatility in expected 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 high-order polynomial (5^{th} degree or more) or employing a local smoothing algorithm (e.g., as shown in Figure 3), which shows a region of flat curve (i.e., $R \approx 1$) in the week before May 1. One possible interpretation is that the IDM models are overfitting. Fitting a curve to include pre-peak data may assuming parameters estimated in part from the early part (before the peak) of the curve are useful in predicting parameters relevant to the later part of the curve, such as physical distancing by the population. Another possibility is that the exponential model is not adequately capturing week-to-week variation in population-level physical distancing. Since the exponential model falls within the confidence interval of the IDM model, this inconsistency may not be relevant for informing policy.

Public policy should be to make as much data available to inform these just-in-time models as can reasonably be balanced against civil liberties and privacy considerations. In particular an understanding of what subpopulations are reporting case counts would considerably improve the value of these models for informing policy. It would be valuable for government agencies to produce and make available data including occupation and living environment to improve these models in a manner that appropriately protects privacy.

The "incidence count half-life ($t_{1/2}$)" metric for post-peak outbreak modeling may be a more useful metric for communicating with the popular press than R_e or R_0. The epidemiologic statistic "reproductive rate (R)" is a mathematical paralog to the physics of radioactive decay statistic "average lifetime (τ)". Indeed, for describing radioactive decay, half-life $t_{1/2}$ has many times greater usage in general and public communication than τ. Furthermore, reproductive rate parameters (e.g., R_e and R_0) depend on many other factors that they are typically incomparable between publications without expert interpretation. Even if the relationship between R and half-life is modeled using a single parameter, the number of infectious days *d*, as in

$$R = e^{(-d * \ln 2 / t_{1/2})}, \qquad [2]$$

there is considerable uncertainty in this number of infectious days. This parameter *d* may vary between individuals, and the intensity may vary considerably over time for each individual. Therefore, there is more uncertainty for estimates of R than with 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, which depends on a number of factors that are neither well observed nor well known for SARS-CoV-2. In particular the length of time an individual is infectious and how individual infectiousness varies over time and circumstances is not known. 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.

REFERENCES

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. [Epub ahead of print] PubMed PMID: 32356944; PubMed Central PMCID: PMC7206929.

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

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.

Inslee. Inslee rolls out COVID-19 risk assessment dashboard with data. https://www.governor.wa.gov/news-media/inslee-rolls-out-covid-19-risk-assessment-dashboard-data. Washington State Governor News & Media. April 29, 2020.

King County. Public Health news and blog. https://www.kingcounty.gov/depts/health/news.aspx. 2020.

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

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

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

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.

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, 2020a.

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. 2020b.

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

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

METHODS AND SUPPLEMENTARY NOTES

Data. Data was acquired by manual transcription of the daily confirmed case counts daily from February 28 through May 11, 2020; neither automated download of these data nor copy/paste functionality was available on this website (PHSKC website, 2020). These raw daily case counts as reported on May 18 are displayed as points in Figure 1. As of 5/18/20, the PHSKC main dataset, used here, only includes results based on PCR, not antibody testing.

Starting Date. This model is robust to choice of initial data date. 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 be slowing, however, uncertainty in these estimates rises as they are derived from fewer data points. For example, the half-life estimated from April 10 data and on is 32 days. If one ignores the last ten days of data, in an effort to replicate the IDM model, the model predicts a half-life of 23 days, corresponding to an R = 0.79. This also suggests a slowing of the rate of decline during the month of April and early May.

Exponential regression. There are at least two reasonable methods to fit an exponential curve to data. One approach is to first log transform the data and then perform linear regression (e.g., `lm(log(Cases)~Date)` in R). Another approach is to directly perform nonlinear regression (e.g., `nls(Cases~a*exp(b*Date))` in R). These two methods produce results with the current data with fits that are very similar (R^2 = 0.80 and 0.84 respectively). However, the second approach (used in the main text) tends to weight large absolute values of residuals slightly more than small absolute values. In this data set, that means that the `nls` curve fit is slightly better to the earlier part of the data and the `lm` curve fit is slightly better for the more recent data. With the `lm` fit rather one obtains a half-life of 25.9 days compared to a half-life of 25.2 days for the `nls` model. This is consistent with a slowing of the rate of decline over time.

Alternative models for regression. For a linear model, R^2 = 0.78. For a quadratic model, R^2 = 0.80. Either of these models would be adequate for most purposes of this paper. The linear model does not capture the apparent gradual slowing of the rate of decline. The quadratic model predicts a future inflection point (parabolic apex) without an intuitive basis. Neither linear nor quadratic models capture the essence of classic models: viral spread through a relatively invariant number of contacts per person per time, multiplied by a relatively invariant probability of disease transmission between susceptible and infectious individuals. Capturing this essence is not required of an empirical model, but since the current best fit is to an exponential curve, this happenstance allows one to capture this essence and more easily relate results to statistics such as R and half-life.

Utility of R as a public health policy tool when it hovers near unity. The half-life statistic is particularly useful for *post-peak* modeling. R is a more versatile and general statistic that has value in informing policy throughout an outbreak. It is particularly useful if the value of R hovers near 1, as there are profound implications for policy depending upon which side of unity the statistic lies. If R is less than but near 1, half-life is near infinity and is inelegant as a reportable metric. Half-life is much more useful if R is consistently somewhat less than 1, and in these circumstances is a key statistic for planning for healthcare demand and tempo of economic and social adjustments.

Estimation of R. A compartmental model was implemented to approximate the SEIR modeling approach of IDM (Thakkar, Burstein, Klein, and Famulare, 2020). The transmission rate was varied until the values of the compartmental model best fit the daily values predicted by the exponential model. This compartmental model is robust for the relationship between half-life and R with respect to most parameters, including population size, percent of population initially immune (as long as it is < about 12%), and non-infectious latent period after infection. However, this relationship, not surprisingly, it depends almost proportionally on the length of time a person is infectious. This is a difficult parameter to estimate or measure precisely, and results in high uncertainty for estimates of R. IDM uses 8 days, so to best compare results this parameter value is used here as well.

Sensitivity to revisions of initial days' reported case counts. It may be possible to assume that the number of cases initially reported for a given day contains enough information to adequately model the outbreak without relying on any subsequent revisions to this first-reported case count. To test how sensitive estimates were to ignoring some or all revisions, an exploration was performed using only the count after one to nine days of

revisions (Table 1). R^2 was only 0.60 for a fit based on first days only, but improved to 0.73 by the fourth days fit, and then gradually approached the best value of 0.84 using the full latest data. Because most of the data points in the model have benefitted from more than 2 days of revisions, the model's predictions would not be substantially impacted if typical revisions to the data were known in advance. Therefore, there is little risk in using the latest data in models to assist in informing public policy.

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 May 18, 2020 by PHSKC. The first date employed for this model is March 26, 2020; the last date used is May 17, 2020, the last reported date as of this writing. There is less uncertainty to this curve fit than for an estimation of reproductive rate (R), because an inference of R requires estimation of additional parameters. The half-life for the incident case count is 25.2 days (95% CI: 22.1—29.0).

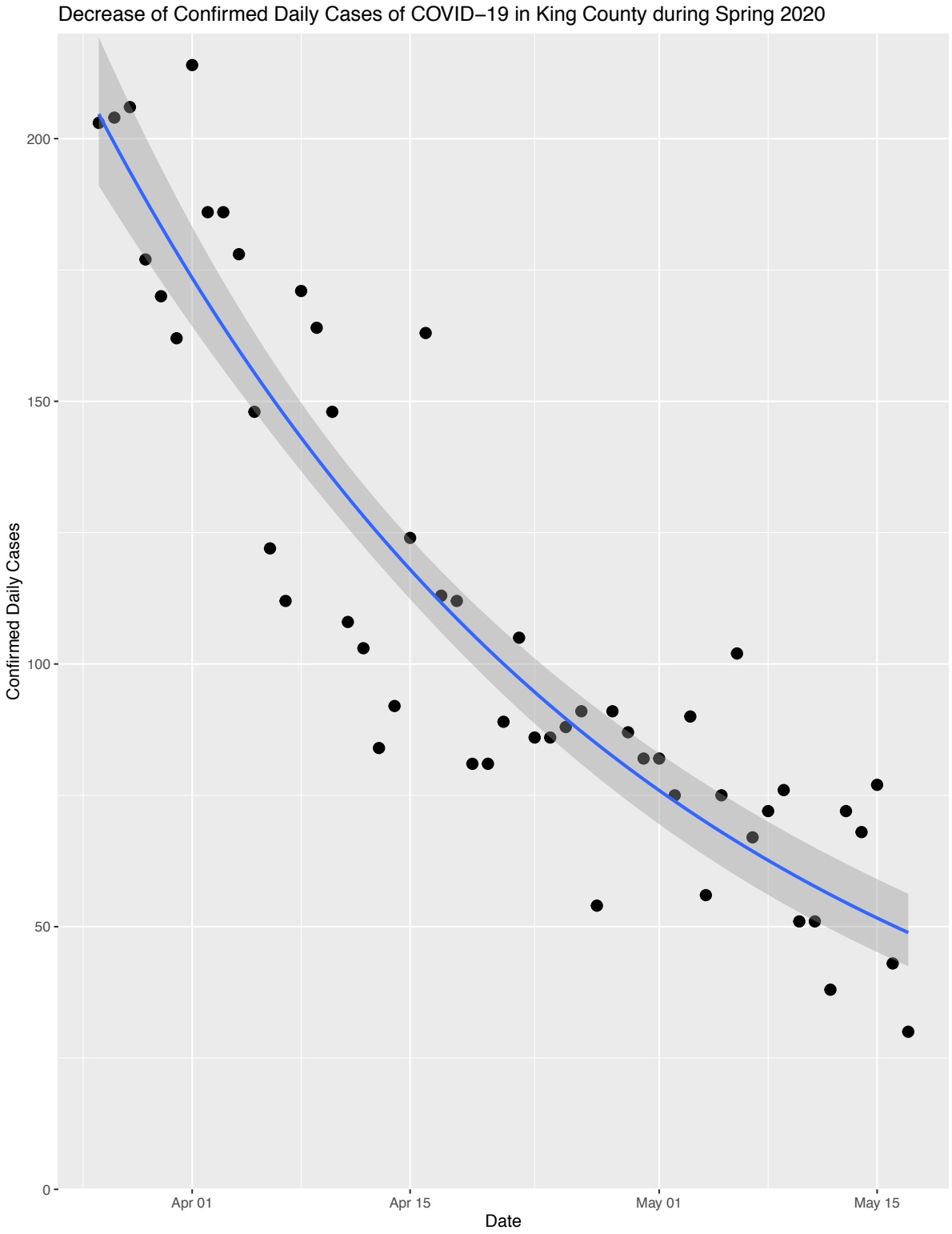

Figure 2. Raw data are adjusted for (1) testing rate and (2) delayed case reporting over the last two days. Testing rate is estimated to have increased 20% since March 26. The most recent days' case reports average 25% less than final counts for that day, and the penultimate days' case reports average 11% less. With this adjustment, the curve fit improves (R^2 = 0.87), and the expected half-life becomes 23 days.

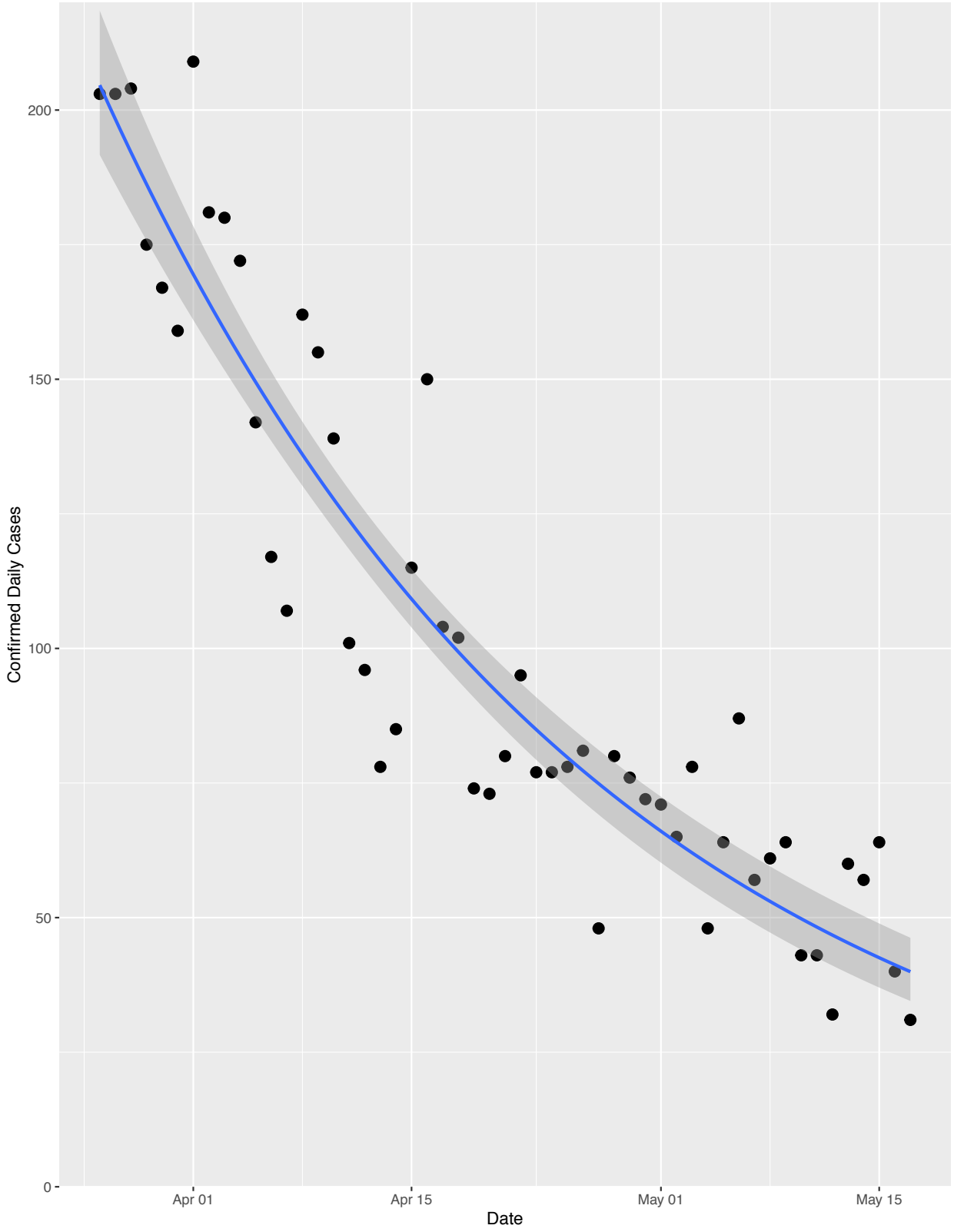

Figure 3. Local curve fitting may either capture weekly dynamics in response to public policy and/or reveal overfitting. With this fit, the decline in case rate appears to stall during the last week of April.

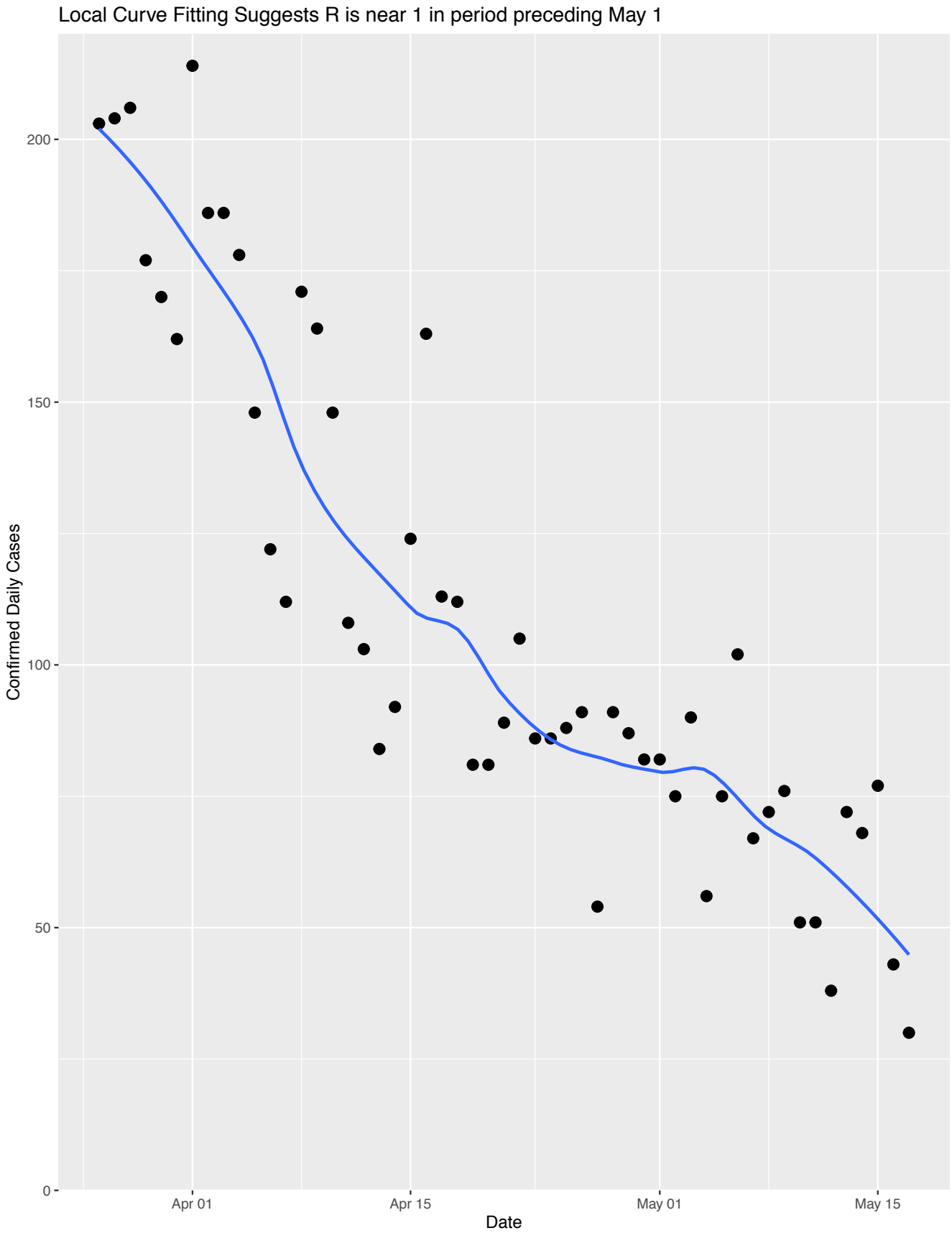

Table 1. *Sensitivity to revisions of initial days' reported case counts.*

Restriction of data revisions to *n*th day after first report	**R^2**	**$t_{1/2}$ (days)**
First (initial)	0.60	31.8
Second	0.62	34.6
Third	0.62	34.5
Fourth	0.73	29.6
Fifth	0.75	29.3
Sixth	0.78	28.4
Seventh	0.80	27.4
Eighth	0.80	26.7
Ninth	0.81	26.0
Tenth	0.81	25.9
Latest	0.84	24.9