An intuitive heuristic for interpreting COVID-19 test positivity and cases per capita

Authors: Roy Burstein, Mike Famulare, Niket Thakkar
Reviewers: Jen Schripsema
Institute for Disease Modeling, Seattle, Washington; covid@idmod.org
Results as of January 19, 2021 10:00 am

What do we already know?

Since April 2020, we have been using compartmental transmission modeling approaches to better understand COVID-19 epidemiology in Washington state. Using models as a platform, we align observed cases, hospitalizations, and mortality into a consistent epidemiology, allowing us to describe the underlying transmission dynamics in terms of population prevalence, cumulative incidence, and effective reproductive numbers. Recently, we have extended this approach to jointly estimate weekly prevalence across 15 regions in Washington state.

What does this report add?

In this report, we build intuitive heuristics for estimating COVID-19 prevalence using more readily available data than that needed to model transmission dynamics. Specifically, we show how test positivity, cases per capita, and testing volume straightforwardly relate to underlying infection rates via case detection rates, hidden quantities of public health interest that generally need to be estimated with more data intensive transmission-model-based inference.

We use data and estimates from Washington state's regions to show that there are generalizable relationships among these quantities that can be used to build a simple mapping between easily observed metrics, such as test positivity and cases per capita, and unobservable measures of transmission risk and burden. We develop a simple, practical, and interpretable prevalence approximation tool and also build intuition for its potential adjustment to adapt to specific local contexts. We tested the simple tool on observed data from all 50 US states, showing high accuracy.

What are the implications for public health practice?

Epidemiological variables, such as prevalence, are not directly observed and typically require complex modeling or well-designed surveillance systems to estimate, neither of which are often available. Test positivity and cases per capita are accessible metrics reported by local health departments in near real-time. We identify four multipliers that can be easily applied to readily available metrics to quickly estimate prevalence. For example, if testing is widespread such that over 1% of the population is being tested each week and nothing else is known about case detection rate, then prevalence can be roughly approximated by multiplying cases per 100,000 in the past seven days by 0.003.

Executive summary

- This report describes an interpretable heuristic that allows you to roughly estimate COVID-19 prevalence using the readily available metrics of test positivity and cases per 100,000. This is a highly simplified tool and only meant for rough approximations and to aid in interpretation. This should not serve as a replacement for more detailed modeling of COVID-19 transmission, if such modeling is available.
- The relationships between prevalence and test positivity and between prevalence and cases per capita are mathematically straightforward, but depend on case detection ratio (cases/infections), which is also unobserved:
    - Prevalence = test positivity * testing volume / case detection ratio
    - Prevalence = cases per capita / case detection ratio
- We have found that these relationships tend to follow expected patterns. In particular, case detection rates stabilize to approximately one detected case representing three infections above a certain level of testing. For example, if a location has had good testing, and found 200 cases in the past week, we would estimate that there have been about 200*3=600 active infections in the past week.
- Thus, we are able to suggest the following simple rules which use test positivity and cases per capita to *approximate* prevalence:
    - If 1% or more of the population is being tested each week (this is the situation in nearly all places today):
        - Current prevalence is approximately test positivity times 0.048.
        - Current prevalence is approximately cases per 100,000 times 0.003.
            - *Alternatively, the number of total infections over the past week is cases multiplied by 3.*
    - If less than 1% of the population is being tested each week:
        - Current prevalence is approximately test positivity times 0.031.
        - Current prevalence is approximately cases per 100,000 divided by 0.0043.
    - If testing volume is lower than 0.6% of the population (6 per 1000), the above rule will yield unstable estimates, and is not recommended for use.
- This ruleset is based on average observed relationships in Washington state. It is only for approximation, but it can be adjusted easily for specific settings if there is strong belief that the average case detection rate differs significantly from the Washington average. For example, if testing is very high, or the contact-tracing effort is known to be effective, case detection ratio will be better than 3. In Washington, for example, we find the case:infection multiplier when testing is >1.8% is closer to 2.3.
- Testing of this tool outside of Washington indicates that the relationships are broadly generalizable, and thus can be used in other locations in the US. It was not tested outside the US, but could be used in locations where case detection rates are expected to be similar.

- This tool can reduce latency in situational awareness arising from lag-time time in reporting total tests, deaths, and hospitalizations, as at minimum it only requires cases, which are often reported in near real-time.

Introduction

Aggregated results of population testing are commonly used to inform policy-makers and the public about the local COVID-19 epidemiological situation. Moreover, metrics such as test positivity (positive tests over total tests) and cases per capita are also used in decision-making around public health control measures. For example, New York City used the '3% rule' as a decision threshold for closing schools, and the recently announced 'Healthy Washington - Roadmap to Recovery' reopening plan uses a test positivity threshold of 10% as part of the criteria to advance phases.

There are many such examples to point to, but generally these metrics are being used as readily available stand-ins for population prevalence, the proportion of the population currently infected with COVID-19, which fundamentally describes community risk and can forewarn possible surges to the healthcare system. Underlying prevalence cannot be measured directly, and requires either complex, data-intensive transmission modeling or large-scale population-based surveillance systems to accurately estimate.

In this report, we use outputs from our transmission model across 15 regions in Washington to better understand the relationship between two readily available metrics, test positivity and cases per capita, and COVID-19 prevalence. The link between these metrics and prevalence is the case detection rate, the fraction of infections that eventually test positive and get reported to the health system, another underlying epidemiological quantity that cannot be measured directly. We find that once testing has exceeded roughly 1% of the population per week, case detection rates stabilize, and accessible metrics become highly predictive of underlying prevalence. Testing is now above this level for most Washington regions and most US states. Along those lines, we show how assumptions about the case detection rate based on observations in Washington can be extrapolated to predict state-level mortality across the US.

Key inputs and assumptions

- We use COVID-19 testing and case data collected by the Washington Department of Health (WA DoH) through the Washington Disease Reporting System (WDRS). Data are aggregated into time series by specimen collection date. We limit this analysis to data collected between May 1, 2020 and November 22, 2020. Before May, testing capacity and utilization were limited in ways specific to the start of the epidemic and are not generalizable to today's situation. Starting in late November, WA DoH temporarily stopped reporting all negative test results—impacting estimates of test positivity and testing volume.
- Weekly prevalence and reporting rate were estimated via the COVID-19 transmission model described in our previous report. Briefly, we fit a compartmental model to testing, hospitalization, mortality, and age-distribution data at the state level to estimate daily infections in Washington as a whole. Those infections are then distributed across 15 geographic regions to

best explain weekly trends in region-specific hospital admissions. Critically, region-specific testing data is not used for model fitting, facilitating the comparisons and inferences presented in this report. We used median posterior estimates from the model.

- As inferences in this report are based on our transmission model, assumptions from our modeling report are relevant here as well. In particular, a key underlying assumption for our model is that population prevalence varies continuously in time. This necessarily creates difficult-to-quantify spatial resolution limitations that apply to our heuristic approach. Our analysis was developed and tested on regions with a population size 100,000 or larger. We expect this approach to work at smaller areas - such at the city scale, but expect it to break down at very small areas.
- Weekly epidemiological estimates are produced for each of the 15 regions defined in our previous report, and merged with each region's weekly case counts, test positivity, and testing volume data from the WDRS. In total, the compiled dataset includes 30 weeks, or 450 region-weeks.
- US state-level data on total tests, cases, and deaths were sourced from the Covid Tracking Project and downloaded here.
- Population estimates for Washington regions and US states were obtained from the US Census Bureau.
- The following terms are used throughout this report. Unless otherwise stated, all are expressed on a weekly time-scale:
    - Prevalence: proportion of the population infected with COVID-19 at PCR-detectable levels. We assume PCR-detection lasts for 12 days directly following a 2-day latent period starting at the time of exposure. Typically expressed as percentage. Model-based estimate.
    - Reporting rate: cases detected per infection. Also sometimes referred to as case-detection rate. Note that the denominator is PCR-detectable infections *per week*, and thus this is not directly interpretable as the proportion of cases that are detected unless adjusted for duration (assumed 12 days). Typically expressed as a percentage. Model-based estimate.
    - Test positivity: the proportion of tests returning a positive result. Typically expressed as a percentage. Directly observed.
    - Testing volume: tests per population. Typically expressed as tests per 1000. Directly observed.
    - Cases per capita: cases per population. Typically expressed as cases per 100,000 or per 100. Directly observed.

Building intuition for the relationship between prevalence and key metrics

It's useful to start by defining the mathematical relationships among the above variables. Specifically, we see that

$$\text{Prevalence} = \frac{\text{Test Positivity} \times \text{Testing Volume}}{\text{Reporting Rate}},$$

or in other words,

$$\frac{\text{Infections}}{\text{Population}} = \frac{\text{Cases}}{\text{Tests}} \times \frac{\text{Tests}}{\text{Population}} \times \frac{\text{Infections}}{\text{Cases}},$$

Thus, the relationship between test positivity and prevalence is modified by testing volume and reporting rate. Furthermore, by rearranging the above, we also see that

$$\text{Prevalence} = \frac{\text{Cases}}{\text{Population}} \times \frac{1}{\text{Reporting Rate}},$$

an even simpler relationship that does not depend on testing volume.

We can use estimates from our transmission model and data on testing to explore how these metrics tend to covary in practice. **Figure 1** shows the pairwise relationships among each of the five metrics (prevalence, test positivity, cases per capita, testing volume, and reporting rate), as observed across the 450 region-weeks in Washington. Test positivity and cases per capita are highly correlated with prevalence (Pearson correlation coefficient: 0.90 and 0.91, respectively), with the remaining variation explained by reporting rate and testing volume. Test volume and reporting rate are also shown to be well correlated (0.65), which indicates that testing volume may serve as a proxy for the unobservable reporting rate, as we would expect intuitively.

From the equations above, we can see that understanding the variation due to reporting rate is critical for developing and interpreting a generalizable heuristic that maps between observable metrics (test positivity and cases per capita) and prevalence. If reporting rate varies unpredictably, then we would expect a highly variable relationship between the observable metrics and prevalence. As implied in Figure 1, this is unlikely the case, as the two observable metrics of interest and prevalence are well-correlated on their own. Despite this, some variation remains, warranting a deeper investigation of reporting rates.

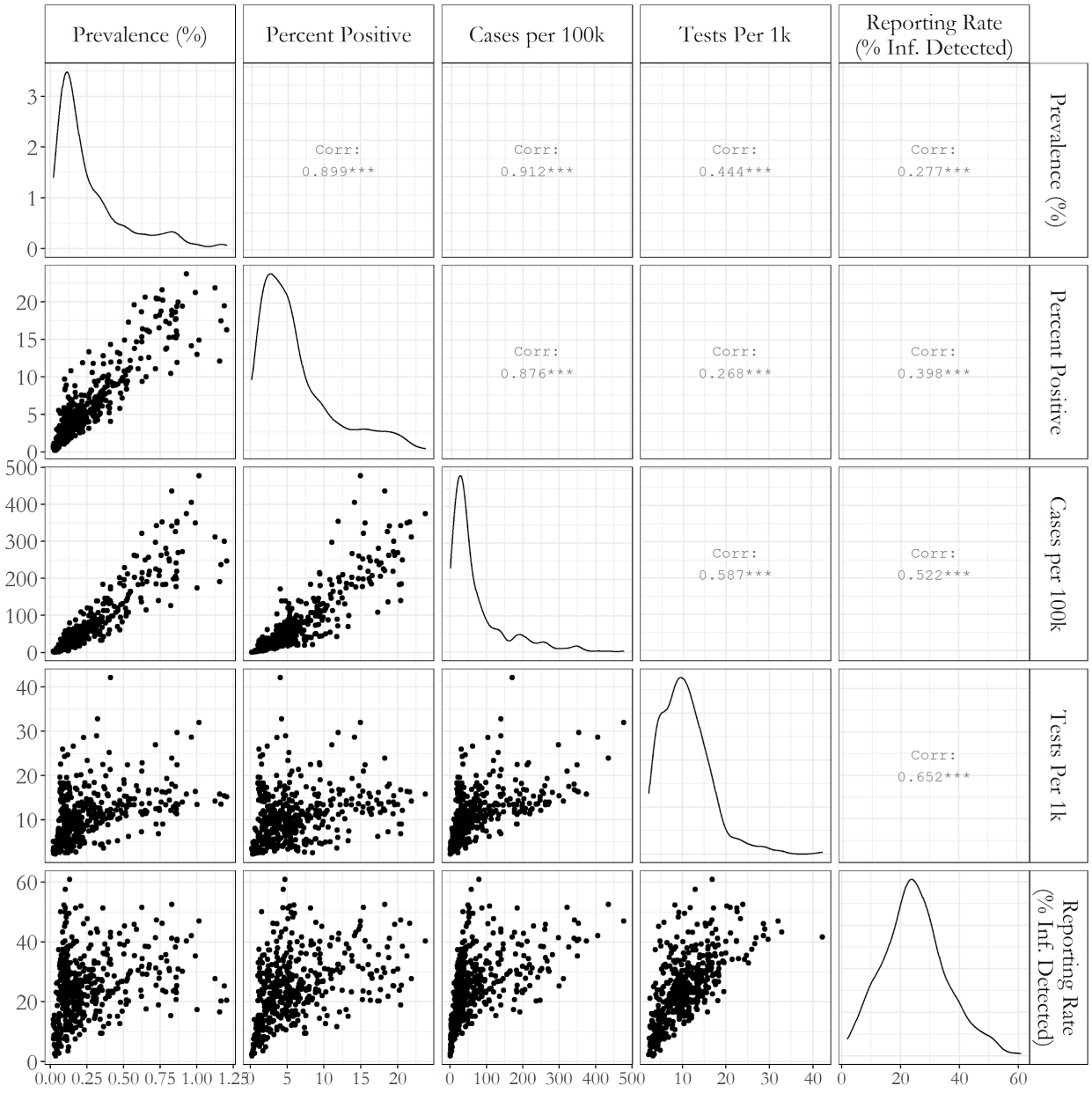

Figure 1: *Relationships between observed and estimated epidemiological metrics across 450 region-weeks in Washington state.*

Figure 2 shows the relationship between infections per case (the inverse of reporting rate) and testing volume. As testing volume increases, so do the proportion of infections caught by the testing system: the average weekly infections per case drops from 14 at a level of 4 or fewer tests per 1000 to about 2.5 at a level of 18 or more tests per 1000, with average gains in reporting rates slowing at about 10 tests per 1000. Furthermore, variation in reporting rates is much higher when testing is below 10 per 1000. As shown in **Figure 3**, this improvement has happened gradually over time in Washington, with the highest testing volume, highest reporting rates, and lowest variation in reporting rates observed in November, at the end of the time period considered in this report. As such, we expect observable metrics to be

more closely tied to prevalence when testing is better, and in Washington's case, for this to have been improving over time.

Figure 4 shows how test positivity and cases per capita respectively relate to prevalence across strata of testing volume. Within each strata, we fit a linear regression with no intercept, and report the slope (β) and explained variation (R^2). We find that linear models fit well within each strata (minimum $R^2$are 0.84 and 0.82, for test positivity and cases per capita, respectively). This confirms that indeed test volume serves well as a proxy for reporting rate. Furthermore, we see that R^2tends to improve as testing volume increases, which is due to narrowing variation in reporting rates as testing improves. The regression slopes represent the modification of each metric within each strata needed to estimate prevalence. For cases per capita, the slope very simply represents the average of the inverse of reporting rate (RR^{-1}= infections per cases) within each strata. For test positivity, the slope represents testing volume per capita / reporting rate. Since reporting rate increases sublinearly with respect to testing volume (Figure 2), the fitted slopes for test positivity increases with increasing testing volume. Both for test positivity and cases per capita, slopes tend to level out above about 10 tests per 1000.

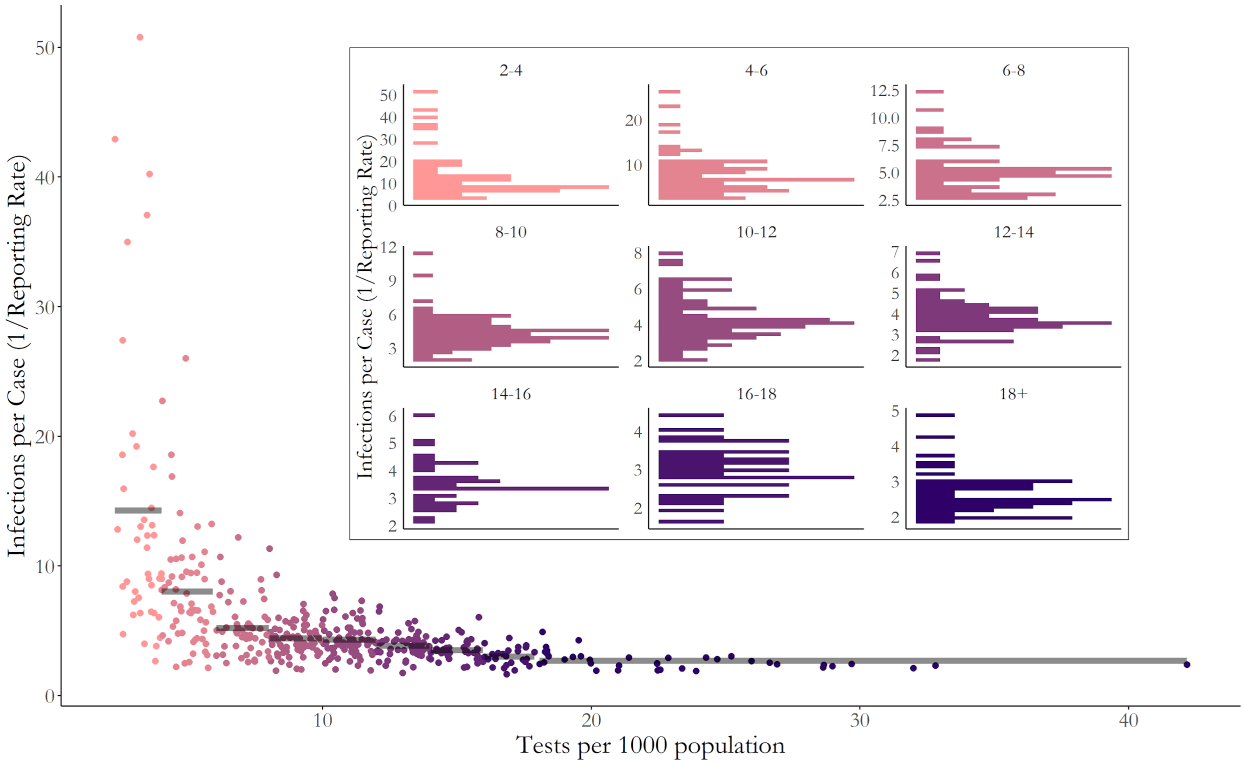

Figure 2: *The relationship between reporting rate and testing volume. Black lines represent mean values at each level of testing volume. Inset shows the distribution of inverse reporting rate at each level of testing volume.*

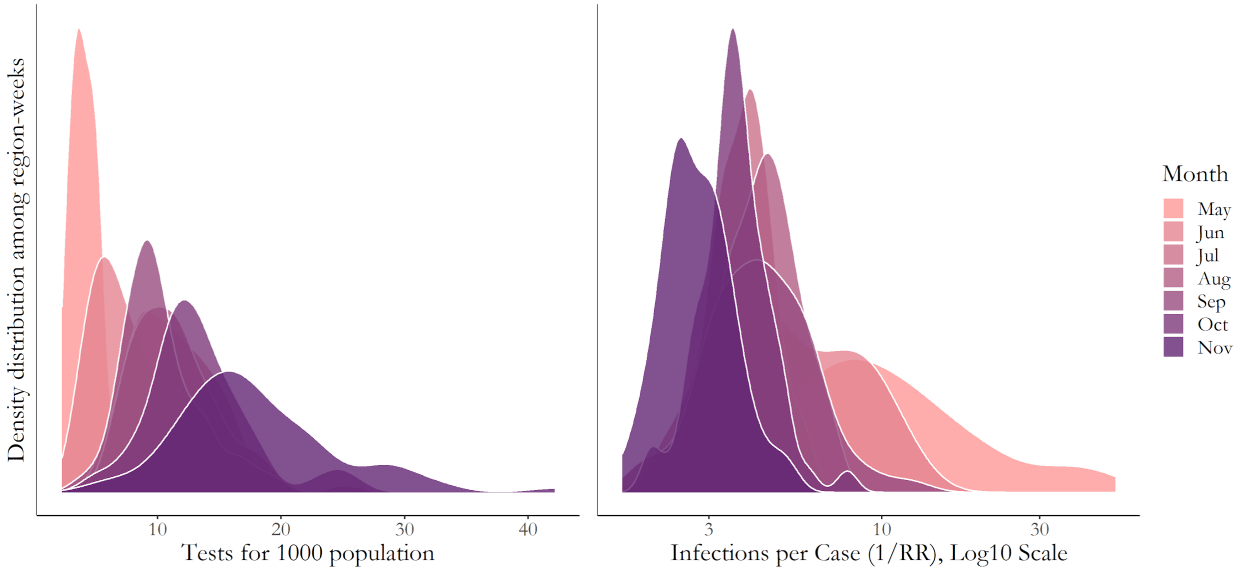

Figure 3: *Changing testing volume and reporting rates over time in Washington. Testing volume has increased and reporting rate has improved over time. Regions with testing volume under 10 per 1000 in November were Clark, North Central, South Central, and South West. Variation in reporting rate has also reduced over time (note that the x-axis is on a log10 scale).*

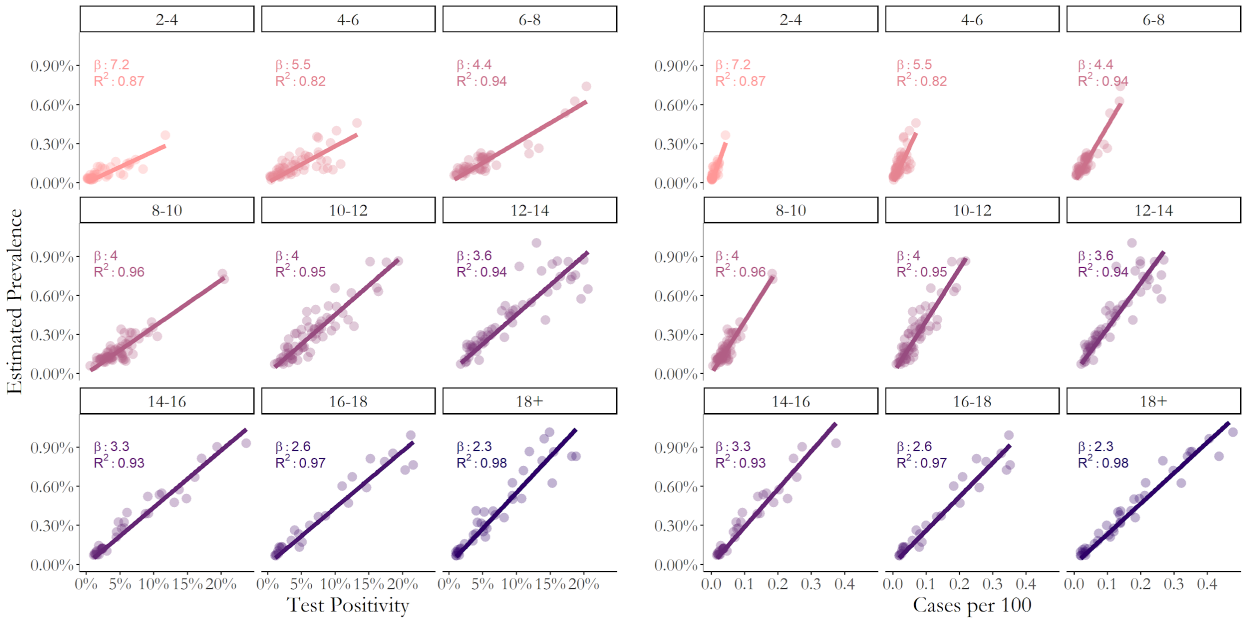

Figure 4: *Showing the relationship between observable metrics (test positivity and cases per capita) across bands of testing volume (per 1000). Lines show individual linear regressions, fitted with no intercept, and annotated with values for slope (β) and proportion of variation explained (R^2).*

A simple tool for interpreting test positivity and cases per capita

Our aim is to develop a predictive model that can be used as a heuristic tool by those interested in inferring underlying epidemiology from readily observable metrics. This effort is enabled by the strong relationships between the two observable metrics, test positivity and cases per capita, and prevalence,

and the stable relationship between testing volume and reporting rate. For both metrics, the needed heuristic will only require a modifying multiplier which depends on level of testing (serving a proxy for reporting rate). One can imagine developing models in this spirit at various levels of complexity, for example, a similar approach, which assumes reporting rate improves over time, is being used to estimate infections in US counties. For our purposes, we wish to distill the intuition described above into a simple, practical, and predictive tool which can be used and understood by users at varying levels of expertise, and calculated easily by hand.

For both test positivity and cases per capita, we tested two regression models: one that allowed for the metric to be modified at finer levels of test volume (as shown in Figure 4), and a simpler one that only allowed for two multipliers, dichotomized at 10 tests per 1000. We found that the loss in $R^2$by moving to the simpler model was minimal: a reduction of 0.006 for percent positive, and 0.033 for cases per capita, and thus proceeded with the simpler model. **Table 1** describes the heuristic tool derived from these regressions, with different multiplicative factors for 'high' or 'low' testing. In Washington, population testing is currently widespread (Figure 3), and most regions are generally in the 'high testing' category. As such, we can simplify the heuristic further into the following statements:

- Prevalence is approximately test positivity multiplied by 0.048.
- Prevalence is approximately weekly cases per 100,000 multiplied by 0.003
    - Alternatively, the number of total infections over the past week is cases multiplied by 3.

For example, if a location has had good testing, and found 200 cases in the past week, we would estimate that there have been 200*3=600 active infections in the past week. This is based on the assumption that one case represents about three infections.

Table 1: *Heuristic tool for converting test positivity and cases per capita to prevalence. Each cell shows the multiplicative factor to use where testing is assumed/known to be either 'high' or 'low'.*

Testing level	Test positivity multiplier (%)	Cases per 100k multiplier
Low, <10 per 1000	0.031	0.0043
High, >10 per 1000	0.048	0.003

The application of these simple multipliers to test positivity and cases per capita among region-weeks in Washington explain 94% and 91% of the variation in prevalence, respectively, as shown in **Figure 5**. While this represents excellent performance from such a simple model, we have traded some accuracy for simplicity, and thus it is important also to understand where such an approach tends to fail. Remaining unexplained variation in our heuristic-based estimates come from main two sources: variation in reporting rates and misspecification of average assumed reporting rates. Fortunately, these sources of variation are somewhat structured and thus predictable.

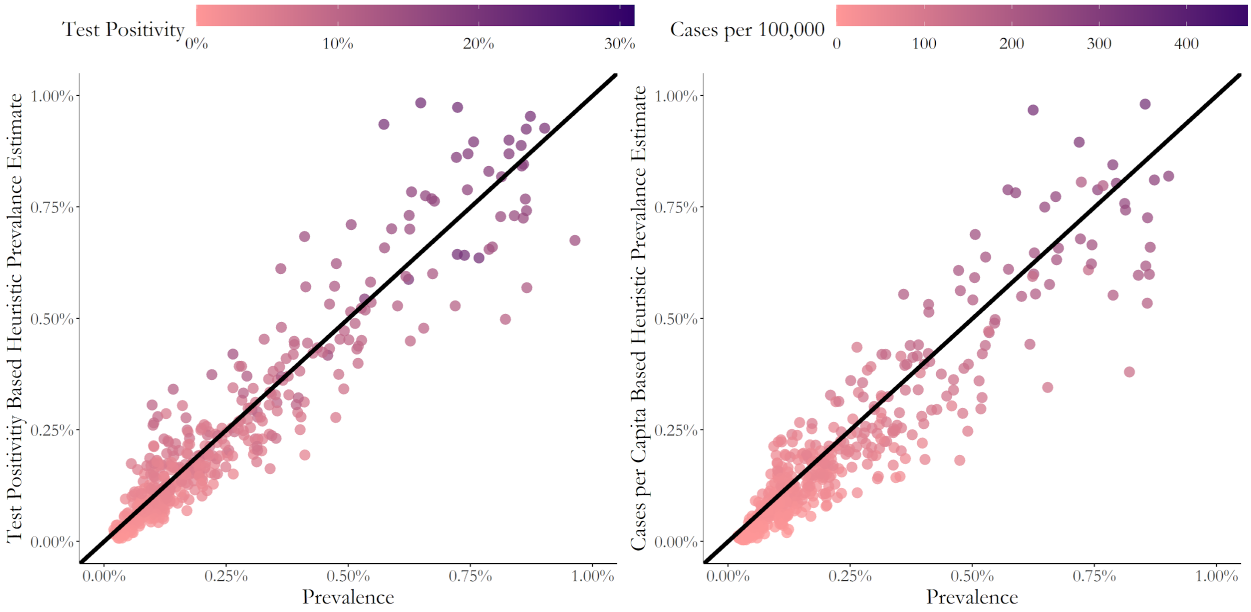

Figure 5: *Comparison of prevalence as estimated from the disease transmission model, to prevalence estimated from the simple heuristic described in Table 1. R^2 for test positivity = 0.94, and R^2 for cases per capita = 0.91.*

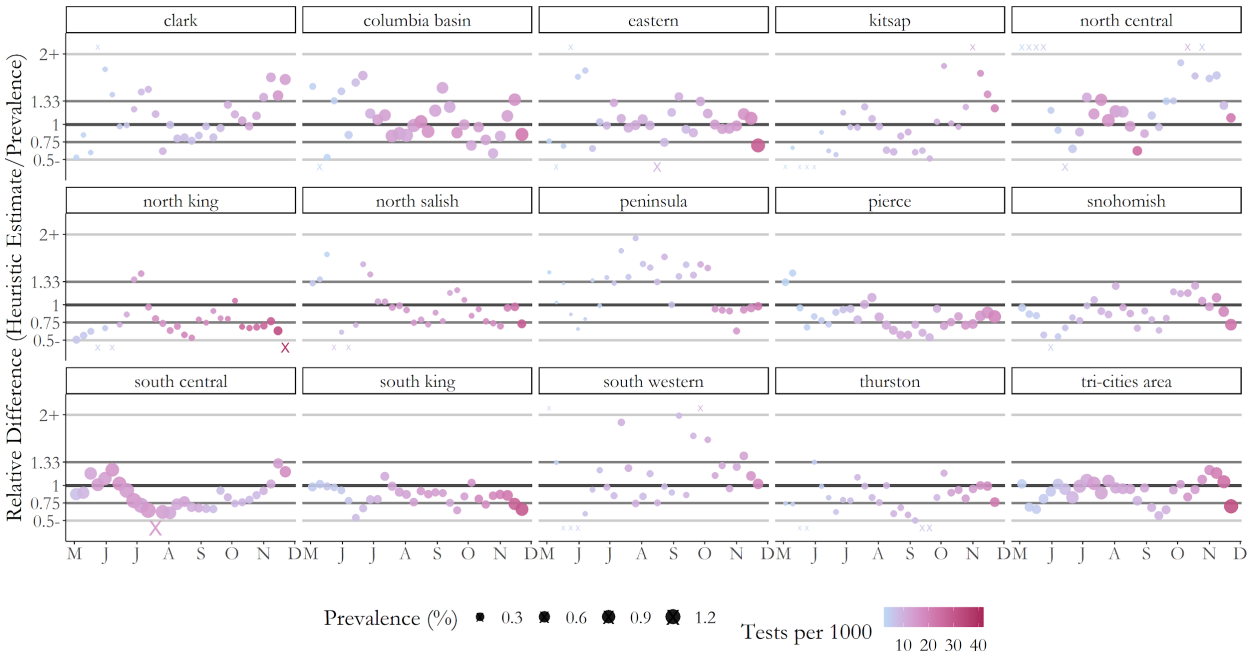

Figure 6: *Relative errors for each region-week in Washington State for the heuristic model derived from test positivity. Relative errors are shown as the ratio of the heuristic prevalence estimate and the transmission model prevalence estimate. The black line at 1 represents no error, the dark grey line at 0.75 and 1.33 represent an error on magnitude of ⅓ the prevalence and the light grey lines represent an error on magnitude of ½ (or two-fold) the prevalence. 60% of region-weeks fell within the ⅓ relative error band, and 93% of region-weeks fell within the two-fold band. Region-weeks outside the two-fold band are represented with an X. The color represents the testing volume in that region-week. Larger relative errors are more common when testing is lower, coinciding with greater variation in reporting rates.*

We illustrate this in **Figure 6**, which shows the relative differences between the 'gold standard' transmission model-based prevalence estimates and the heuristic-based estimates across each region-week (note, for conciseness we only show the test positivity-based heuristic, but results for cases per capita are similar). Overall, large relative errors on the magnitude of two-fold (<0.5x and >2x) occur in 7% (33 out of 450) of region-weeks. Of those, 28 (85%), are region-weeks where testing volume was lower than 10 per 1000, mostly from spring and summer. As such, we expect such a heuristic to yield less reliable results when testing is low, due to higher variability in reporting rate at those levels. Furthermore, we see that certain regions, such as north King County have a bias toward lower prevalence in the heuristic-based estimate. This is a result of misspecification of the assumed reporting rate, as the heuristic is based on averages while testing volumes and reporting rates in north King County have consistently outperformed the rest of the state. A more appropriate high-testing multiplier for north King County would be closer to 0.063, since average weekly testing volume is about 20 per 100,000 and average reporting rate is 33%.

The north King County example emphasizes the point that the specific numbers in Table 1 are derived from mean relationships observed in Washington. Table 1 is a quick and practical tool, but it is also interpretable and thus should also serve as a starting point to be improved based on local knowledge. The test positivity multiplier is the average test volume divided by reporting rate, and could be adjusted up or down if testing improves or degrades, respectively. The case per capita multiplier is simply the inverse of the reporting rate (infections per cases), so if one has a strong belief that the reporting rate has improved (if, for example, improved contact tracing or population coverage of testing had been implemented), the multiplier can be adjusted upwards. Furthemore, if testing volume is low and/or erratic (particularly under 6 tests per 1000), reporting rates become unpredictable, and such heuristics should be used with strong caution or avoided altogether.

Despite the above caveats, it remains instructive to explore the validity of applying the empirically-based heuristic tool in Table 1 outside of Washington state, in order to understand how generalizable we expect the relationships observed in Washington to be. To do so, we used reported cases per capita from the Covid Tracking Project to estimate weekly COVID-19 prevalence for each US state, starting July 1, 2020. Since prevalence is not directly observable, we based our comparison on reported deaths from COVID-19 in each state-week. To estimate deaths we used age-dependent infection fatality ratio (IFR) estimates published by CDC and state-specific age pyramids to make state-specific IFRs, and applied these to three-week lagged estimates of new exposures (derived from week-to-week changes in our heuristic-based prevalence estimates). **Figure 7** compares observed and estimated deaths from 1100 state-weeks from all 50 states between July and end of November 2020. Population normalized $R^2$is 0.87, with a median weekly error of 3.9 deaths, which implies that the relationship between cases per capita and underlying infections observed throughout Washington are broadly generalizable across the US. Test positivity was less predictive than cases per capita, with $R^2$is 0.58, with a median weekly error of 13.5 deaths, this is likely due to inconsistency in test positivity reporting across states due to differing definitions and reporting quality of negative tests.

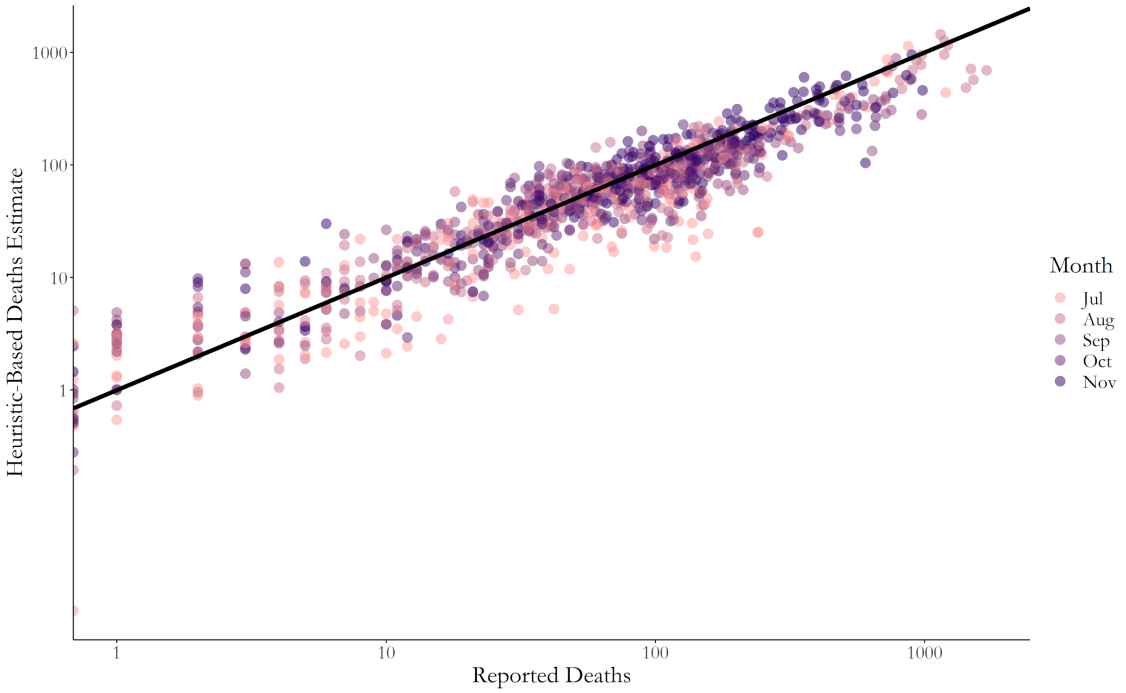

Figure 7: *Comparison across all 50 US states of reported deaths versus the cases per capita heuristic-based estimates of deaths, derived from average reporting rates in Washington state.*

Conclusions

Test positivity and cases per capita are readily accessible metrics which have been publicly reported in near real time throughout the COVID-19 epidemic in the US. In this report, facilitated by data-intensive transmission-model-based estimates of prevalence in Washington, we confirmed that these metrics can be used on their own as predictive signals of the underlying epidemiology.

We discovered that above a certain level of testing volume (about 6 per 1000), the reporting rate tends to remain relatively stable. This stability translates into a predictable relationship between the observable metrics and underlying population prevalence. We distilled this relationship down into a simple rule-based heuristic derived from average reporting rates observed in Washington state. Critically, the straightforward interpretability of the tool lends itself to be adjusted for local contexts, where stronger priors on reporting rate could be applied to make more accurate estimates. That said, however, even the parameters obtained from Washington's data performed well when applied to all 50 states.

From our observations, we believe that cases per capita is a preferred metric over test positivity. Test positivity depends on consistent reporting of total tests, which has a history of changing erratically and is differentially reported across states, and that reporting on the number of negative tests tends to lag over more timely reporting of cases. Furthermore, about 8% of negative tests are not associated with a region in Washington, inflating test positivity. Within our self-consistent sample of Washington region-weeks, cases per capita and test positivity performed about equally, and cases per capita performed

better at predicting deaths across US states. Furthermore, the relationship between cases per capita and prevalence is extremely intuitive: prevalence is cases per capita divided by reporting rate.

Reflecting for a moment as professional infectious disease modelers: One key learning for us from this pandemic has been the value of easily communicated, widely usable tools for interpreting data as it's being reported, not just for the general public but for policy-makers, scientists, and health officials as well. Transmission models offer valuable insight, but they generally don't fill this need, particularly in settings where certain data (like COVID-19 mortality and hospitalization) are difficult to collect and report. The approach here, using a transmission model as a platform for developing more easily generalizable tools, leverages our knowledge in settings where we have large amounts of information and context towards the goal of understanding situations where we have significantly less of both. As we continue to work towards that end, we hope tools like these can help people better understand COVID-19 in their own local context.