Prevalence Tracking Mechanisms for SARS-CoV-2

Jacob Steinhardt
UC Berkeley
jsteinhardt@berkeley.edu

Andrew Ilyas
MIT
ailyas@mit.edu

Abstract

In many areas, prevalence of SARS-CoV-2 will have decreased substantially in the next few weeks due to shelter-in-place and other lockdown orders. Based on Google Mobility Reports and growth rates in New York, California, and British Columbia, we estimate that mobility could in the best case be increased to around 70% of baseline (from a current range of 33%-53%) without leading to a resurgence in infections. As more effective suppression measures such as contact tracing are implemented, the mobility could be increased further. However, given the current uncertainty in these estimates, this can only be done safely if there are measures in place to track the prevalence of COVID-19. We identify shortcomings of existing tracking measures, and show that improved tracking could substantially decrease uncertainty about prevalence. Such measures, if put in place early, could lead to faster re-openings and better-informed public health decisions.

Acknowledgements

Thanks to Heather Youngs, Philip Stark, Jasjeet Sekhon, Ruby Lai, Steve Yadlowsky, Nigam Shah, and Zachary Sunberg for many helpful comments.

1 Introduction

Current shelter-in-place measures across the world prevent COVID-19 from overwhelming medical systems, but come with a large economic cost. Complete relaxation of these measures will likely need to wait until a number of alternative suppression strategies, such as contact tracing, are put in place. However, we provide evidence that shelter-in-place measures could be partially relaxed soon, potentially recovering substantial economic value. Our evidence is based on comparing mobility data and growth rates in several areas, and showing that mobility can potentially be increased while keeping growth rates below zero, even without new suppression strategies. As additional effective strategies are deployed, mobility can be further increased.

We describe here a schematic of what such a gradual re-opening might look like, and discuss infrastructure that would be needed to do this safely. Most importantly, we need methods for tracking the prevalence of SARS-CoV-2 to accurately target mobility to the maximum safe level. Tracking will be based on data sources such as random virological testing, testing in medical facilities, online symptom surveys, and viral prevalence in wastewater. The time lag together with the quality of these tracking methods will affect whether and how quickly we can increase mobility. We can decompose quality into three factors: the threshold (minimum prevalence at which the method is viable), the daily variability, and the mobility confound (how much changes in mobility would affect the output of the tracking).

In a typical metropolitan area in the U.S., we estimate that tracking methods would need to have a threshold of below 0.02% prevalence in order to be used for targeting mobility levels, although methods with a worse threshold could still be useful for other purposes. Tracking at this low level of prevalence is challenging but possible. For instance, random testing, without further improvements, would need a false positive rate (FPR) below 0.02% and at least 20,000 tests performed per day in each area. The number of tests could be decreased by using non-uniform sampling probabilities, and the FPR could be improved by re-running tests or combining with other data sources. Testing in emergency departments (EDs) is more economical and can tolerate a higher FPR, but currently has very high variability due to changes in testing

criteria over time and delay in test results. We show that simple bookkeeping and slight changes in testing could decrease this variability substantially.

In addition to the prevalence threshold, time lag plays a key role in usefulness for policy. Assuming a 1-2 day delay for test results, virological testing in emergency departments has an overall time last of 8-11 days, since patients take 5 days to develop symptoms and more for symptoms to become severe. Random testing in the population has a shorter 4-5 day time lag due to the ability to capture some presymptomatic cases. We also consider other methods, such as tracking viral prevalence in wastewater (4+ day lag) and online surveys of self-reported symptoms (5-6 day lag). Early implementations of these methods probably have prevalence thresholds above 0.02%, but will likely improve and probably approach or surpass the desired threshold.

Importantly, all methods except for random testing have at least moderate daily variability, and have moderate to strong mobility confounds, which will make it challenging to use them for informing policy in real-time. Characteristics of each method are summarized in Table 1, and we discuss the variability and confounding issues in more detail later.

Method	**Lag**	**Current Threshold**	**Future Threshold**	**Variability**	**Confounds**
Vir. Tests (Rand)	4-5 days	FPR (unknown)	$< 0.01\%$ possible	Low	Low
Vir. Tests (ED)	7-10 days	$< 0.01\%$	$< 0.01\%$ likely	Med	Low-High
Symptoms	5-6 days	0.08% $[0.03\%, 0.2\%]$	$< 0.01\%$ possible	Med	Low-Med
Wastewater	4-? days	0.1% $[0.06\%, 2.9\%]$	$< 0.01\%$ possible	Med-High	Med-High

Table 1: A summary of the surveillance strategies discussed in this work.

2 Strategy for Gradual Reopening

There is evidence that in some areas, shelter-in-place measures have reduced the effective reproduction number R_t of SARS-CoV-2 below 1 [Steinhardt, 2020]. In these areas, once prevalence has been brought low enough, it would be possible to partially increase mobility without resuming exponential growth in COVID-19 cases. For instance, we estimate below that in San Francisco, mobility is currently at 39% of normal levels but could be increased to 41-70% while keeping the growth rate below 0. However, in order to do this safely, we need systems in place to monitor the prevalence of COVID-19, in order to ensure that R_t remains below 1 and to potentially re-trigger shelter-in-place measures if prevalence goes above a certain threshold.

One way this might be implemented is to have public health officials issue a daily risk level, color coded as green-yellow-orange-red-magenta-black, similar to air quality advisories. Daily activity guidelines would depend on the risk levels; black would correspond to a full shelter-in-place while green might correspond to best-effort social distancing and good hygiene practices. The risk level should be based not just on current prevalence but projections of future prevalence; health officials should consider consulting with control theorists and other experts on dynamical systems to determine an appropriate policy.

In conjunction with this partial re-opening, we can implement other suppression strategies as they become available. Examples include contact tracing, regular temperature checks, mask-wearing, frequent cleaning of public spaces, applying antimicrobial coating (such as copper tape) to high-touch surfaces, and identifying individuals who are at high risk of spreading SARS-CoV-2 and paying them to work from home. Analyzing the plausibility and efficacy of each of these measures is an important avenue of ongoing work [Ferretti et al., 2020].

We note that in many areas, existing data is consistent with mobility being either below or above the safe threshold. Therefore, even ignoring economic benefits, better tracking of prevalence is necessary for public health.

3 Model Assumptions and Estimation Strategies

Below we calculate growth rate, overall prevalence, and reproduction number, which each rely on model assumptions. For growth rate, we follow the method of Yadlowsky et al. [2020], which models new counts on

Region	Los Angeles	San Francisco	New York	British Columbia
Mobility	53%	39%	33%	61%
Shelter-in-place	March 19	March 17	March 20	N/A
Mobility decrease	March 12-21	March 7-20	March 12-19	March 15-22
Estimated Growth (Pre-SIP)	16.7 [9.8, 24.1]	14.2 [7.6, 21.1]	30.2 [22.5, 38.5]	15.0 [5.1, 25.8]
Pre-SIP R_t	2.1 [1.6, 2.7]	1.9 [1.4, 2.4]	3.3 [2.6, 4.3]	1.9 [1.3, 2.9]
$R_t \times$ Mobility	1.1 [0.84, 1.5]	0.73 [0.56, 0.95]	1.1 [0.86, 1.4]	1.2 [0.78, 1.8]
Proj. Safe Mobility (Absolute)	48% [37%, 64%]	53% [41%, 70%]	30% [38%, 23%]	51% [35%, 78%]
Proj. Safe Mobility (Increase)	0.69x-1.19x	1.05x-1.79x	0.7x-1.16x	0.57x-1.3x

Table 2: The relationship between mobility statistics and transmission dynamics in the four regions of interest we study: Los Angeles, San Francisco, New York County, and British Columbia.

each day as an exponential growth process with overdispersed Poisson observation noise, with overdispersion parameter $\alpha = 0.15$. In some of our data, cumulative counts do not increase on some days, and we assume this was due to missed updates and treat these days as censored and impute their counts.

For estimating prevalence, we do not rely on confirmed case counts because these likely suffer from under-testing. Instead, we use hospital and death counts, assuming a 10-day lag from exposure to hospitalization and 21-day lag from exposure to death. Since hospital counts usually measure confirmed cases, we assume an additional 3-day lag from testing delays. Finally, we assume a hospitalization rate of 4% and fatality rate of 1% [Ferguson et al., 2020]. Therefore, prevalence on day t is estimated by taking cumulative hospitalizations on day $t + 13$, divided by 0.04, or by taking cumulative deaths on day $t + 21$, divided by 0.01. Sometimes these counts are not available and we impute them from nearby days based on growth rates. When reliable estimates from both hospitalizations and deaths are available, we report their average.Finally, we use the epidemiological model proposed in Ferretti et al. [2020] to estimate R_t from reported growth rate. In contrast to SEIR models, the model in Ferretti et al. [2020] directly estimates the generation time from available data, obtaining a Weibull distribution with mean 5.0 and variance 1.9. We believe this model is likely to be more accurate than an SEIR model, since it explicitly models the infection dynamics from data.

4 How Much Mobility Can We Allow?

Shelter-in-place (SIP) measures decrease interaction, primarily by decreasing mobility. We measure this decrease using Google's COVID-19 Community Mobility Reports. These reports measure mobility across six categories: retail, grocery, workplace, transit, parks, and residential. In most areas with COVID-19 outbreaks, mobility in the first four categories has decreased substantially since early March, while residential mobility has increased slightly. Mobility in parks has either increased or decreased depending on region.

We consider mobility in Los Angeles, San Francisco, New York City, and British Columbia. To form a single summary metric we take a weighted average of the 6 mobility categories, placing weight 1 on retail, grocery, workplace, and transit, 0.5 on parks, and 0 on residential. The numbers are summarized in Table 2, together with the date of initial mobility decrease and the date of announced shelter-in-place measures in that region.

In addition, we estimate the daily growth rate in SARS-CoV-2 prevalence before mobility reductions in each region. We estimate growth using the method in Yadlowsky et al. [2020], preferring hospital and death time series to confirmed case counts. We report weighted average estimates combining multiple data sources, weighting by the inverse squared uncertainty on the estimates. For Los Angeles we average state-wide and county-level deaths through April 2nd. For San Francisco we average state-wide deaths and Bay Area deaths through April 2nd. These gave similar growth rate estimates of $r = 16.7\%$/day in LA and 14.2%/day in SF, and agree with the estimate of 15%/day growth from Santa Clara County hospital data reported by Yadlowsky et al. [2020]. For New York City, we use hospitalizations through March 30th and deaths through April 2nd, yielding $r = 30.2\%$/day. We use state-level data since around 70% of state-wide deaths occurred in New York City. Finally, for British Columbia we average confirmed case counts through March 15th and deaths through April 5th, yielding $r = 15\%$/day. In all cases we pick date ranges that are

long enough for accurate estimation, but end before we expect to see leveling off from mobility restrictions. In British Columbia, there is also enough data to estimate the decline in confirmed cases since March 22nd, estimated at $r = -6.3\%$/day.

These estimates are reported in Table 4, along with confidence intervals. Using the Ferretti et al. [2020] model, we also report the derived R_t values. Finally, we can compute a projected post-SIP reproduction number R_t by making assumptions about how mobility affects the reproduction number. A first-pass assumption would be that R_t is quadratic in mobility, since R_t should depend roughly on the total number of pairs of interactions, which is quadratic in movement. A more conservative assumption is that R_t is linear in mobility, as may be the case if pre-SIP spread was primarily through essential workers whose behavior is only partially affected by SIP orders. We use this linear assumption throughout our analysis, computing a projected post-SIP R_t and a safe mobility threshold, which under the linear assumption is the inverse of the pre-SIP R_t.

Overall, we see that in the best case, mobility could be increased substantially: by 80% in San Francisco, 30% in British Columbia, and lower but still meaningfully in New York and Los Angeles. However, in the worst case, we may need to decrease mobility even further than presently. Importantly, it may be some time before the truth can be ascertained from present data. Consider Los Angeles, where we project R_t from 0.84 to 1.5 post-SIP. The lower bound corresponds to growth rate -3.3%/day and the upper bound to 8%/day, with a central estimate of 2.2%/day. These cannot be distinguished from the existing hospital and case data, which have confounds and noise that hide such small differences.

A general lesson is that large differences in R_t lead to only small differences in r, and so we need precise characterizations of the growth rate to correctly target mobility levels. An illustrative example is the 6.3%/day decline in British Columbia. This seemingly small daily decrease corresponds to $R_t = 0.72$, substantially below 1.

5 Disease Surveillance Measures

In a typical metropolitan area, we estimate that keeping prevalence below 0.43% would be sufficient to prevent undue strain on the medical system. In order to be employed towards this end, we estimate that a surveillance system would need to be capable of detecting a prevalence of around 0.02%.

We consider four different forms of surveillance, based on random virological testing (e.g. with swabs), virological testing in emergency departments, surveys of reported symptoms, and viral titre in wastewater. We find that early implementations of wastewater and symptom tracking have thresholds of around 0.2% and 0.08%, respectively. Wastewater tracking may be improved substantially, while we estimate that symptom tracking can only be improved moderately, but potentially below 0.02%. Virological testing can achieve almost arbitrarily low thresholds, limited primarily by the availability of tests and false positive rate. For swab-based testing of ED patients, the false positive rate is unlikely to become a bottleneck, but it may be a bottleneck for tracking low prevalence among a random sample of the population.

5.1 How Sensitive do Surveillance Need to Be?

We estimate here the approximate prevalence threshold that a tracking mechanism would need in order to usefully inform real-time advisory levels. We do this by first estimating an upper bound on prevalence that can be sustained without overloading hospital systems, then further account for tracking delays and the need to set smoothly-changing advisory policies. Our estimate is meant to provide an approximate rule of thumb. In ongoing work we provide a simulation that more directly quantifies the value of different achievable thresholds.

Regarding hospital capacity, we take Los Angeles county as a benchmark. Hospitalizations peaked on April 7th with 2601 total reported COVID-19 hospitalizations (suspected plus confirmed), according to CA Open Data. Los Angeles County has a population of 10.04 million, and assuming a hospitalization rate of 4% this yields a prevalence of 0.65%. Note that this is more accurately the total incidence of new cases over an 8-day window (assuming an 8-day hospital duration), which is a good approximation to total prevalence under fast exponential growth but not if growth is flat.

The Los Angeles prevalence of 0.65% appears to be a level at which medical demand is manageable. We base this on the lack of reports of ICU shortages or other severe signs of medical strain. In contrast, New

York City has 3.9% prevalence by a similar calculation and news reports suggest substantial strain on the city's medical system. Finally, San Francisco has 0.58% prevalence by this calculation, and there are reports of UCSF personnel volunteering in New York City, suggesting that San Francisco is well under its capacity. We take the LA prevalence of 0.65% and multiply by 2/3 to add an additional buffer, suggesting a target prevalence threshold of 0.43%.

In order to be used to set policy, a tracking system must be sensitive at levels substantially below 0.43%. First, it is unsafe to ever reach 0.43% exactly, because there would be additional cases even after a complete lockdown due to within-household infection, accounting for approximately one additional doubling. In addition, health advisories should ideally change slowly over time, and tracking mechanisms typically have some lag. Pre-lockdown growth rates have ranged from 10% to 35%, and post-lockdown growth rates will presumably be smaller than this under a careful scale-up of mobility. Assuming a 7-day lag in tracking and wanting to handle 2 weeks of growth at 10%/day, we would need an additional factor of $1.121 = 7.4$ in sensitivity. Alternately, handling at least 4 days of growth at 25%/day would require a factor of $1.2511 = 11.6$. Taking the greater 11.6 number and combining with the extra doubling, we would like tracking mechanisms to be sensitive at a prevalence of $0.43\%/(2 \cdot 11.6) = 0.019\%$.

In conclusion, a good rule of thumb is that tracking mechanisms should be sensitive to a prevalence of around 0.02%, although the exact number depends on the mechanism's time lag together with local assumptions about growth rate and hospital capacity.

5.2 Random Virological Testing

The most straightforward strategy for estimating prevalence is to test a random sample of the population each day. It is important that this sample is random and not self-selected (the latter being true for drive-through testing and most other current large-scale efforts). To achieve a threshold of 0.02% via uniform random sampling, we would need to perform at least 20,000 tests per day in a given area. This may pose a challenge given limited test availability. Moreover, the false positive rate (FPR) must be very low (less than 0.02%) to avoid skewing results at low prevalence.

The false positive issue may be addressable by storing multiple samples for each patient and re-testing positive results (although this might increase the false negative rate). However, we do not yet know how correlated false positives are upon re-testing. It is also possible that the FPR is already below 0.02%, although limited data exists so far.

The scalability issue could also be addressed by employing non-uniform sampling. Rather than sampling everyone with equal probability, we could skew the sample towards individuals who are more likely to be infected, for instance by targeting certain age groups or occupations. If we are, for instance, able to identify individuals that have 5 times greater propensity to be infected on average, then we would only need 4,000 tests per day and could tolerate a higher FPR of 0.1%.

A drawback is that, if we identify the wrong subgroup to test, we might miss growth in some other subgroup where we have more limited testing. It could take many days or weeks before we notice this. However, given the benefits and our likely ability to identify high-propensity individuals, such non-uniform sampling may still be attractive.

Mobility confounds for random testing are very low, assuming a high-quality random sample. Variability is also low, aside from statistical randomness in the sampling. An advantage over other methods is that this variability decreases at higher prevalences, while for most other methods the variability is due to daily changes in behavior that would remain even at high prevalence.

Finally, random testing has a short time lag. If we use virological tests, then we can identify even presymptomatic patients, probably around 3 days after initial exposure. Assuming a 1-2 day time lag for test results, the total lag is 4-5 days. We note that in contrast, serological tests probably cannot be used for prevalence tracking, as they do not yield positive results until near the end of infection and so have too long of a time lag.

5.3 Emergency Department Testing

An alternative to random testing is testing in emergency departments (EDs). This may be substantially more feasible if tests are limited, since a large fraction of COVID-19 patients will show up to EDs and other medical

facilities; a lower bound is the hospitalization rate, which we estimate at approximately 4% [Ferguson et al., 2020]. Blanket testing of all ED cases with given symptoms and risk factors should therefore be sensitive to even a couple hundred total cases in a population, which is well under the 0.02% threshold in most areas. If blanket testing is infeasible, we could also test a random subsample of cases with given symptoms.

Unfortunately, current test results from EDs and other medical facilities cannot easily be used to detect prevalence. One reason is that decisions about who to test change over time as tests become more available. Another is that in some areas, shortages of testing supplies may delay results substantially, and in a non-uniform way.

We can easily account for both of these with appropriate bookkeeping. To account for testing delays, we should record and report the date of admission and date of testing in addition to the date of result. Reported time series should by default aggregate by date of test or date of admission, and report the number of pending tests from each date. We note that most popular websites, such as the COVID Tracking Project, do not currently do this, probably because the data is not usually available.

To account for changes in testing decisions, we should identify a set of criteria such that patients meeting those criteria will very likely be tested, both currently and in the future. Then, we should add a protocol that all such patients actually are tested; this requires slightly increasing the total number of tests. Finally, we should record the number of positive and negative results each day from patients meeting this criterion. Doing so removes variability due to changes in testing decisions over time.

A final issue is that patients' decisions to show up to a testing facility may change over time and this may be correlated with COVID-19 prevalence. For this reason we recommend obtaining the above counts from EDs or hospitals rather than drive-through tests, as we believe that ED patients' behavior will be mostly dictated by symptoms and less susceptible to behavioral changes. However, there is some evidence that even decisions to report to emergency departments may be affected by COVID-19. For instance, ED visits in New York City for diarrhea, vomiting, and asthma have all decreased by a factor of 4 to 5 since early March based on syndromic surveillance data. It is not clear if this is from decreases in infections due to the shelter-in-place, changes in behavior, or both.

Overall, testing in EDs has a low prevalence threshold, well under 0.02%. We estimate a time lag of 7-10 days, since it takes around 5 days for symptoms to develop and another few days for them to worsen enough to report to an ED. Daily variability is moderate—for instance there are strong day-of-week effects in ED admissions, although these could be partially accounted for. Finally, as discussed above, there could be confounds that are correlated with COVID levels and/or mobility. These confounds could range from mild to large (up to a factor of 5 based on the syndromic surveillance data).

5.4 Wastewater

Wastewater surveillance tracks the level of SARS-CoV-2 viral titre in wastewater. By pooling samples from many individuals, fewer RT-PCR tests can determine prevalence. Similar surveillance systems have been used to track poliovirus and other viruses [Lago et al., 2003], and efforts are currently under way to refine these systems for the current coronavirus epidemic. We make a preliminary estimate of the prevalence threshold of these tests based on two published studies in Norway [Medema et al., 2020] and Massachusetts [Wu et al., 2020], though we stress that this only measures the threshold of crude first-pass systems and the eventual threshold will likely be far better after further refinement.

In the Netherlands, SARS-CoV-2 RNA was detectable in the water supply on March 4th, and we estimate that prevalence on March 4th in the Netherlands was 0.2%, based on looking at hospital counts and deaths and accounting for lag. The Netherlands test does not attempt to quantify RNA levels, and the previous sample was taken on February 6th. Therefore, we can only conclude that the test detected RNA at a prevalence of 0.2% or lower.

In Massachusetts, tests on March 18th detected 49 genome copies per milliliter (gc/mL). At that time we estimate the prevalence was 0.42% based on lagged hospital and death data. In addition, the standard deviation upon retesting the same sample was 28 gc/mL. Finally, wastewater data likely tracks cases on at least a 3-day lag (discussed below), and the doubling time in Massachusetts was approximately 3 days at the time. Accounting for these two factors yields a threshold of 0.42% * (28/49) / 2 = 0.12%. For comparison, the original paper estimates prevalence of 0.1%-5% from the sample they take, which would correspond to thresholds of 0.06% to 2.9%.

In summary, an early implementation of wastewater surveillance in Massachusetts seems capable of quantifying counts of approximately 28 genome copies / mL, corresponding to a threshold of approximately 0.1%. There is, however, substantial room for improvement on these early methods. An approximate lower bound comes from quantification thresholds in aquatic environments [Farkas et al., 2017, Sui et al., 2019, Farkas et al., 2020], which range from 1 to 60 genome copies per microliter (approximately 1000 times better). However, this lower bound is likely to be optimistic as wastewater quantification suffers from greater interference of proteases and nucleases relative to aquatic environments. Another datum is that systems for tracking poliovirus in Havana, Cuba had a threshold of below 1% [Lago et al., 2003], with a lower bound containing 0.

The time lag for wastewater data is hard to characterize from existing data. It is unlikely to be less than 4 days (3 days to develop viral shedding and 1 day for testing delays). However, in contrast to nasopharyngeal and sputum samples, it is unclear when viral shedding peaks in feces, with one patient peaking 17 days or longer after exposure [Tang et al., 2020]. On the other hand, wastewater also contains secretions from nasal drip and coughing, which would make the time lag shorter. We avoid providing an upper bound on wastewater time lag given the limited current evidence.

Finally, there are some challenges to achieving reliable prevalence estimates from wastewater data, such as accounting for the variable viral loading in feces during the time course of the infection, dilution from surface runoff, and other daily variations in wastewater content. Many sewer systems are separated from surface runoff, in which case the first issue would be minimized. However, daily variation may be confounded by mobility, since both the location and behavior of individuals will change as SIP measures are lifted. For instance, people may cook at home less, and they may return to a workplace that is connected to a different sewer system than their place of residence.

5.5 Symptom Surveillance

We next examine tracking based on reported symptoms. We incorporate two lines of evidence: the observed efficacy of New York City's Syndromic Surveillance system at tracking the increase in COVID-19, and a first-principles calculation using the baseline prevalence of flu and our likely ability to distinguish COVID-19 from flu based on symptoms. The syndromic surveillance data suggests a threshold of 0.034% to 0.19%, while the first principles calculation suggests a lower bound of 0.025% from symptoms alone, although this could likely be improved with additional information.

The New York City Syndromic Surveillance system tracks cases of respiratory, flu-like, and other illnesses based on text scraping of doctor's notes from ED visits. We viewed the time series of reported respiratory cases and saw a clear uptick in cases on March 9th, with the first observable uptick in cases on March 2nd. We estimate that ED visits likely lag onset by around 7 days as calculated above. Therefore, these upticks correspond to prevalence on March 2nd and February 24th, for which we estimate prevalences of 0.19% and 0.034% based on lagged death data. We take the geometric mean, 0.08%, as our central estimate for the prevalence threshold of this system. An important caveat is that ED usage has changed substantially over time as a result of COVID-19: for instance, ED visits for unrelated illnesses have dropped by a factor of 4-5. We would need to account for these confounds before using syndromic surveillance to track COVID-19 prevalence.

Large-scale symptom tracking is likely to be implemented via online surveys rather than syndromic surveillance in emergency departments. These may less accurately assess symptoms compared to doctor's notes, but they will be less confounded by changes in ED visits and so may be more reliable overall. In addition, all forms of symptom tracking will improve in the summer months as flu becomes less prevalent: flu levels in June through August are about half that in March.

We can alternately estimate the threshold of any symptom-based tracking method by considering prevalence of flu, together with our ability to distinguish flu from COVID-19. Since COVID-19 is not strongly distinguishable from flu based on symptoms, the threshold of symptom-based tracking will not be much lower than baseline prevalence of flu. Based on a reported 1 flu hospitalization / 100,000 people in a typical week in April and hospitalization rate of 1%, the baseline prevalence of flu in a given week is 0.1%.

We expect that the achievable likelihood ratio of flu vs. COVID-19 based purely on symptoms is not much greater than 4. One line of evidence for this is recent data on symptoms in Stanford hospital [Callahan et al., 2020] among patients testing positive and negative for COVID-19. The greatest likelihood ratio from any

single symptom was 2.7, and we probably cannot improve substantially from this given high co-occurrence of symptoms. Separately, the difference in hospital rates between flu and COVID-19 is roughly 4, so symptom severity also does not strongly distinguish flu vs. COVID-19. Based on this, we expect that symptom-based tracking will not have a prevalence threshold much below 0.025% (one-fourth the prevalence of flu).

However, surveys may be able to improve further by asking information about flu immunization status, age, occupation, and recent behavior. It seems possible that these combined with symptom data could achieve the target threshold of 0.02%.

Finally, surveys likely have moderate daily variability. For instance, if they are administered via online ads, daily patterns of internet usage will affect surface results. Since internet usage is also affected by mobility, online surveys will have mobility confounds as well.

References

Alison Callahan, Jason A. Fries, Saurabh Gombar, Birju Patel, , and Nigam H. Shah. Profiling presenting symptoms of patients screened for sars-cov-2. 2020. URL `https://medium.com/@nigam/an-ehr-derived-summary-of-the-presenting-symptoms-of-patients-screened-for-sars-cov-2-910ceb1b22b9`.

Kata Farkas, Dafydd E. Peters, James E. McDonald, Alexis de Rougemont, Shelagh K. Malham, and Davey L. Jones. Evaluation of two triplex one-step qrt-pcr assays for the quantification of human enteric viruses in environmental samples. *Food and Environmental Virology*, 9(3):342–349, Sep 2017. ISSN 1867-0342. doi: 10.1007/s12560-017-9293-5. URL `https://doi.org/10.1007/s12560-017-9293-5`.

Kata Farkas, Finn Mannion, Luke S. Hillary, Shelagh K. Malham, and David I. Walker. Emerging technologies for the rapid detection of enteric viruses in the aquatic environment. *Current Opinion in Environmental Science & Health*, 16:1 – 6, 2020. ISSN 2468-5844. doi: https://doi.org/10.1016/j.coesh.2020.01.007. URL `http://www.sciencedirect.com/science/article/pii/S2468584420300088`.

Neil M Ferguson, Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma Cucunuba, Gina Cuomo-Dannenburg, Amy Dighe, Ilaria Dorigatti, Han Fu, Katy Gaythorpe, Will Green, Arran Hamlet, Wes Hinsley, Lucy C Okell, Sabine van Elsland, Hayley Thompson, Robert Verity, Erik Volz, Haowei Wang, Yuanrong Wang, Patrick GT Walker, Caroline Walters, Peter Winskill, Charles Whittaker, Christl A Donnelly, Steven Riley, and Azra C Ghani. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. 2020.

Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. In *Science*. American Association for the Advancement of Science, 2020.

Pedro Mas Lago, Jr Gary, Howard E, Luis Sarmientos Perez, Victor Caceres, Julio Barrios Olivera, Rosa Palomera Puentes, Marite Bello Corredor, Patricia Jímenez, Mark A Pallansch, and Roberto Gonzalez Cruz. Poliovirus detection in wastewater and stools following an immunization campaign in Havana, Cuba. In *International Journal of Epidemiology*, 2003. URL `https://www.ncbi.nlm.nih.gov/pubmed/14559748`.

Gertjan Medema, Leo Heijnen, Goffe Elsinga, and Ronald Italiaander. Presence of SARS-Coronavirus-2 in sewage. 2020. URL `https://www.medrxiv.org/content/10.1101/2020.03.29.20045880v1`.

Jacob Steinhardt. Is R0 less than one in New York and California? 2020.

Zhiwei Sui, Siyuan Liu, Sizhang Liu, Jing Wang, Lei Xue, Xiaoxia Liu, Bin Wang, Shaopeng Gu, and Yi Wang. Evaluation of digital pcr for absolute and accurate quantification of hepatitis a virus. In *International Conference on Biomedical Sciences and Information Systems*, 2019.

An Tang, Zhen-dong Tong, Hong-ling Wang, Ya-xin Dai, Ke-feng Li, Jie-nan Liu, Wen-jie Wu, Chen Yuan, Meng-lu Yu, Peng Li, and Jian-bo Yan. Detection of novel coronavirus by rt-pcr in stool specimen from asymptomatic child, china. *Emerging Infectious Disease journal*, 26(6):1337, 2020. ISSN 1080-6059. doi: 10.3201/eid2606.200301. URL `https://wwwnc.cdc.gov/eid/article/26/6/20-0301_article`.

Fuqing Wu, Amy Xiao, Jianbo Zhang, Xiaoqiong Gu, Wei Lin Lee, Kathryn Kauffman, William Hanage, Mariana Matus, Newsha Ghaeli, Noriko Endo, Claire Duvallet, Katya Moniz, Timothy Erickson, Peter Chai, Janelle Thompson, and Eric Alm. Sars-cov-2 titers in wastewater are higher than expected from clinically confirmed cases. *medRxiv*, 2020. doi: 10.1101/2020.04.05.20051540. URL `https://www.medrxiv.org/content/early/2020/04/07/2020.04.05.20051540`.

Steve Yadlowsky, Nigam Shah, and Jacob Steinhardt. Estimation of SARS-CoV-2 infection prevalence in Santa Clara county. 2020.