SCAN Data Results and Technical Report #2

Friday, May 22, 2020

Summary

- Between March 23 and May 9, 2020, SCAN collected and tested 12,482 samples. This testing has uncovered 102 SARS-CoV-2 positive results.
- Late March was likely the peak of prevalence in King County, with prevalence declining since. For the most recent data period ending on May 9, we estimate prevalence is between 15 and 46 in 10,000.
- After a period of fast-declining prevalence in early April (which is consistent with an effective reproductive number lower than one), there has been an attenuation of the decline since late April, wherein prevalence was no longer declining with high certainty. This is consistent with findings from a disease transmission model based on testing and mortality data from the Washington Disease Reporting System (WDRS), where the slowing decline in the daily rate of cases is also apparent.
- Consistent with case reports for PHSKC, south King County has a higher proportion of positive test results relative to north King County in SCAN samples. The parts of the county with a higher proportion of positive results are also the same ones that are underrepresented in SCAN.
- Participants living in larger households are more likely to test positive.
- We are seeing a number of households with more than one household member participating in SCAN. In fact, 39.5% of SCAN participants live with another SCAN participant.
- Thirteen cases have been detected in households through testing following return of an initial positive result in the household.
- SCAN has launched a priority code system that facilitates enrollment of children, a group so far underrepresented in SCAN's sample.
- A large majority -- 87% -- of respondents with a positive result had not sought in-person clinical care before enrolling in SCAN.
- As of May 12, 2020, SCAN's testing of home-based, self-collected samples for COVID-19 and return of results is paused. We are working with the Food & Drug Administration and Washington State Department of Health to resume operations as soon as possible.

General updates on SCAN

The greater Seattle Coronavirus Assessment Network, or SCAN, is a public health surveillance (disease monitoring) program for SARS-CoV-2 (the virus that causes COVID-19) infection in greater Seattle and King County. SCAN is designed to help us better understand the COVID-19 outbreak and, with other

sources of data, inform public health decisions. The SCAN platform launched on March 23, 2020 with an initial focus on testing individuals comprising a broad representation of the greater Seattle and King County region using the at-home sample collection with a self-swabbing kit developed by the Seattle Flu Study (SFS). Please see our first technical report for more background on SCAN.

Please note that as of May 12, 2020, SCAN's testing of home-based, self-collected samples for SARS-CoV-2 with return of results is paused. The team is working with the Food & Drug Administration and Washington State Department of Health to resume testing with return results to participants as soon as possible. Please visit www.scanpublichealth.org for more information and program updates.

Sample representativeness

To achieve better representation by age, geographic region, race/ethnicity, income, and primary language relative to the King County population, SCAN is partnering with community-based organizations (CBOs) and others who serve populations underrepresented in SCAN to provide 'priority codes' that will facilitate enrollment. A defined number of these codes will be made available to CBOs to distribute to their members over time. The first of these codes has been deployed to increase enrollment of children, and future codes will be aimed at increasing enrollment of other underrepresented groups. We expect adoption of the codes will result in a gradual closing of the gaps in representation, especially in conjunction with orientation to SCAN's objectives and training for CBO staff to act as 'navigators' to enroll their members. In addition, as of the week of May 11th, the SCAN website, surveys, emails, and test kit materials are available in English, Spanish, simplified and traditional Chinese, Vietnamese, Somali, Korean, Russian, Amharic, Tigrinya, and Tagalog. Outreach to communities where these languages are spoken is also planned.

Table 1 summarizes SCAN enrollment numbers across several demographic characteristics in the month since our last technical report, while **Figure 1** compares these to the demographic characteristics of the overall King County population. Enrollment of those aged under 20 and over 80 still fall below their proportions in the county as a whole. There is also an over-representation of whites as well as those with household incomes greater than $150,000. More than half of whites, those with household incomes greater than $150,000, and those in the 60-79 year age range enrolled without reporting CLI symptoms on the screener, while for other race/ethnicity, income, and age groups, most of those who enrolled reported CLI symptoms. Whether this reflects a true difference in the proportion of people symptomatic in these groups or inequity in screener access is unclear.

Table 1: Characteristics of SCAN participants between April 10 and May 9 (the period since our last technical report). Note that the numbers and ratios of those who did vs. did not report CLI in this table reflect the population tested and not the total population screened for participation in SCAN.

	Total (% of Total)	Reported CLI on screener	Did not report CLI on screener
All Participants	8119	4272	3847
Age			
0-4	143 (1.8%)	78 (1.8%)	65 (1.7%)
5-9	132 (1.6%)	62 (1.5%)	70 (1.8%)
10-19	290 (3.6%)	161 (3.8%)	129 (3.4%)
20-29	1217 (15%)	671 (15.7%)	546 (14.2%)
30-39	2083 (25.7%)	1218 (28.5%)	865 (22.5%)
40-49	1591 (19.6%)	850 (19.9%)	741 (19.3%)
50-59	1224 (15.1%)	638 (14.9%)	586 (15.2%)
60-69	983 (12.1%)	410 (9.6%)	573 (14.9%)
70-79	396 (4.9%)	152 (3.6%)	244 (6.3%)
80+	60 (0.7%)	32 (0.7%)	28 (0.7%)
Sex at Birth			
Female	4529 (55.8%)	2355 (55.1%)	2174 (56.5%)
Male	3568 (43.9%)	1904 (44.6%)	1664 (43.3%)
Other	2 (0%)	2 (0%)	0
Unknown	20 (0.2%)	11 (0.3%)	9 (0.2%)
Race and Ethnicity			
Amer. Indian or Alaska Native	28 (0.3%)	19 (0.4%)	9 (0.2%)
Asian, not Hispanic	1322 (16.3%)	840 (19.7%)	482 (12.5%)
Black, not Hispanic	192 (2.4%)	141 (3.3%)	51 (1.3%)
Hispanic or Latino, any Race	485 (6%)	287 (6.7%)	198 (5.1%)
Native Hawaiian or Pacific Islander	31 (0.4%)	21 (0.5%)	10 (0.3%)
Other or multi-racial, not Hispanic	461 (5.7%)	245 (5.7%)	216 (5.6%)
White, not Hispanic	5423 (66.8%)	2607 (61%)	2816 (73.2%)
missing	177 (2.2%)	112 (2.6%)	65 (1.7%)
Household Income			
< $25k	507 (6.2%)	346 (8.1%)	161 (4.2%)

$25k - $49k	763 (9.4%)	440 (10.3%)	323 (8.4%)
$50k - $74k	902 (11.1%)	529 (12.4%)	373 (9.7%)
$75k - $99k	874 (10.8%)	446 (10.4%)	428 (11.1%)
$100k - $124k	806 (9.9%)	405 (9.5%)	401 (10.4%)
$125k - $149k	732 (9%)	375 (8.8%)	357 (9.3%)
>= $150k	2397 (29.5%)	1081 (25.3%)	1316 (34.2%)
Prefer not to say	1015 (12.5%)	569 (13.3%)	446 (11.6%)
Don't know	123 (1.5%)	81 (1.9%)	42 (1.1%)
Sought Care			
No	7242 (89.2%)	3508 (82.1%)	3734 (97.1%)
Yes; Doctor's /Urgent Care	170 (2.1%)	140 (3.3%)	30 (0.8%)
Yes; Pharmacy	19 (0.2%)	19 (0.4%)	0 (0%)
Yes; Telemedicine	685 (8.4%)	607 (14.2%)	78 (2%)
Yes; Hospital/ED	33 (0.4%)	27 (0.6%)	6 (0.2%)
Yes; Other	31 (0.4%)	26 (0.6%)	5 (0.1%)
Underlying conditions**			
Chronic heart disease	94 (1.2%)	52 (1.2%)	42 (1.1%)
Chronic lung disease	200 (2.5%)	121 (2.8%)	79 (2.1%)
Diabetes	270 (3.3%)	156 (3.7%)	114 (3%)
Immunosuppressed	294 (3.6%)	193 (4.5%)	101 (2.6%)
None	7334 (90.3%)	3806 (89.1%)	3528 (91.7%)

*CLI = self-reported new COVID-like illness symptoms (cough, fever, shortness of breath) in the past 7 days, as reported on the enrollment screener; **Individuals can have more than one underlying condition.

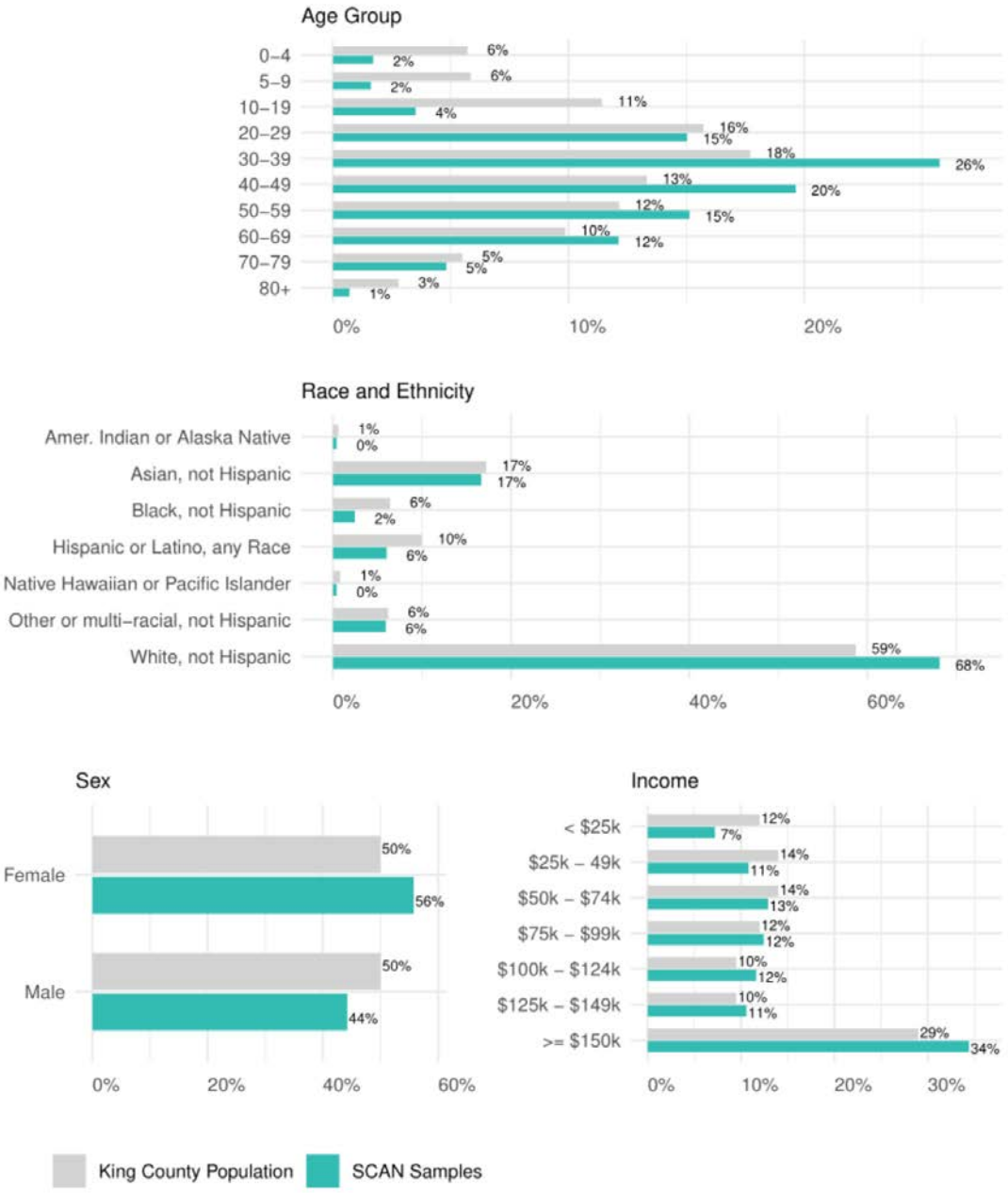

Figure 1: Distribution of SCAN participants between April 10 and May 9 (the period since our last technical report) across age, race and ethnicity, sex, and household income, compared to the distribution in King County.

SCAN test results & estimated prevalence of COVID-19 in King County (March 23 - May 9, 2020)

Table 2 describes data on testing for SARS-CoV-2 among those who reported COVID-like illness (CLI) symptoms on the enrollment screener (including new or worsening cough, fever, or shortness of breath). For samples collected through May 9, 2020, a total of 12,482 conclusive tests have been returned: 7106 from respondents reporting CLI on the enrollment screener, and 5376 from those not reporting CLI on the enrollment screener. Of those reporting CLI, 97 (1.4%) returned a positive result. As of yet, we do not observe a difference in the proportion testing positive across ages or by sex at bith.

Five samples from people who screened into the non-CLI enrollment group returned a positive result. However, all five of these respondents reported some symptoms in the detailed illness questionnaire. These included headaches (N=4), muscle or body aches (N=3), fatigue (N=3), and chills or shivering (N=3), among others. An additional 21 individuals who screened into the CLI enrollment group and tested positive reported symptoms in the detailed illness questionnaire that were inconsistent with their answer to the screener question, with 6 reporting no symptoms and 15 reporting symptoms other than cough, fever, or shortness of breath.

Of those with a positive result, 87% (N=89) did not seek clinical care in a physical location prior to enrolling in SCAN, including 78 who did not seek any care, and 11 who had a telemedicine appointment. An additional 11% (N=11) did go to a doctor's office, urgent care, or other source of care; one reported going to the hospital or emergency department; and one did not respond to the question. This indicates that the majority of respondents returning positive results through SCAN would not have been captured otherwise and missed case investigations and/or contact-tracing by public health.

Table 2: Positive results among those who reported covid-like illness (CLI) symptoms on the enrollment screener. Additionally, 5 participants who did not report CLI symptoms on the enrollment screener tested positive.

	Number with positive tests / total tests	% positive (95% CI)*
Total	97/7106	1.4% (1.1% - 1.7%)
Age		
0-4	3/100	3% (1% - 8.5%)
5-19	3/331	0.9% (0.3% - 2.6%)
20-59	75/5725	1.3% (1% - 1.6%)
60+	16/949	1.7% (1% - 2.7%)
unknown	0/1	
Sex at Birth		
Female	54/3914	1.4% (1.1% - 1.8%)
Male	42/3163	1.3% (1% - 1.8%)
Other	0/4	
Unknown	1/25	

*95% confidence intervals. These intervals likely under-estimate uncertainty as they assume random binomial sampling.

Using a model that combines the test results above with SCAN survey and census data to estimate prevalence across the community, we infer that prevalence declined from 51 [95% CI: 32 - 78] per 10,000 in the time period between March 23 and March 28 to 27 [15 - 46] per 10,000 in the most recent period, between May 4 to May 9. During the most recent period, this is equivalent to between 3 and 10 thousand active infections in King County. The time trend in community prevalence is shown in **Figure 2**, with the underlying test data by date of enrollment, which includes 11,559 tests and 86 positive results (note that some samples collected are not used in prevalence estimation, as discussed below). Given the data and model, we estimate with high certainty (98% posterior probability) that by early May, COVID-19 prevalence in King County had declined since peak prevalence in late March. These gains were made during the first half of April, when prevalence declined rapidly, which is consistent with an effective reproductive number (*Re*) below the critical threshold of 1. However, it appears that progress had begun to stall by mid to late April, as the decline in prevalence has attenuated and stabilized at around 30 per 10,000 since the period starting April 16th. This finding is largely consistent with concurrent findings of increasing transmission from a disease transmission model based on testing and mortality data in the Washington Disease Reporting System (WDRS), which estimated that *Re* was below 1 in early April, but rising to be statistically indistinguishable from 1 later in the month. More recent findings from the transmission model, using case data reported since the pause of SCAN in May, indicate that prevalence has since continued to decline albeit at a slower pace than seen in early April (see **Appendix 1** for more details). Together, all data indicate that peak prevalence likely occurred in late March, around the time that SCAN began.

Further details on the methods that we use to estimate prevalence is given in **Appendix 2** of this report. In brief, we fit a statistical model at the individual level that allows us to predict community-level prevalence over time while adjusting for some known biases in the data. Currently, we adjust for skewed geographical sampling and deliberate screening and oversampling of respondents reporting COVID-like illness symptoms.

As we discussed in the first SCAN technical report, SCAN relies on volunteers who self-select to participate. This can introduce a number of biases because these volunteers may not be representative of the general King County population in ways that are unmeasurable. As such, estimates of prevalence arising from a self-selected sample like SCAN must be interpreted with caution. Despite this, the trend in prevalence estimated from the SCAN sample aligns well with recent research on community spread of COVID-19 in King County (ref1, ref2). We are continuing to work to better understand the data collected by SCAN. Some nuances of these data impact the inferences we are able to draw from them. In the sections that follow, we discuss some of these nuances in more detail.

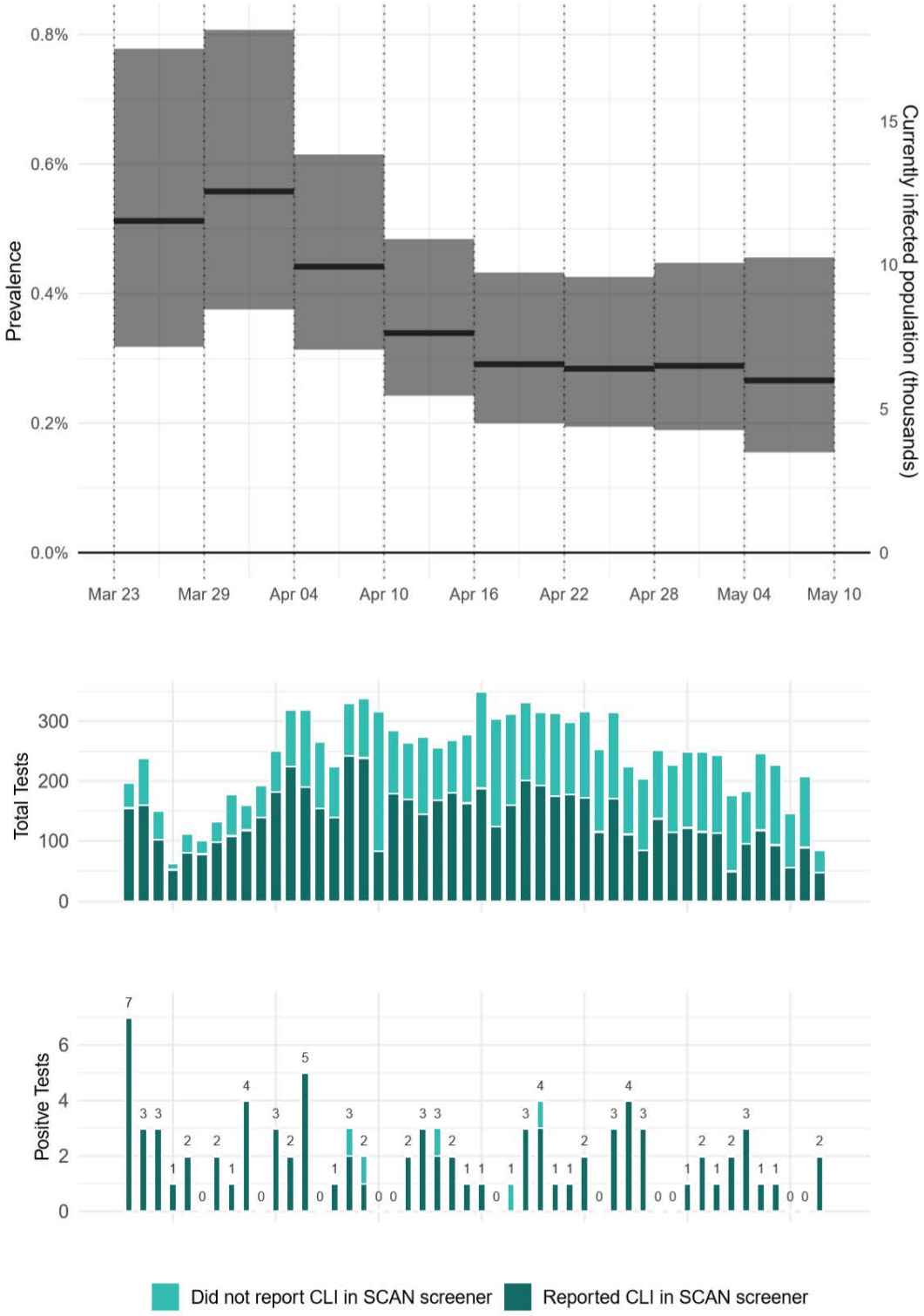

Figure 2: Estimated prevalence and underlying completed test data as returned through May 9. TOP: Community prevalence estimate. Prevalence estimates are made using a self-selected sample and should be interpreted in light of their potential limitations, as discussed in this report. MIDDLE: Total tests by date of collection, stratified by reporting of COVID-like illness (CLI) symptoms in the SCAN enrollment screener. BOTTOM: Total positive tests by date of collection, stratified by reporting of COVID-like illness (CLI) symptoms in the SCAN enrollment screener.

Within-household risk and testing behavior

It stands to reason that individuals living in households with more members may be at higher risk for transmission, as living with people increases the number of contacts each individual has (ref, ref). Of SCAN participants, 14% live alone, while 38% live in a household of 2, 19% live in a household of 3, and 29% live in a household of 4 or more. Indeed, the proportion of COVID-19 positivity in the SCAN sample increases as household size grows. A participant reporting living in a household of 4 or more has 3.0 [95% CI 1.5 - 6.0] times higher odds of testing positive than a participant who lives alone, though those living in a household of 2 or 3 did not show significantly higher risk (1.0 [95% CI 0.4-2.1] and 1.1 [95% CI 0.5 - 2.6], respectively).

Multiple participants from the same household can enroll in SCAN, and 39.5% of SCAN participants live with at least one other SCAN participant. The odds of a positive test for a participant living in a household with 4+ SCAN participants is 7.0 [95% CI 4.1 - 11.9] times higher than for an individual living without other SCAN participants. Since SCAN participants self-select for enrollment, this observation suggests that people are more driven to participate if they know or suspect one of their household members of being infected. One observable way this may manifest is in *de facto* contact-testing behavior within households of SCAN participants. This would occur, for example, if an individual received a positive test, and then more members of their household subsequently enrolled. We find evidence of this in 17 households, shown in **Figure 3**. In total, 30 tests taken after one household member tested positive yielded 11 positive tests (37%), a markedly higher proportion positive than we see in SCAN overall.

This self-selected contact-testing behavior provides an efficient route for identifying individuals at high risk of COVID infection, and it provides hints of the rich value that expanded contact-tracing initiatives will have in intercepting COVID transmission. But it also introduces bias into the data informing community prevalence estimation by mixing in clustered samples with elevated household attack rates. Removing the participants who enrolled after another household member received any result (n = 704, 13 positive), reduces the odds ratio of infection for a participant living in a household with 4+SCAN participants to 4.3 [95% CI 2.1 - 8.7]. In removing these observations from the prevalence model, we reduced total sample size by 5.7%, and total positive samples by 13.1%. In order to minimize the impact of this behavior on community prevalence estimation, we are currently excluding all samples exhibiting onward household testing from the sample when estimating prevalence. These samples were removed prior to our generating the data panels in **Figure 2**. We recognize that while this behavior is observable at the household level, there could be other forms of respondent-driven sampling behavior that may be occuring, for example, among friends or colleagues, which we are not able to observe and adjust for.

In addition to removing these observations, we include a random effect term for household membership in the prevalence model (see **Appendix 2**), which is meant to account for clustering of SCAN samples in households. We plan to return to this topic, and further explore epidemiologically relevant household-level risk factors in a future technical report.

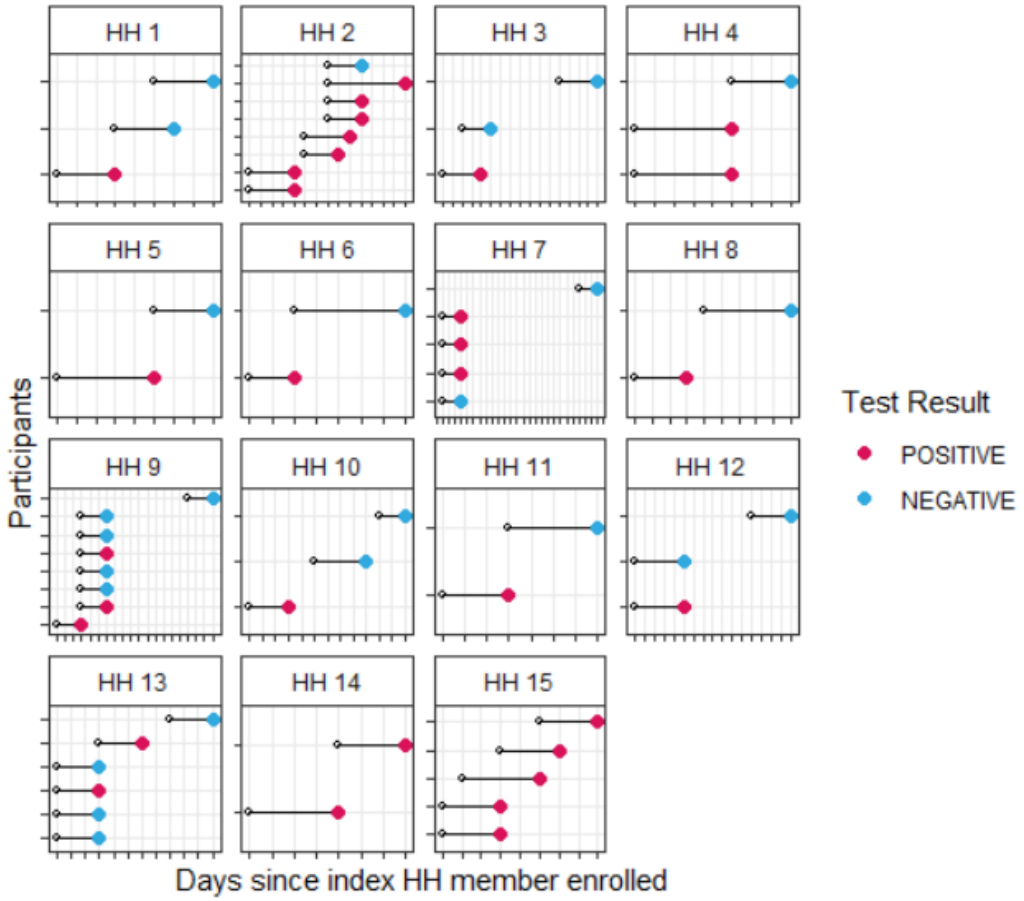

Figure 3: Each plot represents a household with respondents who enrolled in SCAN after one of their household members tested positive. Hollow circles indicate the date of enrollment, and the red or blue circles represent a positive or negative test result, respectively.

Geographic stratification

Figure 4 summarizes the estimated relative risk of receiving a positive test result for a given geographic area within King County, compared to the average risk across the county. We find that geography is correlated with positivity of SARS-CoV-2 in samples collected though SCAN. We describe geographic variation at the level of 2010 Census Public-Use Microdata Areas (PUMAs). PUMAs are collections of census tracts with populations of at least 100,000; King County is divided into 16 PUMAs. **Figure 4** shows odds ratios -- values greater than one are associated with relatively higher odds of testing positive in a PUMA, while values lower than one are associated with a lower number of positive results in the PUMA, relative to average positivity throughout the county. PUMAs in Seattle and parts of the Eastside (Redmond and Kirkland) had lower overall values, while PUMAs in the south part of the county, including Federal Way, Des Moines, Auburn, Maple Valley, and Tukwila, had relatively more positive results. This broad north-south geographic risk gradient is similarly identified through case reports, as reported by Public Health Seattle & King County (link).

The relative differences across PUMAs are marked with substantial uncertainty. For example, the PUMA with the lowest odds ratio, Northeast Seattle, has a 95% uncertainty interval that crosses 1 (meaning it is possible that it is indistinguishable from the county mean). We represent this uncertainty in the top panel of **Figure 4**, where red or blue dots represent the spread of possible values for each PUMA's odds ratio. PUMAs with mostly red dots indicate more confidence in heightened risk.

At the moment, and in light of this high degree of uncertainty, the connection between geography and testing positive in SCAN is at best a descriptive feature of the data, and we emphasize that it does not tell us that a person's location in King County determines their risk for getting COVID. Better understanding of infection risk would require a more detailed understanding of the routes of transmission and the connections between individuals in the population. As we learn more about the epidemiology in King County, we may be able to better understand the underlying causes of the association between geography and testing positive that we see in SCAN data.

It is also likely that positivity within PUMAs is changing over time. For example, the epidemic in King County began with a cluster of cases in Kirkland in late February, before SCAN started; in SCAN data, Kirkland shows lower overall risk throughout a period which began in late March. Currently, SCAN has not collected enough samples from across the county to determine time trends at the PUMA level.

Due to the strong association with geography and skewed SCAN sampling across PUMAs (see next section), we adjust for geography when estimating prevalence (see **Appendix 2**).

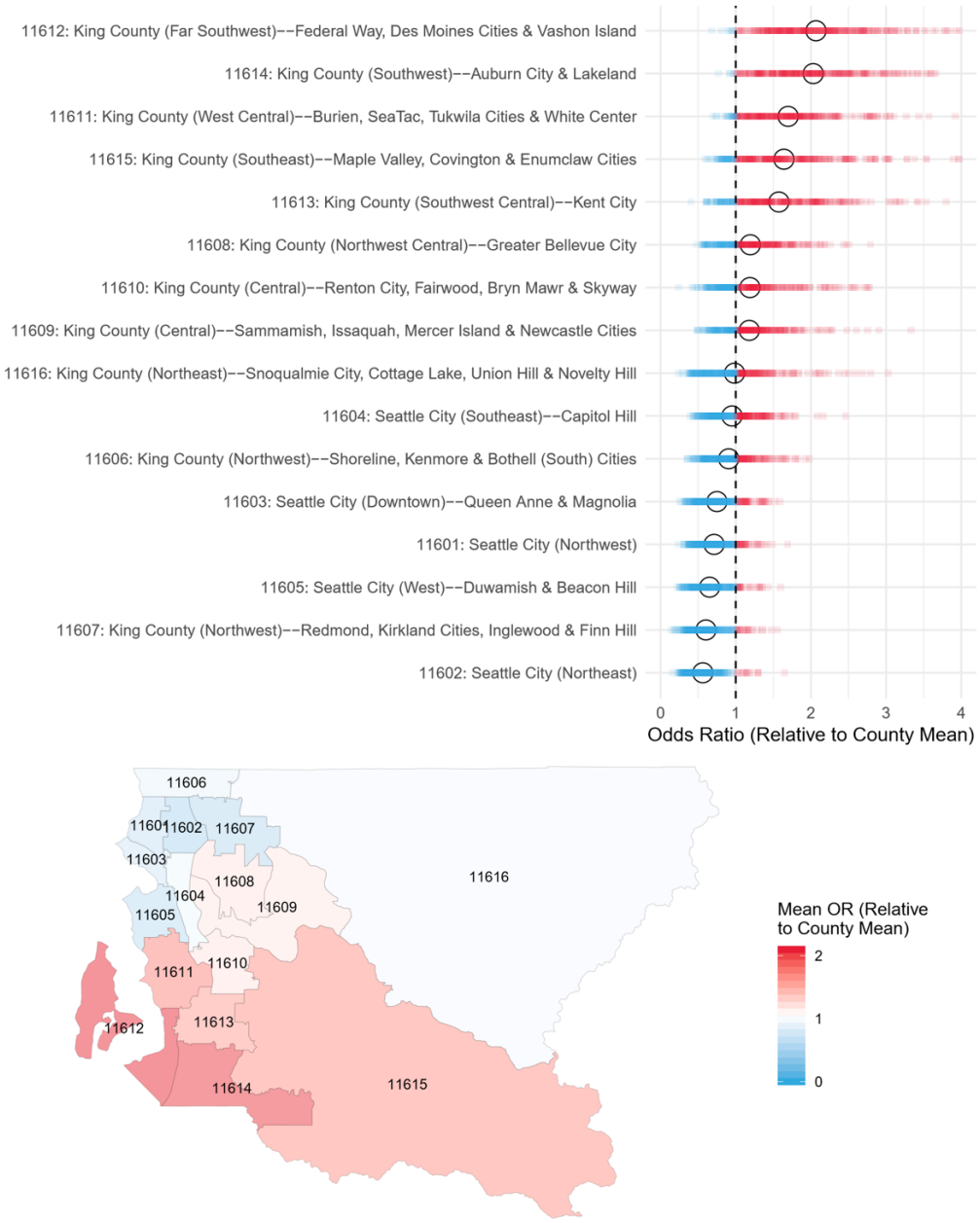

Figure 4: Odds ratios for PUMAs in King County. A value greater than one indicates a higher odds of infection relative to the county mean, and a value between zero and one represents relatively lower risk. The top panel shows uncertainty in these values by plotting 300 posterior draws from the model, more draws on either side of one indicate a relatively higher or lower posible value. The map on the bottom panel shows the mean odds ratio for each PUMA - the same value plotted with a hollow black circle on the top.

Weighting prevalence estimates for imbalance in sample characteristics

As discussed in the first section of this report, SCAN is striving to consistently recruit a sample that represents the general population of King County in several demographic characteristics. As shown in **Figure 1**, the distribution of sample coverage across several demographic strata does not yet completely align with the population distribution in the county. In this section we will discuss adjustments for sample imbalance in geography, age, and reported symptom status.

SCAN enrollment has not yet been evenly distributed across PUMAs. **Figure 5** shows the ratio of the proportion of SCAN samples to the proportion of population living in each PUMA. A ratio of 1 would indicate even representation in the sample. We generally find undersampling in the southern half of the county and oversampling in the northern half. Often, the same parts of the county that have higher positivity in SCAN are also under-sampled. To adjust for this, we give samples from under-represented PUMAs higher weight when estimating prevalence. See **Appendix 2** for more details on post-stratification weighting.

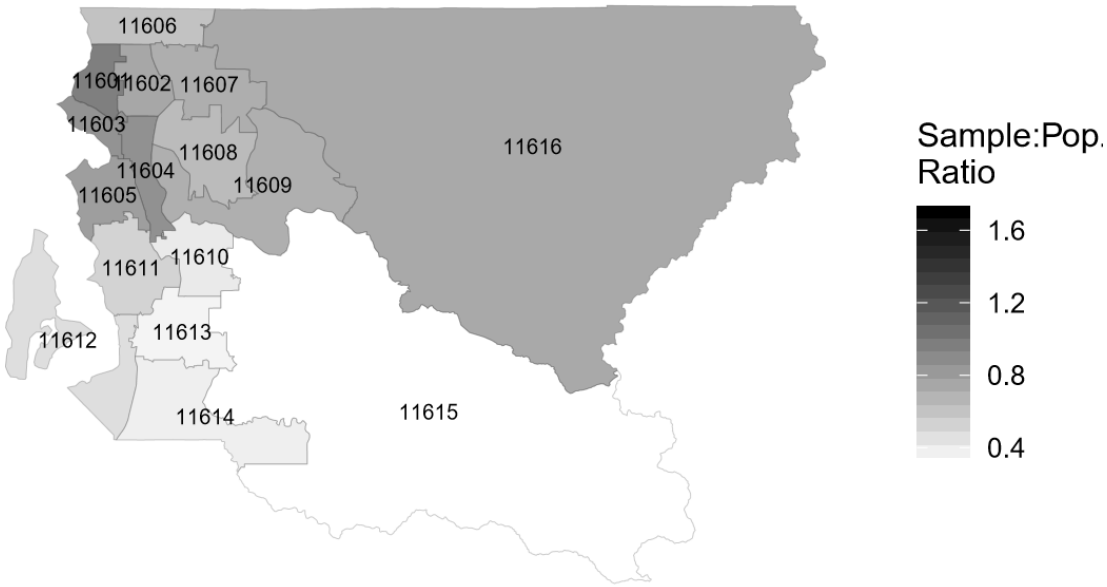

Figure 5: Map showing the ratio of the proportion of SCAN samples to proportion of King County population by PUMA. An evenly distributed sample of King county would mean every PUMA had a ratio of 1. The map indicates undersampling in the southern half of the county and oversampling in the northern half of the county.

Figure 1 shows that SCAN is also under-sampling those under 20 years old and those over 80 years old. While the role of age in infection and transmission is an active topic of scientific inquiry, we did not find evidence to support a relationship between age and infection in SCAN data. Relative to 20-59 year-olds, the odds of infection for 0-4, 5-19, and 60+ year-olds were 2.4 [95% CI 0.7 - 6.9], 0.7 [95% CI 0.2 - 2.1], and 0.9 [95% CI 0.5 - 1.5] times higher. Since age is not significantly correlated with positivity in SCAN data so far, we do not adjust for it when we model prevalence. This may be in part a function of small sample sizes in the younger and older age groups, and is something we plan to monitor for future technical reports.

SCAN was designed to collect community samples from those with and without COVID-like illness (CLI) symptoms. To enroll in SCAN, participants respond to a screener that asks their age, zip code, and the following yes/no question regarding CLI: “In the past week, have you been sick with a new fever, a new or worsening cough, or a new or worsening shortness of breath?”. Enrollment is capped at different levels for those reporting CLI or not. While the majority of the population will not be experiencing CLI at any given time, SCAN is designed to sample more from those reporting CLI. Since sampling is purposely stratified by self-report of CLI, we must account for this when estimating prevalence.

In order to better understand how our CLI-stratified sample relates to the broader population, the SCAN website hosts a short survey which is open to anyone regardless of participation in SCAN swab collection. The survey asks the same CLI symptom question as the screener. This survey operates while the screener is off and not accepting enrollments for the day. This gives us an estimate of how many people would self-report CLI and be screened into each arm were they able to enroll. This short survey has received 28,903 responses, yielding an overall CLI proportion of 15.4%. **Figure 6** shows how this value has changed over time, declining from about 19.4% in early April to about 13.5% in early May. Concurrently the proportion of enrollees who self-reported CLI in the screener has declined over the same period, from about 70% of samples in early April to about 40% in early May.

To account for this source of non-representativeness in the SCAN sample, we first assume that we can divide the population into two mutually exclusive groups on any given day, one that reports CLI, and one that does not. We use the results from the open CLI survey to weight the sample according to the population. For example, if on a given day 15% of the population would self-report CLI, and 60% of the sample screened in by reporting CLI, the weight for a CLI-screened individual would be 0.15/0.60 = 0.25 and the weight for a non-CLI screened individual would be 0.85/0.40 = 2.15. In other words, people without CLI symptoms are much more common in the general population but under-represented in the sample than people with CLI symptoms, so one participant not reporting CLI symptoms on this given day is given the same weight as 2.15/0.25 = 8.6 people reporting CLI. In practice, we calculate these weights as averages over six-day periods. Since SCAN launched, the ratio of non-CLI to CLI has ranged from 11.6 in the period from March 23 to March 28 to 5.5 in the period from May 4 to May 9.

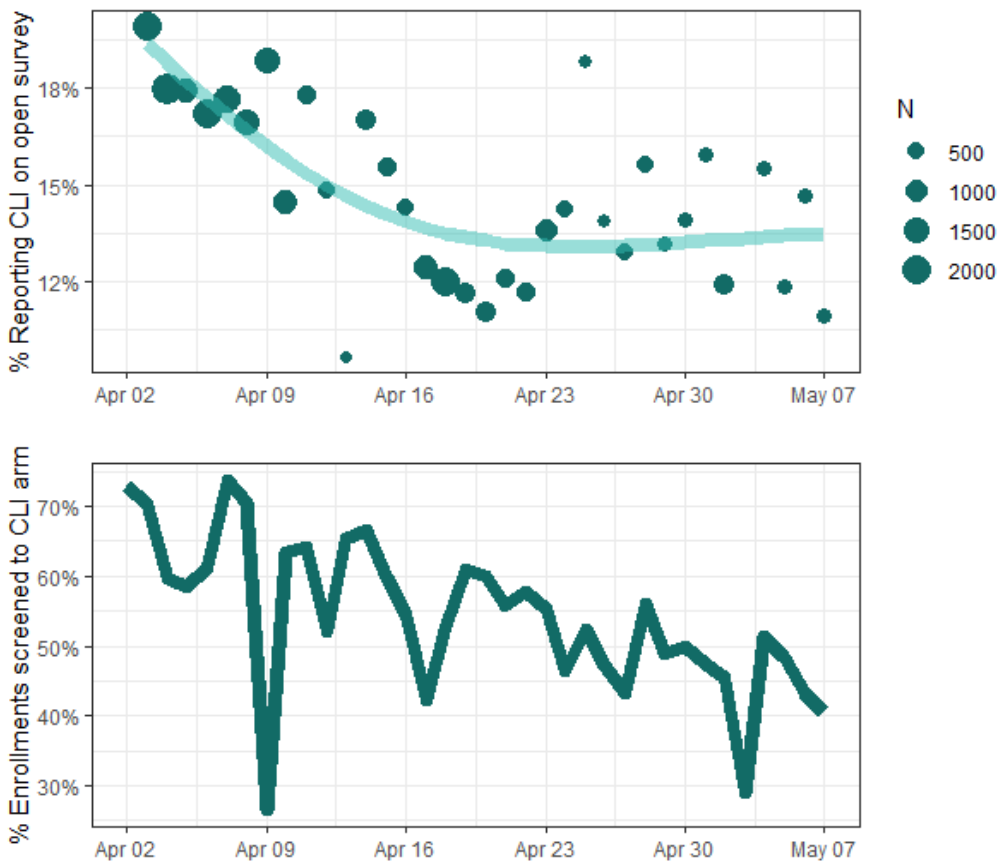

Figure 6: TOP: Proportion responding to the open CLI survey. Daily proportions are plotted as points, with size determined by the number of daily responses. The light green line is a fitted smooth natural cubic spline curve. BOTTOM: Proportion of enrollees screening in with self-reported CLI by day. The ratio of the numbers in these two panels is used to weight samples in each screened arm.

Appendix 1: Consistency with other prevalence estimates and inferred epidemic attributes

The prevalence estimate from SCAN is in close agreement with recent estimates based on transmission modeling of COVID-19 in King County. Since the transmission model is fit to cases and mortality reported to the Washington Disease Reporting System, which includes COVID-19 positives from King County hospitals and clinics, agreement between the two estimates lends confidence to our overall understanding of the epidemiological situation in King County. For a detailed description of the transmission model, see the associated technical report.

Estimates from the two sources are compared in **Figure S1**'s top panel. The beginning of SCAN enrollment in late March coincided with the transmission model's estimated peak in prevalence. Both the transmission model and SCAN show a subsequent decline, steep at first, but slowing significantly by the end of April. In the transmission model, this attenuation implies that the effective reproductive number (R_e), a measure of the transmission rate, was below the critical $R_e = 1$ threshold for declining transmission for some time in late March to mid April, but by late April R_e rose to be statistically indistinguishable from 1, leading to an attenuated decline in prevalence over time. The slow-down is somewhat more pronounced in SCAN estimates compared to those from the transmission model, but estimates remain consistent within each models' respective uncertainty bounds. Transmission modelling utilizing more recent case data indicates that transmission has continued to decline through May, albeit at a slower pace than observed in early April. The model estimates that by May 27th the prevalence of active infections was 16 [95%CI 2 to 33] per 10,000. The top panel of **Figure S1** compares SCAN prevalence estimates with those from the model.

Consistency between SCAN and the transmission model gives us confidence in the transmission model's assumptions and findings associated with other aspects of COVID epidemiology in King County. Some examples are shown in **Figure S1**'s bottom panel. Since the transmission model projects infections in King County since the start of the local epidemic, it can be used to estimate cumulative incidence over time, and according to the model, on May 27th, 2.7% [95% CI: 1.0% - 6.0%] of King County's population had already been infected with COVID-19. This estimate of total incidence further implies that only 16.0% [95% CI: 6.0%-35.2%] of infections were eventually reported to the WDRS.

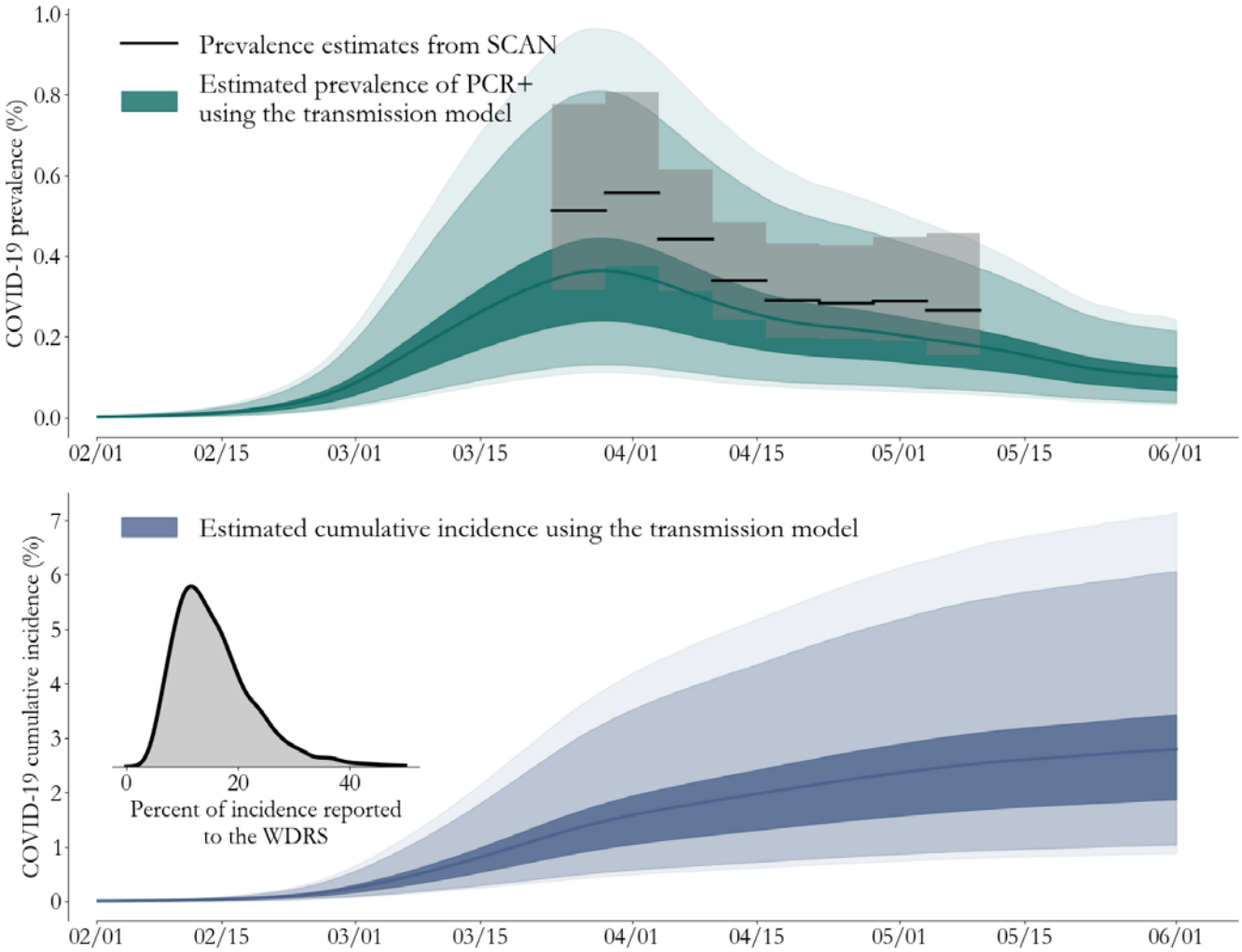

Figure S1: TOP: SCAN prevalence estimates (mean in black, 95% CI in grey) agree with estimates from a transmission model fit to test and mortality data from Washington Department of Health (purple, 50%, 95%, and 99% CIs shaded). BOTTOM: The transmission model can be used to estimate cumulative incidence (blue) and the case detection rate (inset) consistent with SCAN prevalence. Note that the prevalence estimate from the transmission model shown in green only includes those infections currently detectable by PCR, a subset of all active infections that excludes those in the latent period immediately following exposure.

Appendix 2: SCAN Prevalence model

Prevalence updates

In this report, we discussed several changes to the prevalence estimation model we have made since our last technical report where we first introduced the community prevalence estimate. These include adjusting for geography, removing the adjustment for age, varying weighting of self-report CLI in time, and removing samples that were collected after another household member had a result returned. We have also implemented the following methodological updates:

- First, we have replaced the use of the date of enrollment with the date of collection, as samples are subject to delays in shipping and in participant collection -- SCAN participants often do not collect their sample on the day they enroll. Participants write the date of swab collection on their sample tube. Most (82%) of SCAN participants collect their samples one day after they enroll online.

- In the previous iteration of our prevalence model, we allowed the effect of time to vary with little constraint. This was implemented using independent fixed effects for each time period. This approach allows for sudden discontinuities in the estimated prevalence time series due to chance -- for example, if very few positive samples happened to be collected in one time period. Meanwhile, based on our understanding of disease transmission, and the known duration of infection for SARS-CoV-2, it is understood that prevalence should vary somewhat smoothly in time, and that sudden discontinuities in a prevalence time series are not possible. We have now updated the model to account for time using a random walk of order 2 (RW2) process which assumes prevalence varies smoothly in time. As such, prevalence estimates in any particular period are also informed by the data from neighboring periods.

- As we discussed in the report, some SCAN participants can enroll using priority codes, thus allowing them to bypass the self-report CLI screener. This recruitment approach is used to help improve representativeness of SCAN. Thus far, priority code targeting has focused on recruiting children: the median age of a participant code enrollee is 7 years. Our current prevalence-estimation strategy requires that all participants be assigned to either the self-report CLI or non-self-report CLI screening arm. To assign priority code participants to a screened arm, we used their responses to the individual symptom questionnaire. If they selected either cough, fever, or shortness of breath, then they were assigned to the self-report CLI arm. One hundred ninety-seven participants have enrolled using priority codes, 21 of whom were assigned to the self-report CLI arm, and 176 of whom were assigned to the no-self-reported CLI arm.

- Finally, we note that 210 samples were excluded from data used in the prevalence model because we could not match them to a PUMA inside King County or identify their address in order to match their household membership with other participants.

Further in this appendix, we show how each of these changes impacts the prevalence estimate.

Prevalence estimation methodology description

Since SCAN does not test a random sample of individuals, we can not estimate prevalence (or the probability of infection P(I)) directly as the proportion of positive tests. Instead, SCAN samples those reporting CLI symptoms and those not reporting CLI symptoms at separate proportions. A direct estimate from each group would yield two conditional prevalences (P(I|S) - the probability of infection given reporting CLI symptoms; and P(I|A) - the probability of infection given not reporting CLI symptoms).

We assume that any given person in the population reports or does not report CLI symptoms, such that a weighted sum of the conditional prevalences yield a population prevalence: P(I) = P(I|S)*P(S) + P(I|A)*P(A). Where P(S) and P(A) are the proportion of the population reporting CLI symptoms or not. We estimate P(S) and P(A) from the open survey described in this report, and allow them to vary in time. We estimate P(I|S) and P(I|A) using a statistical model that can further account for biases in the raw data, for example as emerging from imbalance sampling across geography and household clustering, as well as to account for changes in prevalence over time. Specifically we fit the following model:

$$C_{i,h,p} \sim Bernoulli(p_{i,h,p})$$

$$logit(p_{i,h,p}) = \beta_c CLI + \eta_h + \nu_p + \phi_t$$

$$\eta_h \sim_{iid} N(0, \sigma_\eta^2); \quad \nu_p \sim_{iid} N(0, \sigma_\nu^2); \; \phi_t \sim RW2(\sigma_\phi^2)$$

Where:

- $C_{i,h,p}$is the test result (1=positive, 0=negative) for individual i living in household h which is located in PUMA p. $C_{i,h,p}$is assumed to follow a Bernoulli distribution with risk of infection$p_{i,h,p}$
- $p_{i,h,p}$is modeled as the logit-transformed sum of the following components:
    - CLI is an indicator for screening into self-report CLI (0=no CLI, 1=CLI) with regression coefficient β_c
    - η_h is a random effect for residence (a.k.a. household)
    - ν_p is a random effect for PUMA of residence
    - $\phi_t$is a random walk of order 2 (RW2) which captures change in time. Time is split into eight 6-day periods.

Analyses were run in R version 3.5.1. The statistical model was fit using the INLA package (version 18.07.12). Default priors were used: $log(1/\sigma^2) \sim loggamma(1,0.00005)$ and $\beta \sim N(0,1000)$

For prediction of prevalence at any given time period, we first drew 1000 posterior predictive samples for each SCAN participant. Individual-level predictions represent inferred individual risk of a positive test result given the factors included in the model. Distribution across posterior samples capture parameter uncertainty from the fitted model. Household-level random effects are not used in model fitting, but not in prediction. Each individual was assigned two weights:

- A CLI weight $w_{1,i}$ which is described in this report as weight as the ratio of the estimated proportion of the population that would report and screen into each CLI arm to the proportion in the SCAN sample in each arm. This accounts for P(S) or P(A) in the above description. This weight varies by 6-day period.
- A PUMA weight $w_{2,i}$. Using population projections for King County PUMAs, the weight is defined as the ratio of the sum of the population in each puma to the sum of $w_{1,i}$ in each PUMA.

Each of the 1000 posterior draws was collapsed across all individuals using a weighted mean with weight $w_{1,i}w_{2,i}$ for each of the eight 6-day time periods. The remaining eight time-specific vectors of 1000 population-weighted samples was summarized to a mean, lower quantile (2.5%), and upper quantile (97.5%).

Prevalence model sensitivity to model choices

As discussed throughout this report, we made a number of model choices in order to arrive at an estimate of prevalence. In the plots to follow, we will show comparisons of prevalence trends with or without each of these choices implemented to illustrate the impact of each choice on the estimate of prevalence.

To start, **Figure S2A** compares the updated methods described in this report with the estimates that would have been generated using the methods described in the first technical report.

The other panels compare the full model from this report with models that don't account for PUMA (**Figure S2B**), include age (**Figure S2C**), use fixed effects over time (**Figure S2D**), do not allow CLI weights to vary in time (**Figure S2E**), or do not remove household contacts that enrolled in SCAN after a household member already received a test result (**Figure S2F**).

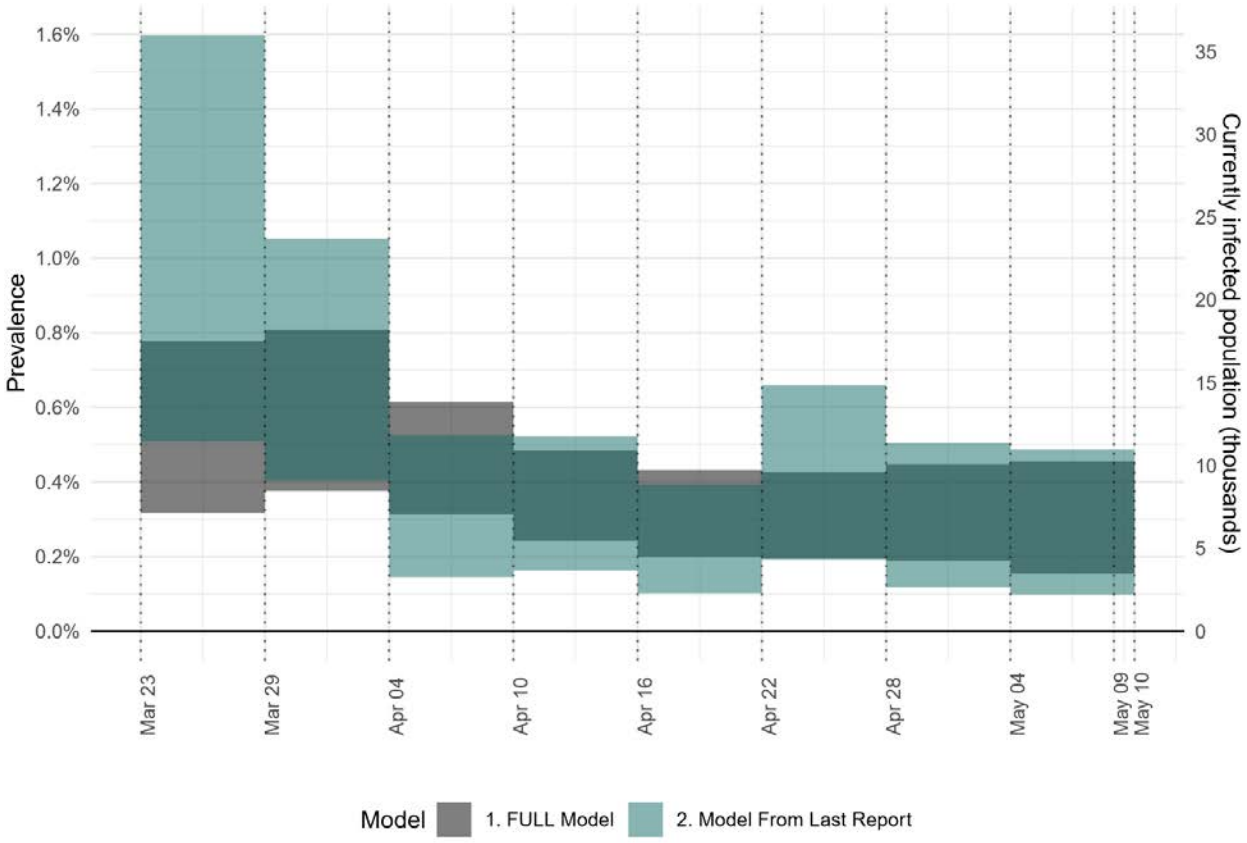

Figure S2A: Comparison of the full model from this report with the model used in the first technical report.

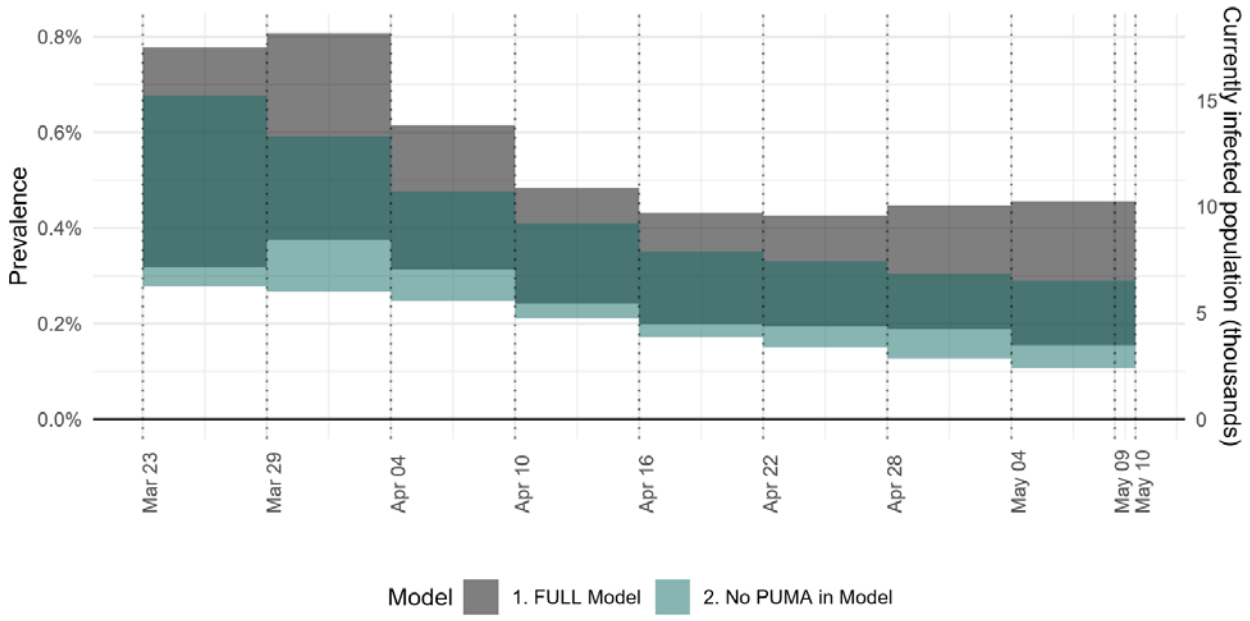

Figure S2B: Comparison of the full model from this report with a model that does not account for PUMA. Including PUMA in the model has the effect of increasing prevalence estimates - since samples from parts of the county where positivity appears to be higher are also generally the areas that are under-represented in the SCAN sample (and they are thus given higher weights in prediction).

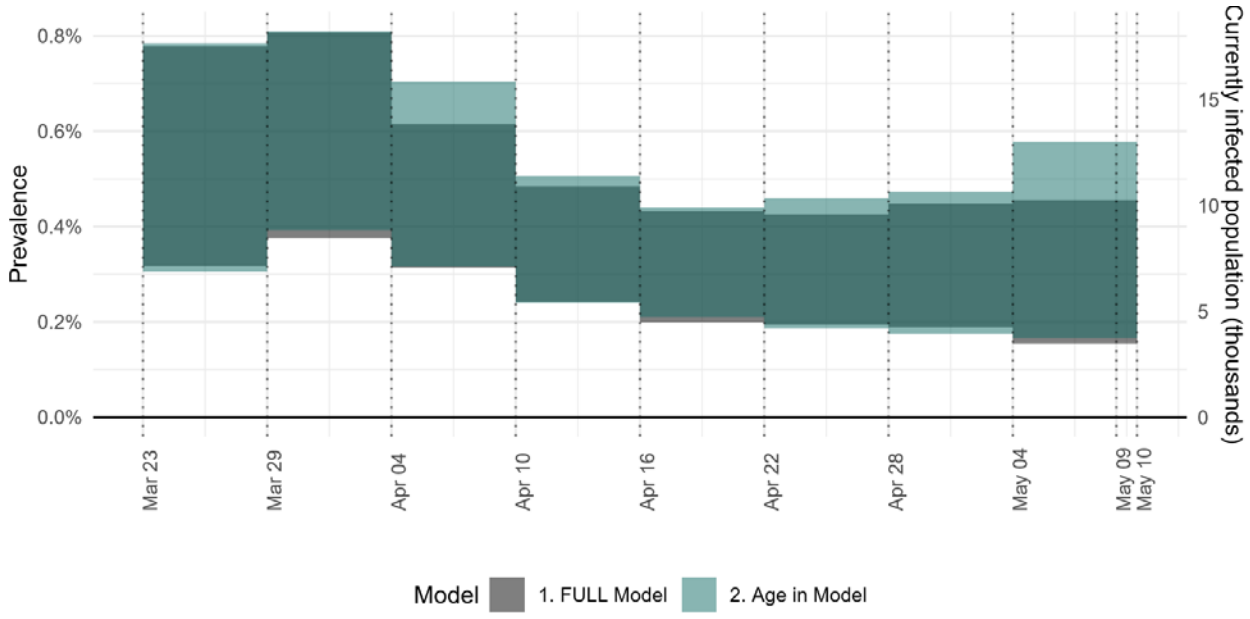

Figure S2C: Comparison of the full model from this report with a model that includes age. We chose not to include age in the current model because there is no effect of age observed in SCAN data. The uncertainty in the model with age is slightly wider because the age effects cannot be estimated with precision.

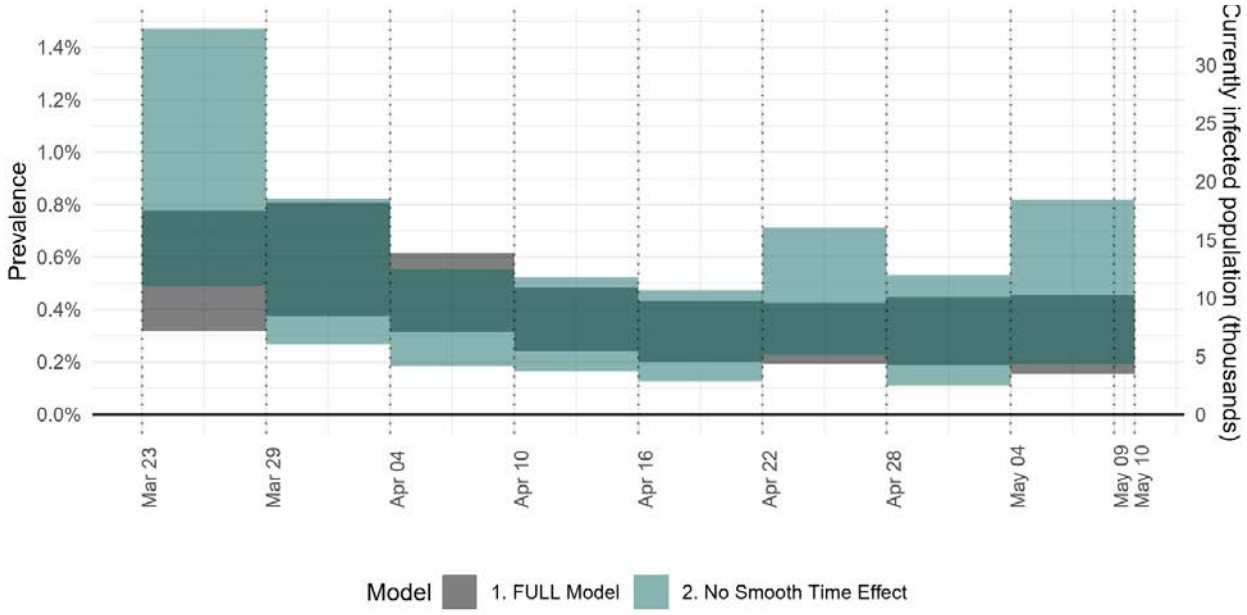

Figure S2D: Comparison of the full model from this report with a model that uses fixed effects for time. The fixed effects allow for a more unconstrained effect of time, including unrealistic disjoints. Since the smooth (RW2) term in the full model uses information from nearby time periods, uncertainty is narrower in the full model.

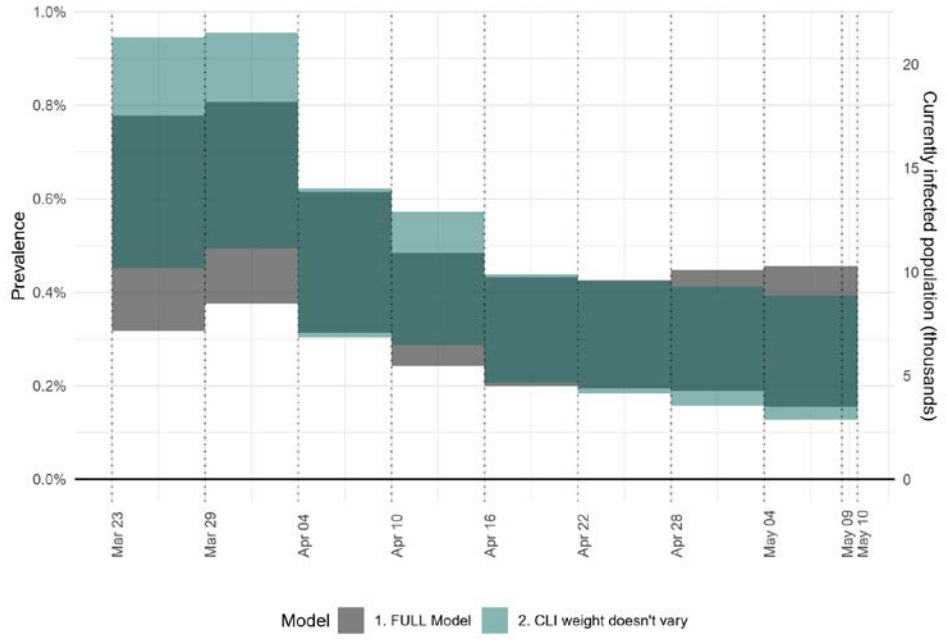

Figure S2E: Comparison of the full model from this report with a model that does not allow CLI weights to vary in time. We assumed the population fraction for CLI was 15.5% - the average CLI response rate from the open survey. Scaling by samples in self-report CLI and non self-report CLI arm, the overall used ratio was 7.25:1, non CLI to CLI. In the earlier period, this leads to over-weighting of the CLI sample, leading to an overestimate of prevalence early on.

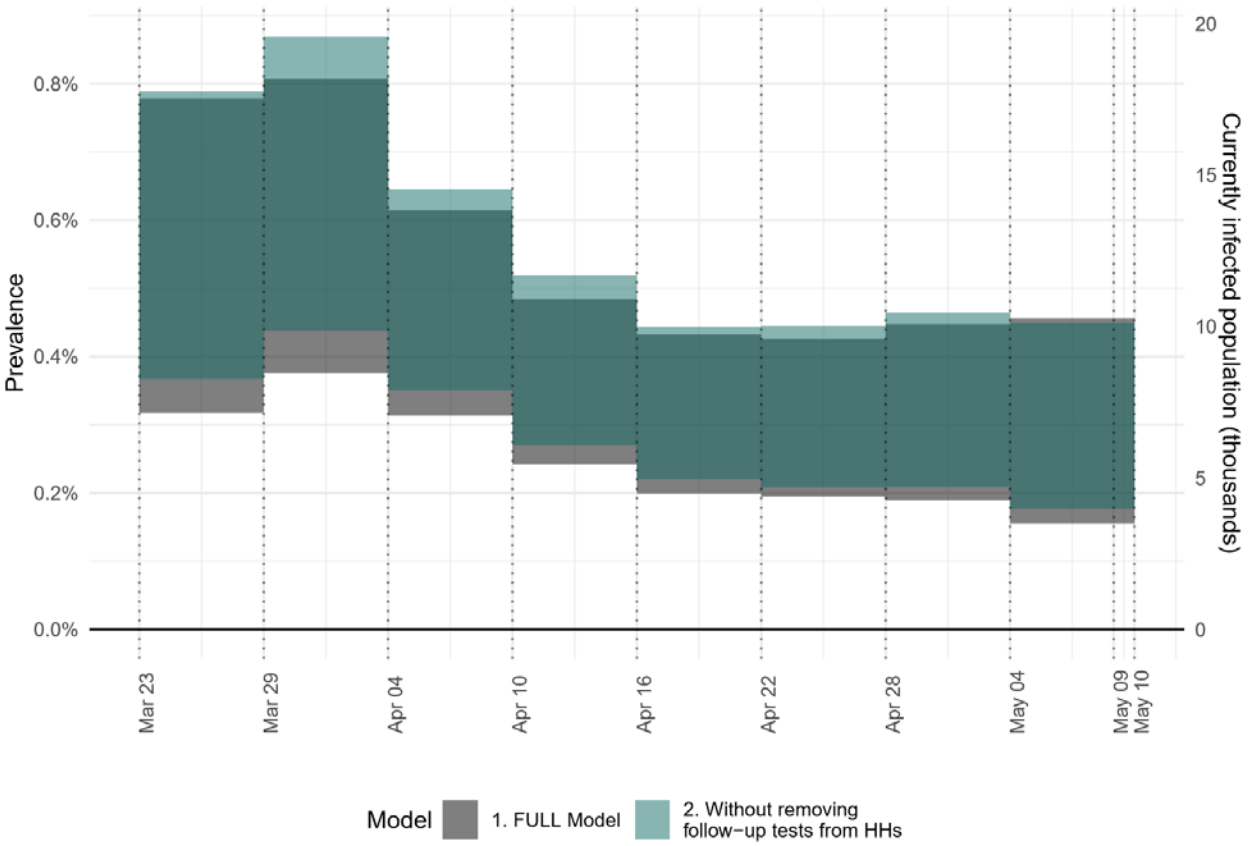

Figure S2F: Comparison of the full model from this report with the same model, but which did not remove household contacts that enrolled in SCAN after a household member already received a test result. As we discussed in this report, some households were utilizing SCAN for within household contact tracing following a positive result. We suspect that including these samples from households with elevated household attack rates in the data informing community prevalence would bias our estimate of community prevalence upward.

SCAN acknowledgements

This technical report was drafted based on SCAN data by Roy Burstein, Karen Cowgill, Michael Famulare and Jay Shendure. Additional thank you to Niket Thakkar for contributions on the disease transmission model.

SCAN is a partnership between the team behind the Seattle Flu Study (SFS) and Public Health — Seattle & King County (PHSKC). SCAN is being executed by the Brotman Baty Institute for Precision Medicine (BBI).

SFS Principal Investigators: Helen Chu, Michael Boeckh, Janet Englund, Michael Famulare, Barry Lutz, Deborah Nickerson, Mark Rieder, Lea Starita, Matthew Thompson, Jay Shendure, and Trevor Bedford

Jeff Duchin (PHSKC) and **Jay Shendure** (BBI) serve as co-leads for managing the SCAN partnership.
Mark Rieder (BBI) serves as SCAN Program Director.
Helen Chu (UW Medicine, BBI), **Janet Englund** (Seattle Children's) and **Michael Boeckh** (Fred Hutchinson Cancer Research Center) co-lead SCAN's clinical operations
Trevor Bedford (Fred Hutchinson Cancer Research Center, BBI) and **Michael Famulare** (IDM) co-lead SCAN data management and epidemiological modeling
Lea Starita (UW Medicine, BBI) and **Deborah Nickerson** (UW Medicine, BBI) co-lead SCAN laboratory operations
Tina Lockwood (UW Medicine, BBI) serves as Clinical Laboratory Director, Northwest Genomics Center
Matthew Thompson (UW Medicine) and **Barry Lutz** (UW) co-lead the SCAN's kit fabrication and community feedback team.
Karen Cowgill (PHSKC) serves as COVID-19 response epidemiologist and PHSKC liaison to SCAN
Stephanie Schrag (CDC) serves as a consultant to SCAN from the CDC US COVID-19 Response team.

SCAN Key Personnel*: Amanda Adler, Elisabeth Brandstetter, Roy Burstein, Shari Cho, Kairsten Fay, Chris Frazar, Rachel Geyer, Peter Han, Jessica Heimonen, Jameson Hurless, Misja Ilcisin, Gernot Kalcher, Ashley Kim, Eric Konnick, Jack Henry Kotnick, Kirsten Lacombe, Jover Lee, Jennifer Logue, Victoria Lyons, Denise McCulloch, Jennifer Mooney, Robin Prentice, Matthew Richardson, Jonathan Rosoff, Jaclyn Ruckle, Thomas Sibley, Sanjay Srinivastan, Melissa Truong, Maggie Van de Loo, Caitlin Wolf. [*not a complete list of individuals contributing to SCAN]

Amazon Care provides infrastructure and logistics capability for this effort in the greater Seattle area, along with other delivery partners.

SCAN is funded by Gates Ventures, the private office of Bill Gates.