Estimating epidemiologic dynamics from single cross-sectional viral load distributions

James A. Hay*[1,2,3†], Lee Kennedy-Shaffer*[1,2,4†], Sanjat Kanjilal[5,6], Marc Lipsitch[1,2,3], Michael J. Mina[1,2,3†]

* These authors contributed equally.

[1] Center for Communicable Disease Dynamics, Harvard T. H. Chan School of Public Health, Boston, MA
[2] Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA
[3] Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA
[4] Department of Mathematics and Statistics, Vassar College, Poughkeepsie, NY
[5] Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, MA
[6] Department of Infectious Diseases, Brigham and Women's Hospital, Boston, MA

[†] Correspondence should be addressed to: jhay@hsph.harvad.edu (J.A.H.), lkennedyshaffer@vassar.edu (L.K.S.), mmina@hsph.harvard.edu (M.J.M.)

Funding
This work is supported by U.S. National Institutes of Health Director's Early Independence Award DP5-OD028145 (MJM, JAH), the Morris-Singer Fund (LKS and ML), U.S. Centers for Disease Control and Prevention Award U01IP001121 (LKS and ML), and U.S. National Institute of General Medical Sciences award U54GM088558 (ML, JAH, LKS).

Data availability
All code is available at https://github.com/jameshay218/ct_dynamics_preprint. The Ct values from Massachusetts will be made available in a later version of this manuscript.

Abstract

Virologic testing for SARS-CoV-2 has been central to the COVID-19 pandemic response, but interpreting changes in incidence and fraction of positive tests towards understanding the epidemic trajectory is confounded by changes in testing practices. Here, we show that the distribution of viral loads, in the form of Cycle thresholds (Ct), from positive surveillance samples at a single point in time can provide accurate estimation of an epidemic's trajectory, subverting the need for repeated case count measurements which are frequently obscured by changes in testing capacity. We identify a relationship between the population-level cross-sectional distribution of Ct values and the growth rate of the epidemic, demonstrating how the skewness and median of detectable Ct values change purely as a mathematical epidemiologic rule without any change in individual level viral load kinetics or testing. Although at the individual level measurement variation can complicate interpretation of Ct values for clinical use, we show that population-level properties reflect underlying epidemic dynamics. In support of these theoretical findings, we observe a strong relationship between the time-varying effective reproductive number, *R(t)*, and the distribution of Cts among positive surveillance specimens, including median and skewness, measured in Massachusetts over time. We use the observed relationships to derive a novel method that allows accurate inference of epidemic growth rate using the distribution of Ct values observed at a single cross-section in time, which, unlike estimates based on case counts, is less susceptible to biases from delays in test results and from changing testing practices. Our findings suggest that instead of discarding individual Ct values from positive specimens, incorporation of viral loads into public health data streams offers a new approach for real-time resource allocation and assessment of outbreak mitigation strategies, even where repeat incidence data is not available. Ct values or similar viral load data should be regularly reported to public health officials by testing centers and incorporated into monitoring programs.

Introduction

Tracking trends in the incidence of infection during an epidemic are vital for deciding on appropriate public health response measures (*1–3*). This information can help decision-makers understand the need for and efficacy of non-pharmaceutical interventions, to plan the deployment of public health resources, and the use of scarce hospital beds and personal protective equipment. In the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) pandemic, estimating the outbreak trajectory and key epidemiological parameters such as the time-varying effective reproductive number, *R(t)*, has generally been done by using daily observed case counts, percent of tests positive, or death counts, ideally confirmed by reverse-transcription quantitative polymerase chain reaction (RT-qPCR) testing. Because of the nonspecific symptoms and variable incubation periods of COVID-19 (*4, 5*), the limited and time-varying availability of testing in many parts of the world, including in the United States, and the delay between infections and reported tests or confirmed deaths (*6*), these approaches are limited in their ability to reliably and promptly detect underlying changes in infection counts (*7*). In particular, whether changes in case counts at different times have stemmed from changes in testing or reflect epidemic growth or decay have been major topics of debate with important economic, health and political ramifications.

Virologic testing is an important method for determining the infection status of an individual and the prevalence of infection in a community. RT-qPCR testing is currently the primary approach for virus detection (*8, 9*). While these tests have quantitative results in the form of cycle threshold (Ct) values, which are inversely correlated with $\log_{10}$ viral loads, they are often reported only as binary results (*10, 11*). Previous work for other infectious diseases has focused on identifying correlations between Ct values and clinical severity and transmissibility (*12–14*). In addition, Ct values or viral loads may be useful in clinical determinations about the need for isolation and quarantine for SARS-CoV-2 (*11, 15*), in determining the phase of an individual's infection (*16, 17*) and predicting disease severity (*16, 18*).

Using viral loads for clinical and epidemiological purposes requires an understanding of the viral load kinetics over the course of infection. For SARS-CoV-2, this has not been fully characterized due to difficulties in quickly identifying and repeatedly testing asymptomatic infections. However, many salient features of the viral load trajectory have been identified using rhesus macaque models and longitudinal human studies with repeated sampling after symptom onset (*19–25*).

From these properties, assessments of infectiousness over the course of infection (*5, 26*) and population-level viral load distributions by time since infection can be reconstructed (*27*).

While individual viral load kinetic trajectories may potentially be used to diagnose the clinical course of infection and determine requirements of test sensitivities (*27*), albeit with caveats from sampling variability and different testing platforms, population-level viral loads can be used for epidemiological assessments of an outbreak. Sample surveys with virologic tests provide cross-sectional estimation of the prevalence of infection, and repeated surveys can provide information on the trends in prevalence and incidence over time (*28*). Population-level signals in measurements obtained under the same sampling protocol and using the same instrument will therefore reflect underlying epidemiological trends. However, a consistently observed but currently unexplained phenomenon is that, as the epidemic declines, the distribution of observed Ct values above the limit of detection appears to systematically decline (*15, 16*).

Previous work has used serologic data to infer unobserved individual-level infection events and population-level parameters of infectious disease spread. In particular, antibody dynamics have been used in tools to identify the infecting strain for multi-strain pathogens (*29, 30*) and to identify the time since infection for an individual (*31, 32*). At the population level, repeated sampling of antibody levels has been used to identify seasonal forcing mechanisms for recurrent epidemics (*33*), estimate an average incidence rate (*34–37*), and identify the force of infection at various time points (*32, 38*). Although virologic data may exhibit more heterogeneity than serological data, they similarly provide single point-in-time observations of underlying within-host infection kinetics. Here, we use virologic data, which can be obtained sooner after infection than antibody titers, to provide real-time monitoring of epidemic dynamics, and demonstrate an approach using a single cross-sectional survey.

In this paper, we show that the changing population distribution of Ct values obtained from positive SARS-CoV-2 samples can be used to infer epidemic dynamics. Early in the epidemic, infection numbers are growing rapidly and the average infection is recent; as the epidemic wanes, however, average time since exposure increases as the rate of new infections decreases—analogous to the average age being lower in a growing vs. declining population of organisms (*39*). Importantly, within infected hosts, Ct values change over time: the viral population undergoes rapid exponential growth after inoculation, succeeded by slower exponential decline as the infection is cleared and low levels of viral RNA persist. Surveillance sampling of infected individuals during epidemic growth is therefore more likely to measure individuals who were

recently infected and therefore in the acute phase of their infection. Conversely, sampling infected individuals during epidemic decline is more likely to capture individuals in the convalescent phase, typically sampling lower quantities of viral RNA. We first demonstrate that rate of transmission is highly correlated with the distribution of detectable Ct values in infected individuals over time using both simulated and real data from Massachusetts. We then formalize this intuition into a novel and robust method able to use viral load measurements (i.e., RT-qPCR Ct values or other forms of virus quantitation) from single cross-sectional virologic test surveys to determine an instantaneous outbreak trajectory, without need for repeated measures.

Results

Relationship between observed Ct values and epidemic dynamics

To first demonstrate how the distribution of observed Ct values changes over the course of an outbreak, we simulate viral load trajectories from infections arising under a deterministic susceptible-exposed-infectious-recovered (SEIR) model (**Fig. 1A**). At each day of the outbreak, Ct values are observed from a random sample of the population using the population-level Ct distribution model described in *Materials and Methods* and shown in **Fig. S1.** Assumed parameter values for the distribution of Ct values for a given time since infection are the point estimates in **Table S1,** capturing individual-level time courses of viral kinetics post infection and the substantial variation in observations resulting from individual variation and sampling differences. By drawing simulated samples for testing from across the population at specific time points, these simulations recreate realistic cross-sectional distributions of detectable viral loads across the course of an epidemic.

Throughout the course of the outbreak, shown in **Fig. 1,** there is a strong relationship between growth rate of new infections at each timepoint, the cross-sectional distribution of time since infection of RT-qPCR-detectable infections at that time point, and distribution of detectable viral loads in the simulated SEIR population at that time point. To infer epidemic growth rate, we would ideally observe the distribution of time since infection of infected individuals, because it is directly related to the average growth rate when those individuals were infected (**Fig. 1A-C**). However, because infections are often unobserved events, we rely on an observable quantity, such as viral load, as a proxy for the time since infection (**Fig. 1D-G)**. Throughout, we assume that there is a single infection time for each individual and ignore re-infection, as it appears to be a negligible portion of infections in the epidemic so far (*40*).

Since the viral load is related to time since infection for each individual (**Fig. 1D**), and the average time since infection varies over the outbreak (**Fig. 1C**), the distributional properties of the measured viral loads (i.e., median and skewness) vary with the growth rate of new cases (**Fig. 1H, Fig. S2**). Across individuals, this relationship holds, even after accounting for significant individual level variation in peak viral loads and growth/clearance kinetics and errors in quantifying viral load. This relationship holds for any observation model (i.e., using a different RT-qPCR instrument or in a different lab); if the distribution of observations is an estimable function of the distribution of times since infection (which it is), then the expected median and skewness of observations at a given point in time are predictable from the growth rate.

Demonstration using simulated data

We used our simulation framework to investigate how properties of the Ct distribution relate to the effective reproductive number, *R(t)*, if observed in a population similar to Massachusetts, USA. Here, *R(t)* is used essentially as a reflection of the growth, or decay, of the epidemic in time. We find that both the median Ct (which would vary with the individual assay) as well as the skewness of the Ct distribution (which would be largely test- and lab-agnostic) show a strong correlation with *R(t)* (**Fig. 2**); as *R(t)* declines, the median Ct of sampled individuals increases and, notably, the distribution becomes more negatively skewed. An important caveat for this comparison is that *R(t)* is the effective reproductive number on that day, whereas Ct values are observed from infections generated across many days in the past.

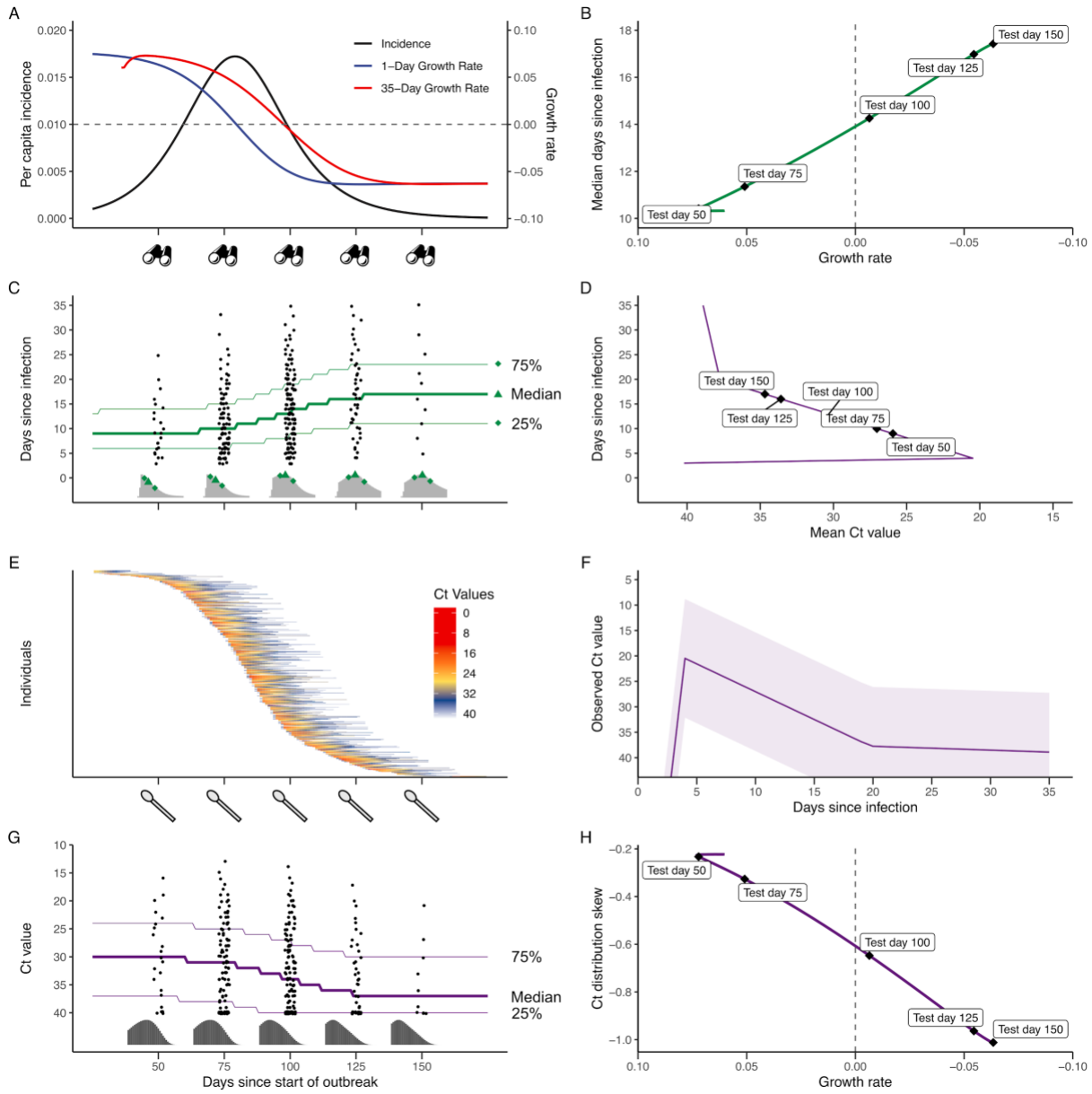

Figure 1. (A) Simulated per capita incidence, daily growth rate, and average growth rate over the preceding 35 days of detectable infections. Binocular symbols represent observation times used in subsequent plots. **(B)** Relationship between growth rate as it varies over the epidemic and the median time since infection. **(C)** Distribution of times since infection for observed, detectable infections at 50, 75, 100, 125, and 150 days after the epidemic start (histograms) and daily median (thick green line) and central 50% percentiles for time since infection. **(D)** Modeled mean viral kinetics on each day post infection. Labels show mean Ct value corresponding to the median time since infection at different points along the epidemic curve. **(E)** Simulated RT-qPCR cycle threshold (Ct) value trajectories of 500 individuals randomly sampled from the epidemic. Swab symbols represent sampling times. **(F)** Modeled viral kinetics showing median Ct value (purple line) and 95% quantiles on simulated observations. **(G)** Distribution of observed, detectable Ct values at each of the five test days (histograms) and median and central 50% percentiles (thick and thin lines, respectively) on Ct values observed over time. **(H)** between growth rate and the skew of the

observable Ct distribution. All results are based on an SEIR model in a population of 1,000,000 with R_0 = 2.5, average latent period of 6.41 days, and average infectious period of 8.79 days, and Ct values were simulated based on inferred post-infection viral kinetics. In **E**, for visual simplicity, each individual's Ct values are assumed to follow the same post-infection trajectory with only a shift in the mean; this assumption is not used elsewhere in this paper.

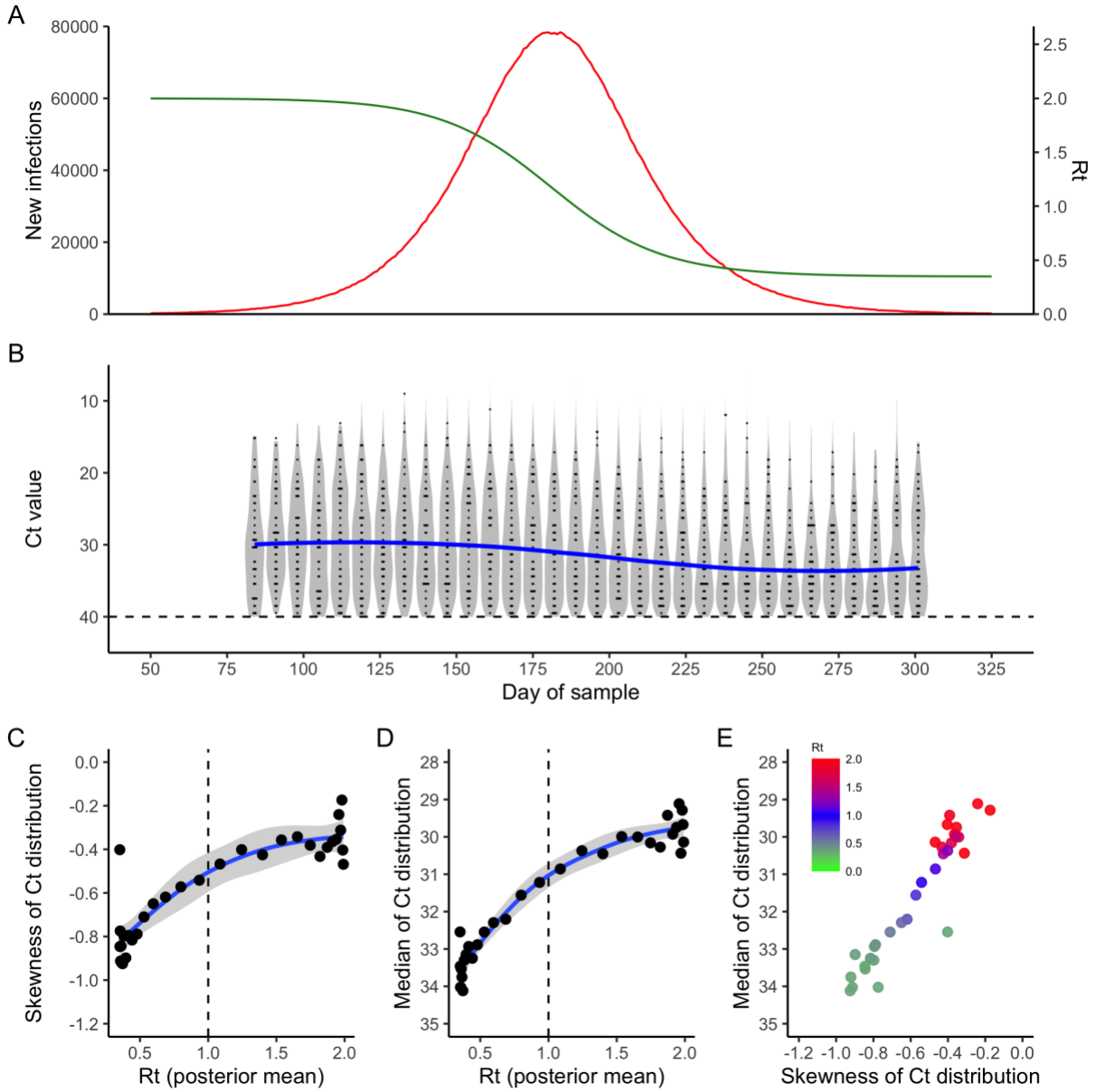

Figure 2. (A) Simulated incidence of new infections (red) and corresponding effective reproductive number over time, *R(t)* (green). **(B)** Weekly distribution of positive (detectable) Ct values. The blue line shows the smoothed median, the dotplot shows a subset of the observed, binned Ct values, and the violin plot shows densities. **(C&D)** Relationship between the skewness and median of the observed Ct distribution and the mean *R(t)* per week. Blue line and shaded region shows fitted smoothing spline +/- 1 standard error. Only weeks with more than 10 detectable Ct values are shown. **(E)** Plotted together, the median and skewness of the observed Ct distribution display a strong correlation with *R(t)* (color of point) per week.

Observation of trends in Ct values from a major tertiary care hospital in Massachusetts

To investigate if our predicted relationship between the distribution of Ct values and *R(t)* is borne out in reality, we used fully anonymized Ct values measured from nearly all hospital admissions into Brigham & Women's Hospital in Boston, MA, between April 3 and August 31, 2020, and aligned them with estimates for *R(t)* based on case counts in Massachusetts. Tests taken prior to April 3 were restricted to symptomatic patients only, while those after April 15 represented universal testing of all hospital admission, regardless of symptomatology. The median and skewness of the Ct distribution both dropped during this period, strikingly similar to our predictions for the peak and subsequent decline of an epidemic. Although the trend was the same, properties of the Ct distribution exhibited higher variation across weeks, likely resulting from changes in transmission intensity (e.g., due to the implementation of and adherence to interventions), small sample sizes in some weeks, and potential sampling bias. In particular, it is important to note that sampling was initially biased in early April towards symptomatic individuals, resulting in more individuals being sampled near the peak of their viral loads (i.e., symptom onset), and therefore skewing the Ct distribution towards lower values. Indeed, we were able to recreate this positive skew by simulating observed Ct values with early biased sampling towards symptomatic individuals (**Fig. S3**). Nevertheless, the relationship held. Importantly for monitoring potential, the relationship demonstrated slight reductions in median Ct (increases in median viral load) and slightly reduced negative skew as cases, and thus estimated *R(t)*, began to increase again in the state in July (**Fig. 3**).

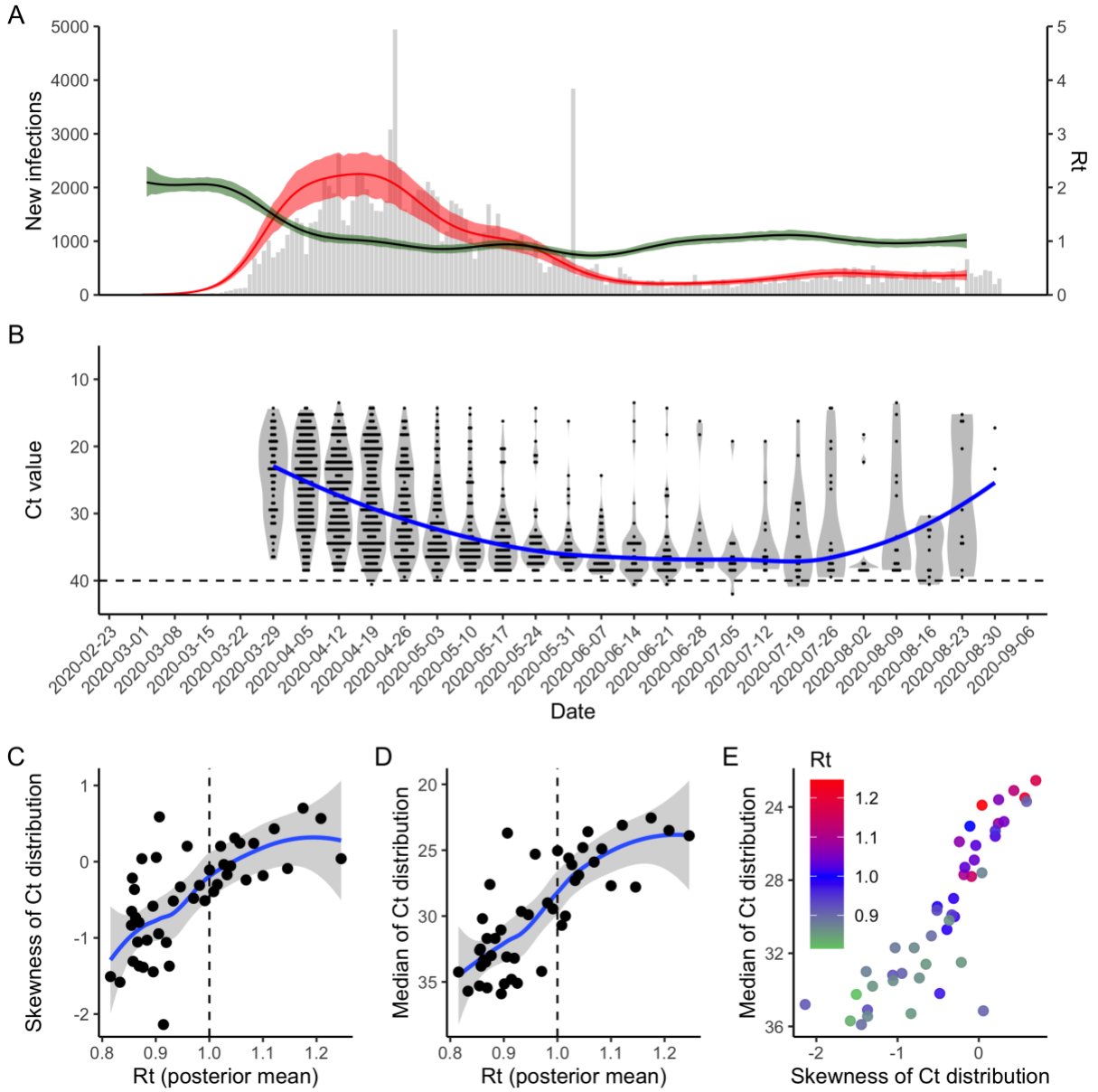

Figure 3. (A) Estimates of the effective reproductive number, *R(t)* (black line and green area), based on the daily incidence of new cases in Massachusetts (grey bars). We used the *EpiNow2* package (see Materials and Methods), which uses a Bayesian approach to first back-calculate incidence by date of infection (red line and region) and subsequently infer *R(t)*. Shaded regions show 95% credible intervals and solid lines show posterior means. **(B)** Weekly distribution of positive (detectable) Ct values from Brigham & Women's Hospital in Boston, MA. The blue line shows the smoothed median, the dotplot shows the observed, binned Ct values, and the violin plot shows densities. **(C&D)** Relationship between the skewness and median of the observed Ct distribution per day against the posterior mean daily *R(t)*. Blue line and shaded region shows fitted smoothing spline +/- 1 standard error. Only days with more than 10 detectable Ct values are shown. **(E)** As in **Fig. 2**, the median and skewness of the observed Ct distribution can be combined to demonstrate a strong correlation with *R(t)* (color of point) per day.

Epidemic growth rates can be inferred based on cross-sectional viral load data

Next, we derive a method to formally infer the epidemic growth rate given a single cross-section of observed, detectable Ct values. If the population-level distribution of Ct values on each day post infection is known, then the distribution of observed Ct values from a largely random sample of infected individuals will reflect their times since infection, which in turn reflects the growth rate when those individuals were infected. The likelihood of the observed viral loads can then be written as a function of the growth rate (*Materials and Methods*), allowing estimation of the average growth rate at any time point in the outbreak. In many cases, especially for a novel infection such as SARS-CoV-2, the population-level dynamics of viral loads and the resulting distribution of Ct values following infection may not be known with certainty. However, available data and qualitative understanding can be sufficient to define Bayesian priors for estimating key viral kinetics parameters. We therefore developed a Bayesian framework to incorporate uncertain prior distributions on the population-level Ct distribution as well as on the growth rate.

Using the same simulated population as in **Fig. 2,** we fit the model to cross-sectional samples at 5 day intervals using the Bayesian priors in **Table S1**. These analyses demonstrate how a single cross-sectional sample of individuals may be used to estimate properties of the outbreak. **Fig. 4A** shows the posterior growth rate estimates alongside the true growth rate at each time point. The credible interval widths reflect the number of detectable Ct values included in each cross-section, and also uncertainty in the viral kinetics and Ct distribution parameters. Larger sample sizes than we drew here and stronger accurate priors would help constrain these estimates. However, with large uncertainties incorporated in this particular version of the framework, the method was still able to discern if the growth rate was positive or negative, even if the point estimates were imprecise (**Fig. 4B**). **Fig. 4C and 4D** demonstrates how the model is fitted to the observed Ct distribution and the corresponding time-since-infection distribution for individuals observed on a particular day.

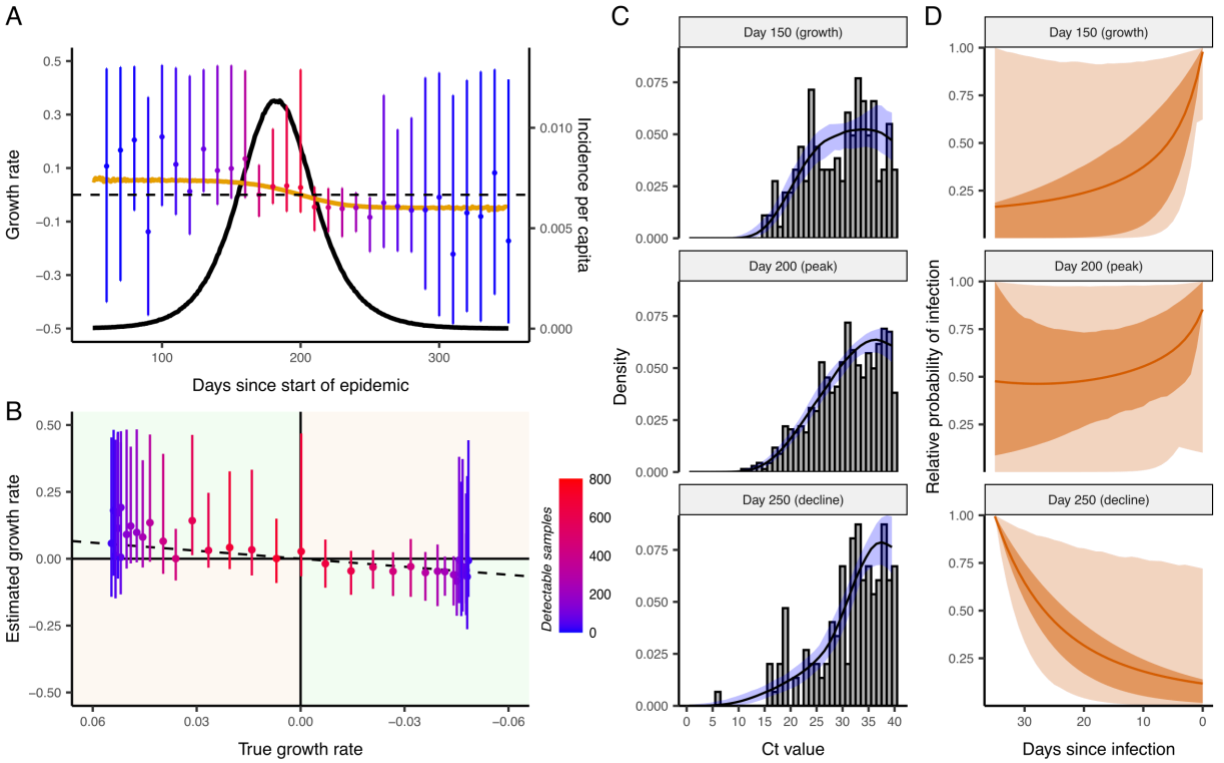

Figure 4. (A) Epidemic growth models independently fit to cross sections at 5-day intervals. The black line shows infection incidence. The orange line shows the average growth rate over the 35 days preceding each day. The pointrange plot shows the posterior median and 95% credible intervals (CIs) of the average growth rate estimated on that day. Points are colored by the number of detectable Ct values used for fitting. **(B)** Posterior growth rate estimates against true 35-day average growth rates. Dashed line shows the 1:1 line. Points and ranges show posterior medians and 95% CIs respectively. **(C)** Model-predicted Ct distributions (line and blue shaded region) fitted to the observed Ct distribution (histogram) during epidemic growth, peak and decline phases. Black line and blue region show posterior median and 95% CIs. **(D)** Posterior estimates for the relative time since infection distribution during epidemic growth, peak and decline phases. Light shaded region shows 95% CIs, dark shaded region shows central 95% CIs, and solid line shows posterior mean. Note that the x-axis is reversed, such that the left hand side shows older infections and the right hand side shows more recent infections.

Discussion

The usefulness of Ct values for public health decision making is currently the subject of much discussion and debate. One unexplained observation which has been consistently observed in many locations is that the distribution of observed Ct values has decreased over the course of the current SARS-CoV-2 pandemic, which has led to questions over whether the virus has lost some amount of fitness (*15, 16, 18*). Our results demonstrate instead that this is expected as a purely epidemiologic phenomenon, without any change in individual-level viral dynamics or testing practices. We find that properties of the population-level Ct distribution strongly correlate with estimates for the effective reproductive number in Massachusetts in line with our theoretical predictions. The method described here uses this phenomenon to estimate a community's position in the epidemic curve, as defined by the growth rate, based on a single cross-sectional survey of virologic test data. We propose a simple method to monitor the skewness and median of the Ct value distribution to estimate changes in the epidemic trajectory. Despite the challenges of sampling variability, individual-level differences in viral kinetics, and limitations in comparing results from different laboratories or instruments, our results demonstrate that Ct values, with all of their quantitative variability for an individual, can be highly informative of population-level dynamics. This information is lost when measurements are reduced to binary classifications.

These results are sensitive to the true distribution of observed viral loads each day after infection. Different swab types, sample types, or instruments may alter the variability in the Ct distribution (*41,42*), leading to different relationships between the specific Ct distribution and the epidemic trajectory. Setting-specific calibrations, for example based on a reference range of Ct values, will be useful to ensure accuracy. Here, we generated a viral kinetics model based on observed properties of measured viral loads (proportion detectable over time following symptom onset, distribution of Ct values from true specimens), and used these results to inform priors on key parameters when estimating growth rates. The growth rate estimates can therefore be improved by choosing more precise, accurate priors relevant to the observations used during model fitting. Results could also be improved if the symptom status of the sampled individuals is known, allowing the inclusion of viral kinetics parameters specific to symptomatic individuals (*16, 43*). The same may be true if demographic features such as age are associated with viral load levels, although there is limited evidence to suggest this so far (*16, 18, 24, 44*). The distribution could also be adjusted to account for any individuals treated with antivirals, such as remdesivir, that may reduce viral loads (*45*). A similar approach may also be possible using serologic surveys, as

an extension of work that has related time since infection to antibody titers for other infectious diseases (*32, 38*).

Our results demonstrate that this method can be used to estimate epidemic growth rates based on data collected at a single time point, and independent of assumptions about the intensity of testing. Comparisons of simulated Ct values and observed Ct values with growth rates and *R(t)* estimates validate this general approach. Results should be interpreted with caution in cases where the observed Ct values are not from a population census or a largely random sample. When testing is based primarily on the presence of symptoms or follow-up of contacts of infected individuals, people may be more likely to be sampled at specific times since infection and thus the distribution of observed Cts would not be representative of the population as a whole. This method may be most useful in settings where representative surveillance samples can be obtained independent of COVID-19 symptoms—and importantly in cities or municipalities that wish to evaluate and monitor, in real-time, the role of various epidemic mitigation interventions, for example by conducting even a single random virologic testing effort as part of surveillance.

This method has a number of limitations. While the Bayesian framework incorporates the uncertainty in viral load distributions into inference on the growth rate, parametric assumptions and reasonably strong priors on these distributions aid in identifiability. If these parametric assumptions are violated, inference may not be reliable. This method may also overstate uncertainty in the viral load distributions if results from different machines or protocols are used to inform the prior. A more precise understanding of the viral load kinetics, and modeling those kinetics in a way that accounts for the epidemiologic and technical setting of the measurements, will help improve this approach and determine whether Ct distribution parameters from different settings are comparable. Because of this, quantitative measures from RT-qPCR should be reported regularly for SARS-CoV-2 cases and early assessment of pathogen load kinetics should be a priority for future emerging infections. The use of a control procedure in the measurements, like using the ratio of detected viral RNA to detected human RNA, could also improve the reliability and comparability of Ct measures.

Future research will evaluate how to incorporate these results into an overall inferential framework for real-time monitoring of epidemic trajectories. In particular, we are currently adapting the framework to combine multiple cross-sectional datasets, include the proportion of negative viral tests in the likelihood, and to use more flexible parametric assumptions for epidemic growth. By using Ct values to determine the growth rate of incident cases and using virologic positivity rates

to assess prevalence of infection, as well as potentially incorporating measured covariates and symptom status, a richer picture of the path and likely future of an outbreak can be realized from one or more virologic surveys.

The Ct value is a measurement with magnitude, which provides information on underlying viral dynamics. Although there are challenges to relying on single Ct values for individual-level decision making, the aggregation of many such measurements from a population contains substantial information. These results demonstrate how population-level distributions of Ct values can provide information on important epidemiologic questions of interest, even from a single cross-sectional survey. Better epidemic planning and more targeted epidemiological measures can then be implemented based on this survey, or use of Ct values can be combined with repeated sampling to maximize the use of available evidence.

Materials and Methods

Model for Population-Level Ct Distribution

We fit a model for the distribution of viral loads at a given day after infection, a. Denote this density by $p_a(x)$, where *x* is the viral load. We denote by ϕ_a the probability that an individual will have a detectable viral load *a* days after infection.

This model is fit based on several key features of the viral load kinetics that have been determined for SARS-CoV-2 infection. In symptomatic cases, detectable viral loads occur prior to the onset of symptoms and generally decline after symptom onset (*19, 46, 47*). Since infectiousness appears to peak before symptom onset, viral load likely does as well (*26*). Challenge studies in rhesus macaques indicate that viral load peaks around 2 days after infection (*22, 23*). Combined with evidence of an incubation period in humans of around 5 days (*5*), these observations suggest rapid exponential rise to a peak viral load during the incubation period and a high viral load upon symptom onset of approximately 6–9 $\log_{10}$ copies RNA per mL (*19, 25*). However, most existing viral load time series begin after symptom onset, and it is therefore difficult to corroborate assumptions for the pre-symptomatic period. Waning of viral loads occurs following symptom onsets, with a majority of patients undetectable around 24 days after onset (*19, 20, 24, 25*). At any point in the infection, there is a considerable amount of person-to-person variation in viral loads (*19, 25*), including a possible difference by symptom status (*20, 48*). This may also reflect protocols used to test the samples and the site and method of sample collection (*25*).

We develop a parametric model for the population-level distribution of Ct values. The measured Ct value *a* days after infection follows a Gumbel distribution with a mean that depends on *a* and a fixed standard deviation *SD*. The mean log-viral load at day *a* follows a two-hinge function that is at the true zero value of VL_{zero} for $a \leq t_{shift}$, rises linearly to a peak log-viral load of VL_{peak} at $a = t_{peak}$, wanes linearly to a log-viral load of VL_{switch} at $a = t_{switch}$, and then wanes at a slower linear rate until it reaches the limit of detection (LOD), a log-viral load of VL_{LOD} at $a = t_{LOD}$. That is, the mean log-viral load is given by:

$$VL_{mean}(a) = VL_{zero} + I_{a>t_{shift}} \frac{VL_{peak} - VL_{zero}}{t_{peak} - t_{shift}}(a - t_{shift}) + I_{a>t_{peak}} \frac{VL_{switch} - VL_{peak}}{t_{switch} - t_{peak}}(a - t_{peak})$$

$$+ I_{a>t_{switch}} \frac{VL_{LOD} - VL_{switch}}{t_{LOD} - t_{switch}}(a - t_{switch}).$$

The log-viral load is converted to a Ct value by: $Ct_{mean}(a) = Ct_{LOD} - log_2(10)\,(VL_{mean}(a) - VL_{LOD})$ where the limit of detection on the Ct scale is Ct_{LOD} = *40* and on the viral load scale is VL_{LOD} = *3*. The Ct value at a days after infection, $Ct(a)$, is then distributed according to $p_a(Ct)$, a Gumbel distribution with mean $Ct_{mean}(a)$ and variance SD^2. This model captures the shape of the observed mean viral load over time and the features described above. However, it overestimates the proportion of infections that will be detectable three weeks or more after infection. To account for this, each day after t_{switch}, we add an additional probability, p_{addl}, that an individual will clear the virus and become undetectable for that day forward, in addition to the probability that the Ct value rises above the limit of detection. So the probability of being detectable on day $a \leq t_{switch}$ is $\phi_a = P[Ct(a) \leq Ct_{LOD}]$ and the probability of being detectable on day $a > t_{switch}$ is $\phi_a = P[Ct(a) \leq Ct_{LOD}](1 - p_{addl})^{a - t_{switch}}$.

We used a least-squares optimization framework to obtain parameter point estimates that gave rise to viral kinetics with the following constraints: the proportion of individuals that are detectable on each day post symptom onset declines in line with existing data (*49*), and the lower 99th percentile of possible Ct values is in line with the lowest observed Ct value in our Brigham & Women's Hospital dataset. We used these point estimates to derive informative priors on key model parameters, as described in **Table S1.** The resulting distribution of Ct values and detectable proportions at each day a after infection are shown in **Fig. S1**. These parameters are used as fixed parameters for the schematic (**Fig. 1**).

Likelihood for Parametric Model of Outbreak Trajectory

Given a known population-level Ct distribution for each day after infection and a known detectable probability, $p_a(x)$and ϕ_a, respectively, we can determine the likelihood of the probability of infection a days prior to the test day, π_a, for all a with a positive probability of detectable viral load, given the observed detectable viral loads $x_1, \ldots, x_n$. Let $\{A_{min}, \ldots, A_{max}\}$ denote the days with a positive probability of detectable viral load, so that an individual with a detectable viral load must have been infected between A_{max} and A_{min} days prior to testing. We further assume that each individual's viral load is independent of the viral load of other individuals in the sample, conditional on the probability of infection for each day.

$$P(X_i = x_i | \pi_{A_{min}}, \ldots, \pi_{A_{max}}) = \sum_{a=A_{min}}^{A_{max}} \frac{p_a(x_i)\,\pi_a}{\sum_{c=A_{min}}^{A_{max}} \phi_c \pi_c}.$$

Thus, the likelihood and log-likelihood, respectively, are given by:

367 $L(\pi_{A_{min}},\ldots,\pi_{A_{max}}|X_1=x_1,\ldots,X_n=x_n) = \prod_{i=1}^{n} \sum_{a=A_{min}}^{A_{max}} \frac{p_a(x_i)\,\pi_a}{\sum_{c=A_{min}}^{A_{max}} \phi_c\pi_c}$, and

368 $l(\pi_{A_{min}},\ldots,\pi_{A_{max}}|X_1=x_1,\ldots,X_n=x_n) = \sum_{i=1}^{n} \left\{log\left[\sum_{a=A_{min}}^{A_{max}} p_a(x_i)\,\pi_a\right] -\right.$

369 $\left. log\left[\sum_{c=A_{min}}^{A_{max}} \phi_c\pi_c\right]\right\}$.

370

371 Assume that the outbreak, over the days prior to testing, is experiencing exponential growth or
372 decline of the daily probability of infection. Then for each day $a \in \{A_{min},\ldots,A_{max}\}$, $\pi_a =$
373 $\frac{e^{-\beta a}}{\sum_{d=A_{min}}^{A_{max}} e^{-\beta d}}$. Note that $\beta > 0$ implies a growing outbreak, $\beta < 0$ a waning outbreak, and $\beta = 0$
374 a flat outbreak with a constant probability of infection across the days prior to testing. Then the
375 likelihood and log-likelihood for β given the observed viral loads are:

376 $$P(X_1=x_1,\ldots,X_n=x_n|\beta) = \prod_{i=1}^{n} \sum_{a=A_{min}}^{A_{max}} \frac{p_a(x_i)\,e^{-\beta a}}{\sum_{c=A_{min}}^{A_{max}} \phi_c e^{-\beta c}}$$

377 $$l(\beta|X_1=x_1,\ldots,X_n=x_n) = \sum_{i=1}^{n} \left\{log\left[\sum_{a=A_{min}}^{A_{max}} p_a(x_i)\,e^{-\beta a}\right] - log\left[\sum_{c=A_{min}}^{A_{max}} \phi_c e^{-\beta c}\right]\right\}$$

378 More flexible parametric models can be used, including assuming piecewise exponential growth
379 or that the probability of infection on each day follows some other function. In these cases, the
380 parametric model for $\pi_a$is simply used in place of $e^{-\beta a}$ in this likelihood and log-likelihood.

381 Additionally, nonparametric estimation can be performed by maximizing the likelihood in terms
382 of π_a directly for $a \in \{A_{min},\ldots,A_{max}\}$. This approach, however, will likely give very high
383 variance unless there is a large sample size and the viral distribution has a small variance
384 compared to the trend in the mean over time.

385 *Bayesian Framework for Estimation and Inference*

386 To incorporate uncertainty in the distribution of viral loads on each day after infection, we
387 construct a Bayesian framework for estimation and inference. A prior distribution is specified for
388 β(or the set of parameters used in the likelihood) as well as for the viral load distributions, $p_a(x)$.
389 The prior for the viral load distributions can be specified in many ways, but we here

parameterize these distributions by a mean and standard deviation for each day a and assume a normal distribution with those parameters, so the prior is specified on the mean and standard deviation. This prior can be based on existing studies of viral load kinetics after infection and can also take into account specific properties of the sample collection procedure, PCR protocols, and machine used to analyze the samples, if known. Estimation and inference can then proceed on the posterior distribution.

The prior distributions used here are listed in **Table S1**. The prior distribution for β is taken to be normal with mean 0 and standard deviation 0.1. Note that $\beta = 0.4$ corresponds to a doubling time of 1.73 days, so value of β above 0.4 or below -0.4 would be highly unusual for SARS-CoV-2.

To assess the outbreak trajectory, interest lies in the marginal posterior distribution of β. A larger positive value of βindicates a faster-growing outbreak and a more negative value of βindicates a faster-waning outbreak. The posterior distribution of the $p_a(x)$ parameters are nuisance parameters for this goal, but may provide useful information for the prior distributions of $p_a(x)$ in future studies in similar settings.

Simulated Data

For **Fig. 1**, outbreaks are simulated according to a deterministic SEIR model with a population of 1,000,000 people, R_0=*2.5*, average latent period of 6.41 days, and average infectiousness period of 8.79 days. The latent and infectiousness periods are based on the best fit SEIR models for the observed prevalences in Massachusetts nursing home data (not shown) and align fairly well with the viral load kinetics implied by the Ct distribution model and with parameters reported elsewhere (*5, 50*).

For **Figs. 2** and **4**, outbreaks are simulated according to a stochastic SEIR model with a population of 6,900,000 people, an initial seed of 100 infected individuals, R_0=*2*, average latent period of 6.41 days, and average infectiousness period of 8.79 days as above. We used the *odin* R package for simulation (https://cran.r-project.org/web/packages/odin/index.html). To simulate cross-sectional sampling, we assumed that 20% of the population was randomly sampled once during the outbreak, which led to approximately 4,000 individuals being sampled and tested per day. Each infected individual is then assigned an observed Ct value on the day of sampling using the viral kinetics model described in *Model for Population-Level Ct Distribution*, with added observation error drawn from a Gumbel distribution with a standard deviation of 6. **Fig. S3** is

generated in the same way, but with additional samples taken from symptomatic individuals. We assumed that 35% of infections became symptomatic, and that symptomatic individuals had a 10% probability of being detected. Detected, symptomatic individuals were observed with some delay post symptom onset drawn from a discretized gamma distribution with shape and rate parameters of 5 and 2 respectively (mean and standard deviation of 2.5 and 1 days respectively).

Brigham & Women's Hospital data, Boston, Massachusetts

Data comes from nasopharyngeal specimens processed on a Hologic Panther Fusion SARS-CoV-2 assay for patients at the Brigham & Women's Hospital in Boston, MA. Testing during the first two weeks in April 2020 were restricted to patients with symptoms consistent with COVID-19 and who needed hospital admission. Following April 15, testing criteria were expanded to include all hospital admissions regardless of symptoms and asymptomatic ER patients who were not admitted. The results were from individuals entering the hospital for non-COVID procedures and thus represent less biased surveillance specimens. Daily data is aggregated by week. Daily confirmed case counts for Massachusetts were obtained from the NYT github page (https://github.com/nytimes/covid-19-data) (*51*).

Estimating the effective reproductive number, R(t)

We used the *EpiNow2* R package for all *R(t)* estimates (https://github.com/epiforecasts/EpiNow2). The package estimates the time-varying reproduction number using the MCMC package *Stan* (https://cran.r-project.org/web/packages/rstan/index.html), incorporating best practices in *R(t)* estimation as described by Gostic et al. (*52–54*). The package takes a time series of confirmed case counts as input, and returns posterior distribution estimates on the daily effective reproductive number. *EpiNow2* simultaneously infers *R(t)* and infection incidence using a method based on the Cori method. We assumed that the confirmation delay distribution was log normally distributed with a mean of 3 (on the linear scale) and standard deviation of 0.5 (on the log scale) days. We assumed that the generation interval and incubation period were uncertain, using the default values provided by the *EpiNow2* package. The mean and standard deviation of the generation interval were assumed to be 3.64 and 3.07 days respectively, with standard deviations for Bayesian priors on these parameters of 0.711 and 0.770 respectively. The incubation period was assumed to be log-normally distributed, with mean and standard deviation (both on the log scale) of 1.621 and 0.418 respectively, with standard deviations for Bayesian priors on these parameters of 0.064

452 and 0.0691 respectively. We used a smoothing window of 7 days and ran 4 chains each for 4000
453 iterations with 1000 iterations of warm up.

References

1. H. V. Fineberg, M. E. Wilson. Epidemic science in real time. *Science* **324**, 987 (2009).

2. Y. Hsieh, D. N. Fisman, J. Wu. On epidemic modeling in real time: An application to the 2009 Novel A (H1N1) influenza outbreak in Canada. *BMC Res. Notes* **3,** 283 (2010).

3. J. Lourenço, R. Paton, M. Ghafari, M. Kraemer, C. Thompson, P. Simmonds, P. Klenerman, S. Gupta. Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic. https://www.medrxiv.org/content/10.1101/2020.03.24.20042291v1 (2020).

4. T. K. Tsang, P. Wu, Y. Lin, E. H. Y. Lau, G. M. Leung, B. J. Cowling. Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study. *Lancet Public Health* **5**, e289–296 (2020).

5. S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, J. Lessler. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann. Int. Med.* **172**, 577–582 (2020).

6. T. Jombart, K. van Zandvoort, T. W. Russell, C. I. Jarvis, A. Gimma, S. Abbott, S. Clifford, S. Funk, H. Gibbs, Y. Liu, C. A. B. Pearson, N. I. Bosse, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, R. M. Eggo, A. J. Kucharski, W. J. Edmunds. Inferring the number of COVID-19 cases from recently reported deaths. *Wellcome Open Res.* **5** (2020).

7. M. Lipsitch, D. L. Swerdlow, L. Finelli. Defining the epidemiology of COVID-19—studies needed. *NEJM* **382**, 1194–1196 (2020).

8. A. Tahamtan, A. Ardebili. Real-time RT-PCR in COVID-19 detection: issues affecting the results. *Expert Rev. Mol. Diagn.* **20**, 453–454 (2020).

9. C. P. West, V. M. Montori, P. Sampathkumar. COVID-19 testing: the threat of false-negative results. *Mayo Clin. Proc.* **95**, 1127–1129 (2020).

10. J. L. Vaerman, P. Saussoy, I. Ingargiola. Evaluation of real-time PCR data. *J. Biol. Regul. Homeost. Agents* **18**, 212–214 (2004).

11. M. R. Tom, M. J. Mina. To interpret the SARS-CoV-2 test, consider the cycle threshold value. *Clin. Inf. Dis.* 10.1093/cid/ciaa619 (2020).

12. J. A. Fuller, M. K. Njenga, G. Bigogo, B. Aura, M. O. Ope, L. Nderitu, L. Wakhule, D. D. Erdman, R. F. Breiman, D. R. Feikin. Association of the C_T values of real-time PCR of viral upper respiratory tract infection with clinical severity, Kenya. *J. Med. Virol.* **85**, 924–932 (2013).

13. S. Bolotin, S. L. Deeks, A. Marchand-Austin, H. Rilkoff, V. Dang, R. Walton, A. Hashim, D. Farrell, N. S. Crowcroft. Correlation of Real Time PCR Cycle Threshold Cut-Off with *Bordetella pertussis* Clinical Severity. *PLoS One* **10**, e0133209 (2015).

14. T. K. Tsang, B. J. Cowling, V. J. Fang, K. H. Chan, D. K. M. Ip, G. M. Leung, J. S. M. Peiris, S. Cauchemez. Influenza A virus shedding and infectivity in households. *J. Inf. Dis.* **212**, 1420–1428 (2015).

15. M. Moraz, D. Jacot, M. Papadimitriou-Olivgeris, L. Senn, G. Greub, K. Jaton, O. Opota. Clinical importance of reporting SARS-CoV-2 viral loads across the different stages of the COVID-19 pandemic. https://www.medrxiv.org/content/10.1101/2020.07.10.20149773v1 (2020).

16. D. Jacot, G. Greub, K. Jaton, O. Opota. Viral load of SARS-CoV-2 across patients and compared to other respiratory viruses. https://www.medrxiv.org/content/10.1101/2020.07.15.20154518v1 (2020).

17. Y. Chen, L. Li. SARS-CoV-2: virus dynamics and host response. *Lancet Inf. Dis.* **20**, 515–516 (2020).

18. T. C. Jones, B. Mühlemann, T. Veith, G. Biele, M. Zuchowski, J. Hoffmann, A. Stein, A. Edelmann, V. M. Corman, C. Drosten. An analysis of SARS-CoV-2 viral load by patient age. https://www.medrxiv.org/content/10.1101/2020.06.08.20125484v1 (2020).

19. K. K. W. To, O. T. Y. Tsang, W. S. Leung, A. R. Tam, T. C. Wu, D. C. Lung, C. C. Y. Yip, J. P. Cai, J. M. C. Chan, T. S. H. Chik, D. P. L. Lau, C. Y. C. Choi, L. L. Chen, W. M. Chan, K. H. Chan, J. D. Ip, A. C. K. Ng, R. W. S. Poon, C. T. Luo, Vi. C. C. Cheng, J. F. W. Chan, I. F. N. Hung, Z. Chen, H. Chen, K. Y. Yuen. Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study. *Lancet Inf. Dis.* **20**, 565–574 (2020).

20. Q. Z. Long, X. J. Tang, Q. L. Shi, Q. Li, H. J. Deng, J. Yuan, J. L. Hu, W. Xu, Y. Zhang, F. J. Lv, K. Su, F. Zhang, J. Gong, B. Wu, X. M. Liu, J. J. Li, J. F. Qiu, J. Chen, A. L. Huang. Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nature Med.* 10.1038/s41591-020-0965-6 (2020).

21. M. Cevik, M. Tate, O. Lloyd, A. E. Maraolo, J Schafers , A. Ho. SARS-CoV-2 viral load dynamics, duration of viral shedding and infectiousness – a living systematic review and meta-analysis. https://www.medrxiv.org/content/10.1101/2020.07.25.20162107v1 (2020)

22. A. Chandrashekar, J. Liu, A. J. Martinot, K. McMahan, N. B. Mercado, L. Peter, L. H. Tostanoski, J. Yu, Z. Maliga, M. Nekorchuk, K. Busman-Sahay, M. Terry, L. M. Wrijil, S. Ducat, D. R. Martinez, C. Atyeo, S. Fischinger, J. S. Burke, M. D. Slein, L. Pessaint, A. Van Ry, J. Greenhouse, T. Taylor, K. Blade, A. Cook, B. Finneyfrock, R. Brown, E. Teow, J. Velasco, R. Zahn, F. Wegmann, P. Abbink, E. A. Bondzie, G. Dagotto, M. S. Gebre, X. He, C. Jacob-Dolan, N. Kordana, Z. Li, M. A. Lifton, S. H. Mahrokhian, L. F. Maxfield, R. Nityanandam, J. P. Nkolola, A. G. Schmidt, A. D. Miller, R. S. Baric, G. Alter, P. K. Sorger, J. D. Estes, H. Andersen, M. G. Lewis, D. H. Barouch. SARS-CoV-2 infection protects against rechallenge in rhesus macaques. *Science* 10.1126/science.abc4776 (2020).

23. J. Yu, L. H. Tostanoski, L. Peter, N. B. Mercado, K. McMahan, S. H. Mahrokhian, J. P. Nkolola, J. Liu, Z. Li, A. Chandrashekar, D. R. Martinez, C. Loos, C. Atyeo, S. Fischinger, J. S. Burke, M. D. Slein, Y. Chen, A. Zuiani, F. J. N. Lelis, M. Travers, S. Habibi, L. Pessaint, A. Van Ry, K. Blade, R. Brown, A. Cook, B. Finneyfrock, A. Dodson, E. Teow, J. Velasco, R. Zahn, F. Wegmann, E. A. Bondzie, G. Dagotto, M. S. Gebre, X. He, C. Jacob-Dolan, M. Kirilova, N. Kordana, Z. Lin, L. F. Maxfield, F. Nampanya, R. Nityanandam, J. D. Ventura, H. Wan, Y. Cai, B. Chen, A. G. Schmidt, D. R. Wesemann,

R. S. Baric, G. Alter, H. Andersen, M. G. Lewis, D. H. Barouch. DNA vaccine protection against SARS-CoV-2 in rhesus macaques. *Science* 10.1126/science.abc6284 (2020).

24. A. T. Xiao, Y. X. Tong, S. Zhang. Profile of RT-PCR for SARS-CoV-2: a preliminary study from 56 COVID-19 patients. *Clin. Inf. Dis.* 10.1093/cid/ciaa460 (2020).

25. R. Wölfel, V. M. Corman, W. Guggemos, M. Seilmaier, S. Zange, M. A. Müller, D. Niemeyer, T. C. Jones, P. Vollmar, C. Rothe, M. Hoelscher, T. Bleicker, S. Brünink, J. Schneider, R. Ehmann, K. Zwirglmaier, C. Drosten, C. Wendtner. Virological assessment of hospitalized patients with COVID-19. *Nature* **581**, 465–469 (2020).

26. X. He, E. H. Y. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. C. Lau, J. Y. Wong, Y. Guan, X. Tan, X. Mo, Y. Chen, B. Liao, W. Chen, F. Hu, Q. Zhang, M. Zhong, Y. Wu, L. Zhao, F. Zhang, B. J. Cowling, F. Li, G. M. Leung. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Med.* **26**, 672–675 (2020).

27. D. B. Larremore, B. Wilder, E. Lester, S. Shehata, J. M. Burke, J. A. Hay, M. Tambe, M. J. Mina, R. Parker. Test sensitivity is secondary to frequency and turnaround time for COVID-19 surveillance. https://www.medrxiv.org/content/10.1101/2020.06.22.20136309v2 (2020).

28. S. Riley, K. E. C. Ainslie, O. Eales, B. Jeffrey, C. E. Walters, C. J. Atchison, P. J. Diggle, D. Ashby, C. A. Donnelly, G. Cooke, W. Barclay, H. Ward, G. Taylor, A. Darzi, P. Elliott. Community prevalence of SARS-CoV-2 virus in England during May 2020: REACT study. https://www.medrxiv.org/content/10.1101/2020.07.10.20150524v1 (2020).

29. H. Salje, D. A. T. Cummings, I. Rodriguez-Barraquer, L. C. Katzelnick, J. Lessler, C. Klungthong, B. Thaisomboonsuk, A. Nisalak, A. Weg, D. Ellison, L. Macareo, I. K. Yoon, R. Jarman, S. Thomas, A. L. Rothman, T. Endy, S. Cauchemez. Reconstruction of antibody dynamics and infection histories to evaluate dengue risk. *Nature* **557**, 719–723 (2018).

30. J. A. Hay, A. Minter, K. E. C. Ainslie, J. Lessler, B. Yang, D. A. T. Cummings, A. J. Kucharski, S. Riley. An open source tool to infer epidemiological and immunological dynamics from serological data: serosolver. *PLoS Comput. Biol.* **16**, e1007840 (2020).

31. B. Borremans, N. Hens, P. Beutels, H. Leirs, J. Reijniers. Estimating time of infection using prior serological and individual information can greatly improve incidence estimation of human and wildlife infections. *PLoS Comput. Biol.* **12**, e1004882 (2016).

32. K. M. Pepin, S. L. Kay, B. D. Golas, S. S. Shriner, A. T. Gilbert, R. S. Miller, A. L. Graham, S. Riley, P. C. Cross, M. D. Samuel, M. B. Hooten, J. A. Hoeting, J. O. Lloyd-Smith, C. T. Webb, M. G. Buhnerkempe. Inferring infection hazard in wildlife populations by linking data across individual and population scales. *Ecol. Lett.* **20**, 275–292 (2017).

33. X. Zhao, Y. Ning, M. I. C. Chen, A. R. Cook. Individual and population trajectories of influenza antibody titers over multiple seasons in a tropical country. *Am. J. Epidemiol.* **187**, 135–143 (2018).

34. D. De Angelis, W. R. Gilks, N. E. Day. Bayesian projection of the acquired immune deficiency syndrome epidemic. *J. R. Stat. Soc. Ser. C Appl. Stat.* **47**, 449–498 (1998).

35. H. E. de Melker, F. G. A. Versteegh, J. F. P. Schellekens, P. F. M. Teunis, M. Kretzschmar. The incidence of Bordetella pertussis infections estimated in the population from a combination of serological surveys. *J. Infect.* **53**, 106–113 (2006).

36. J. Simonsen, K. Mølbak, G. Falkenhorst, K. A. Krogfelt, A. Linneberg, P. F. M. Teunis. Estimation of incidences of infectious diseases based on antibody measurements. *Stat. Med.* **28**, 1882–1895 (2009).

37. P. F. M. Teunis, J. C. H. van Eijkeren, C. W. Ang, Y. T. H. P. van Duynhoven, J. B. Simonsen, M. A. Strid, W. van Pelt. Biomarker dynamics: estimating infection rates from serological data. *Stat. Med.* **31**, 2240–2248 (2012).

38. D. A. Helb, K. K. A. Tetteh, P. L. Felgner, J. Skinner, A. Hubbard, E. Arinaitwe, H. Mayanja-Kizza, I. Ssewanyana, M. R. Kamya, J. G. Beeson, J. Tappero, D. L. Smith, P. D. Crompton, P. J. Rosenthal, G. Dorsey, C. J. Drakeley, B. Greenhouse. Novel serologic biomarkers provide accurate estimates of recent Plasmodium falciparum exposure for individuals and communities. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E4438–E4447 (2015).

39. J. Wallinga, M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B* **274**, 599–604 (2007).

40. M. Gousseff, P. Penot, L. Gallay, D. Batisse, N. Benech, K. Bouiller, R. Collarino, A. Conrad, D. Slama, C. Joseph, A. Lemaignen, F.-X. Lescure, B. Levy, M. Mahevas, B. Pozzetto, N. Vignier, B. Wyplosz, D. Salmon, F. Goehringer, E. Botelho-Nevers. Clinical recurrences of COVID-19 symptoms after recovery: viral lapse, reinfection or inflammatory rebound? *J. Infect.* 10.1016/j.jinf.2020.06.073 (2020).

41. R. Niehus, E. van Kleef, A. Turlej-Rogacka, C. Lammens, Y. Carmeli, H. Goossens, E. Tacconelli, B. Carevic, L. Preotescu, S. Malhotra-Kumar, B. S. Cooper. Quantifying antibiotic impact on within-patient dynamics of extended-spectrum beta-lactamase resistance. *eLife* **9**, e49206 (2020).

42. D. Rhoads, D. R. Peaper, R. C. She, F. S. Nolte, C. M. Wojewoda, N. W. Anderson, B. S. Pritt. College of American Pathologists (CAP) Microbiology Committee perspective: caution must be used in interpreting the cycle threshold (Ct) value. *Clin. Infect. Dis.* 10.1093/cid/ciaa1199 (2020).

43. Z. Zhang, T. Xiao, Y. Wang, J. Yuan, H. Ye, L. Wei, H. Wang, X. Liao, S. Qian, Z. Wang, L. Liu. Early viral clearance and antibody kinetics of COVID-19 among asymptomatic carriers. https://www.medrxiv.org/content/10.1101/2020.04.28.20083139v1 (2020).

44. L. Held. A discussion and reanalysis of the results reported in Jones et al. (2020). https://osf.io/bkuar/ (2020).

45. W.-C. Ko, J.-M. Rolain, N.-Y. Lee, P.-L. Chen, C.-T. Huang, P.-I. Lee, P.-R. Hsueh. Arguments in favour of remdesivir for treating SARS-CoV-2 infections. *Int. J. Antimicrob. Agents* **55**, 105933 (2020).

46. M. M. Arons, K. M. Hatfield, S. C. Reddy, A. Kimball, A. James, J. R. Jacobs, J. Taylor, K. Spicer, A. C. Bardossy, L. P. Oakley, S. Tanwar, J. W. Dyal, J. Harney, Z. Chisty, J. M. Bell, M. Methner, P. Paul, C. M. Carlson, H. P. McLaughlin, N. Thornburg, S. Tong, A. Tamin, Y. Tao, A. Uehara, J. Harcourt, S. Clark, C. Brostrom-Smith, L. C. Page, M. Kay, J. Lewis, P. Montgomery, N. D. Stone, T. A. Clark, M. A. Honein, J. S. Duchin, J. A.

Jernigan. Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. *NEJM* **382**, 2081–2090 (2020).

47. J. Y. Kim, J. H. Ko, Y. Kim, Y. J. Kim, J. M. Kim, Y. S. Chung, H. M. Kim, M. G. Han, S. Y. Kim, B. S. Chin. Viral load kinetics of SARS-CoV-2 infection in first two patients in Korea. *J. Korean Med. Sci.* **35**, e86 (2020).

48. Y. Liu, L. M. Yan, L. Wan, T. X. Xiang, A. Le, J. M. Liu, M. Peiris, L. L. M. Poon, W. Zhang. Viral dynamics in mild and severe cases of COVID-19. *Lancet Inf. Dis.* **20**, 656–657 (2020).

49. B. Borremans, A. Gamble, K. C. Prager, S. K. Helman, A. M. McClain, C. Cox, V. Savage, J. O. Lloyd-Smith. Quantifying antibody kinetics and RNA shedding during early-phase SARS-CoV-2 infection. https://www.medrxiv.org/content/10.1101/2020.05.15.20103275v1 (2020).

50. S. T. Ali, L. Wang, E. H. Y. Lau, X.-K. Xu, Z. Du, Y. Wu, G. M. Leung, B. J. Cowling. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. *Science* 10.1126/science.abc9004 (2020).

51. The New York Times. Covid in the U.S.: Latest Map and Case Count. *New York Times* 2020. Accessed at https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html on September 8, 2020.

52. S. Abbott, J. Hellewell, R. N. Thompson, K. Sherratt, H. P. Gibbs, N. I. Bosse, J. D. Munday, S. Meakin, E. L. Doughty, J. Y. Chun, Y. W. D. Chan, F. Finger, P. Campbell, A. Endo, C. A. B. Pearson, A. Gimma, T. Russell, CMMID COVID modelling group, S. Flasche, A. J. Kucharski, R. M. Eggo, S. Funk. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Res.* [awaiting peer review] **5**, 112 (2020).

53. B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, A. Riddell. Stan: a probabilistic programming language. *J. Stat. Softw.* **76**, v076i01 (2017).

54. K. M. Gostic, L. McGough, E. B. Baskerville, S. Abbott, K. Joshi, C. Tedijanto, R. Kahn, R. Niehaus, J. Hay, P. M. De Salazar, J. Hellewell, S. Meakin, J. Munday, N. I. Bosse, K. Sherrat, R. N. Thompson, L. F. White, J. S. Huisman, J. Scire, S. Bonhoeffer, T. Stadler, J. Wallinga, S. Funk, M. Lipsitch, S. Cobey. Practical considerations for measuring the effective reproductive number, Rt. https://www.medrxiv.org/content/10.1101/2020.06.18.20134858v3 (2020).

Supplementary Figures and Tables

Table S1. Parameters in the viral load and cycle threshold (Ct) value distribution. Point estimates shown are used in the simulations for Figs. 1, 4, S1, and S2. Bayesian priors are used in Fig. 4.

Parameter	Description	Prior distribution	Point estimate (mean)	Standard deviation
β	Exponential growth rate	Normal	0	0.10
t_{shift}	Time (in days) from infection to start of growth of viral load	Fixed	2.00	NA
t_{peak}	Time (in days) from infection to peak viral load	Fixed	4.00	NA
t_{switch}	Time (in days) from infection to start of reduced waning viral load rate and additional probability of becoming undetectable	Normal	16.0	1.00
t_{LOD}	Time (in days) from t_{switch} until the mean viral load is equal to the limit of detection (VL_{LOD})	Normal	30.0	1.50
VL_{zero}	Mean viral load (in $\log_{10}$ copies RNA per mL) for $a \leq t_{shift}$	Fixed	−3.00	NA
VL_{peak}	Mean peak viral load (in $\log_{10}$ copies RNA per mL) at $a = t_{peak}$	Normal	9.00	1.00
VL_{switch}	Mean viral load (in $\log_{10}$ copies RNA per mL) at $a = t_{switch}$	Normal	3.70	0.125
SD	Standard deviation (for Gumbel	Normal	6.00	0.50

	distribution) of cycle threshold (Ct) values around mean at any day *a*			
p_{addl}	Probability of becoming permanently undetectable each day for $a > t_{switch}$	Beta	0.12	0.05
Ct_{LOD}	Limit of detection for Ct values	Fixed	40.0	NA
VL_{LOD}	Limit of detection for viral loads (in $\log_{10}$ copies RNA per mL)	Fixed	3.00	NA

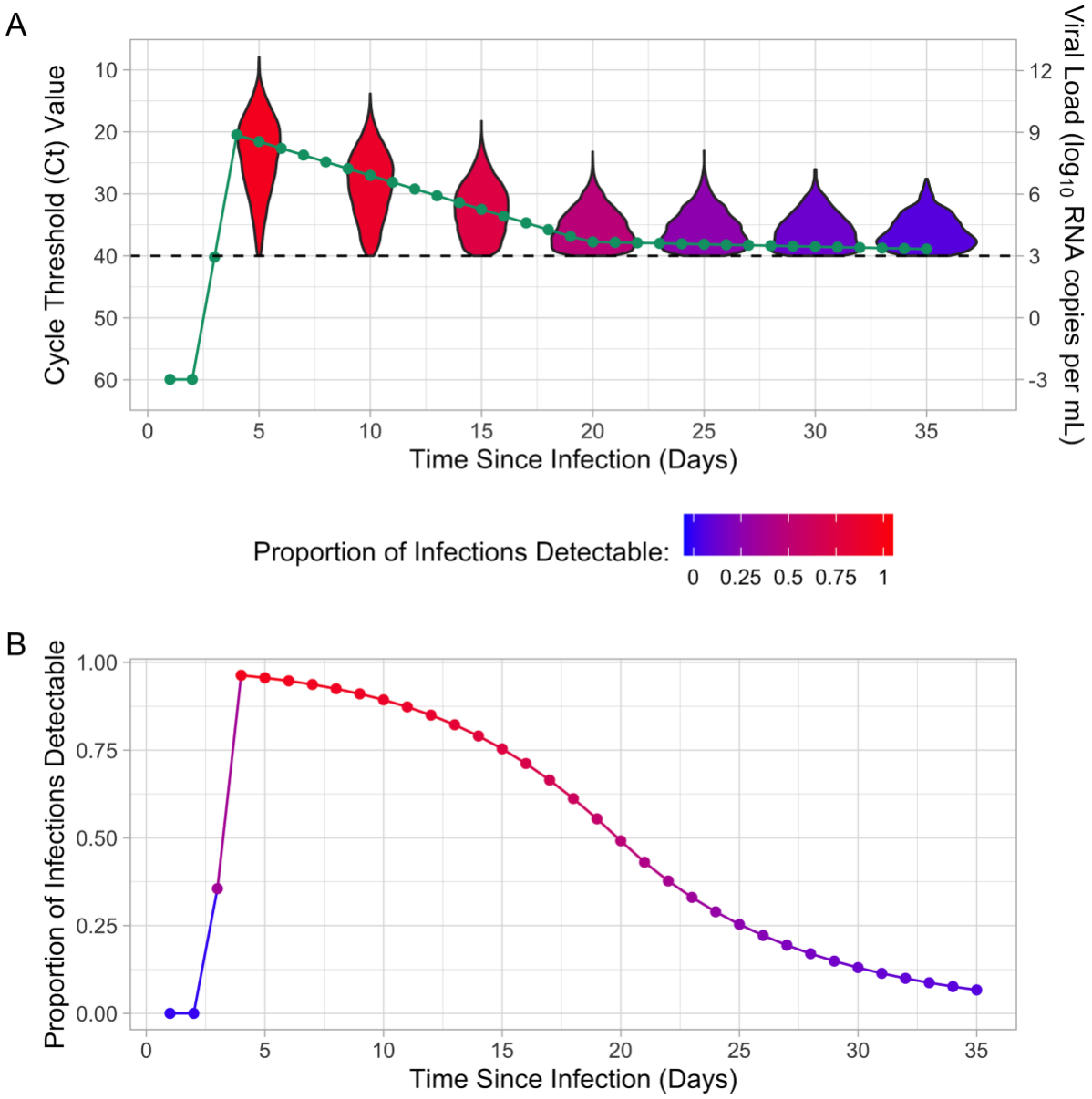

Figure S1. **(A)** Mean cycle threshold (Ct) value, mean viral load, and distribution of Ct values for detectable infections by time since infection. **(B)** Proportion of infections that are detectable by time since infection for the population-level Ct distribution. The proportion of infections that are detectable is indicated by the color of the violin plot and the proportion detectable line. The dashed line indicates the limit of detection (Ct value of 40 or viral load of 3).

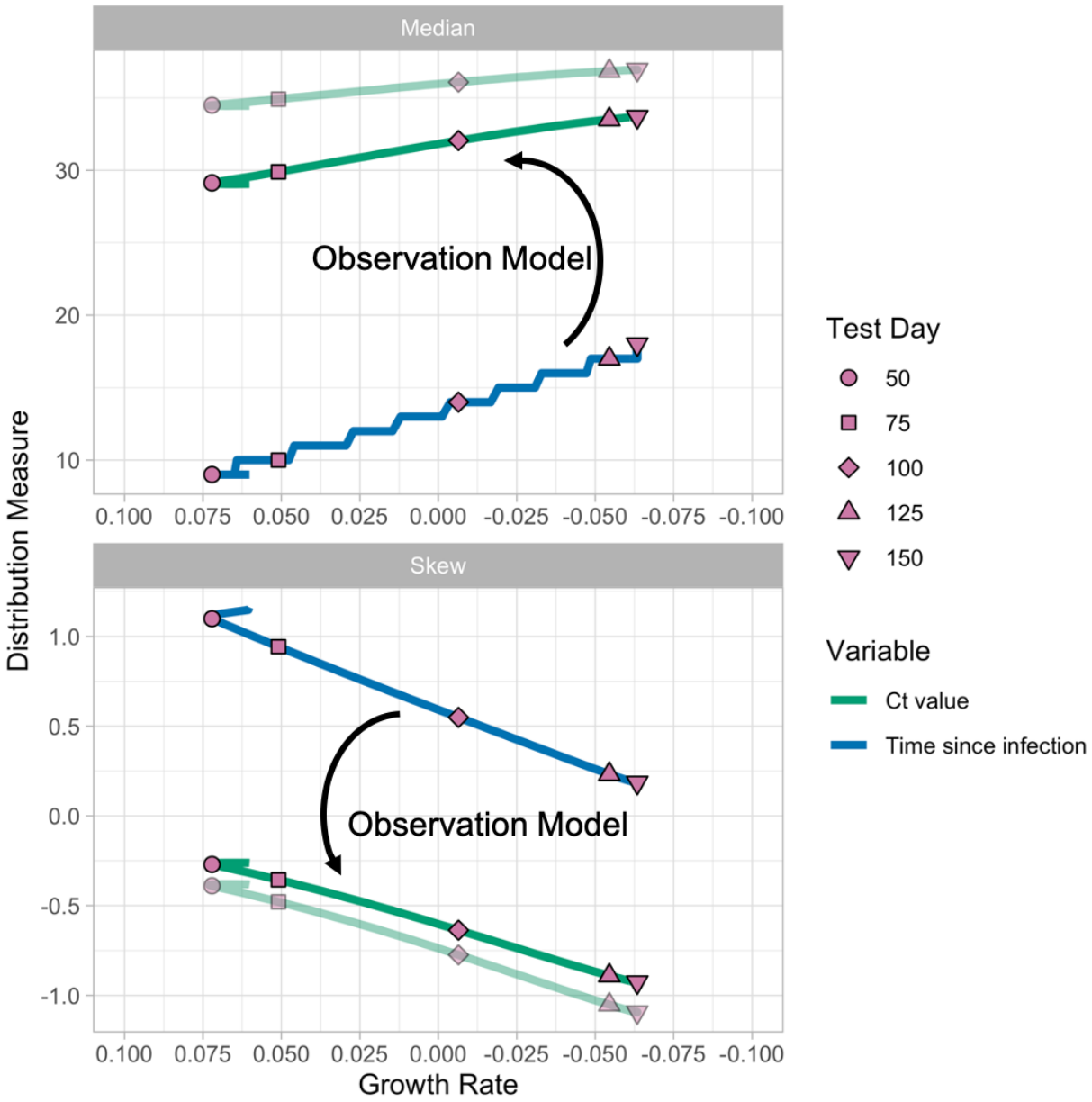

Figure S2. Relationship between time since infection and detectable cycle threshold (Ct) distribution and growth rate. The median and skew of the time since infection distribution are directly related to growth rate. Although the time since infection distribution (blue lines indicate median, top, and skew, bottom, of time-since-infection) is not observed directly, Ct values can be used as a proxy observation model. The solid green line shows the observation model assumed for Fig. 1, whereas the faint green line shows how a relationship would still exist for an alternative observation model, the two differing, for example, by different tests in different labs. The test days shown in Fig. 1 are indicated by points. All results are based on an SEIR model in a population of 1,000,000 with R_0 = 2.5, average latent period of 6.41 days, and average infectious period of 8.79 days, and Ct values are distributed according to the population-level Ct value distribution.

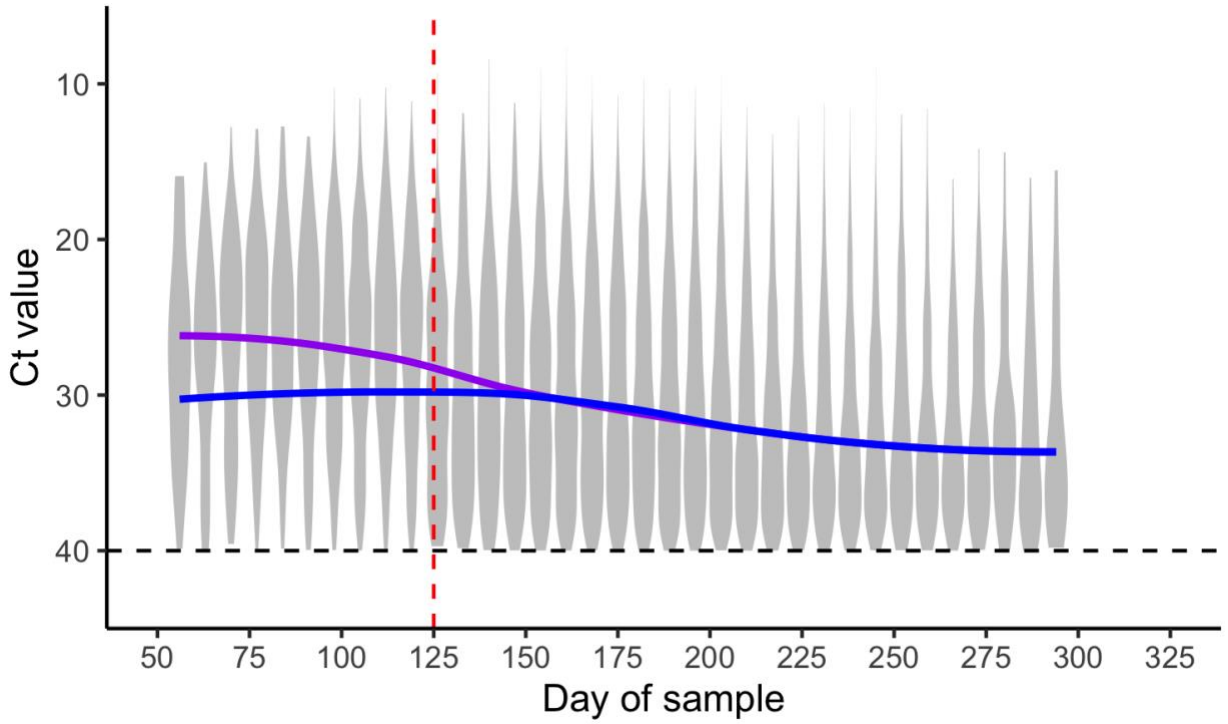

Figure S3. Weekly distribution of positive (detectable) Ct values as in **Fig. 2**, but up to day 125 sampled individuals are a combination of randomly sampled cross-sections and individuals sampled in the days immediately after symptom onset, assuming that 35% of infections become symptomatic. The blue line shows the smoothed median from only cross-sectionally sampled individuals, whereas the purple line shows the smoothed median from combined cross-sectional and symptomatic sampled individuals. The violin plot shows densities of all sampled individuals. For these simulations, all symptomatic individuals had a 10% probability of being detected up to day 125. For detected individuals, we added a delay between symptom onset and sampling time drawn from a discretized gamma distribution with a mean of 2.5 days and standard deviation of 1 day.