# Nowcasting epidemic trends using hospital- and community-based virologic test data

Tse Yang Lim[1]*, Sanjat Kanjilal[2,3], Shira Doron[4], Jessica Penney[4], Meredith Haddix[5], Tae Hee Koo[5], Phoebe Danza[5], Rebecca Fisher[5], Yonatan H. Grad[1,6†], James A. Hay[1,7*†]

*Correspondence to: tseyanglim@hsph.harvard.edu, james.hay@ndm.ox.ac.uk

†These authors jointly supervised the work.

[1] Center for Communicable Disease Dynamics, Harvard T.H. Chan School of Public Health, Boston, MA
[2] Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, MA
[3] Department of Infectious Diseases, Brigham and Women's Hospital, Boston, MA
[4] Division of Geographic Medicine and Infectious Diseases, Tufts Medical Center, Boston, MA
[5] Disease Control Bureau, Los Angeles County Department of Public Health, Los Angeles, CA
[6] Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA
[7] Pandemic Sciences Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK

# Abstract

Epidemiological surveillance typically relies on reported incidence of cases or hospitalizations, which can suffer significant reporting lags, biases and under-ascertainment. Here, we evaluated the potential of viral loads measured by RT-qPCR cycle threshold (Ct) values to track epidemic trends. We used SARS-CoV-2 RT-qPCR results from hospital testing in Massachusetts, USA, municipal testing in California, USA, and simulations to identify predictive models and covariates that maximize short-term epidemic trend prediction accuracy. We found SARS-CoV-2 Ct value distributions correlated with epidemic growth rates under real-world conditions. We fitted generalized additive models to predict log growth rate or direction of reported SARS-CoV-2 case incidence using features of the time-varying population Ct distribution and assessed the models' ability to track epidemic dynamics in rolling two-week windows. Observed Ct value distributions accurately predicted epidemic growth rates (growth rate RMSE ~ 0.039-0.052) and direction (AUC ~ 0.72-0.78). Performance degraded during periods of rapidly changing growth rate. Predictive models were robust to testing regimes and sample sizes; accounting for population immunity or symptom status yielded no substantial improvement. Trimming Ct value outliers improved performance. These results indicate that analysis of Ct values from routine PCR tests can help monitor epidemic trends, complementing traditional incidence metrics.

# Introduction

Epidemic monitoring and outbreak surveillance are vital public health functions, providing early warning of emerging threats, informing healthcare capacity planning and transmission control policies, and helping to evaluate the effectiveness of interventions[1–4]. A common approach to epidemic monitoring, exemplified during the COVID-19 pandemic, is to track the incidence of reported positive diagnostic tests, clinical cases[5,6], or deaths[7]. These data can inform key statistics such as the epidemic growth rate or effective reproductive number[8–11] and are fundamental to nowcasting and forecasting an epidemic's trajectory[12–14]. However, these data streams can be substantially lagged, biased, and incomplete due to testing delays, capacity limitations, cost, and changing test-seeking behavior[15,16]. Thus, there has been growing interest in alternative data sources, such as wastewater surveillance[17,18], internet search trends[19], and digital contact tracing[20], that do not depend on large-scale testing of individuals,.

A novel data source for epidemic monitoring described during the COVID-19 pandemic is the population-level distribution of viral loads among infected individuals, approximated using cycle threshold (Ct) values from reverse-transcription quantitative polymerase chain reaction (RT-qPCR) testing[21–24]. For certain acute respiratory viruses such as SARS-CoV-2, a low Ct value (high viral load) typically suggests that an individual was sampled early in their infection, whereas a high Ct value (low viral load) measurement suggests sampling later in infection[25–27]. Thus, a population-level sample of predominantly low Ct values (high viral loads) indicates that most sampled infections are of recent onset, corresponding to a growing epidemic, whereas a sample of predominantly high Ct values (low viral loads) corresponds to a declining epidemic consisting of mostly late infections and post-infectious viral persistence[21]. Unlike count-based surveillance methods, estimating epidemic growth rate based on the distribution of measured viral loads does not depend on the number of positive tests.

Multiple studies have reported on the feasibility of using population-level Ct values to track SARS-CoV-2 epidemic trends,[28–38] though it remains unclear under which conditions they are a practical source of epidemiological information. While the relationship between sampled viral loads, viral kinetics, and epidemic dynamics can be described mathematically under ideal conditions, in practice there are several factors which complicate its application as a practical epidemic monitoring tool. Measured Ct values are determined by a combination of biological factors, such as immunological history and variant causing the infection[39–41], and practical factors such as whether individuals are tested at a random point in their infection (e.g., asymptomatic screening) or around the time of peak viral load prompted by symptom onset[42], demography of the tested population[43], sample type[44], and RT-qPCR platform[45]. Whether these factors are prohibitively confounding when using Ct value distributions for epidemic monitoring has yet to be explored.

Here, we investigated the real-world feasibility of using SARS-CoV-2 Ct values to nowcast epidemic trajectories over three years of the COVID-19 pandemic. We first used synthetic datasets to benchmark nowcasting model performance and examined biological and logistical factors that might impede or improve nowcast accuracy. We then applied the same models to three real SARS-CoV-2 RT-qPCR testing datasets, collected across multiple geographic areas in the United States and under different population sampling strategies, to assess and inform the use of this approach in real-time estimation of epidemic growth rates.

# Results

*Correlation between epidemic growth rates and Ct value statistics using synthetic datasets*

To understand how biological and practical factors might affect Ct-based nowcasting performance, we created several synthetic Ct value datasets using real population-level reported incidence curves for Massachusetts, USA, combined with a viral kinetics model parameterized by longitudinal SARS-CoV-2 viral kinetics data, and sampling regimes representing a mixture of symptom-driven testing and asymptomatic screening (see Methods). Using synthetic datasets in this way allowed us to incorporate or exclude the effects of certain confounding factors on observed population-level Ct value distributions, in addition to the effect of the epidemic trajectory itself (see **Table S1**).

We first simulated an ideal dataset assuming: 1) highly asymmetric viral kinetics, with a very short growth phase and longer clearance phase; 2) low variation in observed viral load/Ct value for a given time-since-infection; and 3) a uniform probability of sampling an individual at any number of days after infection or symptom onset. We varied each of these factors in turn, resulting in four alternative scenarios with either: 1) increased symmetry in viral kinetics, with a more similar growth and clearance phase duration; 2) moderate variation in observed viral load/Ct value for a given time-since-infection; 3) a low-variance, gamma-distributed delay between infection or symptom onset and sampling; and 4) a realistic baseline scenario combining all three factors (see **Supplementary Text 1**).

All the synthetic datasets showed a clear negative correlation between the 7-day rolling average epidemic growth rate of cases and 7-day rolling average mean Ct value from the simulated symptomatic and asymptomatic samples, though the realistic baseline scenario showed the weakest correlation (**Figure S1**). Ct values from both symptom-based and random testing showed a relationship with epidemic growth rate (**Figure S1**), though Ct values observed through symptom-based testing were typically lower and exhibited less variation.

With each synthetic dataset, we fit generalized additive models (GAM) with smoothing splines to predict growth rates of cases as a non-linear function of daily mean and skewness of observed Ct values. We also fit corresponding logistic GAMs to predict the epidemic direction, i.e., whether incidence is growing or declining. We assessed in-sample fits of model-predicted vs. observed growth rates and direction across the entire dataset, based on RMSE and AUC respectively. We then refit the models using separate training and testing subsets of the data. To approximate a realistic application of the Ct-based approach in an ongoing epidemic, we fit the models using only the first 16 weeks of data and then performed rolling nowcasts with a two-week time horizon, using the fitted model to estimate the epidemic growth rate and direction daily over the next two weeks based on the Ct values reported during that time. At the end of each two-week window, we re-fit the model using all Ct values and incidence data up to that time point, then nowcast the next two-week window, and so on. As a sensitivity analysis, we compared RMSE and AUC with a fixed train-test split date at the end of 2021 (**Table S2**).

With the ideal synthetic dataset, the GAMs closely tracked observed growth rates using Ct value means and skew (**Figure S2 & Figure S3**; in-sample RMSE = 0.0191, approximately 10% of the range in observed log incidence growth rates), as well as accurately predict epidemic direction (in-sample AUC = 0.916). Nowcast accuracy over a rolling two-week window was slightly worse than the in-sample predictive performance (mean across all nowcast windows, RMSE = 0.0206, AUC = 0.867) but was still able to accurately track the epidemic over the full time period (**Figure 1**).

Model predictive performance was worse when using the realistic baseline synthetic dataset (**Figure 1, Figure S2 & Figure S3**; in-sample RMSE = 0.0319, AUC = 0.78; nowcast RMSE = 0.042, AUC = 0.698). The three factors examined individually had similar impacts on model performance; asymmetry of viral load trajectories caused the greatest increase in RMSE but only a slight decrease in AUC, while the distribution of delays between infection and sampling caused the

126 smallest increase in RMSE but the largest decrease in AUC (**Figure 1**, **Table 1**). When these
127 models were applied to nowcasting growth rates in two-week increments, the greatest perfor-
128 mance reduction occurred when increasing the individual variation in Ct values for a given time-
129 since-infection (**Table 1**).

130 **Table 1.** Predictive performance of GAMs using synthetic datasets, predicting per-day growth
131 rates from daily Ct value statistics.

| Dataset | RMSE, in-sample | RMSE, nowcast | AUC, in-sample | AUC, nowcast |
|---|---|---|---|---|
| Ideal condition | 0.0191 | 0.0206 | 0.916 | 0.867 |
| Realistic kinetics | 0.0245 | 0.0286 | 0.889 | 0.841 |
| Realistic variation | 0.024 | 0.0302 | 0.878 | 0.822 |
| Realistic sampling | 0.0237 | 0.027 | 0.865 | 0.824 |
| Baseline condition | 0.0319 | 0.042 | 0.78 | 0.698 |

132

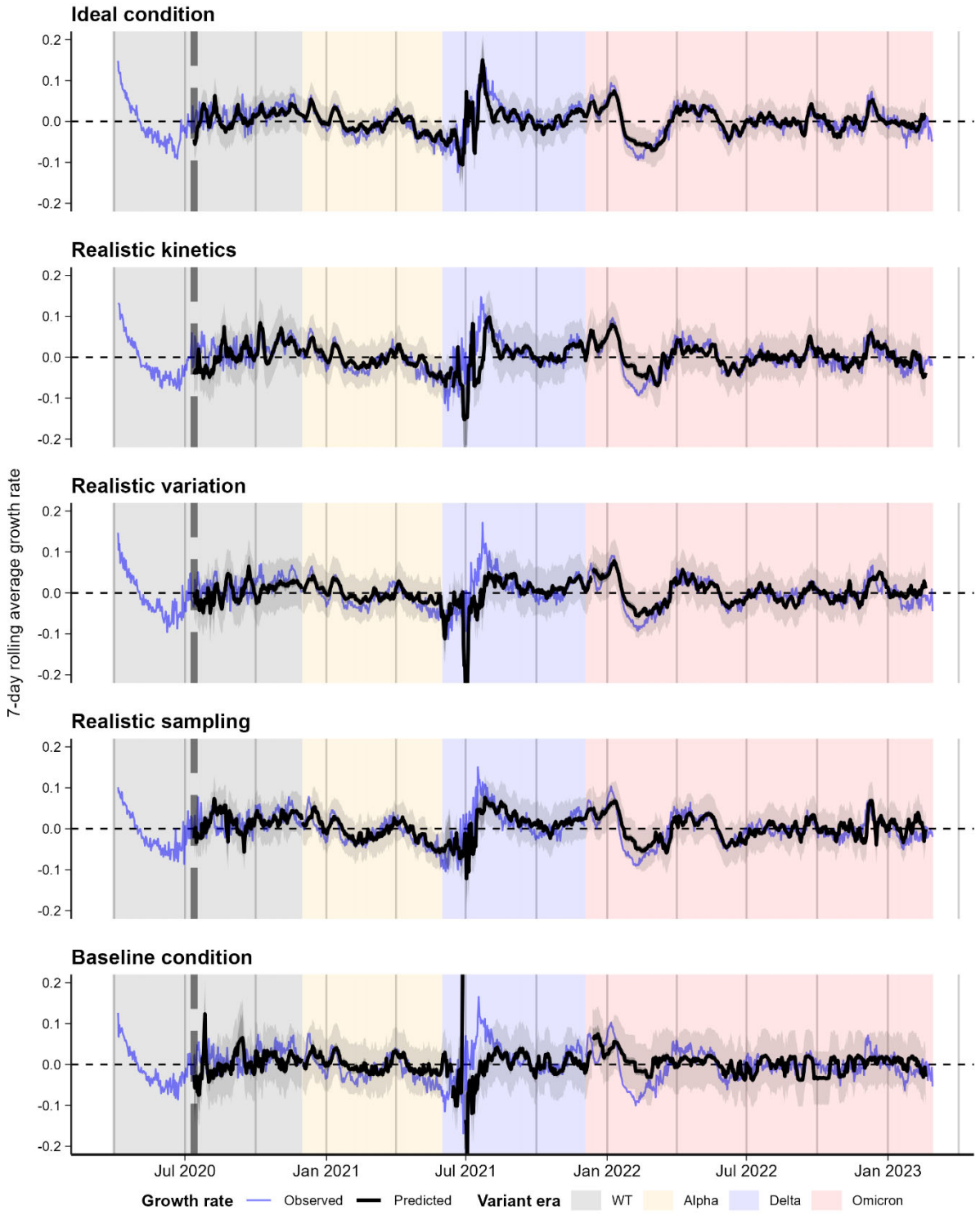



**Figure 1.** Model-predicted (black) vs. observed (blue) log incidence growth rates for the five synthetic datasets, with model-predicted 95% confidence intervals (dark shading) and 95% prediction intervals (light shading). Predictions are from 2-week rolling nowcast, concatenated into a single time series. Vertical dashed line denotes the end of the initial training period. Variant era was included for comparability with later models; the synthetic datasets do not include any impact of viral variant on kinetics. WT=wild type.

*Real-world relationship between observed Ct-value statistics and epidemic trajectories*

Having established a baseline for model nowcasting performance using the synthetic data, we next tested the nowcasting models on two RT-qPCR datasets: 1) routine hospital testing data from the Mass General Brigham hospital system in eastern Massachusetts (MGB), spanning Mar 2020-Feb 2023, and 2) municipal testing data from Los Angeles County, California (LAC), spanning May 2020-Jul 2021 and Jan-Sep 2022. The MGB data came largely from mandatory screening testing of outpatient, inpatients and emergency room admissions, while the LAC data were primarily symptom-driven voluntary testing (see Methods and **Table S3**). Both datasets contained specimen collection dates and Ct values for SARS-CoV-2 positive results; LAC data also included vaccination status, symptom status, and symptom onset dates. MGB Ct values came from seven platform/assay combinations, while Ct values from LAC data came from one PCR platform with two possible assays (see Methods).

We limited our analysis to tests reporting Ct values, using the first available recorded Ct value for each infection episode (see Methods). The final analyzed sample included 104,534 (MGB) and 279,492 (LAC) Ct values. We also applied our method to a third, smaller set of testing data from the Tufts Medical Center, also in eastern Massachusetts, with a total sample of 10,214 Ct values. We compared these Ct values against reported COVID-19 incidence for Massachusetts (MGB and Tufts) and Los Angeles County (LAC). We segment the data into four 'variant eras' based on the SARS-CoV-2 variant known or believed to be dominant in the U.S. during different approximate time periods, to allow for differences in viral kinetics by variant (see Methods).

Ct value distributions from both MGB and LAC datasets showed substantial variation over the course of the pandemic (**Figure 2A & Figure 3A**). Reported COVID-19 incidence in both locations varied over time as well (**Figure 2B & Figure 3B**), with large infection waves in the winters of 2020-21 and 2021-22, though the pattern of incidence was not synchronized across both settings.

While absolute incidence varied widely, incidence growth rates remained largely between ±0.2 throughout the course of the pandemic (**Figure 2B & Figure 3B**).

We found the mean and skewness of observed Ct value distributions (calculated daily over a seven-day moving window and excluding days with fewer than 10 Ct values reported) correlated with the growth rate in reported incidence (**Figure 2C & Figure 3C**). Analysis of cross-correlation functions found Ct value distributions lagged incidence growth rate in the MGB data, with strongest correlations at around 19-days lag (autocorrelation function, ACF= -0.462), and led incidence growth rates for the LAC data, with strongest correlations at around 10-days lead (ACF = -0.062) (**Figure S4 & Figure S5**). However, for real-time nowcasting, we focused on the relationship between same-day Ct values and incidence (i.e., lag=0 days; **Figure 2C & Figure 3C**), which still showed high correlation. Higher incidence growth rates corresponded with lower same-day average Ct values (Spearman's correlation coefficient: MGB Rho = -0.43, LAC Rho = -0.22) and with positively skewed Ct distributions (MGB Rho = 0.35, LAC Rho = 0.43).

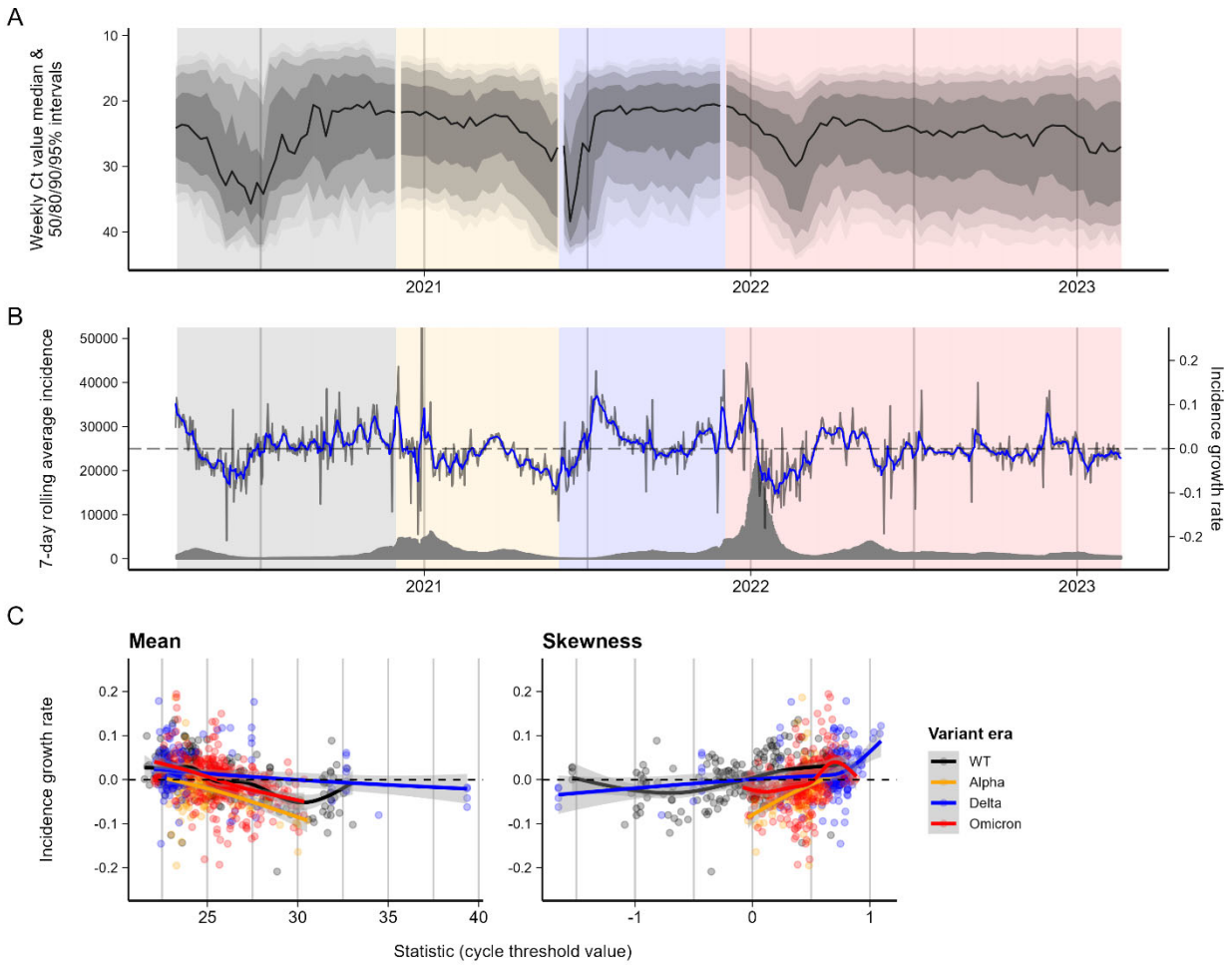



**Figure 2.** Ct values from the Mass General Brigham hospital system and corresponding reported COVID-19 incidence in Massachusetts, USA. (**A**) Weekly Ct value quantiles over time, showing weekly median Ct value and 50/80/90/95% quantiles. (**B**) 7-day rolling average reported incidence (grey bars), growth rate in 7-day rolling average reported incidence (grey line), and smoothed growth rate (blue line). Background is shaded by time periods of different variant dominance. Vertical dashed line demarcates the test-train split. **(C**) Incidence growth rate compared to smoothed daily mean and skewness of Ct value distributions. Colored lines and shaded grey regions show fitted cubic spline GAMs with 95% confidence intervals, stratified by period of variant dominance.

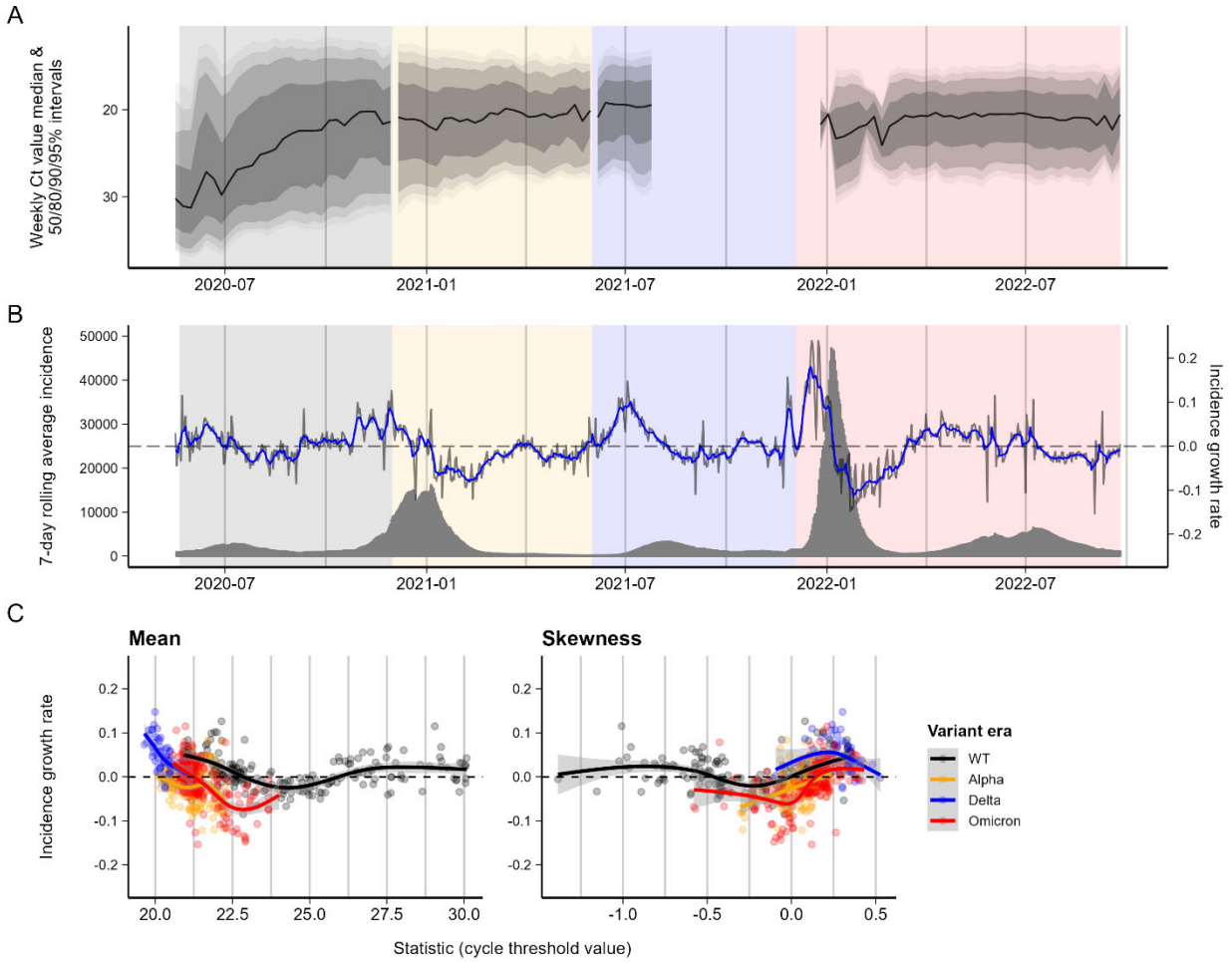



**Figure 3.** Ct values from Los Angeles County and corresponding reported COVID-19 incidence. (**A**) Weekly Ct value quantiles over time, showing weekly median Ct value and 50/80/90/95% quantiles. (**B**) 7-day rolling average reported incidence (grey bars), growth rate in 7-day rolling average reported incidence (grey line), and smoothed growth rate (blue line). Background is shaded by time periods of different variant dominance. Vertical dashed line demarcates the test-train split. (**C**) Incidence growth rate compared to smoothed daily mean and skewness of Ct value distributions. Colored lines and shaded grey regions show fitted cubic spline GAMs with 95% confidence intervals, stratified by period of variant dominance.

*Nowcasting epidemic growth rates using Ct values in Massachusetts, USA and Los Angeles County, USA*

We next re-trained the same GAM models used with synthetic data to the MGB and LAC dataset using smooth functions of mean and skewness of Ct values to predict log incidence growth rates, with corresponding logistic models to predict epidemic direction. Model predictions were compared against observed values first in-sample across the entire dataset then over a rolling two-week nowcast window, as well as with a single fixed train-test split date at the end of 2021.

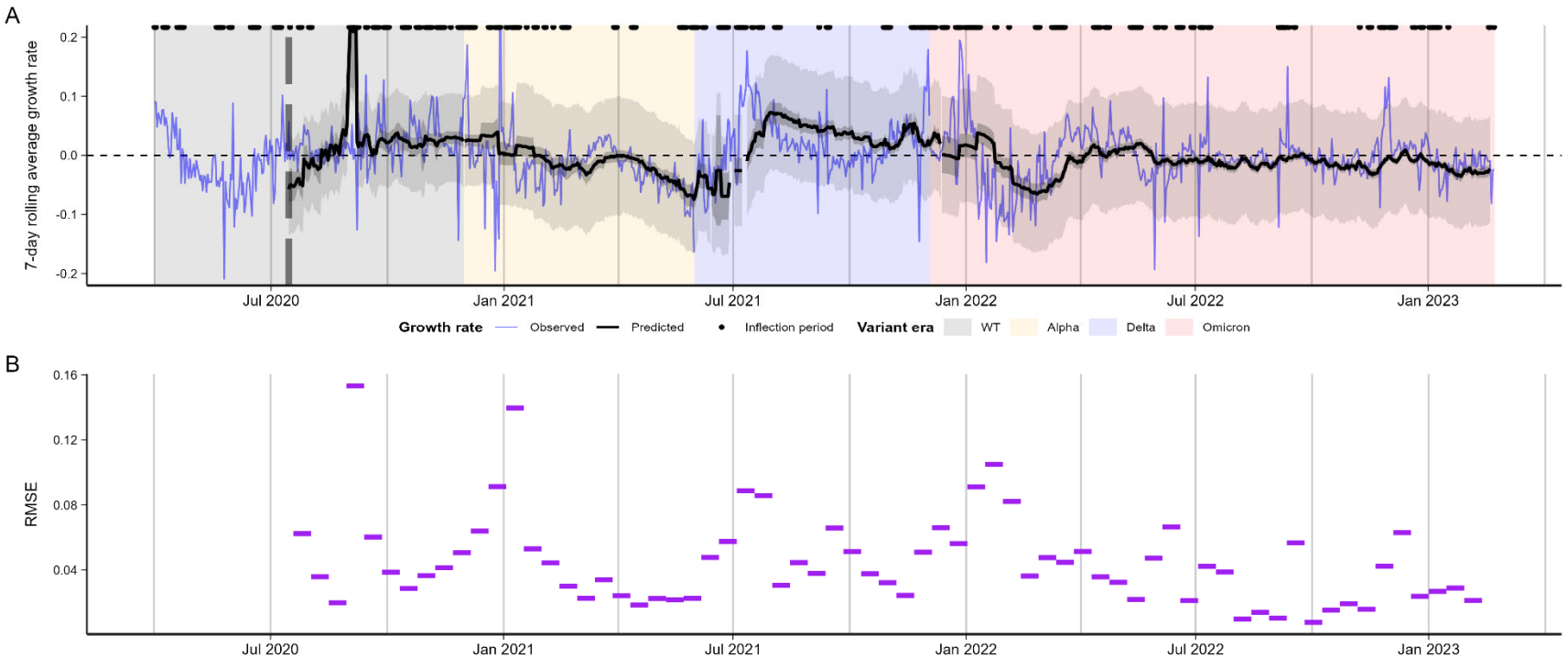
In both datasets, this simple model achieved in-sample prediction accuracy for incidence growth rate only slightly worse than performance on the realistic synthetic data, with relatively small absolute errors (MGB RMSE = 0.0451; LAC RMSE = 0.0335, see **Figure S6-S9**, **Table S4**, and **Table 2**). Corresponding logistic regression models successfully discriminated growing from declining incidence (Area under the curve: MGB AUC = 0.785, LAC AUC = 0.843).

The models were able to nowcast growth rates, in two-week increments with models periodically refitted to more recent data, with accuracy slightly worse than in-sample model fits (MGB RMSE = 0.0523, LAC RMSE = 0.039) (**Figure 4A & Figure 5A**). This level of nowcast accuracy was likewise only slightly worse than nowcasting performance with realistic synthetic data. While average prediction error was relatively small, comparable to in-sample model error and to prediction error with realistic synthetic data, accuracy was highly variable from one two-week window to the next (**Figure 4B & Figure 5**B). Nowcast accuracy was comparable to model performance over a fixed multi-month prediction window, slightly better for one dataset and worse for the other (MGB RMSE = 0.047, LAC RMSE = 0.0458) **(Table S2)**. Nowcast predictions of epidemic direction were slightly worse than in-sample ones (MGB AUC = 0.723, LAC AUC = 0.784) and outperformed the directional discrimination test with realistic synthetic data. In addition, over all two-week nowcast windows combined, model-predicted growth rates correlated moderately well with observed ones (Spearman's Rho: MGB Rho = 0.398, LAC Rho = 0.556).

220 **Table 2.** Predictive performance of the selected GAM using data from MGB and LAC, predicting
221 per-day growth rates from daily Ct value statistics.

| ***Dataset*** | **RMSE** | | | **AUC** | | |
|---|---|---|---|---|---|---|
| | In-sample | Nowcast | Periods of rapid change in growth rate | In-sample | Nowcast | Periods of rapid change in growth rate |
| *MGB* | 0.0451 | 0.0523 | 0.0645 | 0.785 | 0.723 | 0.722 |
| *LAC* | 0.0335 | 0.039 | 0.0471 | 0.843 | 0.784 | 0.772 |

222




223

224 **Figure 4.** **(A)** Model-predicted (black) vs. observed (blue) log incidence growth rates for MGB data, with 95% confidence intervals
225 (dark shading) and 95% prediction intervals (light shading). **(B)** RMSE of predicted vs. observed log incidence growth rates for each 2-
226 week nowcasting window. "Inflection periods" refer to times when the absolute smoothed log incidence growth rate exceeded 0.025
227 over a one-week period, marked with points above each subplot.

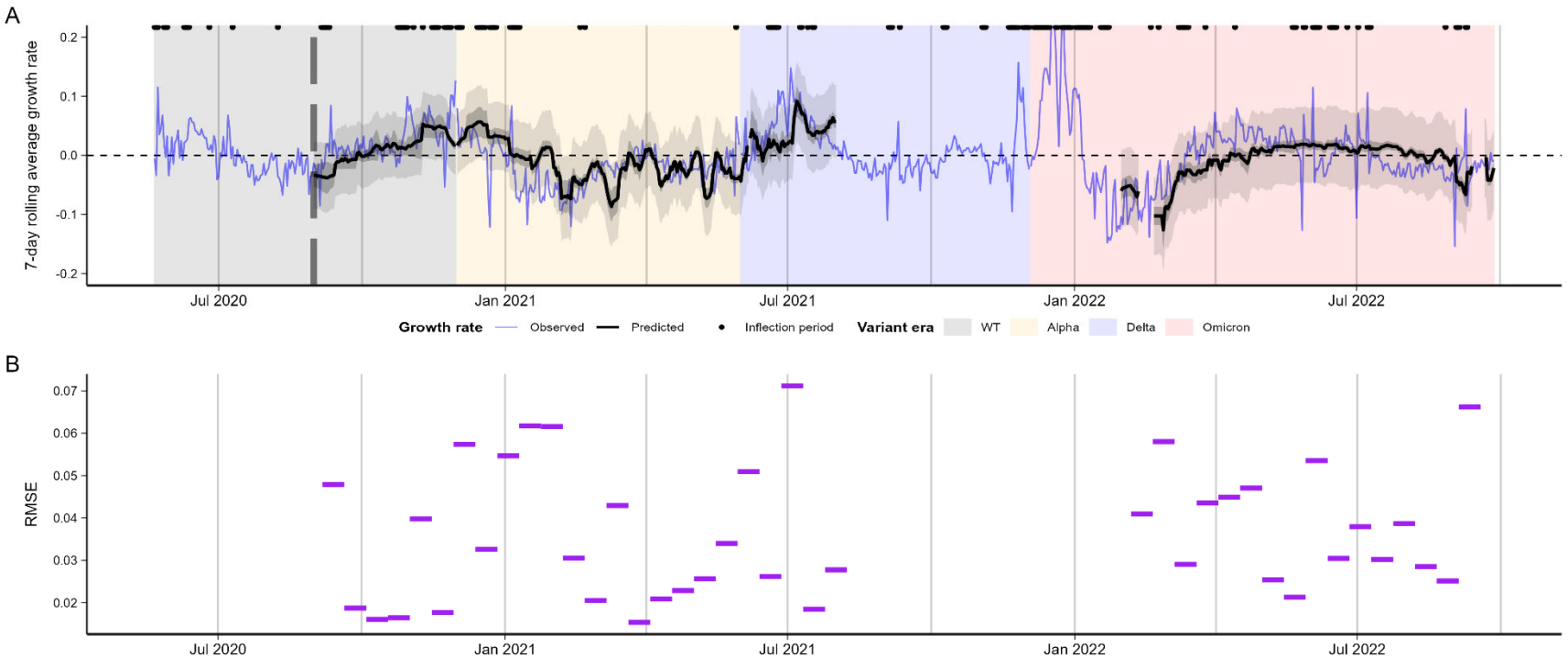



**Figure 5. (A)** Model-predicted vs. observed log incidence growth rates and RMSEs for LAC data. Model-predicted (black) vs. observed (blue) log incidence growth rates for MGB data, with 95% confidence intervals (dark shading) and 95% prediction intervals (light shading). **(B)** RMSE of predicted vs. observed log incidence growth rates for each 2-week nowcasting window. "Inflection periods" refer to times when the absolute smoothed log incidence growth rate exceeded 0.025 over a one-week period, marked with points above each subplot.

*Nowcasting performance during time periods of rapid change in growth rate*

To assess nowcasting performance during periods of rapid change in the epidemic trajectory, we identified times when the absolute smoothed incidence growth rate exceeded 0.025 over a one-week period. This definition captured 30.1% of nowcast dates for the MGB data (284/944 days) and 17.5% for LAC (98/560 days). We then recalculated in-sample and out-of-sample prediction accuracy for growth rate and epidemic direction during just these periods.

Across both datasets, prediction error over periods of rapid change was greater than over the whole nowcast period (MGB RMSE = 0.0645 [change] vs. 0.0523 [nowcast], LAC RMSE = 0.0471 [change] vs. 0.039 [nowcast]; see **Table 2**). However, directional prediction accuracy was comparable between periods of rapid change and the whole nowcast period (MGB AUC = 0.722 vs. 0.723, LAC AUC = 0.772 vs. 0.784).

*Nowcasting performance with variable sample size and outlier removal*

We assessed the sensitivity of nowcasting performance to sample size both by randomly downsampling the MGB dataset (100 random draws) and by analyzing a third, smaller dataset from Tufts Medical Center using the same response variable (i.e., log incidence growth rates for Massachusetts) but with approximately 10% of the total sample size of the MGB data (see Methods; **Figure S10, S11**). In most cases, prediction accuracy for incidence growth rate was comparable with the downsampled datasets and the equivalent full datasets (**Figure 6**; see also **Table S5**). Only with 10% of the full dataset (but not with the Tufts dataset) did nowcasting accuracy degrade appreciably; with 50-75% downsampling or a daily maximum of 25 positive samples, accuracy improved compared to baseline. Likewise, directional prediction accuracy was generally similar between downsampled and full datasets, with substantially worse accuracy only for the 10% downsample. Improved accuracy may reflect reduced influence of outliers – downsampling the full dataset tends to exclude the days with smallest sample sizes, which are otherwise given equal weight in model training to days with more observations, while sub-sampling each day's

observations reduces the impact of outliers on each day's observed Ct value distribution. To test this, we examined model performance with trimming of outlier Ct values from each day's observed data. Trimming outliers reduced prediction error with 2.5%, 5%, and 10% trims (**Figure 6**, **Table S5**), while 2.5% and 5% trims also improved directional prediction accuracy.

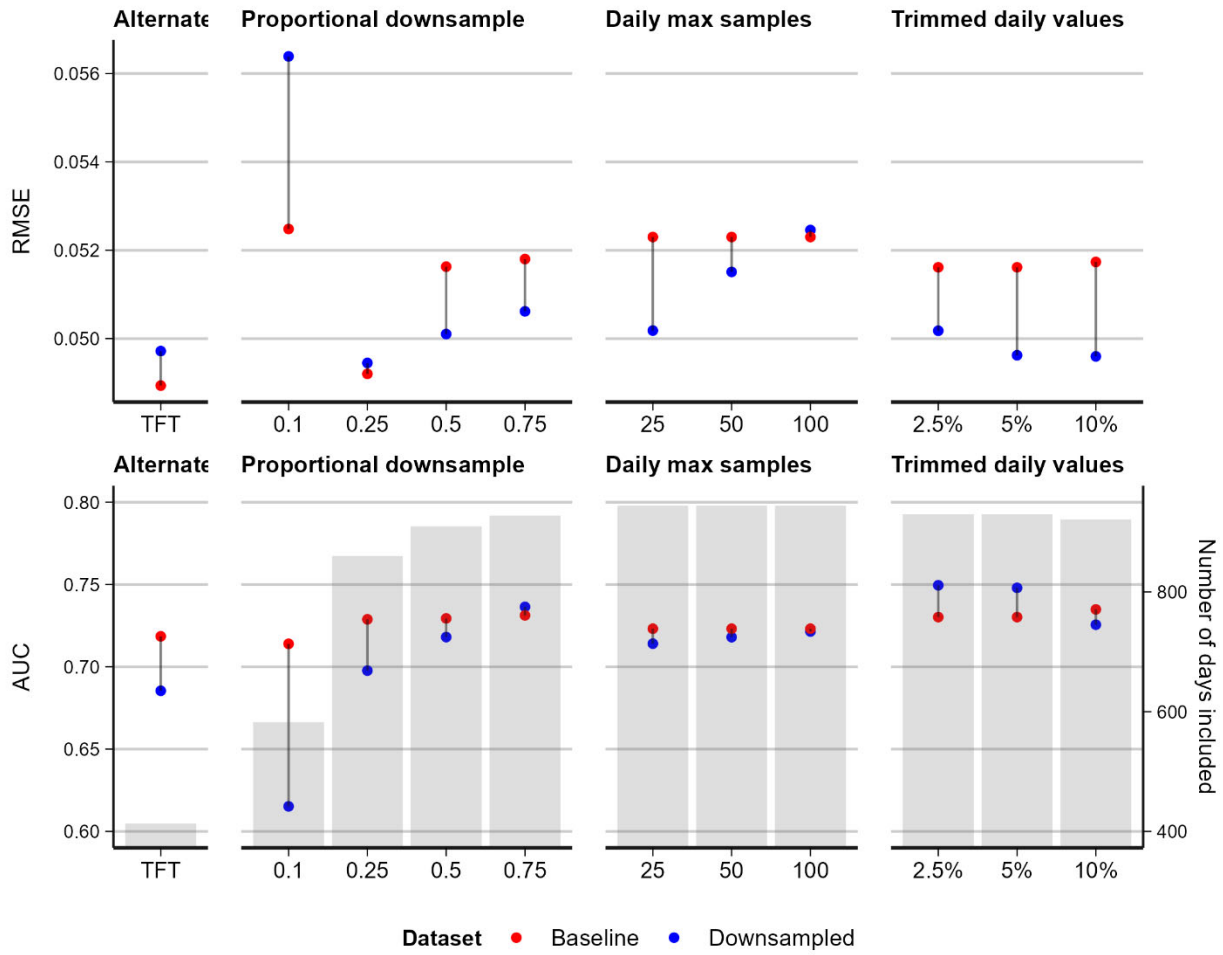



**Figure 6.** Model performance for downsampled MGB and full Tufts datasets. Baseline comparison metrics are re-calculated for only the days included in each downsampled dataset's nowcast. For proportional and daily max downsampling, both downsampled and baseline performance are averaged over 100 random draws (and their corresponding days included). Trim percentages indicate quantiles trimmed from each end of daily Ct value distributions (i.e., 5% trim yields the 5-95 percentile range of Ct values).

*Sensitivity analysis*

The reason for testing (e.g., symptom driven testing vs. screening asymptomatic outpatients) is expected to result in different distributions of observed Ct values due to variation in when individuals are tested during their infection; therefore, the relationship between Ct values and epidemic growth rate is expected to differ correspondingly. In addition, in the MGB data, individuals were swabbed differently and tested on different PCR platforms depending on their reason for seeking healthcare, including a mixture of patients tested as outpatients, inpatients and in the emergency room. To understand the impact of these factors on the modelled relationship between Ct values and growth rate, we assessed performance of GAMs using only 1) MGB data from routine outpatient screening, the majority of whom were sampled in the same way and tested on the same PCR platform **(Figure S12)**; 2) LAC data stratified by symptom status (symptomatic vs. asymptomatic vs. no known symptom status); 3) LAC data from tests conducted on asymptomatic individuals and those without known symptom status; and 4) LAC data from unvaccinated individuals with no known previous SARS-CoV-2 infection. In all cases, we compared performance to the base model for the respective data source. The relationship between Ct values and growth rate appeared to differ when subsetting or stratifying by these variables (**Figure S13**), but including these stratifications in the model did not always improve predictive performance. Restricting to outpatient tests only improved prediction error compared to baseline (nowcast RMSE = 0.0494 vs. 0.0523 base), whereas incorporating symptom status or immune history slightly worsened prediction error (nowcast RMSE = 0.454 for symptom-stratified, 0.0415 for asymptomatic/no symptom status only, 0.0401 for immunologically naïve only, vs. 0.039 base, see **Table S6**).

# Discussion

Under real-world conditions, simple generalized additive models using the mean and skewness of recorded Ct values could nowcast (log) incidence growth rates with prediction errors (RMSE) of approximately 0.04-0.05. Across both settings (Massachusetts and Los Angeles County),

295 growth rates generally varied between approximately ±0.2, so this level of accuracy in modelled
296 estimates, while not highly precise, is nonetheless informative. These models are also able to
297 identify if incidence is growing or shrinking with AUC greater than 0.7, substantially better than
298 chance.

299 Nowcast accuracy over two-week time horizons is slightly worse than the quality of in-sample
300 model fits, especially early into the emergence of new dominant viral variants whose effect cannot
301 yet be accurately estimated. During periods of rapid change in incidence growth rate (e.g., just as
302 a new outbreak wave is developing), nowcast accuracy for growth rate is slightly worse, possibly
303 due to larger absolute growth rates during such periods. Crucially, however, directional predictions
304 remain moderately accurate during those times.

305 Our results support the theoretical expectation that epidemic dynamics influence population-level
306 viral load distributions, and therefore can be inferred from them[21]. They also corroborate the find-
307 ings from other settings, where Ct values have been used successfully to infer epidemic growth
308 rates or reproduction numbers[28–38]. Our analysis builds on these studies with one of the largest
309 empirical tests of this nowcasting approach to date using data from two locations in the USA over
310 a three-year period. Epidemic growth rates and directions were accurately nowcasted using both
311 datasets, despite showing different Ct value trends and capturing different populations, highlight-
312 ing the generalizability of this approach. Furthermore, these data covered a long-time window and
313 included periods of different variant dominance and population immunity, suggesting Ct values
314 could continue to augment infectious disease surveillance as SARS-CoV-2 epidemiology contin-
315 ues to change.

316 In practice, several factors can confound the relationship between Ct values and epidemic dy-
317 namics (measured here as growth rate of case incidence), including testing delays, sampling re-
318 gimes (i.e., community-based random testing vs. testing patients in hospital), symptomatic (diag-
319 nostic) vs. asymptomatic (screening) testing, immunological history, and the inherent individual-

level variability in SARS-CoV-2 viral kinetics. Our synthetic data analyses help disambiguate these confounding factors by comparing degradation in predictive performance between different synthetic datasets. Predictive performance was slightly worse in the real datasets compared to a 'realistic' synthetic dataset. One key contributor is that Ct values from the real datasets were collected using multiple RT-qPCR assays and/or platforms and were not standardized and may generate different Ct values for the same underlying viral load, limiting the comparison of Ct values across platforms and assays (see Methods)[45–47]. Additionally, the data-generating model for our 'realistic' synthetic dataset did not incorporate the impact of vaccination or past infection which affect individual viral load trajectories[40,48], potentially contributing to the differences in performance between models with empirical vs. synthetic data.

Our synthetic data analysis also highlights the importance of considering the delay between infection and sampling an individual in determining the population-distribution of Ct values. Fundamentally, the relationship between population-level epidemic dynamics and viral load distributions arises because individuals' viral loads reflect times since infection[21], and hence cross-sectional distributions of viral loads (or Ct values) reflect the distribution of times-since-infection among currently infected individuals, similar to the relationship between incidence and prevalence. This relationship can be readily described mathematically if individuals are randomly sampled, with a uniform probability of sampling any time after infection. Random cross-sectional samples capturing infections at random points in their infection are rare (see [22,31,36,49]) but are reasonably well approximated in our datasets by routine screening of hospital outpatients. However, a more realistic sampling delay distribution – such as if individuals tend to be tested shortly after suspected exposure or developing symptoms – biases the probability of sampling over time since infection and dilutes the signal of infection age. Symptom-driven testing where individuals are tested due to recent symptom onset beginning at around the same time as peak viral load, is the most common source of data used for epidemiological surveillance, which reduces any epidemic signal in

the population-level Ct distribution. In the extreme, if individuals were sampled with the same delay following infection, then any observed variation in viral loads would arise from random individual variation at a single time-since-infection rather than reflecting a distribution of times-since-infection among current infections. Changes in public health recommendations around testing and screening algorithms, such as recommendations around pre-travel testing or hospital admissions screening, may therefore change the relationship between population Ct values and epidemic dynamics, which may bias Ct-based epidemiological estimates if not accounted for.

PCR platform differences and nonrandom sampling regimes are both addressable challenges, at least in principle. Ct value data could be calibrated across platforms and assays using standardized samples. Random surveillance sampling could reduce the bias in testing delay found with symptom-driven testing. True random sampling may be important, as voluntary testing by asymptomatic individuals may still show some bias in testing delays (**Table S6**). When we approximated these changes by subsetting one of our datasets to only results from outpatient screening tests, which were largely collected and analyzed the same way (**Figure S12**), we found small improvements in model predictive performance compared to using the full, mixed dataset (**Table S6**). While random surveillance sampling at low prevalence may yield very few infections detected, nowcasting accuracy was not severely degraded even with substantially reduced sample sizes (**Figure 6**). Both these changes would improve the accuracy of simple Ct-based nowcasting models. Even absent such logistical solutions, however, we found the simple statistical heuristic of trimming outliers (2.5-5%) from daily observed Ct values improves nowcasting accuracy (**Figure 6**).

Beyond confounding factors, it is plausible that growth rate of reported COVID-19 cases may not be the most accurate benchmark against which to compare Ct value distributions. First, symptomatic cases occur and are reported with a lag relative to infections, and may be affected by changes in testing behavior, for example with the increased availability of home-based rapid

antigen tests. Alternative benchmarks, such as growth rate in hospitalizations, mortality, or wastewater viral loads, may therefore yield stronger relationships (possibly with some time-shifting); investigating these relationships would be a fruitful avenue for further research. In addition, geographically aggregated incidence may mask heterogenous outbreak trajectories at finer scale, e.g., city or even neighborhood level. Such finer-scale incidence data may yield cleaner relationships with Ct value distributions, especially if matched to the catchment areas for the Ct value data collection process.

Another challenge for modeling Ct value dynamics is the choice of mathematical model to capture the relationship between observed Ct values and underlying epidemic growth rates. The link between epidemic dynamics and viral loads observed through random cross-sectional surveillance can be described precisely based on the convolution of the infection incidence curve and viral kinetics curve[21,31,36]. In contrast, viral loads observed through non-random or convenience samples, such as symptom-driven testing, arise from complex data generating processes which are difficult to describe mathematically, and thus past studies, including ours, tend to favor regression models to estimate epidemic dynamics from observed Ct values[30,33,34]. Future work should focus on more complex statistical methods that take into account the time-series nature of the data[37], the non-linear and potentially non-monotonic relationship between Ct values and growth rates, and combine multiple data streams to provide more accurate predictions of epidemic dynamics[36].

Tracking epidemic growth rates in near-real-time remains an important challenge for public health surveillance. Our analyses show that simple Ct-based models can accurately track SARS-CoV-2 epidemic growth rates, highlighting their potential use in augmenting infectious disease surveillance systems. Ultimately, their greatest strength lies in their speed and simplicity. The models presented here are conceptually straightforward and computationally lightweight, easy to implement even in resource-constrained settings, and, unlike wastewater testing, are reliant only on data already routinely collected as part of screening or diagnostic testing. Our analyses show that

they retain their accuracy even with limited sample sizes or during periods of rapid change in epidemic trajectories, such as during the transition from the end of one epidemic wave to the start of the next one, and so could provide rapid situational awareness as outbreak waves emerge. Further research could examine how Ct-based estimates of epidemic trajectories complement other, orthogonal indicators such as wastewater surveillance, as well as potential applications to different viral pathogens with well-characterized viral kinetics such as influenza or RSV[50,51].

# Methods

*Study settings & data sources*

*Massachusetts*

Massachusetts Ct value data comes primarily from testing in 16 hospitals in the Mass General Brigham hospital system, with a catchment area largely in eastern Massachusetts. The full dataset comprises 2,671,041 SARS-CoV-2 test results, with specimen collection dates ranging from 3 Mar 2020 to 23 Feb 2023, of which 161,273 were positive. There were 3531 individuals who appeared to experience repeat infections (defined as >60 days between positive results), of which 72 individuals had 2 or more repeat infections. As we could not rule out long COVID or other idiosyncratic viral kinetics, we drop these 72 individuals from the final dataset. Limiting to results reporting Ct values and first reported Ct values for each confirmed case yields the final sample of 104,534 Ct values used in this analysis (**Table S3**), of which the earliest specimens were collected on 31 Mar 2020.

Samples are from a combination of routine outpatient (77,700; 74.3% of samples) and inpatient (7,311; 7.0%) screening and diagnostic tests, as well as ER patient testing (19,523; 18.7% of samples); while not entirely random nor representative, routine screening tests suffer less self-selection bias than symptom-based or voluntary testing. We did not have access to information on patients' vaccination or infection history, infecting variant, or symptom status.

The final sample includes specimens collected from nasal and nasopharyngeal swabs (approx. 2:1 ratio). Specimens were processed using seven different RT-qPCR platform/assay combinations (**Table S3**), variously targeting E/N/N1/N2/ORF1ab genes. For the main analysis here, Ct values were pooled across platforms/assays; where a single result reported Ct values for multiple target genes, the lowest value was used.

Daily confirmed case counts for Massachusetts were obtained from the Massachusetts Department of Public Health COVID-19 dashboard[52].

We also analyzed a secondary dataset of Ct values from Tufts Medical Center in Boston, Massachusetts for comparison. This dataset comprised 84,848 test results with collection dates ranging from 18 Feb 2021 to 31 Oct 2022, of which 10,338 were positive. Filtering the reported test results using the same criteria as used for the MGB data yielded a final sample of 10,214 Ct values used here. **Figure S11** summarizes the reported Ct value distributions over time and compares these to reported COVID-19 incidence.

*Los Angeles County*

LAC Ct value data comes from municipal COVID-19 testing sites operated by the LAC Department of Public Health and Department of Health Services, comprising approximately 10% of all municipal testing conducted in LAC during the sample period. The full dataset comprises 330,034 SARS-CoV-2 positive test results, with specimens collected over two time periods – 21 May 2020 to 27 Jul 2021, and 30 Dec 2021 to 29 Sep 2022. (Note: data were unavailable for the intervening period.) The data contain an infection episode identifier; limiting to the first reported Ct value for each infection episode yields the final sample of 279,492 Ct values used in this analysis.

The final sample includes specimens collected through nasal, nasopharyngeal, and oral swabs, and analyzed by Fulgent Genetics using an in-house platform and ThermoFisher QuantStudio™ 6 and 7 PCR systems. Two RT-qPCR assays were used; before mid-Nov 2020, analyses used exclusively LOINC 94531-1 targeting N1 and N2 genes, while subsequently the majority of analyses used LOINC 94533-7 targeting the N gene. Where a single result reported Ct values for multiple target genes, the lowest value was used. Symptom status was reported for approximately 75% of the sample, of which in turn approximately 75% (56% of the full sample) are reported as symptomatic for COVID-19 (**Table S3**). For symptomatic cases, most specimens were collected

1-10 days after symptom onset (modal delay of 3 days). The sample also included vaccination status, with approximately 24% of results coming from vaccinated (partially, fully, or boosted) individuals (**Table S3**).

Daily confirmed case counts were obtained from the LAC DPH COVID-19 dashboard[53].

*Synthetic datasets*

We built on a previously published model to simulate realistic Ct value distributions that would be expected under testing and sampling schemes similar to real-world data[21]. Full details of the simulation framework are given in **Supplementary Text 1**. First, we parameterized a viral kinetics model describing the expectation and distribution of Ct values over all days following infection using previously published longitudinal SARS-CoV-2 testing data (**Figure S15, Table S7**)[48]. This is a piecewise linear model governed by a set of control points determining the time from infection to peak viral load, time from peak viral load to an inflection point at a high Ct value, and a longer-term clearance rate with a daily probability of full clearance. Second, we simulated approximately 2 million infections with infection times distributed based on the reported incidence of COVID-19 cases in Massachusetts between 5 March 2020 and 25 Feb 2023. Third, we simulated a surveillance system as a mixture of random testing (i.e., symptom-independent) and symptom-based testing (individuals are tested with a random delay following a randomly generated incubation period). Combining these three simulation steps gave a synthetic dataset of Ct values for a mixture of asymptomatic and symptomatic individuals tested at various times post infection and over a multi-wave SARS-CoV-2 epidemic (**Figure S1**). Different scenarios were captured by changing the parameters used either for the viral kinetics model or sampling delay distribution (**Figure S16**).

*Statistical methods*

We calculated daily incidence-based growth rates as the natural log-transformed ratio of 7-day moving average new reported cases for each day to the 7-day moving average for the preceding day:

$$y_t = \ln \frac{\sum_{k=0}^{6} f_{(t-k)}}{\sum_{k=1}^{7} f_{(t-k)}}$$

where $y_t$ is incidence growth rate and $f_t$ is daily incidence at time $t$. We defined epidemic direction as growing when $y_t > 0$ and declining when $y_t \leq 0$.

*Classifying time periods of rapid incidence change*

To identify periods of rapid change in incidence growth rate, we first smoothed the daily incidence growth rate (as defined above) using a centered 7-day moving average:

$$y'_t = \frac{1}{7} \sum_{k=-3}^{3} y_{(t+k)}$$

We then identified times when the absolute change in smoothed log incidence growth rate $y'_t$ equals or exceeds 0.025 over a one-week period, denoting the midpoint days of those weeks as periods of rapid change. That is, time $t$ is defined as having rapid change in incidence if and only if $\left|y'_{(t+3)} - y'_{(t-3)}\right| \geq 0.025$.

*Growth rate & epidemic direction models*

We modeled incidence growth rate using a generalized additive model (GAM) incorporating the mean and skewness of Ct values:

$$\ln y_t = \beta_0 + s_{\bar{x}}(\bar{x}_t) + s_g(g_t) + \beta_v v_t$$

Where $s_{\bar{x}}$ and $s_g$ are smoothing functions fitted using cubic regression splines[54], and $\bar{x}_t$ and $g_t$ are the 7-day rolling averages at time $t$ of the daily mean and skewness respectively of Ct values

from samples collected or over the window from time $t$ to $t-6$, excluding days with fewer than 10 Ct values reported. $v_t$ is a categorical variable identifying the SARS-CoV-2 variant known or believed to be dominant in the U.S. during different approximate time periods. For our datasets, we designated four such variants / time periods: wild type (up to 30 Nov 2020), Alpha (01 Dec 2020 to 31 May 2020), Delta (01 Jun 2020 to 03 Dec 2021), and Omicron (04 Dec 2021 onwards). We used this rough approximation rather than relying on more direct and detailed observations, e.g. sequencing data linked to our datasets, to better represent a realistic use case for the Ct-based method such as a small municipal public health department. In such cases, resources for extensive sequencing may not be available, necessitating reliance on broader national trends. When encountering new variant[s] in a nowcasting or testing period not present in training data, our models use a realistic decision rule of making predictions based on the last known variant from training data.

We model epidemic direction using logistic regression models equivalent to the GAMs used for incidence growth rate.

To determine our choice of model, we tested a series of log-linear regression models and GAMs, using different predictors (daily Ct mean, standard deviation, and skewness), functional forms (log-linear vs. cubic regression splines), and variant interaction terms. We fitted these models to the baseline synthetic dataset and compared their AIC as well as in-sample and nowcasting performance (see below). There was a clear bias-variance tradeoff between models; more flexible model specifications yielded better AIC and in-sample fit, at the cost of worse out-of-sample or nowcasting performance (see **Table S2** and **Figure S14**). We ultimately selected the final model using mean and skewness with a cubic spline, as the theoretical relationship between cross-sectional Ct values and epidemic growth rates is non-linear and depends on the *distribution* of Ct values observed; short of fitting the growth rate model to the entire distribution of observed values,

using mean and skewness provides a parsimonious way to include information about the shape of the distribution in the model.

*Evaluating model performance*

To evaluate the performance of Ct-based nowcasting models, we conducted two model validation tests. First, we fitted the main models to each dataset using only data up to 31 Dec 2021, then used the fitted model to predict incidence growth rates and epidemic direction for the remainder of each dataset (from 01 Jan 2022 onwards), based on observed Ct values reported in each dataset. We assessed prediction performance using RMSE between model-predicted and observed incidence growth rates, as well as AUC for directional predictions from the logistic regression model.

Next, we conducted a 'rolling' nowcast test, intended to simulate a realistic application of this approach. For each dataset, we trained the main models on the first 16 weeks of available data, using the models thus fitted to predict incidence growth rates and epidemic direction over the following 2-week period using only reported Ct value statistics. We then re-fit the models incorporating those two weeks of incidence data (i.e., up to 18 weeks) and predict the *subsequent* 2-week period, repeating this re-fitting and prediction procedure in 2-week increments up to the end of each dataset. We report prediction performance as RMSE or AUC across all 2-week prediction periods concatenated into a single prediction time series for each dataset and model, while detailed period-by-period performance is reported in the online repository at [53].

*Impact of reduced sample size and outliers on Ct-based growth rate estimation*

As a sensitivity analysis, we repeated the rolling nowcast analyses using artificially down-sampled datasets. We generated downsampled versions of the dataset in two ways: 1) by randomly drawing 10/25/50/75% of the total test results available, or 2) by limiting the maximum number of positive test results for each day to 25/50/100, discarding any additional tests. We then reassessed

nowcasting performance on each of these downsampled datasets. We repeated this analysis with 100 different randomly downsampled datasets for each size, taking the mean of model performance metrics over the 100 draws at each size. We also compared nowcasting performance with a similar analysis using a third, smaller dataset from Tufts Medical Center, which uses the same response variable data as the MGB dataset (i.e., log incidence growth rates for Massachusetts) but has approximately 10% of the total sample size. The downsampling process can result in some days being excluded from the downsampled dataset model's nowcast. Nowcasting performance can vary considerably from day to day, with outlier days having disproportionate impact. To ensure fair comparison of the impact of downsampling on model accuracy, rather than the impact of certain days being excluded as an indirect result of the downsampling process, we recalculated performance metrics for the baseline model's nowcasts based on just the days included in any given downsampled model's nowcasts, once again taking the mean of model performance metrics over the 100 different baseline subsets included for each sample size.

To assess the impact of outliers on nowcasting performance, we trimmed daily observed Ct value distributions by 2.5/5/10% (yielding 95/90/80% ranges) before calculating Ct value distribution statistics, using the trimmed data for both training and nowcasting. Repeat draws were not required as the trimming is deterministic. As with the downsampling analysis, we recalculated baseline model performance metrics for only days included at each trim level.

*Data & code availability*

Data and analysis code are available online at https://github.com/gradlab/ct-nowcasting [NOTE: we will update this to a Zenodo DOI before publication].

*Acknowledgements & financial disclosures*

JAH is supported by a Wellcome Trust Early Career Award (grant 225001/Z/22/Z). This work was supported in part by the Francis P. Tally, MD, Fellowship in the Division of Geographic Medicine and Infectious Disease (JAP). This project has been funded in part by contract 200-2016-91779

with the Centers for Disease Control and Prevention (CDC). Disclaimer: The findings, conclusions, and views expressed are those of the authors and do not necessarily represent the official position of the CDC. The authors also thank Jason Cheng and Hanlin (Harry) Gao of Fulgent Genetics for assistance with data for the analysis.

All authors declare no competing interests. No authors nor our institutions received any payments or services in the past 36 months from a third party that could be perceived to influence, or give the appearance of potentially influencing, the submitted work.

*Ethics guidelines*

The authors declare that all relevant ethical guidelines have been followed and all necessary IRB and/or ethics committee approvals have been obtained.

*Author contributions*

TYL, YHG, and JAH conceptualized the project. TYL and JAH designed the analyses, developed the code, and created the visualizations. TYL, SK, JP, MH, THK, and PD prepared data. SK, SD, RF, and YHG provided resources and contributed to analysis design and interpretation. YHG provided primary supervision and funding support. TYL and JAH wrote the first draft. All authors provided critical review and revision of the text and approved the final version.

# References

1. Bhatia, S. *et al.* Lessons from COVID-19 for rescalable data collection. *Lancet Infect. Dis.* **23**, e383–e388 (2023).
2. Cori, A. *et al.* Key data for outbreak evaluation: building on the Ebola experience. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, 20160371 (2017).
3. Lipsitch, M. *et al.* Infectious disease surveillance needs for the United States: lessons from COVID-19. *arXiv [cs.CY]* (2023).
4. Lipsitch, M. *et al.* Improving the evidence base for decision making during a pandemic: the example of 2009 influenza A/H1N1. *Biosecur. Bioterror.* **9**, 89–115 (2011).
5. https://ourworldindata.org/covid-cases.
6. UKHSA data dashboard. https://ukhsa-dashboard.data.gov.uk/.
7. Flaxman, S. *et al.* Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).
8. Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.* **178**, 1505–1512 (2013).
9. Parag, K. V., Thompson, R. N. & Donnelly, C. A. Are epidemic growth rates more informative than reproduction numbers? *J. R. Stat. Soc. Ser. A Stat. Soc.* (2022) doi:10.1111/rssa.12867.
10. Abbott, S. *et al.* Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Res.* **5**, 112 (2020).
11. Charniga, K. *et al.* Updating reproduction number estimates for mpox in the Democratic Republic of Congo using surveillance data. *Am. J. Trop. Med. Hyg.* **110**, 561–568 (2024).
12. Charniga, K. *et al.* Nowcasting and forecasting the 2022 U.S. mpox outbreak: Support for public health decision making and lessons learned. *Epidemics* **47**, 100755 (2024).

13. Günther, F., Bender, A., Katz, K., Küchenhoff, H. & Höhle, M. Nowcasting the COVID-19 pandemic in Bavaria. *Biom. J.* **63**, 490–502 (2021).
14. Reich, N. G. *et al.* A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 3146–3154 (2019).
15. Rahmandad, H., Lim, T. Y. & Sterman, J. Behavioral dynamics of COVID-19: estimating underreporting, multiple waves, and adherence fatigue across 92 nations. *Syst. Dyn. Rev.* **37**, 5–31 (2021).
16. Tsang, T. K. *et al.* 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–e296 (2020).
17. Huisman, J. S. *et al.* Wastewater-based estimation of the effective reproductive number of SARS-CoV-2. *Environ. Health Perspect.* **130**, 57011 (2022).
18. Tisza, M. J. *et al.* Virome sequencing identifies H5N1 avian influenza in wastewater from nine cities. *bioRxiv* 2024.05.10.24307179 (2024) doi:10.1101/2024.05.10.24307179.
19. Stolerman, L. M. *et al.* Using digital traces to build prospective and real-time county-level early warning systems to anticipate COVID-19 outbreaks in the United States. *Sci. Adv.* **9**, (2023).
20. Kendall, M. *et al.* Drivers of epidemic dynamics in real time from daily digital COVID-19 measurements. *Science* **385**, (2024).
21. Hay, J. A. *et al.* Estimating epidemiologic dynamics from cross-sectional viral load distributions. *Science* **373**, eabh0635 (2021).
22. Walker, A. S. *et al.* CT threshold values, a proxy for viral load in community sars-cov-2 cases, demonstrate wide variation across populations and over time. *Elife* **10**, (2021).
23. Penney, J., Jung, A., Koethe, B. & Doron, S. Cycle threshold values and SARS-CoV-2: Relationship to demographic characteristics and disease severity. *J. Med. Virol.* **94**, 3978–3981 (2022).

24. Sala, E. *et al.* Systematic review on the correlation between SARS-CoV-2 real-time PCR cycle threshold values and epidemiological trends. *Infect. Dis. Ther.* **12**, 749–775 (2023).
25. Kissler, S. M. *et al.* Viral dynamics of acute SARS-CoV-2 infection and applications to diagnostic and public health strategies. *PLoS Biol.* **19**, e3001333 (2021).
26. Tsang, T. K. *et al.* Influenza A Virus Shedding and Infectivity in Households. *J. Infect. Dis.* **212**, 1420–1428 (2015).
27. Brint, M. E. *et al.* Prolonged viral replication and longitudinal viral dynamic differences among respiratory syncytial virus infected infants. *Pediatr. Res.* **82**, 872–880 (2017).
28. Harrison, R. E. *et al.* Cycle Threshold Values as Indication of Increasing SARS-CoV-2 New Variants, England, 2020-2022. *Emerg. Infect. Dis.* **29**, 2024–2031 (2023).
29. Musalkova, D. *et al.* Trends in SARS-CoV-2 cycle threshold values in the Czech Republic from April 2020 to April 2022. *Sci. Rep.* **13**, 6156 (2023).
30. Lin, Y. *et al.* Incorporating temporal distribution of population-level viral load enables real-time estimation of COVID-19 transmission. *Nat. Commun.* **13**, (2022).
31. Aguilar Ticona, J. P. *et al.* Extensive transmission of SARS-CoV-2 BQ.1* variant in a population with high levels of hybrid immunity: A prevalence survey. *Int. J. Infect. Dis.* **139**, 159–167 (2024).
32. Andriamandimby, S. F. *et al.* Cross-sectional cycle threshold values reflect epidemic dynamics of COVID-19 in Madagascar. *Epidemics* **38**, 100533 (2022).
33. Alizon, S. *et al.* Epidemiological and clinical insights from SARS-CoV-2 RT-PCR crossing threshold values, France, January to November 2020. *Euro Surveill.* **27**, (2022).
34. Yin, N. *et al.* Leveraging of SARS-CoV-2 PCR cycle thresholds values to forecast COVID-19 trends. *Front. Med. (Lausanne)* **8**, (2021).
35. Khalil, A. *et al.* Weekly Nowcasting of New COVID-19 Cases Using Past Viral Load Measurements. *Viruses* **14**, (2022).

655 36. Sharmin, M. *et al.* Cross-sectional Ct distributions from qPCR tests can provide an early
656 warning signal for the spread of COVID-19 in communities. *Front. Public Health* **11**, 1185720
657 (2023).
658 37. Ahuja, V., Bowe, T., Warnock, G., Pitman, C. & Dwyer, D. E. Trends in SARS-CoV-2 cycle
659 threshold (Ct) values from nucleic acid testing predict the trajectory of COVID-19 waves.
660 *Pathology* **56**, 710–716 (2024).
661 38. Moro, A. *et al.* Trends in SARS-CoV-2 cycle threshold values in Bosnia and Herzegovina—A
662 retrospective study. *Microorganisms* **12**, 1585 (2024).
663 39. Kissler, S. M. *et al.* Viral dynamics of SARS-CoV-2 variants in vaccinated and unvaccinated
664 persons. *N. Engl. J. Med.* **385**, 2489–2491 (2021).
665 40. Russell, T. W. *et al.* Combined analyses of within-host SARS-CoV-2 viral kinetics and
666 information on past exposures to the virus in a human cohort identifies intrinsic differences
667 of Omicron and Delta variants. *PLoS Biol.* **22**, e3002463 (2024).
668 41. Fryer, H. R. *et al.* Viral burden is associated with age, vaccination, and viral variant in a
669 population-representative study of SARS-CoV-2 that accounts for time-since-infection-
670 related sampling bias. *PLoS Pathog.* **19**, e1011461 (2023).
671 42. Hay, J. A., Kennedy-Shaffer, L. & Mina, M. J. Viral loads observed under competing strain
672 dynamics. *medRxiv* 2021.07.27.21261224 (2021).
673 43. Jones, T. C. *et al.* Estimating infectiousness throughout SARS-CoV-2 infection course
674 Downloaded from. *Science* (2021) doi:10.1126/science.abi5273.
675 44. Wyllie, A. L. *et al.* Saliva or nasopharyngeal swab specimens for detection of SARS-CoV-2.
676 *N. Engl. J. Med.* **383**, 1283–1286 (2020).
677 45. Rhoads, D. *et al.* College of American pathologists (CAP) microbiology committee
678 perspective: Caution must be used in interpreting the cycle threshold (ct) value. *Clin. Infect.*
679 *Dis.* **72**, e685–e686 (2021).

680 46. Arnaout, R. *et al.* The limit of detection matters: The case for benchmarking severe acute
681 respiratory syndrome Coronavirus 2 testing. *Clin. Infect. Dis.* **73**, e3042–e3046 (2021).
682 47. Cuypers, L. *et al.* Nationwide harmonization effort for semi-quantitative reporting of SARS-
683 CoV-2 PCR test results in Belgium. *Viruses* **14**, 1294 (2022).
684 48. Hay, J. A. *et al.* Quantifying the impact of immune history and variant on SARS-CoV-2 viral
685 kinetics and infection rebound: A retrospective cohort study. *Elife* **11**, (2022).
686 49. Elliott, P. *et al.* Rapid increase in Omicron infections in England during December 2021:
687 REACT-1 study. *Science* **375**, 1406–1411 (2022).
688 50. Brainard, J. *et al.* Comparison of surveillance systems for monitoring COVID-19 in England:
689 a retrospective observational study. *Lancet Public Health* **8**, e850–e858 (2023).
690 51. Mellor, J. *et al.* Understanding the leading indicators of hospital admissions from COVID-19
691 across successive waves in the UK. *Epidemiol. Infect.* **151**, e172 (2023).
692 52. https://www.mass.gov/info-details/covid-19-reporting.
693 53. http://dashboard.publichealth.lacounty.gov/covid19_surveillance_dashboard/.
694 54. Mgcv: Mixed GAM computation vehicle with automatic smoothness estimation.
695 *Comprehensive R Archive Network (CRAN)* https://cran.r-
696 project.org/web/packages/mgcv/index.html.
697 55. Singanayagam, A. *et al.* Community transmission and viral load kinetics of the SARS-CoV-2
698 delta (B.1.617.2) variant in vaccinated and unvaccinated individuals in the UK: a prospective,
699 longitudinal, cohort study. *Lancet Infect. Dis.* **22**, 183–195 (2022).
700 56. lazymcmc. Preprint at https://github.com/jameshay218/lazymcmc.

701