Supplementary information

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

Tse Yang Lim*, Sanjat Kanjilal, Shira Doron, Jessica Penney, Meredith Haddix, Tae Hee Koo, Phoebe Danza, Rebecca Fisher, Yonatan H. Grad†, James A. Hay*†

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

†These authors jointly supervised the work.

Table S1. Factors confounding Ct value inference

Factor	Explanation	Scenario				
		1	2	3	4	5
Biological factors						
Inter-individual variation in viral kinetics	Individuals' viral loads and viral load trajectories vary substantially, even accounting for immune history, demographic factors, etc.	0	0	X	0	X
Symmetry of viral load trajectory	While the viral growth phase is generally much shorter than the clearance phase, any single viral load measurement could come from either phase[41].	0	X	0	0	X
Impact of immune history on viral kinetics	Past exposure through previous infection, vaccination, or both may result in faster viral clearance[40,48].	0	0	0	0	0
Impact of demographic factors on viral kinetics	Older individuals generally have higher viral loads and slower viral clearance than younger ones[43,48].	0	0	0	0	0
Impact of viral variant	Different SARS-CoV-2 variants may be associated with different viral load trajectories[40,48].	0	0	0	0	0
Logistical factors						
PCR platform / assay	Ct values are not typically standardized across different PCR platforms and assays[46]; differences in individual testing protocol (e.g. location swabbed) could contribute further differences in measured viral loads.	0	0	0	0	0
Testing behavior and sampling regime or delay distribution	Cross-sectional Ct value distributions reflect the convolution of the distribution of true infection ages and the sampling delay distribution; if sampling delays are highly clustered (e.g. mostly 3-5 days after infection), observed Ct distributions will reflect primarily individual-level random variation rather than informative variation in infection ages[21]. Sampling regime (e.g. representative random sampling, contact-tracing based sampling, voluntary testing, hospital outpatient screening) would influence the sampling delay distribution – random sampling theoretically results in a uniform delay distribution, while e.g. symptom-driven voluntary testing results in highly clustered sampling delays.	0	0	0	X	X
Synthetic data scenarios are numbered as follows: 1) Ideal condition, 2) realistic asymmetry in viral kinetics, 3) moderate individual-level variation, 4) clustered sampling delay distribution, 5) realistic baseline						

Table S2. Comparison of the 24 different models used to model the relationship between daily reported Ct value statistics and epidemic growth rates, fitted to the realistic baseline synthetic dataset. We tested various combinations of three Ct-value statistics (mean, standard deviation and skewness), incorporation of a variant 'era' interaction term (intercept only or intercept and coefficient), and different functional forms of the model (log-linear regression or cubic regression splines).

			RMSE			AUC		
Spline	**Model**	**AIC**	**In-s.**	**Now.**	**Inf.**	**In-s.**	**Now.**	**Inf.**
None	Mean (no variant era)	-4151	0.034	0.0349	0.0434	0.743	0.701	0.66
None	Mean + variant era	-4175	0.0335	0.0374	0.0471	0.752	0.667	0.636
None	Mean * variant era	-4230	0.0326	0.0383	0.0507	0.763	0.694	0.602
None	Mean + st.dev. (no variant era)	-4159	0.0338	0.0354	0.0442	0.75	0.696	0.651
None	Mean + st.dev. + variant era	-4186	0.0333	0.0379	0.0479	0.763	0.666	0.647
None	(Mean + st.dev.) * variant era	-4256	0.032	0.0396	0.0525	0.78	0.686	0.605
None	Mean + skew (no variant era)	-4149	0.034	0.0351	0.0434	0.743	0.695	0.66
None	Mean + skew + variant era	-4174	0.0335	0.0376	0.0472	0.752	0.66	0.638
None	(Mean + skew) * variant era	-4227	0.0325	0.0395	0.0519	0.764	0.683	0.604
None	Mean + s.d. + skew (no variant era)	-4157	0.0338	0.0358	0.0446	0.751	0.689	0.654
None	Mean + s.d. + skew + variant era	-4184	0.0333	0.0383	0.0482	0.763	0.659	0.646
None	(Mean + s.d. + skew) * variant era	-4255	0.0319	0.0412	0.0534	0.783	0.677	0.604
Cubic	Mean (no variant era)	-4229	0.0326	0.0356	0.0466	0.755	0.715	0.668
Cubic	Mean + variant era	-4258	0.0321	0.0413	0.062	0.78	0.705	0.635
Cubic	Mean * variant era	-4296	0.0313	0.553	1.24	0.796	0.707	0.612
Cubic	Mean + st.dev. (no variant era)	-4238	0.0324	0.036	0.0476	0.763	0.709	0.662
Cubic	Mean + st.dev. + variant era	-4272	0.0318	0.0414	0.0619	0.789	0.703	0.64
Cubic	(Mean + st.dev.) * variant era	-4344	0.0304	0.552	1.23	0.811	0.691	0.649
Cubic	Mean + skew (no variant era)	-4232	0.0325	0.0366	0.0479	0.755	0.706	0.664
Cubic	Mean + skew + variant era	-4261	0.0319	0.042	0.0627	0.78	0.698	0.633
Cubic	(Mean + skew) * variant era	-4339	0.0303	0.554	1.24	0.8	0.682	0.602
Cubic	Mean + s.d. + skew (no variant era)	-4240	0.0323	0.0371	0.0491	0.762	0.705	0.67
Cubic	Mean + s.d. + skew + variant era	-4274	0.0317	0.0422	0.0629	0.789	0.697	0.638
Cubic	(Mean + s.d. + skew) * variant era	-4379	0.0296	0.0545	0.0926	0.814	0.672	0.649
Key: In-s. = in-sample; Now. = nowcast; Inf. = inflection point								

Table S3. Summary characteristics of the SARS-CoV-2 testing datasets.

	MGB	**LAC**	**Tufts**
Sample size	2,671,041 total 161,273 positive 104,534 included	330,034 positive 279,463 included	84,848 total 10,338 positive 10,214 included
Dates	Mar 2020-Jan 2023 (1022 days)	May 2020-Jul 2021, Jan-Sep 2022 (680 days)	Feb 2021-Oct 2022 (496 days)
Testing modality	Hospital outpatient, inpatient, ER	Voluntary outpatient testing	Hospital outpatient, inpatient, ER
Platforms/assays	7 platforms: Broad in-house assay; Cepheid SARS-CoV-2; Cepheid multiplex SARS-CoV-2/influenza/RSV; Hologic Fusion; Roche Cobas SARS-CoV-2; Roche Cobas multiplex SARS-CoV-2/influenza; Roche Liat multiplex SARS-CoV-2/influenza (see Figure S12)	Fulgent Genetics platform using ThermoFisher QuantStudio™ 6 and 7 PCR system, with LOINC 94531-1 (primarily to Nov 2020) and LOINC 94533-7 (primarily after Nov 2020)	Alinity single-plex and Alinity multiplex
Symptom status known?	No	Yes (approx. 55% symptomatic from Sep 2020 onward)	Yes (approx. 65% symptomatic)
Vaccination status known?	No	Yes (approx. 25% of all included results; >70% of 2022 results)	No

Dataset	AIC	BIC	RMSE				Spearman's Rho				95% PrI coverage				AUC			
			In-s.	Now.	Inf.	FS	In-s.	Now.	Inf.	FS	In-s.	Now.	Inf.	FS	In-s.	Now.	Inf.	FS
MGB	-3493	-3433	0.0451	0.0523	0.0645	0.047	0.523	0.398	0.333	0.44	0.949	0.929	0.873	0.938	0.785	0.723	0.722	0.754
LAC	-2660	-2585	0.0335	0.039	0.0471	0.0458	0.649	0.556	0.394	0.573	0.953	0.912	0.837	0.888	0.843	0.784	0.772	0.724
TFT	-1691	-1649	0.0415	0.0497	0.0591	0.0695	0.455	0.266	0.149	0.554	0.944	0.864	0.791	0.801	0.754	0.685	0.584	0.796
Key: In-s. = in-sample; Now. = nowcast; Inf. = inflection point; FS = fixed train-test split																		

Table S4. Summary of the performance of the chosen model in predicting epidemic growth rates using Ct values for MGB, LAC, and Tufts datasets, including in-sample fits, nowcast performance, inflection period performance, and fit over the testing period with a single fixed train-test split. Metrics reported are RMSE of predicted vs. observed log incidence growth rates, Spearman's rank-order correlation coefficient for predicted vs. observed growth rates, proportion of observed growth rates falling within the 95% prediction interval, and AUC for epidemic direction predictions.

Table S5. Summary of model performance metrics for the downsampled MGB and external comparison (Tufts) datasets, for nowcast performance and comparable baseline nowcast performance. Metrics reported are RMSE of predicted vs. observed log incidence growth rates, Spearman's rank-order correlation coefficient for predicted vs. observed growth rates, proportion of observed growth rates falling within the 95% prediction interval, and AUC for epidemic direction predictions.

Dataset	Days included	RMSE		Spearman's Rho		95% Prl coverage		AUC	
		Now.	Comp.	Now.	Comp.	Now.	Comp.	Now.	Comp.
Tufts	413	0.0497	0.0489	0.266	0.348	0.864	0.937	0.685	0.719
10% downsample	582.54	0.0564	0.0525	0.137	0.321	0.927	0.932	0.615	0.714
25% downsample	860.02	0.0494	0.0492	0.323	0.389	0.939	0.942	0.698	0.729
50% downsample	909.84	0.0501	0.0516	0.373	0.4	0.936	0.936	0.718	0.729
75% downsample	927.64	0.0506	0.0518	0.399	0.41	0.935	0.935	0.736	0.731
25/day max samples	944	0.0525	0.0523	0.39	0.398	0.93	0.931	0.721	0.723
50/day max samples	944	0.0502	0.0523	0.375	0.398	0.935	0.931	0.714	0.723
100/day max samples	944	0.0515	0.0523	0.384	0.398	0.932	0.931	0.718	0.723
2.5% trimmed	930	0.0502	0.0516	0.431	0.416	0.933	0.935	0.75	0.73
5% trimmed	930	0.0496	0.0516	0.436	0.416	0.938	0.935	0.748	0.73
10% trimmed	921	0.0496	0.0517	0.397	0.413	0.933	0.936	0.726	0.735
Key: Now. = nowcast; Comp. = comparison data									

	Model	AIC	BIC	RMSE				Spearman's Rho				95% PrI coverage				AUC			
				In-s.	Now.	Inf.	FS	In-s.	Now.	Inf.	FS	In-s.	Now.	Inf.	FS	In-s.	Now.	Inf.	FS
MGB	Base model	-3493	-3433	0.0451	0.0523	0.0645	0.047	0.523	0.398	0.333	0.44	0.949	0.929	0.873	0.933	0.785	0.723	0.722	0.754
	Outpatient only	-3432	-3358	0.0439	0.0494	0.0653	0.041	0.546	0.336	0.234	0.462	0.947	0.941	0.903	0.95	0.804	0.724	0.696	0.763
LAC	Base model	-2660	-2585	0.0335	0.039	0.0471	0.0458	0.649	0.556	0.394	0.573	0.954	0.912	0.837	0.884	0.843	0.784	0.772	0.724
	Symptom stratified	-2005	-1914	0.0326	0.0454	0.0507	0.0648	0.728	0.45	0.369	0.313	0.949	0.846	0.744	0.577	0.907	0.829	0.765	0.5
	Asymptomatic only	-1986	-1916	0.0335	0.0415	0.0497	0.0487	0.693	0.55	0.399	0.535	0.932	0.866	0.8	0.856	0.88	0.809	0.784	0.5
	Immunologically naive only	-2556	-2500	0.0344	0.0401	0.0474	0.055	0.62	0.489	0.42	0.31	0.944	0.896	0.821	0.776	0.819	0.761	0.754	0.475

Key: In-s. = in-sample; Now. = nowcast; Inf. = inflection point; FS = fixed train-test split

Table S6. Summary of sensitivity analyses comparing performance of the base model in predicting epidemic growth rates using Ct values for LAC against models using only asymptomatic / unknown symptom status individuals, and using only immunologically naïve (no known vaccination or previous SARS-CoV-2 infection) individuals, including in-sample fits, nowcast performance, inflection period performance, and fit over the testing period with a single fixed train-test split. Metrics reported are RMSE of predicted vs. observed log incidence growth rates, Spearman's rank-order correlation coefficient for predicted vs. observed growth rates, proportion of observed growth rates falling within the 95% prediction interval, and AUC for epidemic direction predictions.

Table S7. Viral kinetics model parameters, descriptions and values used in each of the 5 synthetic datasets. Cells are shaded grey where assumed values differ to estimated values. Values in "Ideal condition" are the maximum posterior probability estimates from fitting the model. Bottom row shows the sampling delay distribution used for each scenario.

Parameter	Description	Ideal condition	Realistic kinetics	Clustered sampling	Realistic variation	Baseline condition (estimated value)
t_p	Days to peak viral load	1.00 days	2.56 days	1.00 days	1.00 days	2.56 days
c_p	Minimum Ct value	15.0	25.0	15.0	15.0	25.0
c_s	Ct value at inflection point	31.6	31.6	31.6	31.6	31.6
t_s	Days from peak to inflection point	14.0 days	8.41 days	14.0 days	14.0 days	8.41 days
t_c	Days from inflection point to full clearance	45.6 days	45.6 days	45.6 days	45.6 days	45.6 days
c_0	Baseline Ct value at time of infection	40.0	40.0	40.0	40.0	40.0
σ_{obs}	Unmodified variance of observed Ct values for a given time-since-infection	2.15	2.15	2.15	4.29	4.29
t_m	Days to reach minimum variance in observed Ct values	12.1 days	12.1 days	12.1 days	12.1 days	12.1 days
s_m	Proportion reduction on variance at minimum (observation variance decreases from σ_{obs} to $s_m\sigma_{obs}$)	0.622	0.622	0.622	0.622	0.622
p_c	Daily probability of full clearance	0.203	0.203	0.203	0.203	0.203
Sampling delay distribution		Uniform(0,7)	Uniform(0,7)	Gamma(2,2)	Uniform(0,7)	Gamma(2,2)

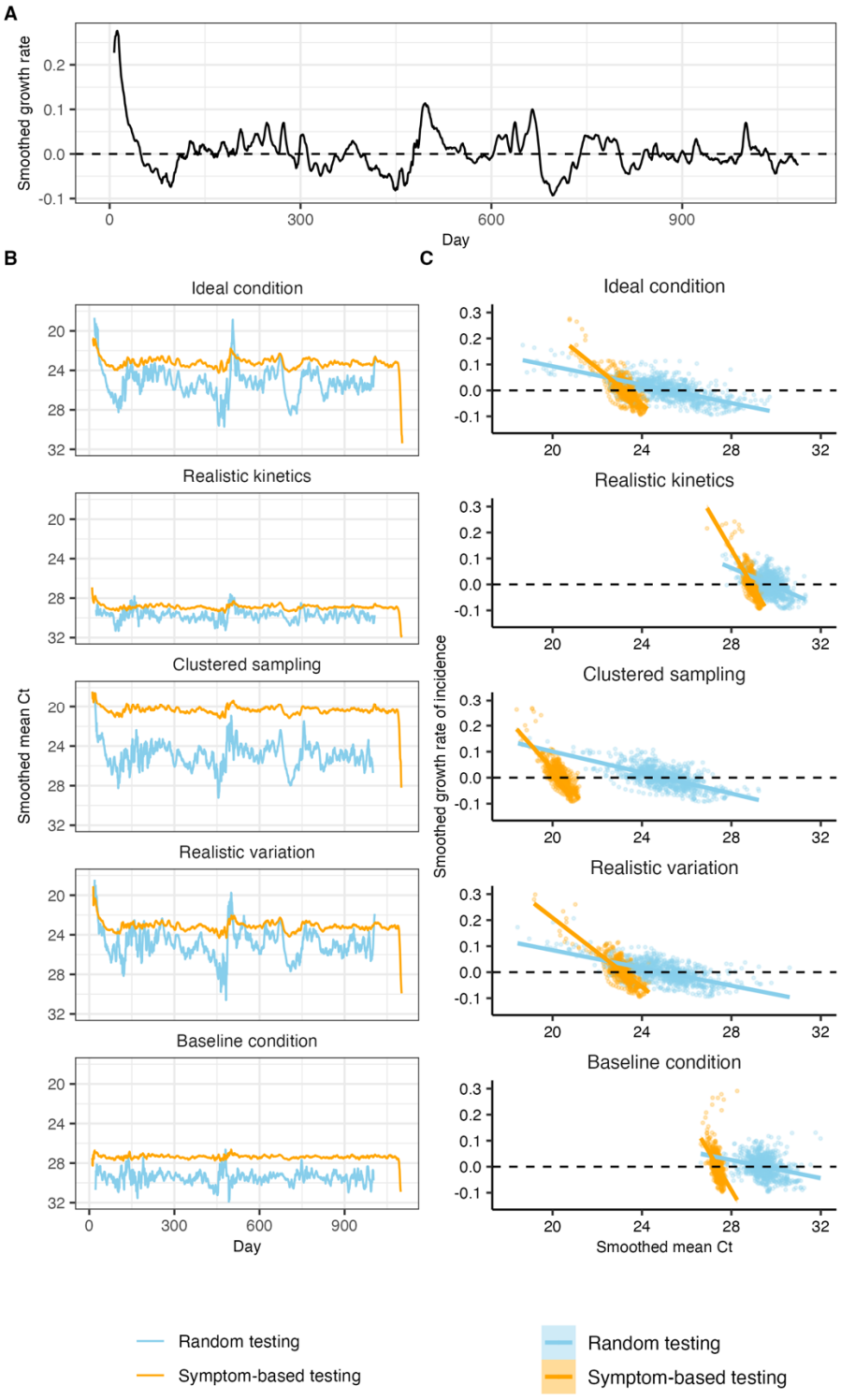


Figure S1. (A) 7-day rolling average growth rate of infections used for the simulations. **(B)** Outputs of each synthetic dataset scenario showing the 7-day average mean Ct value reported through the two surveillance strategies. **(C)** Scatterplot showing the relationship between 7-day rolling average growth rates and 7-day rolling average mean Ct value for each scenario, stratified by surveillance strategy. Solid lines show fitted linear regression models.

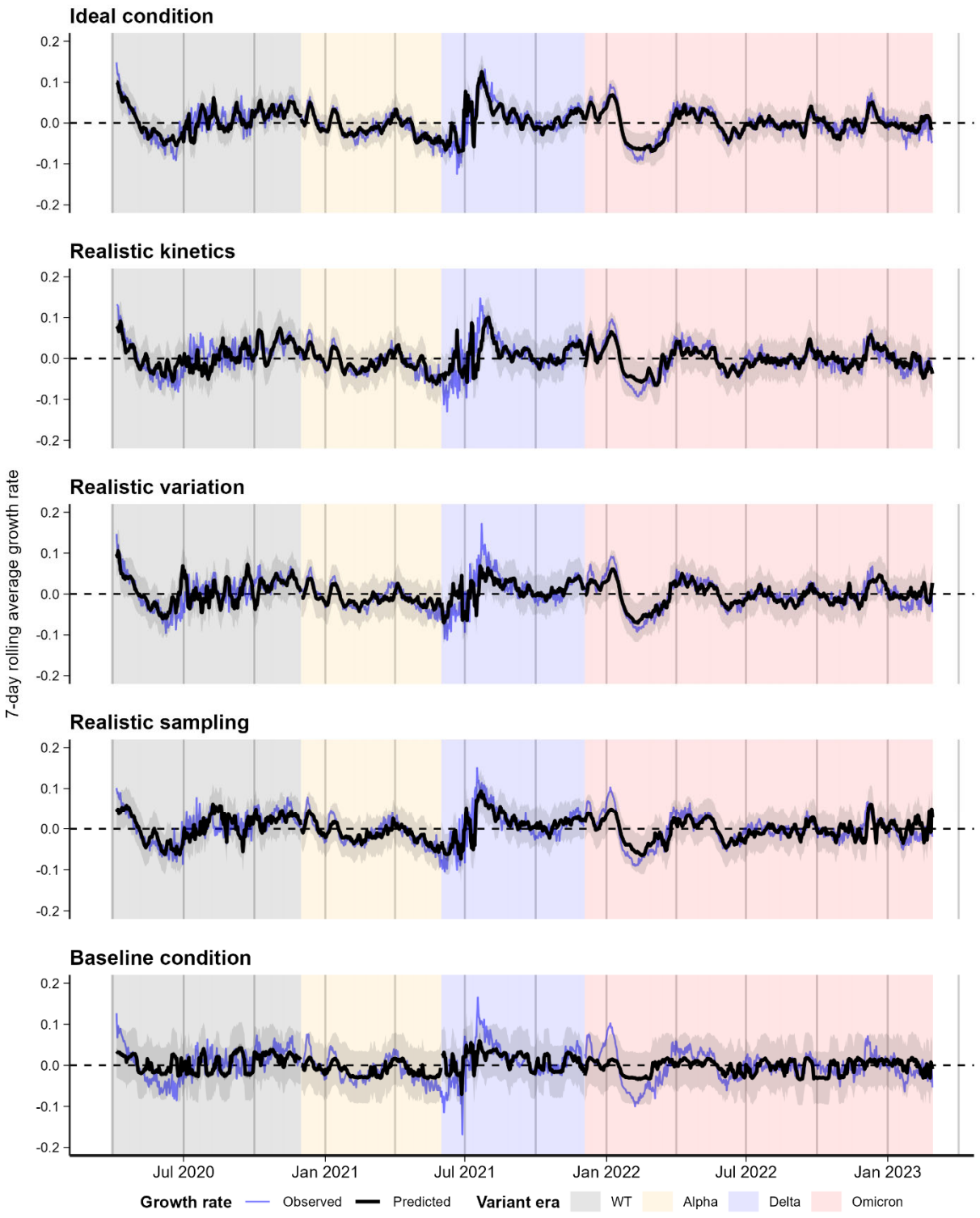


Figure S2. In-sample fits of the best-performing GAM model predicting epidemic growth rates over time using only reported Ct values using the 5 synthetic datasets. Blue line shows true growth rate of infections used for the simulation. Black lines and shaded region show model-predicted growth rates and 95% confidence (dark shading) / prediction (light shading) intervals.

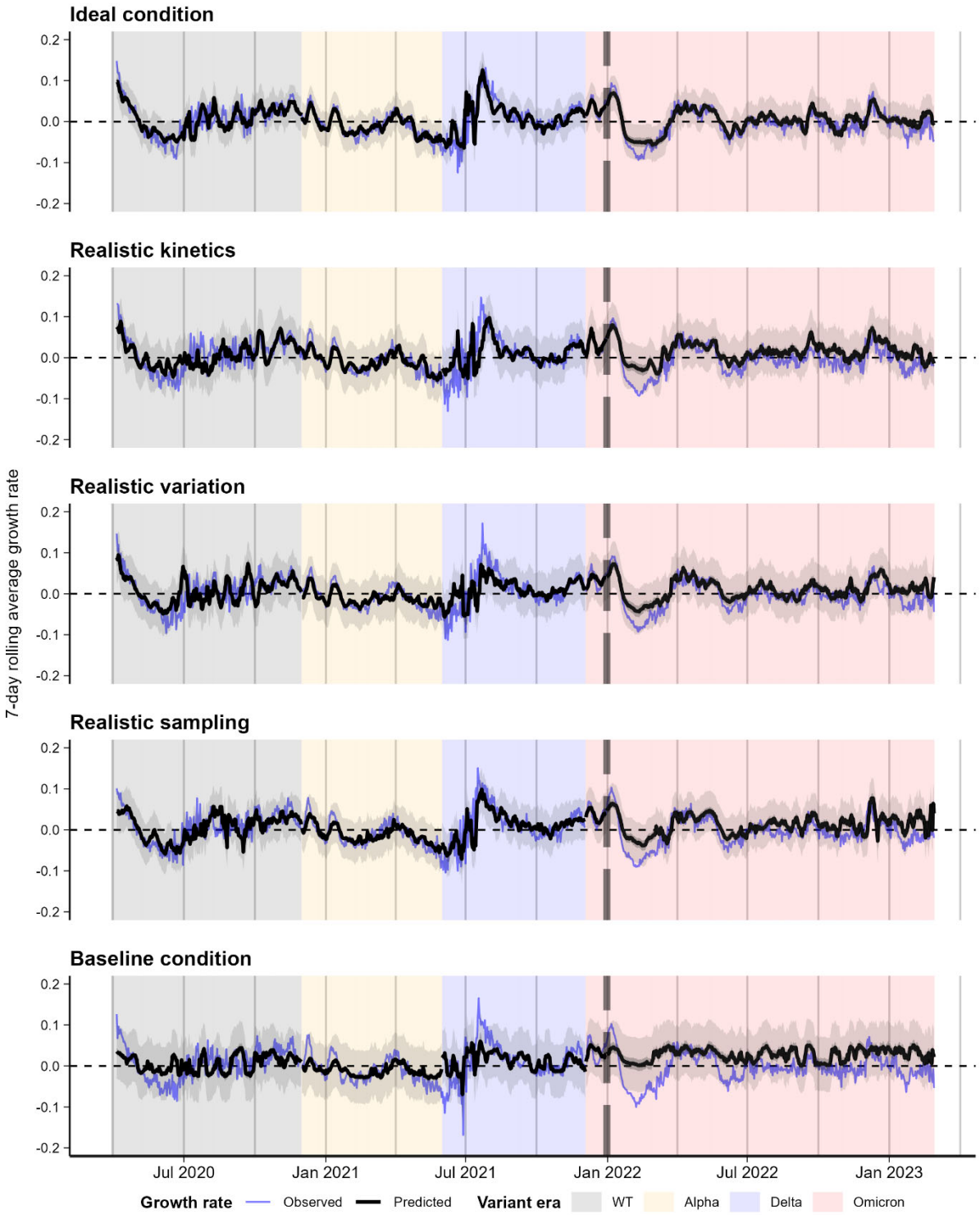


Figure S3. Training dataset fits (up to vertical dashed line) and test dataset predicted epidemic growth rates over time using only reported Ct values using the 5 synthetic datasets. Results shown are from the best-performing GAM model. Blue line shows true growth rate of infections used for the simulation. Black lines and shaded region show model-predicted growth rates and 95% confidence (dark shading) / prediction (light shading) intervals.

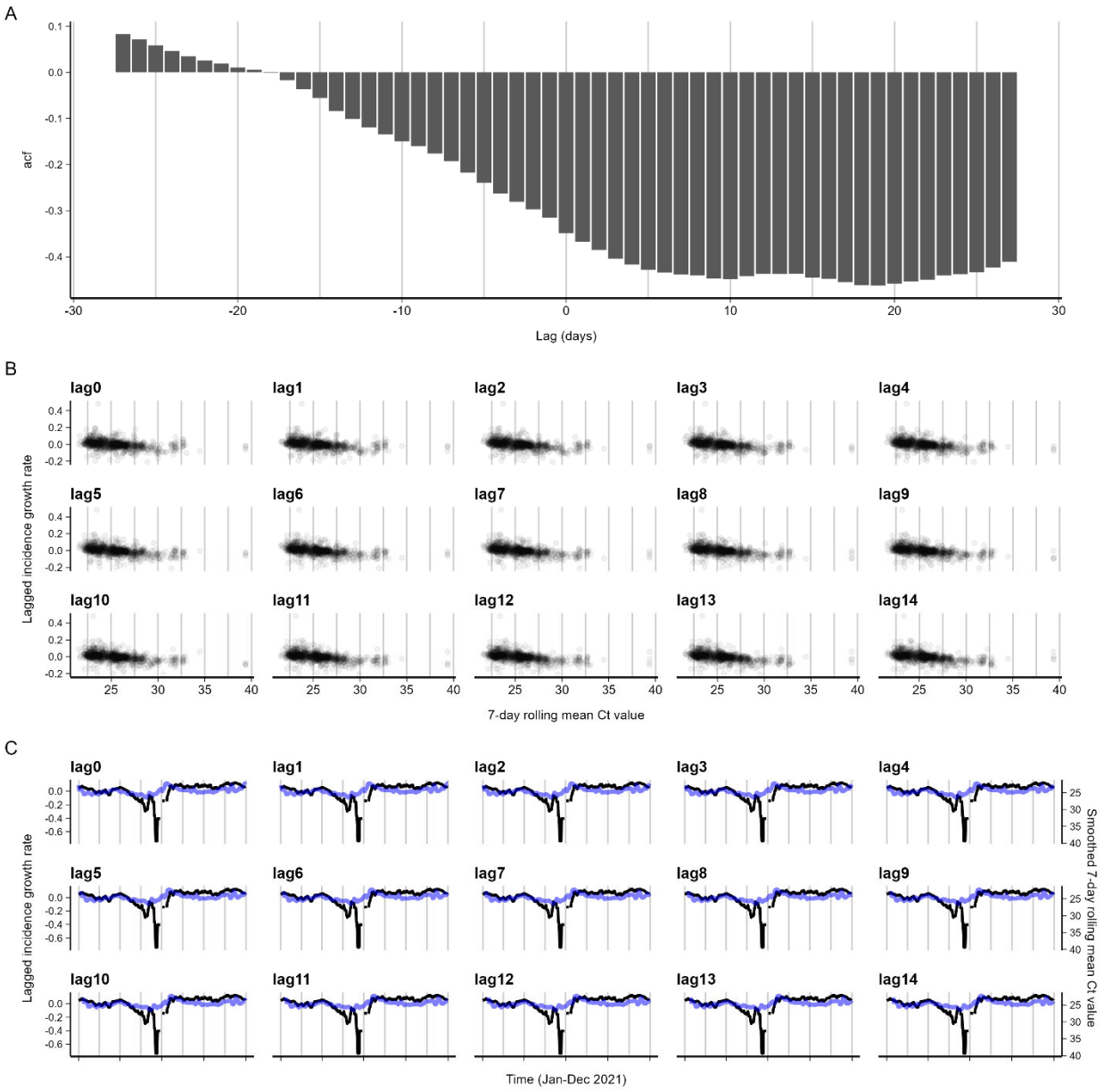


Figure S4. Cross-correlations between log incidence growth rates and mean Ct values at different lead/lag times, for MGB data. **(A)** correlation strength, **(B)** scatterplots of daily observed growth rate and mean Ct value at different lags, **(C)** growth rate and mean Ct value over time with different time shifts for the Ct value curve.

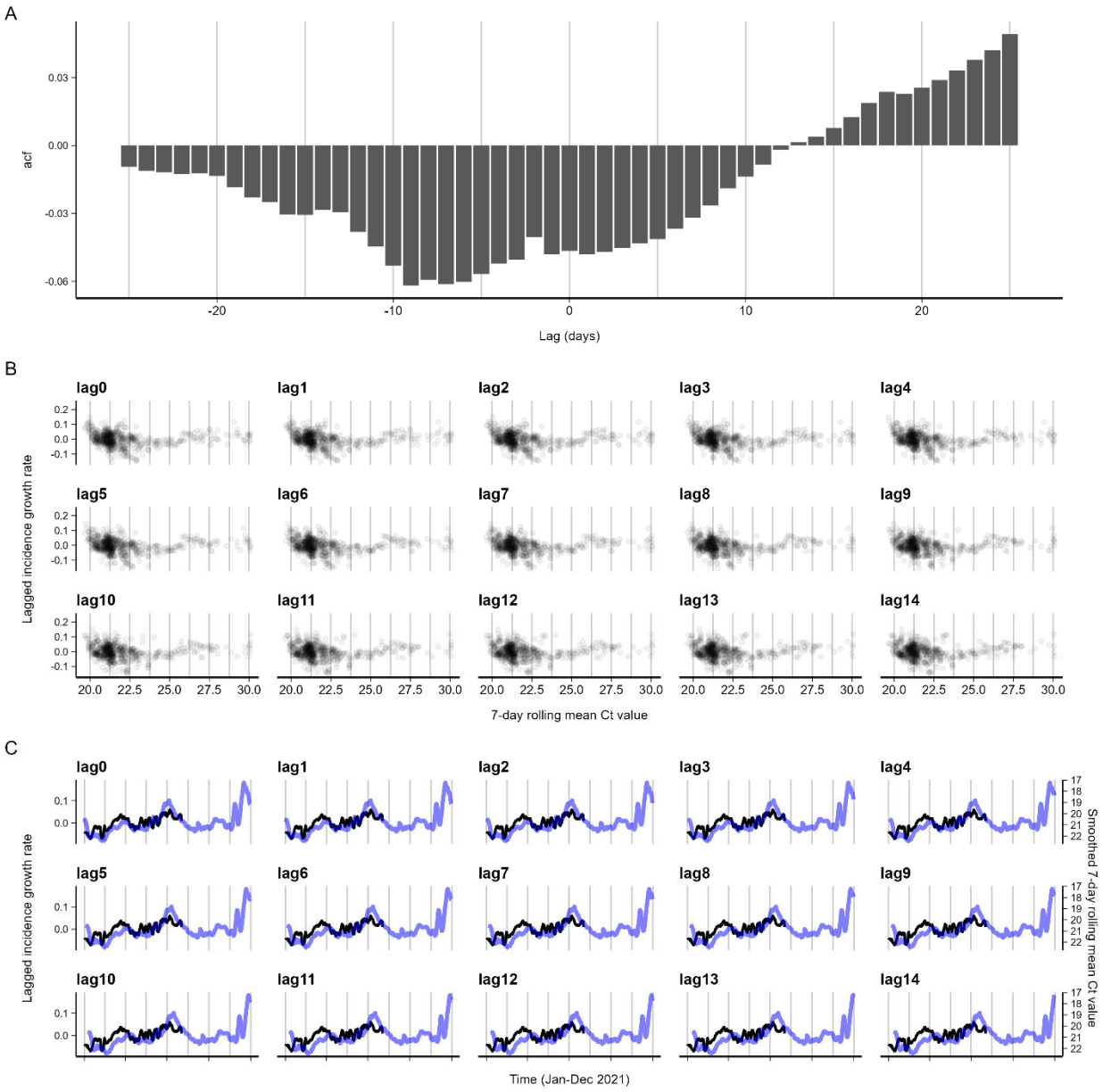


Figure S5. Cross-correlations between log incidence growth rates and mean Ct values at different lead/lag times, for LAC data. **(A)** correlation strength, **(B)** scatterplots of daily observed growth rate and mean Ct value, **(C)** growth rate and mean Ct value over time with different time shifts for the Ct value curve.

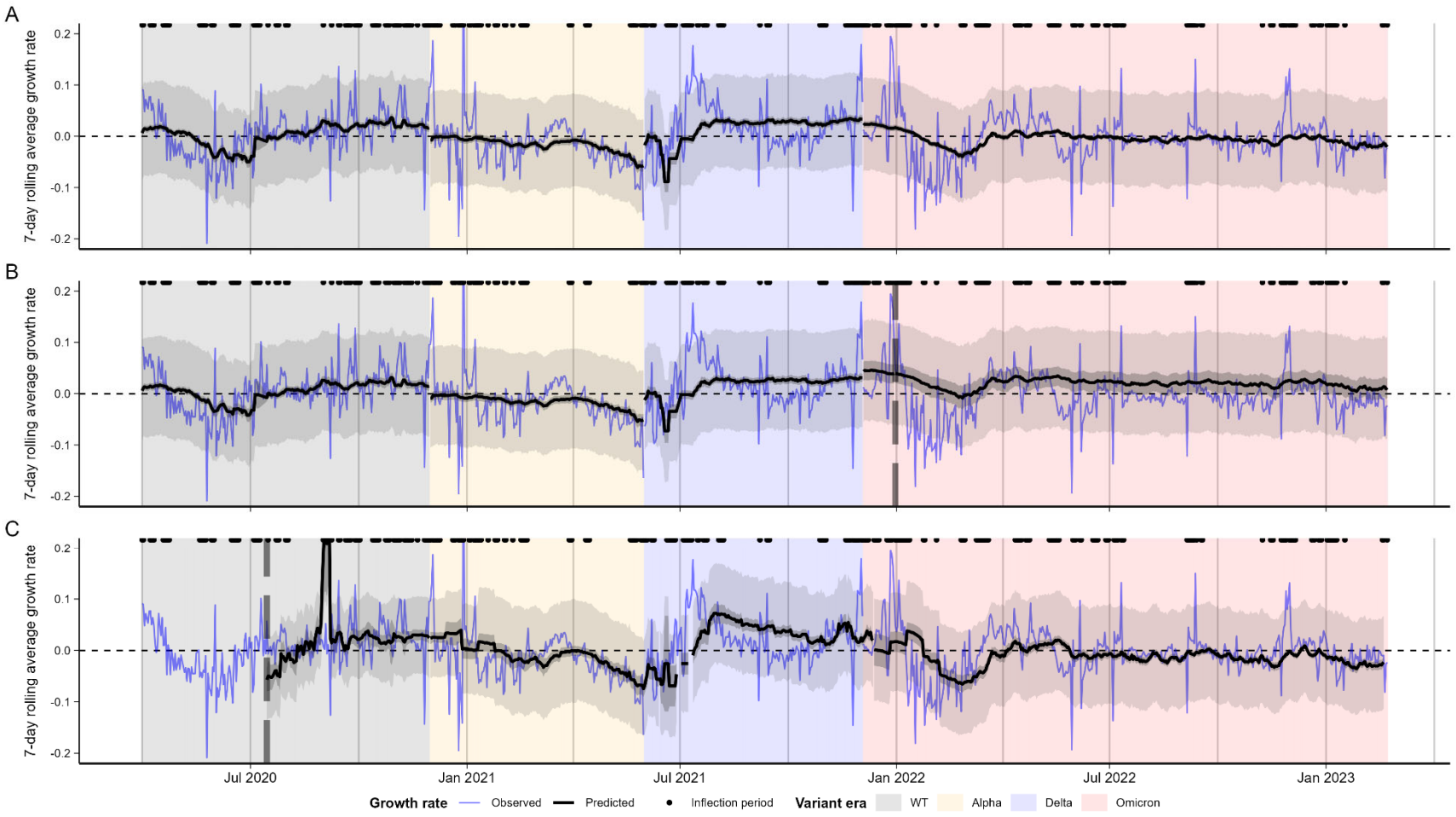


Figure S6. Model-predicted vs. observed log incidence growth rate for models fitted using the MGB data, showing **(A)** in-sample fits , **(B)** fit over the testing period with a single fixed train-test split shown by the vertical dashed line, and **(C)** two-week rolling nowcast fits, starting at the vertical dashed line.

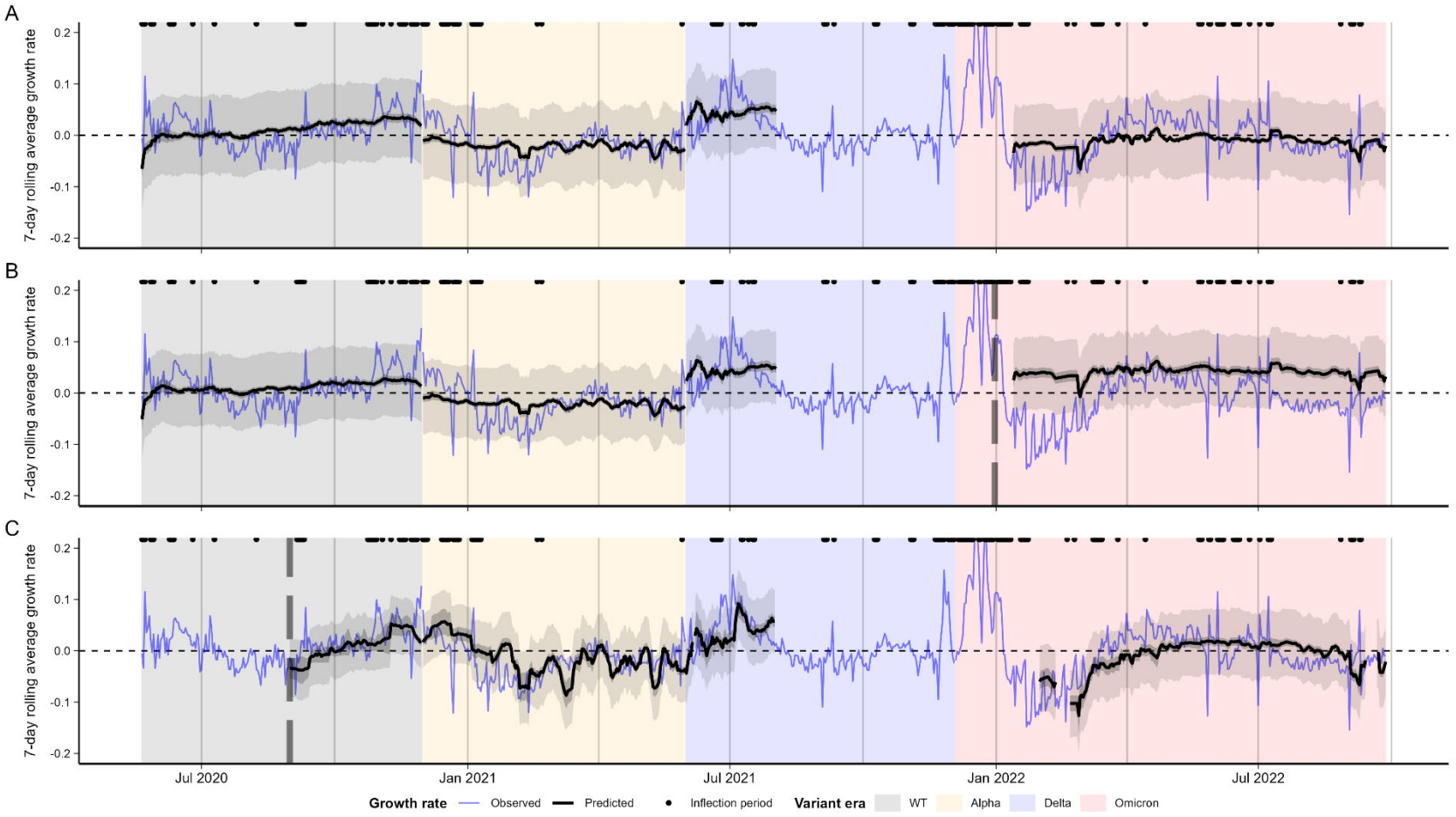


Figure S7. Model-predicted vs. observed log incidence growth rate for models fitted using the LAC data, showing **(A)** in-sample fits, **(B)** fit over the testing period with a single fixed train-test split shown by the vertical dashed line, and **(C)** two-week rolling nowcast fits starting at the dashed vertical line.

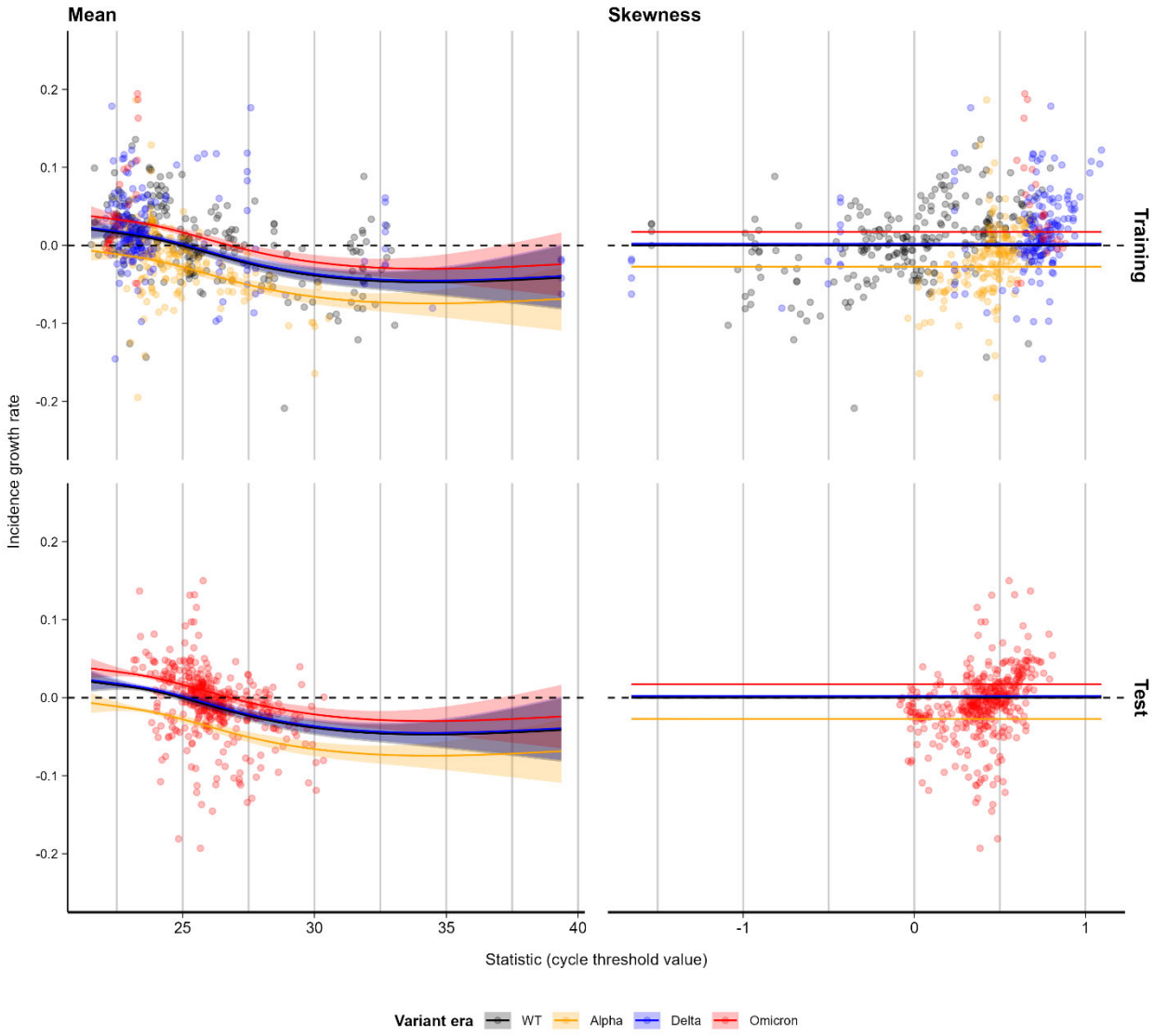


Figure S8. Best-fit lines (with standard errors) showing the modeled relationship between mean (left) and skewness (right) in Ct values, by variant era, against growth rate for MGB data. Upper panels show fit to training data with a fixed cutoff at 31 Dec 2021, while lower panels show the fit of the trained model to data in the subsequent test period (01 Jan 2022 onward).

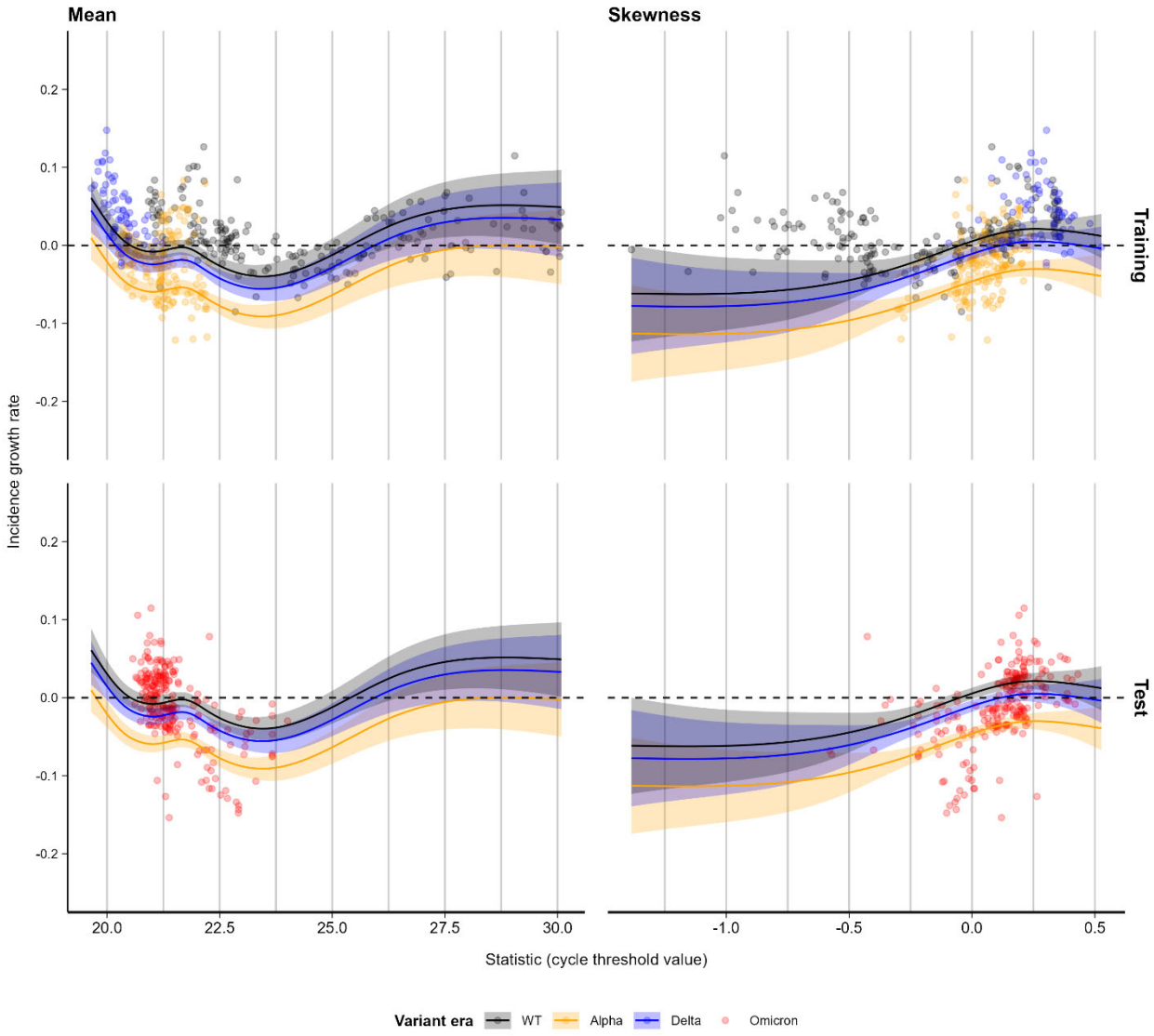


Figure S9. Best-fit lines (with standard errors) showing the modeled relationship between mean (left) and skewness (right) in Ct values, by variant era, against growth rate for LAC data. Upper panels show fit to training data with a fixed cutoff at 31 Dec 2021, while lower panels show the fit of the trained model to data in the subsequent test period (01 Jan 2022 onward). Note the absence of Omicron-era data from the training period.

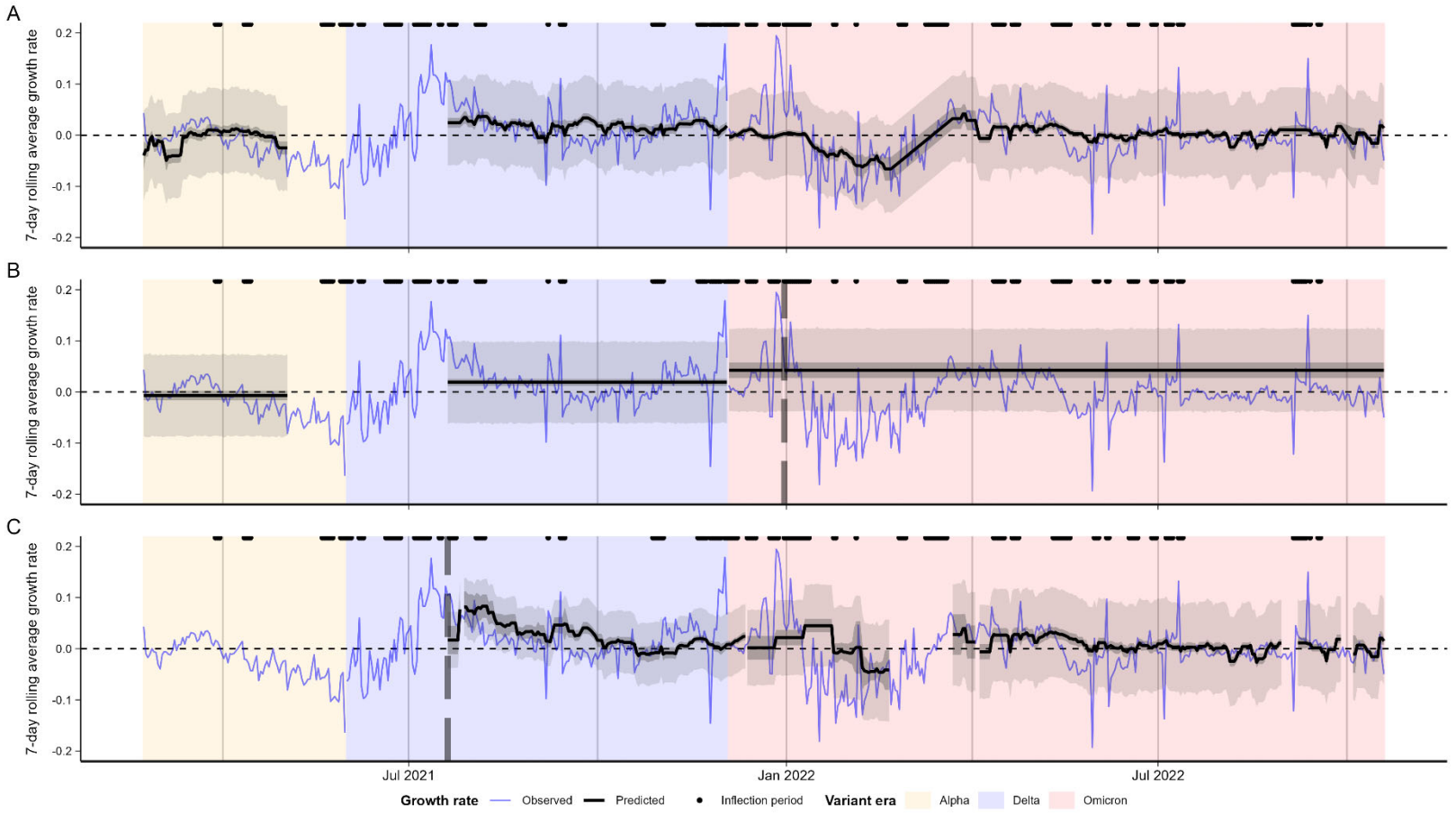


Figure S10. Model-predicted vs. observed log incidence growth rate for models fitted using the Tufts data, showing **(A)** in-sample fits, **(B)** fit over the testing period with a single fixed train-test split shown by the vertical dashed line, and **(C)** two-week rolling nowcast fits starting at the dashed vertical line.

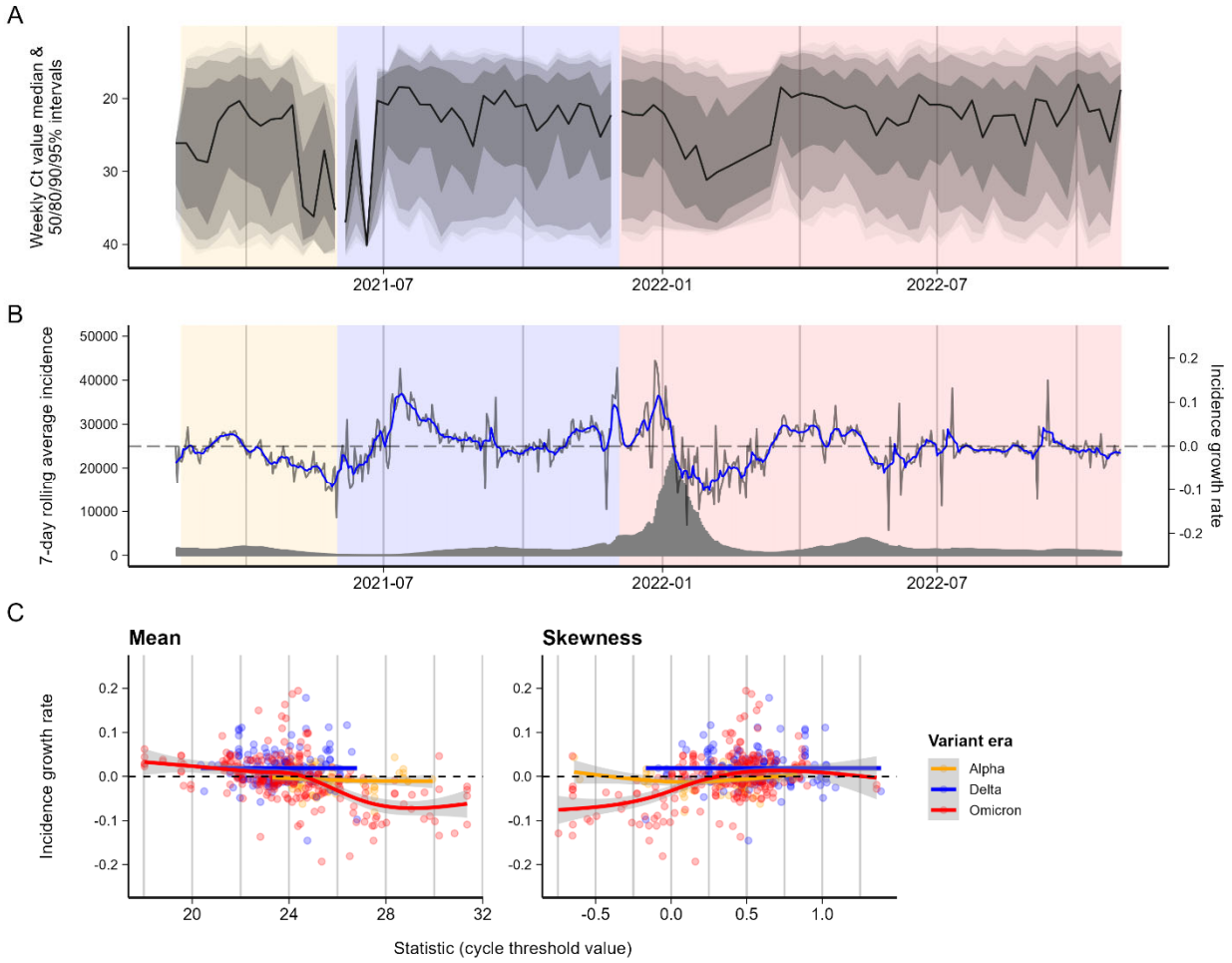


Figure S11. (**A**) Ct value distributions by week for the Tufts dataset. Solid line shows mean, shaded ribbons show 50% and 95% quantiles. (**B**) Incidence of COVID-19 cases and corresponding epidemic rates for Massachusetts, USA. Grey line shows growth rate of cases, whereas blue line shows 7-day rolling mean growth rate. (**C**) Relationship between Ct value means and skewness against epidemic growth rates.

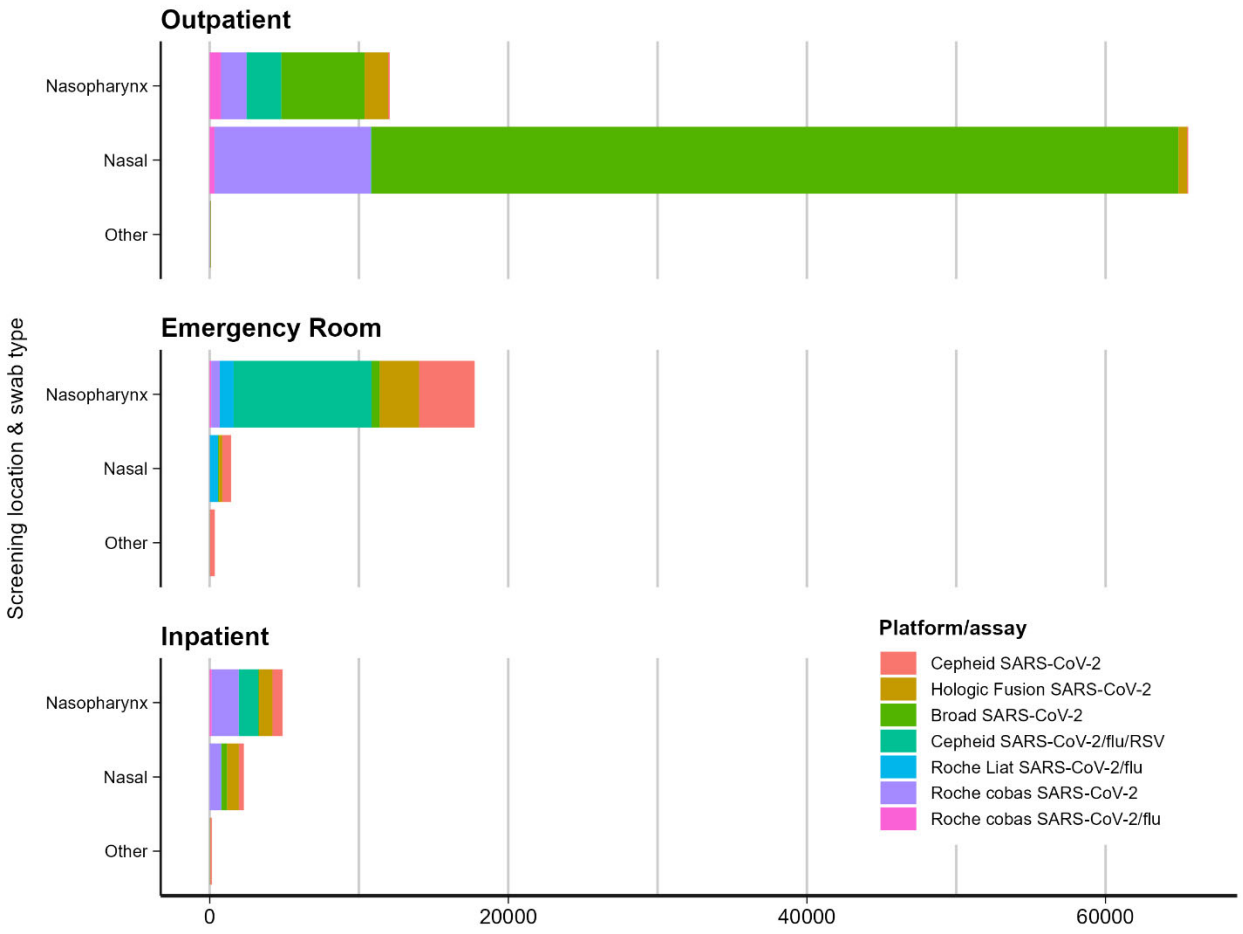


Figure S12. Distribution of test results included in the MGB dataset, broken down by screening location (outpatient pre-procedural screening, ER testing, inpatient testing), swab type (nasopharyngeal vs. nasal vs. other), and PCR platform / assay used for analysis.

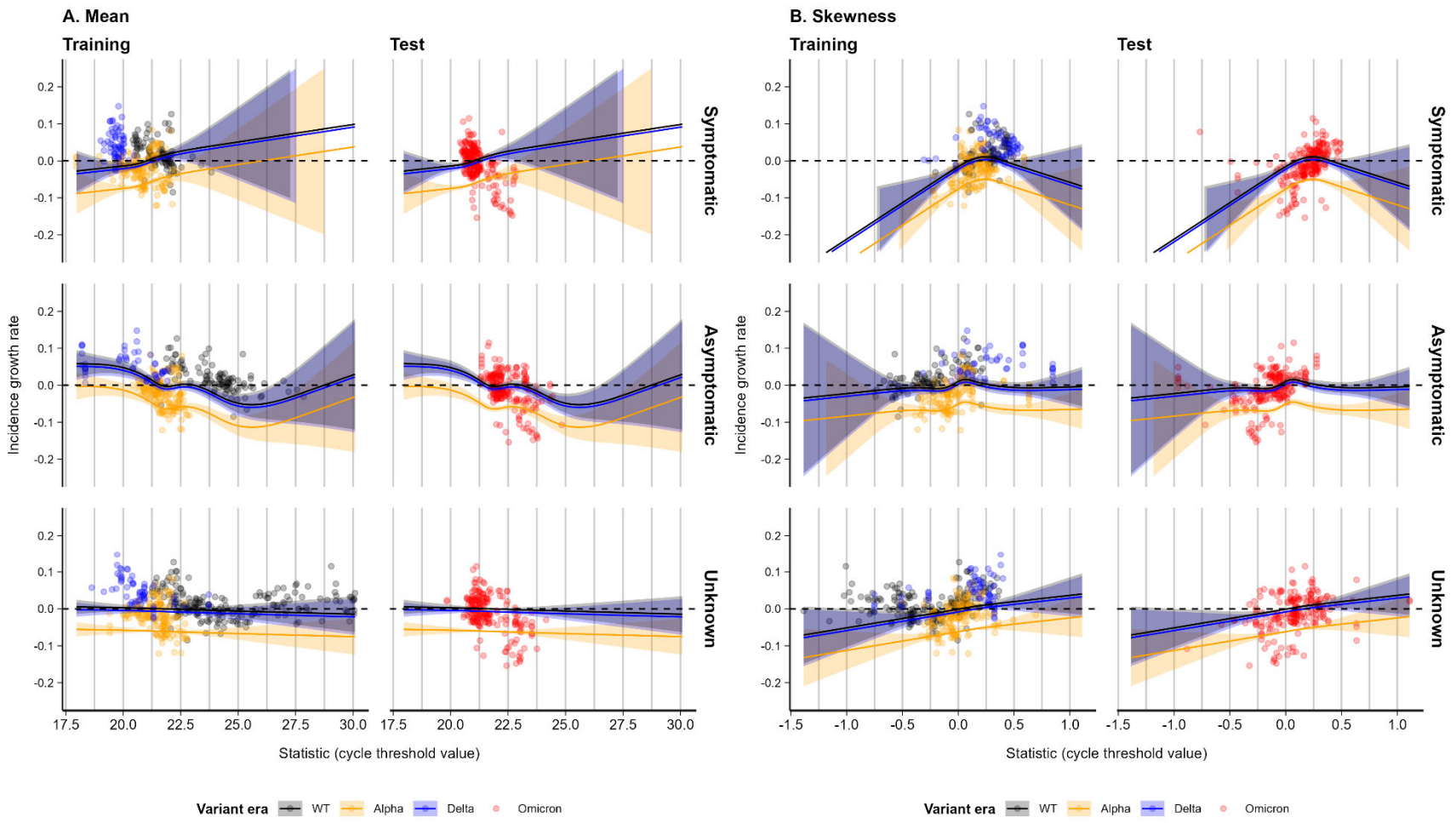
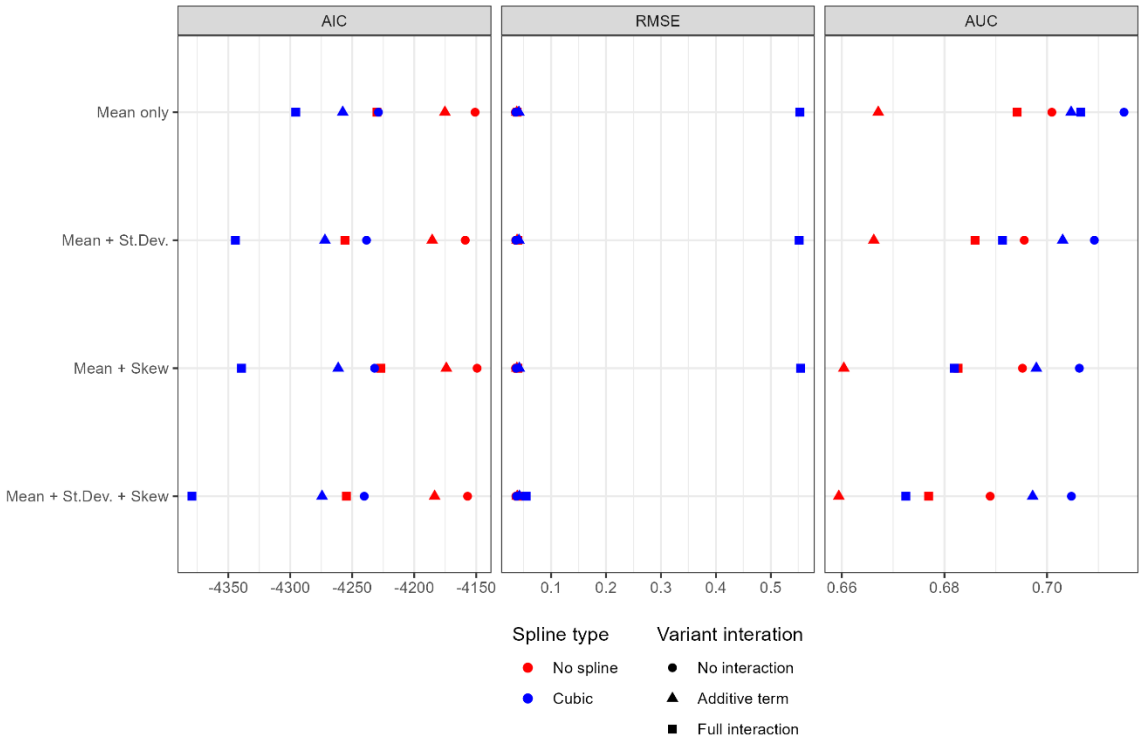


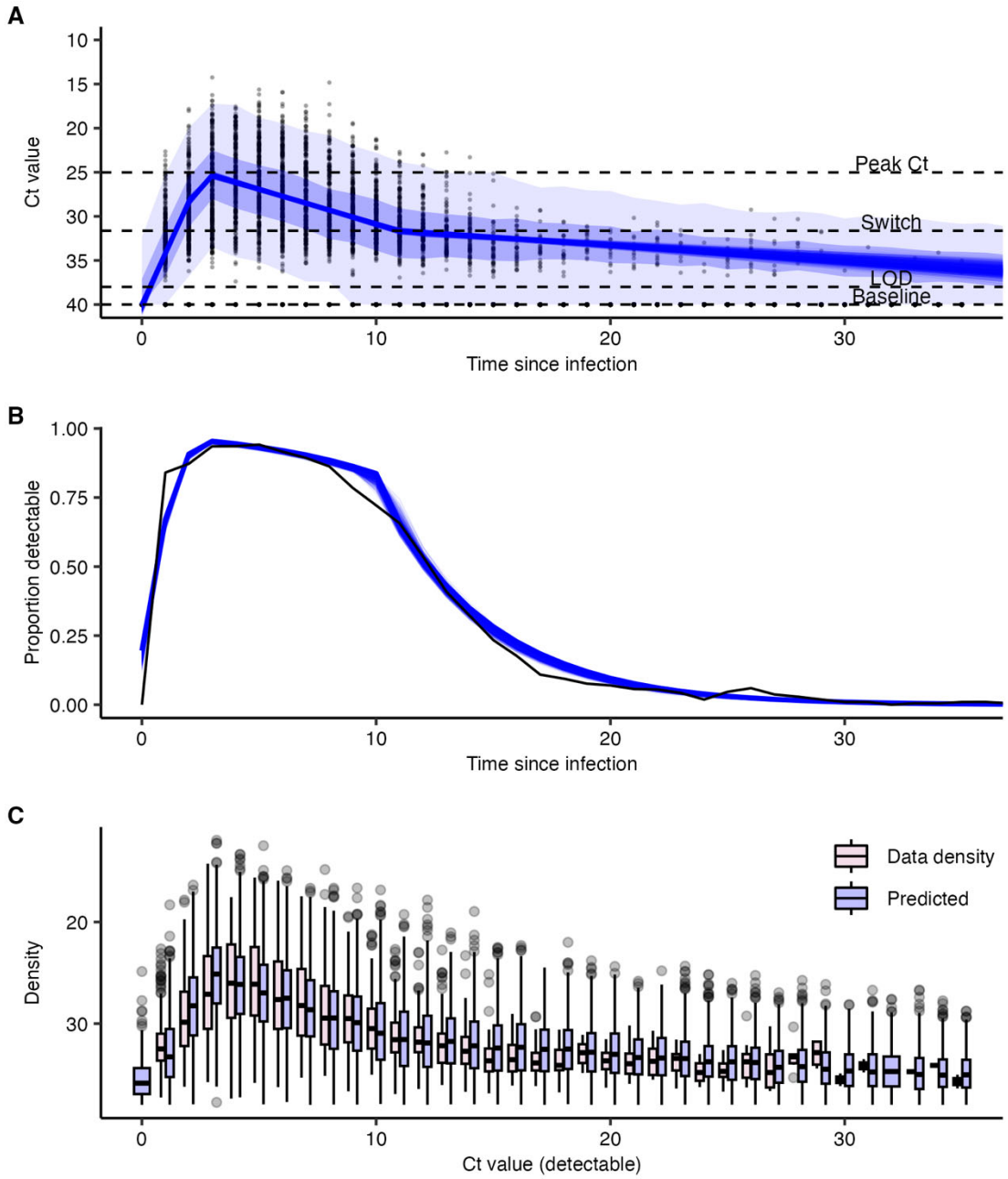
Figure S13. Best-fit lines (with standard errors) showing the modeled relationship between mean (left) and skewness (right) in Ct values against growth rate, stratified by reported symptom status, for LAC data, up to 31 Dec 2021 (training period) and after (test period). Note the differing relationships by symptom status stratum.



113

114 **Figure S14.** Comparison of predictive performance (RMSE and AUC) for the 24 tested models.

115



116

117 **Figure S15.** Viral kinetics model fitted to longitudinal SARS-CoV-2 RT-qPCR testing data
118 following a previous negative test. **(A)** Solid blue lines show 1000 posterior draws of the mean Ct
119 value over time. Shaded envelope shows 50% and 95% quantiles. Horizontal lines show control
120 points used to parameterize the model. **(B)** Solid blue lines show 1000 posterior draws for the
121 model-predicted proportion of positive tests over time-since infection overlaid on empirical
122 proportion detectable. **(C)** Box plots for the posterior distribution of observed Ct values on each
123 day post infection (blue) compared to distribution of raw data (red).

124

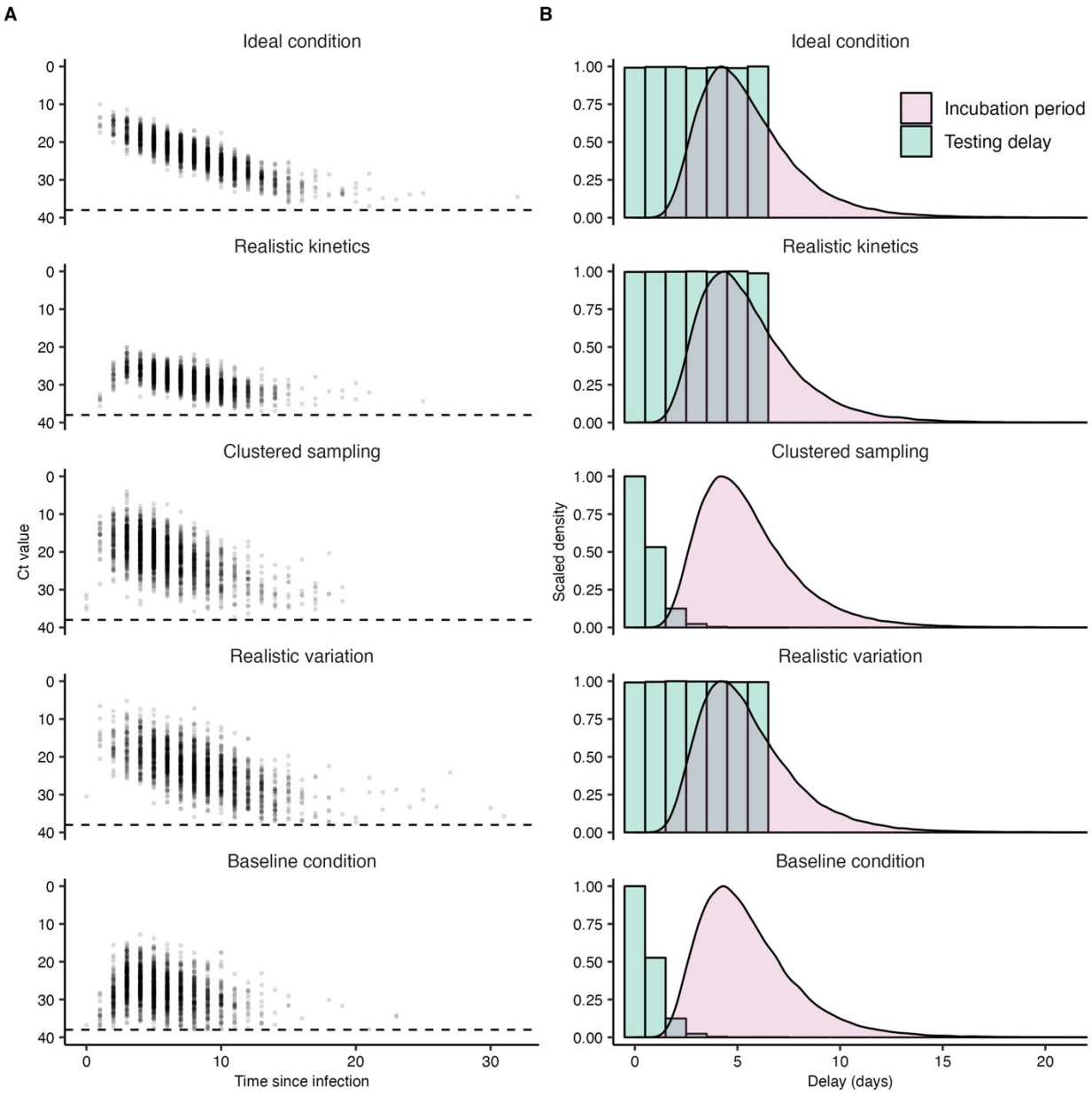


Figure S16. (A) Randomly simulated Ct values for the 5 synthetic data scenarios. "Baseline condition" shows the Ct value distribution as estimated from fitting to the longitudinal testing data, whereas the other scenarios show assumed Ct value distributions after changing some parameters. **(B)** Assumed incubation period distribution (red) and sampling delay distribution (green) for each scenario.

Supplementary Text 1 – Synthetic datasets

1. Viral kinetics model

We adapted a previously published viral kinetics model describing the mean and distribution of Ct values over time-since-infection[21]. Our simulations do not need to track individual infections, and thus unlike other published viral kinetics model[25,40,48,55], our model describes only the mean and variance of Ct values for all infections given time-since-infection rather than modeling each individual's viral trajectory. Model parameters and interpretation are shown in **Table S6**.

We used a piecewise linear model of the form:

$$f(t) = \begin{cases} c_0, & t \leq 0 \\ \mu t + c_0, & 0 < t \leq t_p \\ c_p - \omega_1(t - t_p) + c_0, & t_p < t \leq t_p + t_s \\ c_s - \omega_2(t - t_p - t_s) + c_0, & t > t_p + t_s \end{cases}$$

Where c_0 is the true baseline Ct value at time of infection; $\mu = \frac{c_p - c_0}{t_p}$ is the Ct value growth rate; c_p is the minimum Ct value; t_p is the time from infection to minimum Ct value; $\omega_1 = \frac{c_p - c_s}{t_s}$ is the initial clearance rate; c_s is the Ct value at which waning switches to a second, slower clearance rate; $\omega_2 = \frac{c_p - c_0}{t_c}$ is the second, slower clearance rate; t_c is the time taken to decay from c_s to c_0.

We model the distribution of observed Ct values around the mean Ct value ($f(t)$) to capture three observations:

1. Ct values are highly varied on a given day post infection.
2. The variance of the Ct distribution is not necessarily constant over time.
3. Most individuals clear their infections quickly, but a small proportion remain detectable at a very high Ct value for many days after infection.

The distribution of observed Ct values on a given day post infection is modeled as a truncated Normal distribution:

$$C(t) \sim N(f(t), \sigma(t))_0^{38}$$

Where N is the normal distribution. We assumed the distribution was truncated between 0 and 38 based on the distribution of observed Ct values in the NBA dataset. $f(t)$ is the mean of the Normal distribution and $\sigma(t)$ is the time-varying variance given by:

$$\sigma(t) = \begin{cases} \sigma_{obs}, & t < t_p + t_s \\ \sigma_{obs}(1 - \frac{1 - s_m}{t_m}(t - t_p - t_{s)}, & t_p + t < t \leq t_p + t_s + t_m \\ s_m \sigma_{obs}, & t > t_p + t_s + t_m \end{cases}$$

σ_{obs} gives the variance. The second term describes a gradually decreasing variance during the second clearance phase, declining at a constant rate over duration t_m before reaching a minimum of $s_m \sigma_{obs}$.

In addition, the probability of a sample having a detectable sample on a given day following infection is the product of two probabilities: the probability of having a Ct value less than the limit of detection, given by the cumulative density of the Normal distribution; and the probability of having not cleared the infection by that day:

$$\phi(t) = P[C(t) < c_0](1 - p_c)^{t - t_p + t_s}$$

The first part of the equation gives the cumulative density of the Normal distribution. The second part describes an additional process, whereby each day from $t_p + t_s$ onwards there is a daily probability, p_c, of becoming fully undetectable, representing clearance of the infection.

2. Parameterizing the base model

We parameterized the viral kinetics model using publicly available longitudinal data from the National Basketball Association (NBA) [48]. These data were a convenience sample from daily testing of NBA players, staff, and other affiliates over the course of the pandemic. Clinical samples were combined anterior nares and oropharynx swabs (collected separately from each anatomical site and combined in a single tube). Samples were tested using the Roche Cobas target 1 assay to give Ct values against the ORF1ab gene target. For this analysis, we used only tests from infections where the first positive sample was preceded by a negative test, intended to capture viral kinetics immediately following infection. Furthermore, as our objective was only to simulate a realistic model for the distribution of Ct values over time-since-infection, we did not stratify the data by covariates such as age group, symptom status, vaccination status or variant. Ultimately, we fit our model to 3,627 positive samples and 8,252 negative samples, representing 403 distinct infection episodes with samples taken between day 0 and 51 following a previous negative test.

We used a Markov chain Monte Carlo algorithm [56] to estimate posterior distributions for the model parameters conditional on the NBA dataset using uninformative uniform priors for the model parameters and a likelihood function based on the model described above. We ran 3 chains for 150,000 iterations, discarding the first 50,000 iterations as burn in. High effective sample sizes (>1000) and $\widehat{R}$ values <1.1 were obtained for all estimated parameters. We used the maximum *a posteriori* estimates as point estimates for the simulations. Model fits are shown in **Figure S15**.

3. Simulated surveillance data

We simulated Ct values observed under a realistic surveillance system using the following algorithm:

1. Infection times were simulated for *N=2,000,000* individuals (the cumulative incidence of cases in Massachusetts) by drawing infection times from the 7-day rolling mean reported incidence of cases from Massachusetts.
2. Two surveillance strategies were simulated giving each individual two possible sampling times:
    a. Random cross-sectional testing, representing detection of symptomatic infections. Uniformly distributed sampling dates were simulated for each infection. The time-since-infection was given as the difference between the sampling date and infection date (thus, many individuals were sampled before they became infected or long after they cleared their infection). All individuals are assigned one random sampling time.
    b. Symptom-based surveillance was simulated by assuming all infected individuals became symptomatic (note that it is not important to reflect the true symptomatic fraction for SARS-CoV-2, as we are only interested in generating a large number of simulated Ct values). Each symptomatic individual was assigned a randomly generated incubation period drawn from a log-normal distribution with mean = 1.621 and standard deviation = 0.41 on the log scale . For symptom-based surveillance, each symptomatic individual was additionally given a sampling delay drawn from a distribution. An individual's sampling date was given as their infection date, plus their incubation period, plus their sampling delay.
3. Expected Ct value at their sampling time were calculated using the viral kinetics model described above combined with the MAP estimates from the model fitting.
4. Time to full clearance was simulated for each individual from a negative binomial distribution with success probability p_c.

5. Finally, observed Ct values were simulated from a normal distribution with mean given by the expected Ct value given time-since-infection and the time-dependent variance as described above. If the simulated Ct value is greater than the limit of detection or the individual had already fully cleared the infection, then the Ct value was set to 40.

4. Synthetic data scenarios

To understand how the relationship between observed Ct values and true growth rate of infection incidence varies across different scenarios, we implemented 5 different scenarios. We start with the “Ideal” scenario, which modifies the simulation parameters to provide an unrealistic scenario where we expect to see a consistent and clear relationship between surveillance Ct values and incidence growth rates. We then return each of these simulation parameters to realistic values one at a time to understand which factors confound our ability to infer growth rates from surveillance Ct values.

Scenarios:

1. Ideal conditions: assuming that viral kinetics are extremely left-skewed (very fast growth phase relative to clearance phase), that there is little variation in observed Ct values given time-since-infection, and that the delay from symptom onset to sampling time is uniformly distributed between 0 and 7 days.
2. Realistic viral kinetics: using the viral kinetics parameter estimates from fitting the model, but with half the variation in observed Ct values given time-since-infection and uniformly distributed sampling delays.
3. Realistic variation: using an extremely left-skewed viral kinetics curve and uniformly distributed sampling delays, but with the variance of the Ct value distribution given time-since-infection based on the model estimates.

4. Realistic sampling: using an extremely left-skewed viral kinetics curve and reduced variance in observed Ct values, but assuming that sampling delays are gamma-distributed with a mean of 4 days and variance of 8 days.
5. Realistic scenario: using the viral kinetics parameter estimates and variance from fitting the model, and assuming that sampling delays are gamma-distributed with a mean of 4 days and variance of 8 days.

Figure S16 shows the assumed viral kinetics models, and incubation and sampling delay distributions used for each of the scenarios. **Figure S1** shows the assumed epidemic growth rate curve, and the resulting mean Ct values over time through the two surveillance systems.