Supplement to *Early risk-assessment of pathogen genomic variants emergence*

December 2022

Contents

1 Supplemental Methods

We describe a general method to estimate multi-strain dynamics and relative fitness advantages by partially pooling information across patches (geographic subunits, i.e. countries) and strains (Main Text Fig 1). This statistical approach leverages a hierarchical mixed-effects Bayesian framework. The model has two levels of hierarchy. In the first level, country-specific variant fitness

advantages are structured such that the fitness advantage of a variant in one location informs the expected fitness advantages of variants in other locations to formalize the assumption that a variant's properties in one location are likely to be similar in another. In the second level, variants' mean fitness advantages, averaged over countries, consist of a shared (hierarchical) normal distribution. This approach shares information across variants to formalize the ecological assumption that most variants will be similarly fit to their recent ancestors and observing extreme deviations in fitness is uncommon. This assumption leads to shrinkage on extreme fitness advantage estimates for novel or otherwise infrequently observed variants for which we might otherwise overfit to noise in the data.

In the main text, we apply the model to SARS-CoV-2 sequences submitted to GISAID **noauthor˙undated-xp** up through July 1, 2022, focusing specifically on BA.4 and BA.5's global emergence.

1.1 Data processing

Line list SARS-CoV-2 sequence metadata was extracted from the GISAID database **noauthor˙undated-xp**. Each row of the dataset contains information on the collection date, submission date, location of sample collection, and the assigned pango lineage **noauthor˙undated-vz** of the sequence submitted to the database. The country name in the location field is mapped to the corresponding three letter ISO country code. Sequence submissions missing complete date information (i.e. month and year instead of day month year) are excluded from the analysis and we assume this missingness is completely at random. Because pango lineage assignments are often delayed following initial submission, any sequences submitted on the reference date are excluded as they will still be labeled as "Unassigned" because they have not yet been assigned a lineage by GISAID's build of the Pangolin assignment tool **noauthor˙undated-ie**. Line list data was aggregated by country, collection date, and pango lineage to get daily counts of the number of lineages observed in each country. In order to make the number of unique lineages tractable for model fitting and analysis, we manually set the lineages we are interested in tracking. During this time period, this corresponded to: BA.1, BA.2, BA.2.12.1, BA.4, and BA.5. All pango lineages except for those labeled as the variants we're tracking and WHO VOCs were aggregated by the first number in their pangolineage assignment. For example, BA.5.1 would be assigned to the BA.5 lineage. We truncated the data to the past 90 days from the reference date for model fitting. Lineages with fewer than 50 observed sequences globally were aggregated into "other" along with the sequences labeled as "Unassigned". For visualization and model evaluation, we further collapse the pangolineage assignments into the "variants we're tracking", with all other pango lineages falling into the "other" category. Summary statistics (i.e. daily and weekly observed variant prevalences and standard errors) are calculated from this aggregation level, for comparison with model outputs.

1.2 Hierarchical generalized linear modeling approach

We model the dynamics of competing variants of a directly transmitted infectious disease. This approach can be used generally for competing strains, but we use here the example of SARS-CoV-2 variants using sequence metadata from GISAID to produce estimates of relative variant growth rates and true proportion of total cases in a given country. Using a hierarchical generalized linear approach, the model shares information across countries and across variants to improve estimates of variant growth advantages and dynamics in settings with sparse sampling.

1.3 Observation process

We consider the observed counts of lineages as drawn from a multinomial distribution. If we consider all sequences observed over the past 90 days (90 days ago $t = 0; t \in \mathbb{W}$), on day t we observe N_t sequences total ($N_t \in \mathbb{W}$), of which n_{it} sequences are of variant lineage i ($n_{it} \in \mathbb{W}$; $N_t = \sum_i n_{it}$). Then the set of observed variant lineages over the time period is $i_t \in \{i, ..., I\}$, where I is the reference variant, which we set as the dominant variant. Although this quantity changes over time due to the emergence of mutations that warrant a new lineage designation, we consider it here to be both fixed and known (i.e. the number of unique Pango lineages known on day $t = 90$). The multinomial probability mass function (PMF) and its unknowns motivate the regression model formulation

$$
\begin{aligned}
Y_{ijt} &\sim multinom(N_{jt}, p_{ijt}) \\
p_{ijt} &= \frac{e^{\eta_{ijt}}}{\sum_{j=1}^{J} e^{\eta ijt}} \\
\eta_{ijt} &= \beta_{0ij} + \beta_{1ij} z_t \\
z_t &= \frac{t - \mu_t}{\sigma_t}
\end{aligned}
$$

Where i indexes variants, j indexes countries, t indexes time, and μ^t and σ^t describe the mean and standard deviation of the vector of timesteps where each timestep is a day (this transformation is discussed more in the Bayesian modeling approach subsection). The parameters $\beta_{0Ij} = \beta_{1Ij} = 0$ are fixed to ensure identifiability. The intercepts β_{0ij} describe the initial variant prevalence on the scale of the linear predictor on day t. The slope coefficients (β_{1ij}) describe the difference in intrinsic growth rates for the variant in the numerator and the variant in the denominator (i.e. $\beta_{1ij} = r_{ij} - r_{Ij}$ where $r_I j$ is the intrinsic/Malthusian growth rate of the dominant variant in country j). However, this model does not produce an estimate of the actual intrinsic growth rates (i.e. r_{ij} or r_{Ij}) because fitting I cases would make the model overdetermined. This is why we refer to the β_{1ij} terms as relative variant growth rates in the text. This formulation accounts for changes in sample size over time (i.e., changes in N_t) and, from these counts, estimates the expected proportion of the population made up of each variant. For interpretability, we present the global and

country-specific estimates of the relative growth rates, β_{1ij} and $\mu_{\beta_{1i}}$ as relative weekly fitness advantages using the relation:

$$f = e^{7\beta_{1ij}} - 1$$

$$f = e^{7\mu_{\beta_{1i}}} - 1$$

As described by Davies et al.**Davies2021-if**

1.4 Hierarchical modeling structure

The hierarchical modeling approach pools information across variants and across countries, strengthening inference in settings with sparse information. This partial pooling addresses two forms of sparsity: variability in information available across lineages because new lineages have only been observed for a short period of time and systematic geographic variability in sequencing availability limiting the amount of information available in specific countries.

In the first layer of hierarchy, the model shares information across variants, making the assumption that most observed variants should be similar to each other. The model likelihood structures expected relative variant growth rates (i.e., $\mu_{\beta_{1i}}$) as drawn from a population distribution of growth rates:

$$\mu_{\beta_{1i}} \sim N(\mu_{\text{hierarchical}}, \sigma_{\text{hierarchical}})$$

The parameter $\mu_{\text{hierarchical}}$ specifies the expected relative variant growth rate of any variant compared to the reference variant I, which we assign to be the dominant variant to increase interpretability and numerical stability. The parameter $\sigma_{\text{hierarchical}}$ specifies the expected amount of variability of mean relative variant growth rates around $\mu_{\text{hierarchical}}$. The parametric assumption of a normal distribution assumes that radically more or less fit variants than the bulk of the population is quite rare (i.e., the population is not leptokurtic). Additional details on model fitting are available in the Bayesian Model Approach subsection.

The model also shares information across countries, allowing variant dynamics in one country to inform estimates of variant dynamics in other countries. Individual countries have country-specific relative variant growth rates β_{1ij}, where j indexes countries. The parameter $\mu_{\beta_{1i}}$ is a global estimate of relative variant fitness, pooled across countries. Systematic variation in growth rates across countries is learned empirically, expressed as a variance-covariance matrix Σ:

$$\beta_{0ij} \sim N(\mu_{\beta_{0}i}, \sigma_{\beta_{0}i})$$

$$\begin{bmatrix} \beta_{11j} \\ \vdots \\ \beta_{1Ij} \end{bmatrix} \sim MVN(\begin{bmatrix} \mu_{\beta_{11}} \\ \vdots \\ \mu_{\beta_{1I}} \end{bmatrix}, \Sigma)$$

$$\Sigma = \begin{bmatrix} \sigma^2_{\beta_{11}} & & & & \\ \sigma_{\beta_{11}}\sigma_{\beta_{12}}\rho_{21} & \sigma^2_{\beta_{12}} & & & \\ \sigma_{\beta_{11}}\sigma_{\beta_{13}}\rho_{31} & \sigma_{\beta_{12}}\sigma_{\beta_{13}}\rho_{32} & \sigma^2_{\beta_{13}} & & \\ \vdots & & & \ddots & \\ \sigma_{\beta_{11}}\sigma_{\beta_{1I}}\rho_{I1} & \cdots & & & \sigma^2_{\beta_{1I}} \end{bmatrix} = D(\sigma_{\beta_1})\Omega D(\sigma_{\beta_1})$$

Country-specific vectors of relative variant growth rates are drawn from this multivariate normal distribution, with a shared vector of global means and empirically learned variability around this mean vector. This approach allows for stable estimates of variant dynamics in countries with limited sequencing capacity or when variants have only been observed in a limited subset of countries.

Systematic correlations between relative variant growth rates in countries (i.e.,β_{1ij}) is modeled by the variance-covariance matrix Σ . This matrix can be decomposed into a correlation matrix Ω and a vector of independent lineage-specific variances $\sigma^2_{beta_{1i}}$:

$$\Omega = \begin{bmatrix} 1 & & & & \\ \rho_{21} & 1 & & & \\ \rho_{31} & \rho_{32} & 1 & & \\ \vdots & & & \ddots & \\ \rho_{I1} & \cdots & & & 1 \end{bmatrix}$$

$$\sigma^2_{\beta_1} = \begin{bmatrix} \sigma^2_{\beta_{11}} & & & \\ & \sigma^2_{\beta_{12}} & & \\ & & \ddots & \\ & & & \sigma^2_{\beta_{1I}} \end{bmatrix}$$

This approach identifies systematic correlations between realizations of relative variant growth rates from this multivariate normal distribution. In other words, it identifies if realizations of variant i in one country are systematically higher or lower when realizations of another variant are higher in that country. The variances of the β_{1ij} realizations around the mean $\mu_{\beta_{1i}}$are independent, allowing differences in local dynamics to reflect changing variability across variants.

1.5 Intercept structure

The model adds parametric structure to country-specific intercepts, enforcing the idea that variants are introduced to new countries at similar times. More formally, the variant-country intercept β_{0ij} is drawn from a hierarchical distribution:

$$\beta_{0ij} \sim t(\mu_{\beta_{0i}}, \sigma^2_{\beta_{0i}}, \nu = 2)$$

Unlike with the relative variant growth rates, there is no additional hierarchical structure on the means or variances of the intercepts β_{0ij}. The means $\mu_{\beta_{0i}}$ are

left independent, to allow for independent and unstructured variant emergence. Likewise the $\sigma^2_{\beta_{0i}}$ parameters are left independent across variants. This allows for independent patterns of between-country variant invasion across variants and allows for different amounts of variability initial prevalence of prevalence on the initial day of the model. The parametric choice of a student t distribution allows for substantial variability in timing between countries (i.e., kurtosis, heavy tails in the distribution).

1.6 Model assumptions and limitations

This approach makes a number of assumptions about the accuracy of the data on competing variants. It assumes that variants are randomly sampled within spatial units (e.g., countries, in this case), that lineages are correctly assigned to the sequences (e.g., the Pangolin tool assigns lineages without error), and that collection dates are correctly reported. However, the model does allow for missing data — lineages can be observed an arbitrary number of times and in an arbitrary number of spatial units.

The model does not currently account for spatial structure in any form — neither in estimated variant prevalences nor in estimated intercepts. It does not take into account which countries are neighbors or in close spatial proximity (i.e., there is no spatial kernel). It also does not account for structure in variant invasion patterns due to, for example, airline mobility patterns. This approach likely loses some spatial information that could improve estimates, but also makes the model conceptually simpler and limits the amount of information needed to run it. The model structure also assumes that relative variant fitness advantage is constant over time and linear on the scale of the linear predictor. This assumption is plausible because of the short time frame over which the model is run — 90 days. Therefore, any systematic changes in the host population that would impact the fitness of particular variants is unlikely to be large enough to lead to large changes in fitness advantage.

1.7 Bayesian modeling approach

We fit this model with a fully Bayesian approach to allow for flexible numerical sampling and directly interpretable parameters. We place informative priors on model parameters:

$$\mu_{\beta_{0i}} \sim t(-5, 5, 3)$$

$$\sigma_{\beta_{0i}} \sim N^{+}(2, 1)$$

$$\Omega \sim LKJ(2)$$

$$\sigma_{\beta_{1i}} \sim N^{+}(0.5, 2)$$

$$\mu_{\text{hierarchical}} \sim N(-1, 0.5)$$

$$\sigma_{\text{hierarchical}} \sim N(1, 0.1)$$

These informative prior distributions are derived from the posterior of a previous, independent model fit. This model was fit to a subset of European countries for a period corresponding to November and December of 2021 with non-informative priors. From that model fit, we use the first and second moments from these marginal posterior distributions for hyperparameters as weakly informative prior distributions except for the prior on Ω and the degrees of freedom on $\mu_{\beta_{0i}}$. These priors are set to general weakly informative priors as recommended by the Stan software creators, both for numerical simplicity and to allow for different correlations because of the different subsets of variants observed **noauthor˙undated-fa**. The time period for the original model fit used to develop the prior distributions does not overlap with any of the observed time points in any of the model fits interpreted here and the set of observed variants and the dominant variant are both different, suggesting that the priors provide reasonable scaling for the numerical sampler without being overfit to the specific time period or variants.

For computational and modeling tractability, the time covariate is centered and scaled:

$$z_t = \frac{t - \mu_t}{\sigma_t}$$

Where μ_t and σ_t are the mean and standard deviation of the vector of timesteps, with one day as one timestep. As a result, the priors on $\mu_{\text{hierarchical}}$, $\sigma_{\text{hierarchical}}$, and $\sigma_{\beta_{1i}}$ are on the scale of one standard deviation in scaled time (i.e. going from 0 to 1 on this scale is about 26 days). All hierarchical distributions are in a non-centered parameterization except for the distribution of which is in the centered parameterization.

The model is fit using Hamiltonian Monte Carlo (HMC) with the No-U-Turn sampler in CmdStan v2.29.2 **noauthor˙undated-um**. Using HMC instead of a more approximate variational approach allows for full characterization of posterior uncertainty. We run the model with 4 parallel chains with 2500 warmup iterations and 500 sampling iterations per chain for a total of 2000 sampling iterations. For all fits, the Gelman-rubin split $\hat{r}$ statistic was less than 1.01, no samples hit the maximum treedepth of 10, there were no divergences, and E-BFMI is above 0.3, suggesting that the numerical sampler converged and was able to perform unbiased sampling.

2 Retrospective validation of country-specific variant prevalence projections

2.1 Processing and fitting of historical datasets

Line-list sequence data from GISAID was downloaded on the following dates, which we refer to as "reference" dates: April 30th, 2022, May 16th, 2022, May 27th, 2022, June 4th, 2022, June 27th, 2022 and July 1st, 2022. A consensus set

of lineages was found by identifying all lineages that exceed the global threshold of 50 or more observed sequences in the past 90 days across any of the 6 reference datasets. These datasets were each processed as described in the Data Processing subsection. We fit the multicountry model to each reference dataset independently using all sequences collected within the 90 days preceding the reference date and the same consensus set of lineages to be estimated. We define the calibration period as the period over which any sequences were collected and had been submitted by the reference date globally in GISAID, with the forecast period defined by any days without any observed sequences globally. To test the model's predictive power, we forecasted 21 days out and compare this to any data observed by the last reference dataset from July 1st, 2022. The model returns the mean variant prevalence estimate and the predicted number of observed sequences of each lineage, for both the 90 day time period of calibration and the 21 day forecast. For the predicted number of observed sequences of each lineage, the model must know the number of total samples sequenced on each day. We fed into the model inputs the number of observed sequences per day by collection date for the final comparison reference dataset from July 1st, 2022. Using the predicted number of sequences of a particular lineage allows us to account for the higher uncertainty in observed lineage prevalence on days with lower number of reported sequences.

2.2 Estimation model comparison: Multicountry vs. Single country stability over time

For each reference dataset, we compare variant prevalence estimations from the multicountry model with estimations from country-independent multinomial models using the nnet package version 7.3.17 in R **Ripley2022-hc**. The nnet package returns the maximum likelihood estimation (MLE) of variant fitness advantages and mean variant prevalences. In order to get an estimate of the uncertainty of these outputs, non-parametric bootstrapping with replacement was performed, generating 100 boot-strapped datasets with time points randomly sampled with replacement from the true data. Each boot-strapped dataset was fit using the nnet multinom function. To get the predicted observed number of sequences of a particular lineage, we used the stats package **noauthor˙undated-eu** in R to simulate from the multinomial probability mass function with the point estimates of each parameter pertaining to each boot-strapped parameter set.

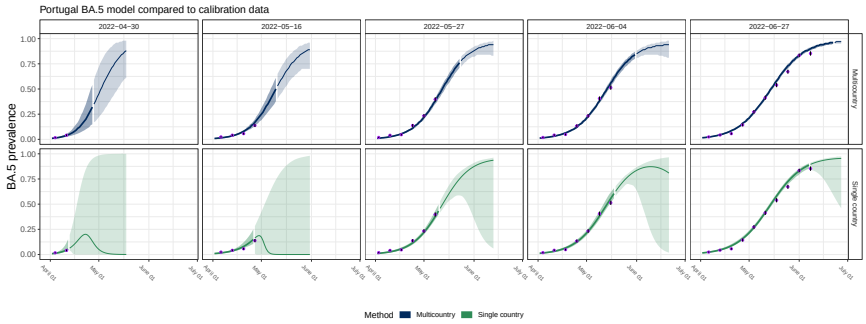


Figure S1: Multicountry (top) and single country (bottom) model estimated variant prevalence compared to the data observed at that time. Each column represents the date that the models were run and estimates were made (reference date). Shading indicates the calibration period (darker) and the forecast period (lighter). Line indicates the median projection. Purple dots indicate the weekly average observed prevalence of BA.5 as of that reference date, with bands indicating the standard error.

Figure S1 shows the model estimated BA.5 proportions in Portugal compared to, in this instance, the observed variant proportions as of that reference date (purple points), for the multicountry model and the single country model. In the main text in Figure 4, we compare the model estimates to the observed variant proportions as of July 1st, 2022. These differ because of the delay from specimen collection to sequence submission, resulting in backfilling of prior proportion estimates as new sequences get added.

In Figure S1, we observe that both models fit the observed data from Portugal relatively well during the calibration period, but that the model that relies only on the data from Portugal is highly erratic, as the refitting to the bootstrapped datasets results in widely variable projected BA.5 proportions a the early reference dates.

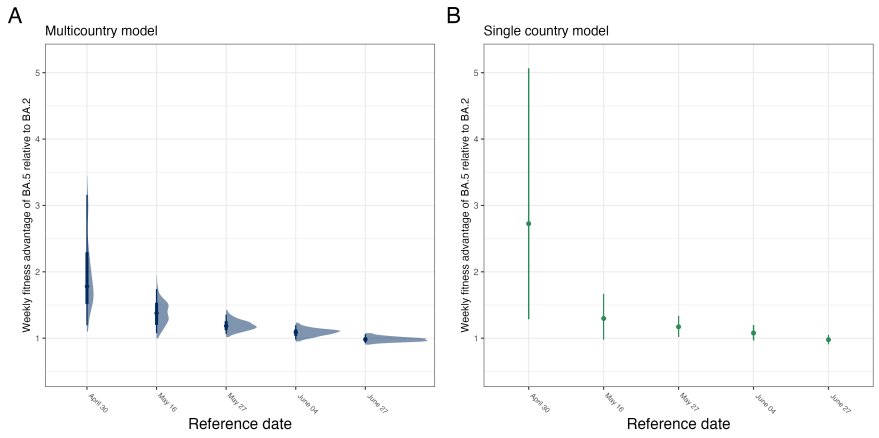


Figure S2: Multicountry (A) and single country (B) estimates of the weekly fitness advantage of BA.5 at each reference date. A. For the multicountry model, the posterior distribution of the estimated weekly fitness advantage is shown as of each reference date. Points indicate the median, bands indicate the 95th and 75th percentiles of the posterior distribution. B. Single country model estimates of the estimated weekly fitness advantage is shown as of each reference date. Points indicate the estimated fitness advantage applied to the observed data, bands indicate the 95% confidence intervals on the single country model using the normal approximation.

In Figure S2, we compare the estimates of the Portugal-specific weekly fitness advantage from the multicountry model and the single country model in order to compare the stability of the estimated variant fitness advantage in a particular region. For the multicountry model, we see that the early estimates do shift over time, however the median estimate is much closer to the later estimate than is the case for the single country model. We hypothesize that the reason for this early overestimate in the single country model could be due to factors such as demographic stochasticity, sampling bias, and overfitting to noisy data when data is sparse. In the main text, we show that the global estimated variant fitness advantage of BA.5, while initially more uncertain, remains relatively stable over the 5 reference dates (Main text, Fig 4B).

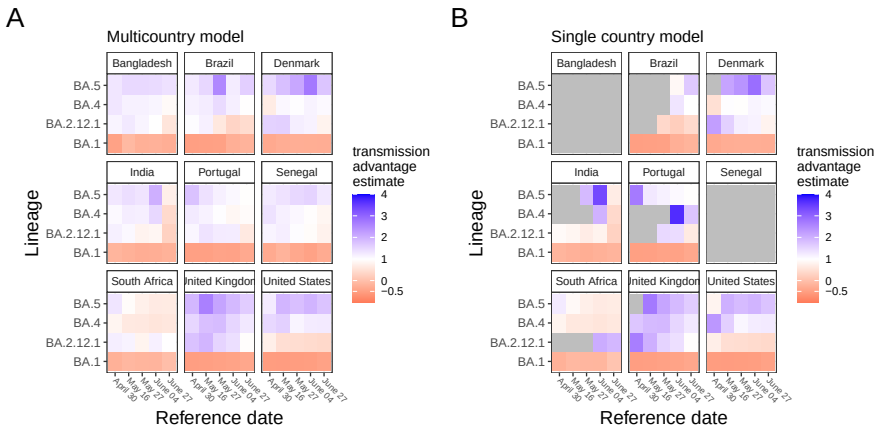


Figure S3: Comparison of the country-specific fitness advantage estimates across reference date and countries. A. Multicountry model estimates across reference dates show relative stability, even at early time points for BA.4 and BA.5. B. Single country model estimates across reference dates show the greater decline in estimated fitness advantage of emerging variants such as BA.5 over time (i.e. in the UK and Portugal), and overall less stability than in the multicountry model. Gray indicates time periods where there was insufficient data for model convergence within the country for that variant (a problem we don't see in the multicountry model because it leverages data for

that variant from other countries).

In Figure S3, we show across a range of countries the estimated country-specific transmission advantage for the multicountry model (S3A) and the single country model (S3B). What we see is that there is greater fluctiation over time (left to right in each heatmap) in the estimates in the single country model compared to the relative stability of the estimates from the multicountry model. Likewise, we are able to make estimates of country-specific fitness advantages for all country-variant combinations in the multicountry model, whereas in the single country model, if that variant has not been observed in that country or there is not sufficient data, the model is unable to converge (gray gaps).

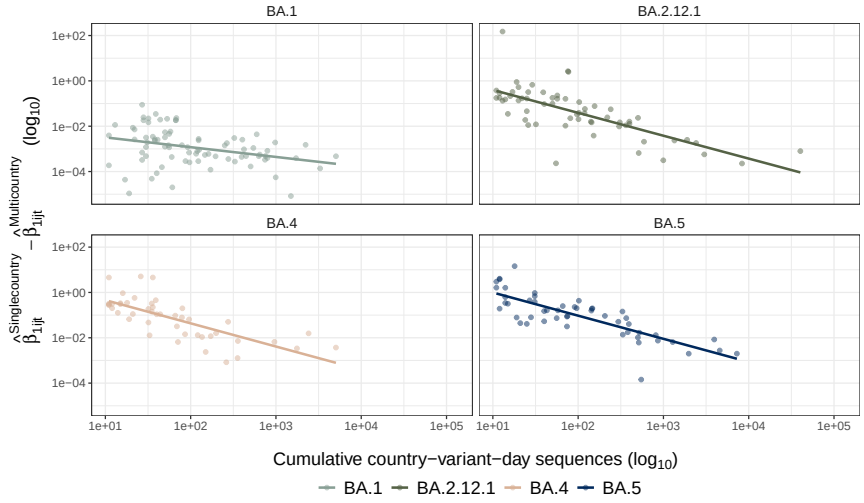


Figure S4: Difference between country-specific fitness advantage from the multicountry model ($\hat{\beta}^{\mathbf{multicountry}_{ijt}}$) and single-country models ($\hat{\beta}^{\mathbf{multicountry}_{ijt}}$). The log differences in the two estimates (y-axis) are plotted against log cumulative sequences of a variant observed in a country on a day (x axis).

In Figure S4, we compare the difference between the multicountry fitness advantage estimate and the single country fitness advantage estimate for four key variants as a function of the number of observed sequences of that variant in that country up until the time the estimation was made. The estimated coefficients from the multicountry model are consistently lower at low sequence counts than those estimated by single country maximum likelihood multinomial regressions. The maximum likelihood estimate coefficients are both higher initially and decline more slowly with the increase in sample size than those from the multicountry model. We note, however, that the comparison in terms of sample size is not one-to-one because the multicountry model uses all global sequences, while the single country is constrained to total within-country sequences (x axis

values). Estimates from models fit on country-variant-days with fewer than 10 cumulative sequences are discarded to remove extreme outliers.

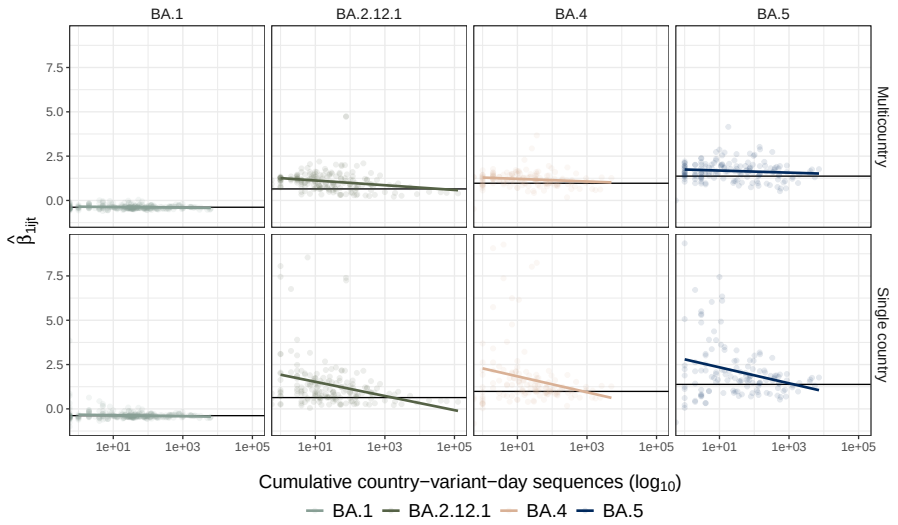


Figure S5: Estimated country-specific fitness advantage ($\hat{\beta}^{\mathbf{multicountry}_{ijt}}$). Horizontal lines are the average of the estimates from the top 10% of cumulative sequences for the variant-lineage bin to visually approximate the asymptote. Estimated fitness advantages greater than 10 are removed for visual clarity, but all such estimates are from the MLE model and would further exaggerate the contrast (n = 44).

In Figure S5, we compare the estimated weekly variant fitness advantage for 4 key variants for the multicountry model (top) and the single country model (bottom) versus the number of observed sequences of that variant in that country up until the time the estimation was made. Maximum likelihood estimates are higher initially and converge more slowly in the single country model than in the multicountry model estimates, highlighting the improved stability of the multicountry model estimates. This is particularly evident in the case of BA.5 (right column) where multicountry estimates are relatively stable while the single country estimates decline significantly as the number of sequences increases.

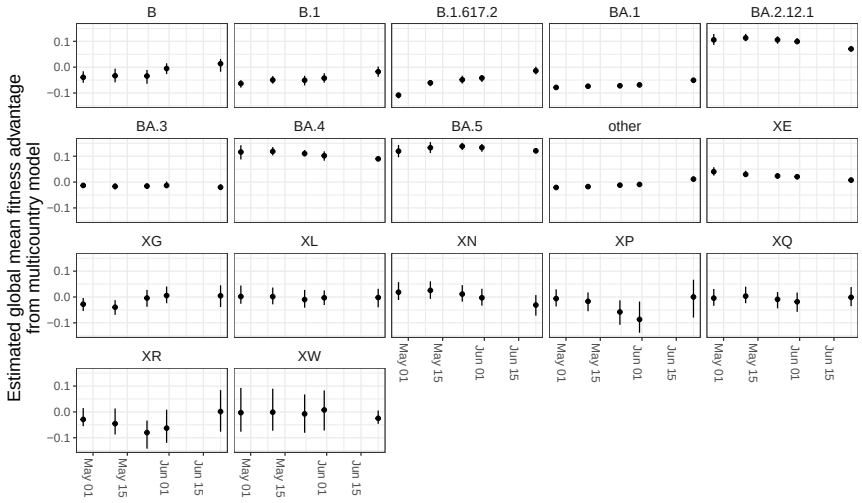


Figure S6: Estimated global fitness advantages from multicountry model over time. Estimated global means of variant fitness advantages estimated across the 5 reference dates.

In Figure S6, when we look at the estimates of global variant fitness advantage over time we see that estimates are stable with no systematic trend across variants. The lack of trend systematic trend across variants suggests that this estimator is appropriate to evaluate the risk posed by an emerging variant.

2.3 Evaluation of model estimates using the Brier score

For both the multicountry and the MLE model, we used 100 draws from the distributions of lineage prevalence estimates to evaluate the accuracy of the model predictions compared to the observed daily lineage prevalence from the July 1st, 2022 reference dataset. For countries that observed no sequences of a particular lineage in the consensus dataset during the 90 day time window, the MLE estimation model does not estimate a prevalence for that lineage. To enable a fair comparison of the two model outputs at the country-level, we collapse all prevalence estimations from the multicountry model for these unobserved lineages into "other" for that country. We use the Brier score to evaluate the accuracy of each draw of the model output. The Brier score is calculated at the country-level for both the calibration period and the forecast period, and the two combined. Because we have a distribution of probabilistic predictions from our model output (variant proportions over time), we get a distribution of the mBrier score at each reference time point as a result of the evaluation process.

To estimate the Brier score for each country and reference dataset from the multinomial model output, we compare the mean estimated probability that a sequence is the ith lineage at the tth time, $(\hat{p}_{t,i})$, to the observed binary outcome

$\Psi_{i,t,k}$ of whether the kth sequence on the tth day is the ith lineage.

$$BS = \frac{1}{\sum_{t=1}^{\tau} N_t} \sum_{t=1}^{\tau} \sum_{i=1}^{l} \sum_{k=1}^{N} (\hat{p}_{t,i} - \Psi_{i,t,k})^2$$

Where N_t is the total number of sequences collected on the tth day, l is the number of possible outcomes, in this case, the number of unique lineages estimated, and τ is the number of time points in the time window of interest. τ will depend on whether we are computing the Brier score for the calibration period, the forecast period, or the combination of the two.

For computational efficiency, instead of lining up the set of 1s and 0s for each lineage for each sequence collected each day, we can use the fact that and $\sum \Psi$ and $\sum \Psi^2$ are sufficient statistics for $\sum (p - \Psi)^2$ to evaluate the Brier score for each lineage at each time point. If we expand the above equation:

$$BS = \frac{1}{\sum_{t=1}^{\tau} N_t} \sum_{t=1}^{\tau} \sum_{i=1}^{l} \sum_{k=1}^{N} (\hat{p}_{t,i}^2 - 2\hat{p}_{t,i}\Psi_{i,t,k} + \Psi_{i,t,k}^2)$$

Which is equivalent to:

$$BS = \frac{1}{\sum_{t=1}^{\tau} N_t} \sum_{t=1}^{\tau} \sum_{i=1}^{l} (N\hat{p}_{t,i}^2 - 2\hat{p}_{t,i} \sum_{k=1}^{N} \Psi_{i,t,k} + \sum_{k=1}^{N} \Psi_{i,t,k}^2)$$

Replacing the sum of the binary (1,0) $\Psi_{i,t,k}$ with the number of sequences of each lineage, $n_{i,t}$, as such, $\sum \Psi_{i,t,k}^2 = n_{i,t}$ we can write the Brier score in terms of the number of observed sequences of the ith lineage ($n_{i,t}$) and the total number of sequences, N_t collected each day.

$$BS = \frac{1}{\sum_{t=1}^{\tau} N_t} \sum_{t=1}^{\tau} \sum_{i=1}^{l} (N_t\hat{p}_{t,i} - 2\hat{p}_{t,i}n_{i,t} + n_{i,t})$$

The Brier score is calculated for each draw from the multinomial output for each country and each time period evaluated.

3 Case studies

3.1 Identifying hemispheric drivers of influenza dynamics

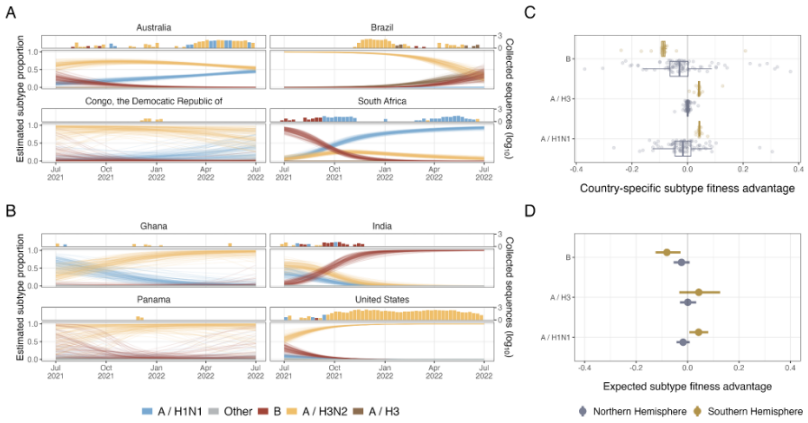


Figure S7: Application of the framework to Influenza dynamics in both hemispheres. (A) Flu variant dynamics for a subset of Southern Hemisphere countries. (B) Flu variant dynamics for a subset of Northern Hemisphere countries. (C) Country-specific fitness advantages relative to A / H3N2 for selected subtypes. (D) Expected fitness advantages relative to A / H3N2

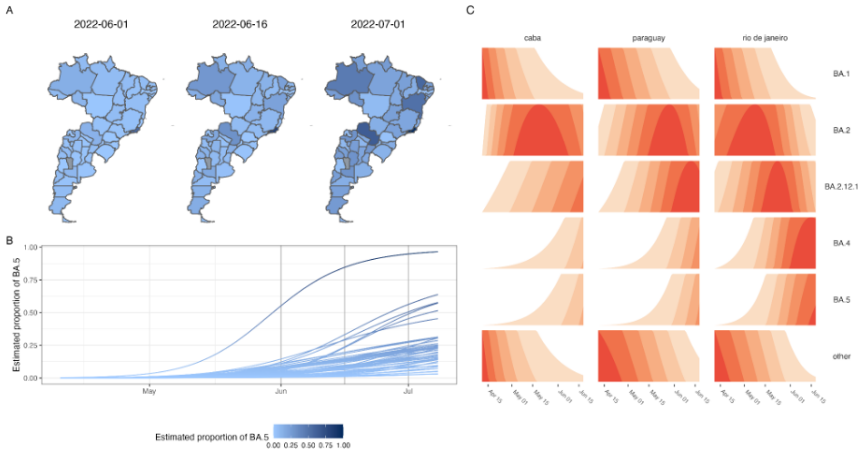
We apply this approach to influenza dynamics, estimating fitness and subtype dynamics separately for the Northern and Southern hemispheres. In Figure S7A, we show estimates of influenza subtype prevalence from July, 2021 to July, 2022, identifying diverse dynamics across the selected countries from the Southern Hemisphere. In Australia and South Africa, prevalence of the B subtype steadily declines over the observed year and is replaced by a combination of A / H1N1 and A / H3N2. These dynamics differ from those of Brazil, where the dominant A / H3N2 subtype is replaced by A / H3 and B subtypes. Dynamics in the DRC are estimated imprecisely, but suggest that the A / H3N2 is dominant on July 1, 2022 — matching overall influenza dynamics in the Southern Hemisphere and the data observed from within the DRC.

In Figure S7B, we present estimated influenza subtype dynamics in the Northern Hemisphere from July 1, 2021 to July 2022. We find the A / H3N2 subtype increased in prevalence in the United States, Ghana, and Panama over the observed year. In all the selected countries, A / H1N1 prevalence declined or remained negligible. Notably, however, India's dynamics were meaningfully different from those of the other selected countries: the relative proportion of the B subtype increased from July, 2021 to July, 2022. It displaced both A / H1N1 and A / H3N2, unlike in the other selected countries. We note that in many of these cases, influenza dynamics in the Northern Hemisphere appear to lag those of the Southern Hemisphere. In much of the Southern Hemisphere, the A / H3N2 and B subtypes are initially dominant and replaced over the observed year. In the Northern Hemisphere, the A / H3N2 and B subtypes are a small

proportion of initial observations and increase in relative proportion throughout the observed year.

In Figures S7C and S7D, we present estimates of country-specific and overall hemispheric mean fitness advantages for selected influenza subtypes. Because separate models are fit to the Northern and Southern Hemispheres, we present estimates for subtype fitness advantage relative to A / H3N2 separately for each hemisphere. We find that estimates of country-specific fitness advantage for the A / H3N2 are more diffuse in the Northern Hemisphere than in the Southern Hemisphere for all 3 subtypes. Except one outlier, estimated fitness advantages for all lineages are tightly clustered in the Southern Hemisphere. The Northern Hemisphere's more diffuse estimates have epidemiological implications: the estimates are scattered on either side of 0, indicating that the variance in country-specific fitness relative to A / H3N2 is high and any naive estimate for an unobserved country will be highly uncertain. The mean fitness advantages are similar for the B, A / H3, and A / H1N1 subtypes across hemispheres. For all three, estimates are either so close that the difference of means is statistically indistinguishable or close enough to not be epidemiologically meaningful. These results indicate that the country-specific differences are more important to influenza dynamics than hemisphere-level differences in subtype fitness.

3.2 Estimating SARS-CoV-2 dynamics at administration level 1.



We apply the method described here to estimate SARS-CoV-2 dynamics at the AL0 and AL1 methods to demonstrate the flexibility of this modeling approach. We aggregate SARS-CoV-2 sequencing counts in GISAID from the states of Brazil and the provinces of Argentina (i.e., AL1 units) as well as

combining sequences from all of Paraguay. These selections demonstrate how one might apply this method to estimate sub-national dynamics, where such data exist, and illustrate that the modeling approach remains coherent when estimating AL1 and AL0 spatial units together.

In Figure S8A and S8B, we show SARS-CoV-2 dynamics in these spatial units across time. We find that the BA.5 variant was uncommon across all the selected spatial units on June 6, 2022. Estimated variant prevalence was below 1% in all the modeled spatial units (Figure S8A). The expected prevalence of BA.5 rose over the course of the time period, increasing steadily through mid-June and into July (Figure S8B). We found, however, marked heterogeneity in the expected prevalence of BA.5 across the spatial units. BA.5 increased most quickly in Rio de Janeiro, approaching fixation by July 1, 2022. Although BA.5 prevalence was expected to have been increasing in all spatial units, BA.5 prevalence was below 50% on July 1, 2022 in most other regions in Brazil and Argentine.

In Figure S8C, we show prevalence over the observed time period for selected variants in three selected spatial units (CABA: Buenos Aires, Paraguay, and Rio de Janeiro). We identify the decline of BA.1 in all three of the spatial units during the time period. We find that BA.2 and BA.2.12.1 peaked in late May and early June and that it beginning to be outcompeted by BA.4 and BA.5. We also find that BA.4 had an earlier foothold in the region than BA.5.

This example demonstrates the ability of this applied method to be applied across heterogeneous spatial units. It identifies dynamics of SARS-CoV-2 variants and allows for variation across the spatial units in its estimates.

3.3 Data/code availability

All code is made publicly available at this Github repository Github repository. We do not include any data in the repository, but all results can be reproduced using the GISAID SARS-CoV-2 and flu metadata for authenticated users.

4 Applications

In addition to developing the method presented here so that others can apply it sequencing data across diseases and geographic regions, The Pandemic Prevention Initiative at The Rockefeller Foundation is committed to applying the methods to real-world data, in real-time, and making all relevant outputs public so that others may leverage them for their specific needs. As a result of and alongside this investigation, two separate data tools have been created and are or will be made public via dashboards – the Next Generation Sequencing Capacity Map and the Global Covid-19 Variant Tracker.

The Next Generation Sequencing (NGS) Capacity Map, developed and hosted by our collaborators at FIND, is a dashboard that has provided real-time global genomic surveillance capacity since 2021. It includes country-level data on SARS-CoV-2 genomic sequencing levels, diagnostic capacity, NGS facilities, and

investments in different aspects of capacity building. Users can look at the current state of diagnostic and sequencing capacity levels in each country and can also assess how these have changed over time globally and by income group. The tool also provides data-driven "archetypes" which provide recommendations around the coordination of capacity building efforts globally. The tool is intended to illuminate the disparities in sequencing capacity across the globe, which give rise to the need for the method proposed in this manuscript. All code used to build the dashboard metrics and corresponding archetypes, as well as a detailed description of the methods, are provided at the PPI NGS Github page.

The second tool, the Global Covid-19 Variant Tracker, is a direct application of the described method here. This live dashboard will provide estimates of the global fitness advantage of novel variants as they emerge throughout the globe, and will highlight key trends throughout different geographic regions for these variants of interest. We envision these estimates, which give a sense of the overall risk posed by emerging variants, will be most useful for the research community, in particular, those interested in identifying the biological characteristics of novel variants (i.e. the degree to which they evade immunity from certain vaccines/prior infections with past variants), as well as guiding in the development of variant-targeted vaccines and antibody therapies. Additionally, the dashboard will also provide country-specific variant prevalence dynamics, including nowcasts of variant prevalence and inferred cases with each variant using OWID case data **Mathieu2020-ka**, for all countries with data available in GISAID in the past 90 days. Additionally, because countries have divergent immune landscapes, we provide an estimate of the country-specific variant fitness advantage so that individual countries can plan for the variants expected to circulate in their area. The goal of these outputs will be to provide situational awareness and inform decision-making for public health officials deciding on treatment recommendations that vary with variants, preparing for variant-driven surges, and catering the development of novel vaccines and therapeutics to the local variants. Additionally, we think this information can be useful to the general public to help them gain situational awareness around the shifting variant landscape. Lastly, we believe the combination of variant prevalence estimates and variant fitness advantages can help guide disease forecasting efforts, providing infectious disease modellers with critical input data needed to project future epidemiological indicators that will be affected by variants. Because of this use-case, we have made all data available on public S3 buckets in the form of machine readable csvs to facilitate their use in subsequent analyses. In the spirit of open science and transparency, all code used in the development of the Global Covid-19 Variant Tracer pipeline is available on the PPI Variant Tracker Github page. We encourage individuals to reach out with any questions regarding the methods or its applications.