Phenotypic evolution of SARS-CoV-2: a statistical inference approach

Wakinyan Benhamou [iD], Sébastien Lion [iD], Rémi Choquet⋆ [iD] and Sylvain Gandon⋆ [iD]

CEFE, CNRS, Univ Montpellier, EPHE, IRD, Montpellier, France
⋆: equal contribution

June 5, 2023

Abstract

Since its emergence in late 2019, the SARS-CoV-2 virus has spread globally, causing the ongoing COVID-19 pandemic. In the fall of 2020, the Alpha variant (lineage B.1.1.7) was detected in England and spread rapidly, outcompeting the previous lineage. Yet, very little is known about the underlying modifications of the infection process that can explain this selective advantage. Here, we try to quantify how the Alpha variant differed from its predecessor on two phenotypic traits: the transmission rate and the duration of infectiousness. To this end, we analysed the joint epidemiological and evolutionary dynamics as a function of the Stringency Index, a measure of the amount of Non-Pharmaceutical Interventions. Assuming that these control measures reduce contact rates and transmission, we developed a two-step approach based on *SEIR* models and the analysis of a combination of epidemiological and evolutionary information. First, we quantify the link between Stringency Index and the reduction in viral transmission. Secondly, based on a novel theoretical derivation of the selection gradient in an *SEIR* model, we infer the phenotype of the Alpha variant from its frequency changes. We show that its selective advantage is more likely to result from a higher transmission than from a longer infectious period.

1 Introduction

In December 2019, acute pneumonias of as yet '*unknown etiology*' were increasingly reported in Wuhan, the capital of the Hubei Province in Central China [30]. Since then, the infectious agent responsible of this emerging zoonosis, a virus of the family *Coronaviridae* named SARS-CoV-2 (Severe Acute Respiratory Syndrome-CoronaVirus-2), has spread worldwide, causing the pandemic COVID-19 (Coronavirus Disease-2019) [49] that is still ongoing today.

The possibility of a rapid SARS-CoV-2 adaptation was initially met with considerable scepticism [16, 40]. Indeed, compared to other single-stranded RNA viruses, the mutation rate of SARS-CoV-2 is relatively low (estimated at the onset of the pandemic around $6.8-9.8\times10^{-4}$ substitution.site^{-1}.year^{-1} [45, 46]). Besides, all the observed mutations in SARS-CoV-2 were initially thought to be neutral or slightly deleterious. The occasional rise of some mutations could be due to demographic stochasticity [15, 9] but the dramatic rise of specific mutations in different regions of the world challenged the hypothesis that none of these mutations were beneficial. In particular, the analysis of the emergence and the spread of several Variants of Concern (VOCs) across the world – e.g. Alpha (lineage B.1.1.7), Delta (lineage B.1.617.2) or Omicron (lineage B.1.1.529) (see for example CoVariants [21] or Nextsrain [17]) – demonstrated that these variants carry adaptive mutations that explain their faster rate of spread in the human population [31]. However, each of these mutations may act on various dimensions of the fitness landscape of the virus and affect different life-history traits. It is therefore much less clear *why* specific variants are favoured. In other words: which phenotypic trait(s) can explain this increase in viral fitness? Viral fitness is governed by multiple life-history traits like the transmission, the virulence or the recovery rates of the virus [9]. It is crucial to understand which traits are involved in the increase in fitness because they may have very different implications for epidemiological dynamics and public health. For instance, an increase in the transmission rate or in the duration of infectiousness both lead to an increase in viral fitness but they may have distinct consequences for the efficacy of Non-Pharmaceutical Interventions (NPIs), implemented to mitigate the epidemic. It is therefore very important to understand and track this adaptation to optimize our control strategies.

In the following, we will focus on the first of these VOCs: the lineage B.1.1.7, categorised as *Variant of Concern 202012/01* and afterwards named "*Alpha variant*". This variant emerged in early fall 2020 in the South-East region of England [36, 47] and then spread rapidly across the country (**Fig. 1**). The reproduction number of the Alpha variant (i.e. its expected number of secondary infections) was estimated to be 40-100% higher than for the previous lineage [5, 47]. Several studies aimed to unravel what phenotypic differences could explain this increased fitness. First, Davies *et al.* [5] explored various underlying biological mechanisms and suggested that a higher transmission rate per contact for the Alpha variant was the most parsimonious explanation, but that a longer duration of infectiousness – merely increasing the number of opportunities of transmission – could also explain the data very well. Blanquart *et al.* [3] developed another methodological approach considering three phenotypic traits: the overall reproduction number, the mean and the standard deviation of the generation time distribution of the infection. They showed that the selective advantage of the Alpha variant was likely to be driven by a higher reproduction number with an unaltered mean generation time.

The present work is a new attempt to characterise the life-history traits of the Alpha variant, for which we consider two phenotypic traits: (i) the transmission rate and (ii) the recovery rate (inverse of the mean duration of infectiousness). We propose a novel approach to estimate these two phenotypic traits based on the analysis of the time-varying fluctuations of the selection coefficient driven by the variability in the intensity of NPIs used to limit the spread of the virus. As pointed out by Otto *et al.* [32], the selection coefficient of the Alpha variant (i.e. the slope of the change in its logit-frequency) varied with the intensity of NPIs, measured by the "*Stringency Index*", a composite score published by the Oxford COVID-19 Government Response Tracker (OxCGRT) [18]. In [9] and [32], control

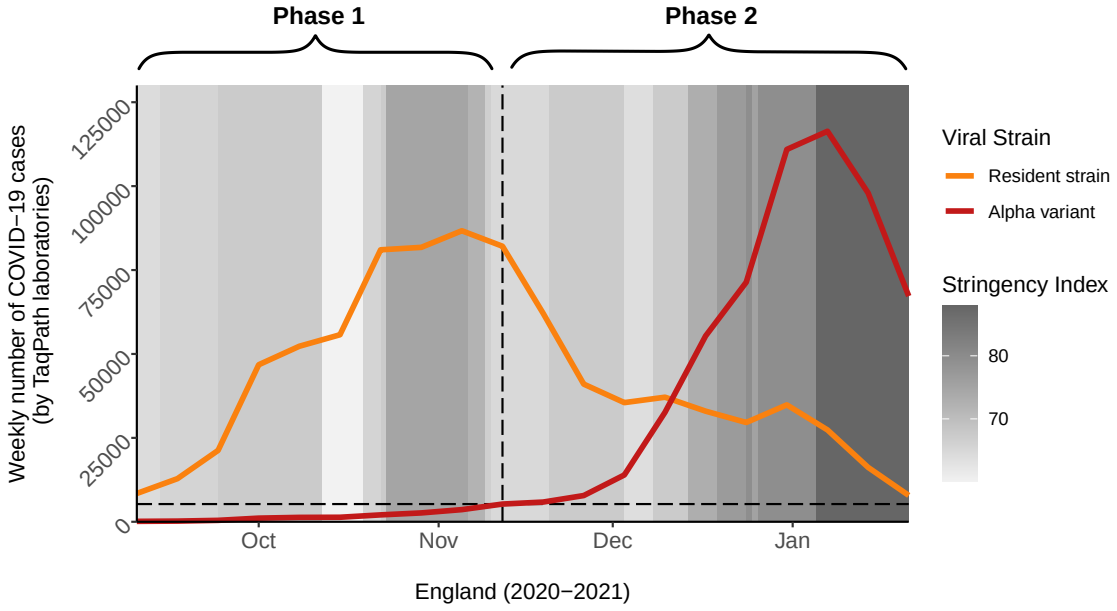

Figure 1: **The two consecutive phases of the analysis of the spread of the Alpha variant.** In phase 1 (before the emergence of the Alpha variant), we assume the epidemic is driven solely by the resident strain; in phase 2 (after the emergence of the Alpha variant), the epidemic results from the joint dynamics of the resident strain and the Alpha variant. In the first step of our analysis, we estimated the impact of the Stringency Index (a measure of the amount of NPIs implemented to mitigate the epidemic, from 0 (no control) to 100 (stringest control)) on the propagation of the resident strain during phase 1. In the second step of our analysis, knowing the impact of NPIs, we estimated the phenotypic differences between the resident strain and the Alpha variant during phase 2. The dates reported on the chart match the middle of each week (Thursday). We set the end of phase 1 when the Alpha variant reached 5% of the cases tested positive at the national scale (horizontal dashed line). For the sake of simplicity, we show data at the national scale but the starting date of phase 2 varied among regions (see **Fig. S1** and **Methods, §4.1**).

measures that reduce contact rates between infectious and susceptible hosts are predicted to reduce the (relative) selective advantage of variants that have a higher transmission rate – in addition to slowing down the spread of the epidemic – but without affecting the selective advantage of variants that have a longer duration of infectiousness. In the following we exploit these contrasting effects of NPIs on the selection coefficient to infer the transmission rate and the mean duration of infectiousness of the new variant.

We use a stepwise approach of two consecutive phases of the epidemic (**Fig. 1**). First, we focus on the analysis of the epidemiological dynamics taking place just before the emergence of the Alpha variant (i.e. just before it reached 5% of the positive cases) and we infer the relationship between the Stringency Index and the effectiveness of the control measures (NPIs) on the viral propagation in the UK. Second, we derive a novel expression for the selection coefficient of a variant in a susceptible-exposed-infectious-recovered ($SEIR$) model. Knowing the impact of NPIs on the viral propagation from the first step, we use our expression of the selection coefficient to infer the effects of the mutations

of the Alpha variant on (i) the transmission rate and (ii) the mean duration of infectiousness from the analysis of the evolutionary dynamics taking place, in each region of England, just after the emergence of the variant (i.e. just after it reached 10% of the positive cases).

2 Results

In the first step of the analysis, we develop an $SEIR$ model (see equations (3) and **Fig. S3**) to capture the effect of the control measures $c(t)$ on the epidemiological dynamics. The effectiveness of these control measures have been quantified and monitored with the Stringency Index [18]. As shown in the methods (§4.1.1), the Stringency Index depends mainly on NPIs that decrease contacts with susceptible hosts, and we therefore assume that NPIs only affect transmission, and not the infectious period. We model the link between the effectiveness of the control measures $c(t)$ and the Stringency Index $\psi(t)$ at each time point t through the following function:

$$c(t) = k\left(\frac{\psi(t)}{100}\right)^a, \tag{1}$$

with k, the maximum achievable effectiveness, and a, a "*shape*" parameter; $\psi(t)$ takes values between 0 (no control) and 100. We generated daily new fatality cases (4), daily new cases tested negative (6) and daily new cases tested positive (7) that we fitted to observed data using weighted least squares (WLS) (see **Methods, §4.2.2**). The best WLS estimates for this model yielded $k = 1$ and $a = 3.78$. The adjusted model seemed to fit the general dynamics of the data even though somewhat locally perfectible (**Fig. S6**). We then quantified the uncertainty of our parameter estimates using wild bootstrap [29, 24]: we reiterated about 2000 non-linear optimizations on perturbed data in order to get 2000 new sets of estimations (cf. **Methods, §4.2.2**). We thus obtained the joint distributions of the estimated parameters (see **Fig. S4**), and in particular parameters k and a that govern equation (1).

In the second step of the analysis we seek to explain the rapid spread of the Alpha variant through an increase in the transmission rate and/or the recovery rate. We developed an $SEIR$ model which takes into account the circulation of both the Alpha variant and the previous lineage, which we will refer to as the resident strain (**Methods, §4.1.2**). This model was used to derive an approximation of the temporal dynamics of the overall frequency $\widetilde{f}_m(t)$ of the Alpha variant. Under the assumptions of weak selection and quasi-equilibrium of fast variables (for more details, see **SI Appendix**), we obtained the following approximation for the selection coefficient s(t) of the Alpha variant:

$$s(t) \;=\; \frac{\mathrm{d}\ \mathrm{logit}(\widetilde{f}_m(t))}{\mathrm{d}t} \;\approx\; \frac{\kappa + \overline{r}(t)}{\kappa + \overline{\gamma}(t) + 2\overline{r}(t)}\left[\frac{\Delta\beta}{\overline{\beta}(t)}\left(\overline{r}(t) + \overline{\gamma}(t)\right) - \Delta\gamma\right], \tag{2}$$

with $\mathrm{logit}(\widetilde{f}_m(t)) = \ln\big(\widetilde{f}_m(t)/\big(1-\widetilde{f}_m(t)\big)\big)$ and where $\Delta\beta$ and $\Delta\gamma$ are the phenotypic differences between the Alpha variant and the resident strain in terms of transmission and recovery, respectively; $\overline{\beta}(t)$ and $\overline{\gamma}(t)$ refer to the average transmission and recovery rates across all genotypes; κ is the transition rate from the exposed state E to the infectious state I. Lastly, $\overline{r}(t)$ is the average growth rate of the

epidemic:

$$\overline{r}(t) = q(t)\left((1-c(t))\overline{\beta}(t)\frac{S(t)}{N} - \overline{\gamma}(t)\right),$$

with $q(t)$, the frequency of infectious individuals among infected hosts (i.e. $I(t)/\big(E(t)+I(t)\big)$). It is important to note that NPIs affect the selection coefficient of the variant (2) through the growth rate of the epidemic $\overline{r}(t)$, which depends on the amount of control $c(t)$. Crucially, this impact is stronger if the Alpha variant is more transmissible (i.e. $\Delta\beta > 0$) (see also [9] and [32]). Interestingly, we found – as in [32] – a negative correlation between the selection coefficient of the Alpha variant in England and the Stringency Index: -0.88 at the national scale (95% CI [-0.98; -0.39]) and between -0.97 (London, 95% CI [-0.99; -0.86]) and -0.81 (South West, 95% CI [-0.97; -0.14]) at the regional level (**Fig. S2**). In the following, we approximated $\overline{r}(t)$ using the quasi-equilibrium expression of $q(t)$, we assumed that the proportion of susceptible hosts remained approximately constant during the second phase of the analysis ($S(t)/N \approx S/N$) and we neglected the effect of the rise in frequency of the variant on the average phenotypic trait values in (2) and $\overline{r}(t)$ (weak selection assumption).

Under these assumptions along with the previous best WLS estimates for the control parameters from the first step of the analysis ($k = 1$, $a = 3.78$), a linear mixed-effects model (MEM) led to the following estimations of the phenotypic differences (per day): $\Delta\beta = 0.15$ (95% CI $[0.033; 0.258]$) and $\Delta\gamma = -0.047$ (95% CI $[-0.099; +0.001]$) (**Fig. 2**). With a significance level of 5%, likelihood-based comparisons of nested MEMs show a significant effect for $\Delta\beta$ but not for $\Delta\gamma$ (although with a p-value very close to the significance threshold) (**Table S4**). In addition, we sought to propagate to the second phase the uncertainty of our estimates of the parameters k and a. Starting from each of the almost 2000 pairs $\{k; a\}$ based on previous wild bootstrap computations, we obtained as many new estimators for $\{\Delta\beta; \Delta\gamma\}$. For $\Delta\beta$, 95% of them were between 0.147 and 0.153 (**Fig. S7-A**), for which each corresponding 95% CI remained positive (**Fig. S8**). In contrast, 95% of these 2000 estimates were between -0.054 and -0.046 for $\Delta\gamma$ (**Fig. S7-B**), among which 61% of the corresponding 95% CIs included 0 (**Fig. S8**). These distributions led us to conclude that the Alpha variant has a higher transmission rate than the resident strain. With these estimates of $\Delta\beta$ and $\Delta\gamma$ and in the absence of NPI, the selection coefficient of the Alpha variant was computed, on average, around 0.77 per week (standard deviation: 0.02 per week).

We also explored the robustness of these estimations by applying $\pm 10\%$ and $\pm 20\%$ perturbations in the fixed parameters of our model (cf. **Table 1**) to investigate how they would affect our results. First, we kept the best WLS estimates for the control parameters ($k = 1$, $a = 3.78$), and we applied the perturbations to the fixed parameters of the second phase of the analysis. Our estimations of $\Delta\beta$ and $\Delta\gamma$ were not very sensitive to these perturbations (cf. **Fig. S10**). Second, we applied the perturbations in the fixed parameters of the first step in order to get new estimates of the control parameters k and a. We used these new estimates in the second phase of the analysis to estimate the phenotypic differences $\Delta\beta$ and $\Delta\gamma$. The parameter k was hardly affected by these perturbations but the parameter a was more sensitive, in particular when varying the transmission rate or the initial proportion of susceptible hosts (cf. **Fig. S11-1**). Next, we reiterated the second step with these new estimations of the pair $\{k; a\}$. All the 95% CIs of the estimates of $\Delta\beta$ remained positive after these

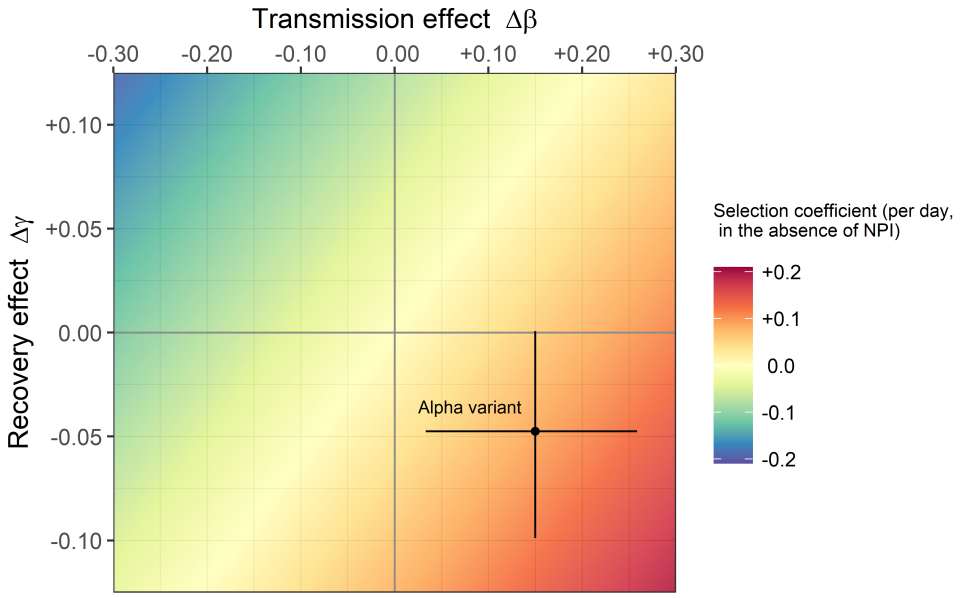

Figure 2: **Phenotypic profile of the Alpha variant (transmission and recovery rates) relative to the resident strain.** By definition, the phenotype of the resident strain is located at the origin of the graph ($\Delta\beta = 0$; $\Delta\gamma = 0$). Linear MEM estimates (black point, expressed per day) of phenotypic differences in transmission $\Delta\beta$ and in recovery $\Delta\gamma$ as well as 95% CIs (black cross) are based on the best WLS estimates of control parameters k and a from the analysis of phase 1 ($k = 1$ and $a = 3.78$). We obtained $\Delta\beta = 0.15$ (95% CI $[0.033; 0.258]$) and $\Delta\gamma = -0.047$ (95% CI $[-0.099; +0.001]$). For the fixed parameters, we set: $S/N = 0.75$, $\kappa = 0.2$, $\beta_w = 0.25$ and $\gamma_w = 0.1$. The colored background represents the values of the selection coefficient (in the absence of NPI) as a function of $\Delta\beta$ and $\Delta\gamma$; the selection coefficient is here around +0.11 per day (or +0.77 per week) for the Alpha variant. Estimates and 95% CIs based on the joint distributions of parameters k and a from wild bootstrap computations are represented in **Fig. S8**.

perturbations. However, some perturbations led to more negative values of $\Delta\gamma$ (i.e. the 95% CIs of $\Delta\gamma$ included only negative values, **Fig. S11-2**). Note that this effect seems to be driven by the variations in the estimation of the parameter a (cf. **Fig. S11**). Taken together, the results of these analyses confirm the conclusion that the Alpha variant has a higher transmission ($\Delta\beta > 0$). An increase in the mean duration of infectiousness ($\Delta\gamma < 0$) seems less likely but cannot be completely ruled out.

3 Discussion

We developed a two-step approach to characterise the phenotypic variation of the Alpha variant relative to the previously dominant lineage. In the first step of the analysis, we focus on the epidemiological dynamics before the emergence of the Alpha variant and we used an $SEIR$ model, a simplified representation of an age-structured model, to infer the effect of the Stringency Index on the reduction of transmission induced by these control measures. This led us to infer a convex increasing function that captures the effect of the Stringency Index on the reduction in the number of contacts with susceptible hosts (**Fig. S5**).

The second step of this approach is based on the analysis of the change in frequency of the Alpha variant after its emergence. Using evolutionary epidemiology theory [8, 7, 9], we derive an expression for the gradient of selection in an $SEIR$ model. The analysis of selection in such a class structured environment (the virus is infecting both the E and the I hosts) is facilitated under the assumption of weak selection and the approximation of quasi-equilibrium for fast variables [27, 14, 28]. We recover a classical result derived in simpler SIR models: the intensity of selection for higher transmission rates depends on the availability of susceptible hosts and the amount of NPIs aiming to reduce contact (e.g. social distancing or face coverings). In contrast, selection for longer durations of infectiousness is much less sensitive to these control measures. Using our independent estimation of the effectiveness of NPIs based on the Stringency Index, we inferred both $\Delta\beta$ and $\Delta\gamma$ of the Alpha variant from the temporal dynamics of its logit-frequency. This analysis suggests that the selective advantage of the Alpha variant was mainly driven by a higher transmission rate. An increase in the mean duration of infectiousness (i.e. a lower rate of recovery) seems less likely but cannot be completely ruled out. Interestingly, recent experimental studies of viral transmission confirm the transmission advantage of the Alpha variant. Viral shedding in breath aerosols were recently found to be higher in individuals infected with the Alpha variant than with previous lineages [25].

Several specific mutations of the Alpha variant could explain these phenotypic differences. Preliminary genomic characterisations detected around 17 non-synonymous substitutions or deletions compared to the previous lineage; about half were associated with the protein S gene, including mutations of immunological significance [38]. In particular, the mutation N501Y, known to increase the affinity of the viral glycoprotein S for the human receptor ACE2 [44], and the mutation P681H, adjacent to a serine protease cleavage site that is required for cell infection [23], are both likely to affect the within-host development of the virus in infected hosts. How this development affects key phenotypic traits like transmission and recovery rates in human host is difficult to explore experimentally. Our analysis can thus provide a complementary approach that may help to link genetic and phenotypic variation.

Yet, it is important to note that this analysis relies on several simplifying assumptions. For instance, we assumed that infectiousness began at the same time as the onset of symptoms – i.e. the latent period and the incubation period coincide perfectly in time. Yet, transmission from a pre-symptomatic state is a distinctive feature of SARS-CoV-2 [41, 9, 20]. Besides, our framework sticks to the $SEIR$ class of models formalised by ODEs, with κ and γ, the (constant) rates of leaving the exposed and infectious compartments, respectively. This implicitly yields sojourn times in the different compartments that are exponentially distributed – and thus, markovian or memoryless [13, 43]. As a result, the generation time follows a hypoexponential distribution (generalized Erlang distribution) with mean $1/\kappa+1/\gamma$ and variance $1/\kappa^2 + 1/\gamma^2$ [48]. Our analysis does not allow the mean and variance of this distribution to change independently but a variation in γ does affect the mean and the variance of the generation time. Several studies, however, have discussed the influence of the shape of the generation time distribution on both the epidemiological and evolutionary dynamics of the pathogen [6, 48, 34, 33, 3, 1]. We show in the **SI Appendix §S7** how to recover our results using the selection on the shape of the generation time distribution used by Blanquart *et al.* [3]. In both analyses, variations in the intensity of NPIs are assumed to impact the effective reproduction number without altering the generation

time distribution (which means they only impact transmission). Nevertheless, some control measures like contact tracing and post-symptomatic isolation may impact the duration of infectiousness, the generation time distribution and the selection on the variant [33].

Data availability and quality are major limiting factors in any statistical inference analysis. The Stringency Index provides a rough approximation of the intensity of control at the national scale. More precise and more local estimations of control would allow us to refine our estimations. In addition, we show in the **SI Appendix §S6** how the availability of data frequency among different types of hosts (i.e. the differentiation between the exposed and the infectious compartments) may provide another way to estimate $\Delta\beta$ and $\Delta\gamma$.

To conclude, we contend that it is important to exploit the joint epidemiological and evolutionary dynamics of SARS-CoV-2 to better understand its phenotypic evolution. This phenotypic evolution is undermining our efforts to control the epidemic. New variants are emerging and are affecting other phenotypic traits. In particular, the ability of new variants (e.g. Omicron) to escape immunity has major impact on the epidemiological dynamics [35]. Inference approaches using both epidemiological and evolutionary analysis could yield important insights on the adaptive trajectories on the phenotypic landscape of SARS-CoV-2, and possibly other pathogens.

4 Methods

4.1 A two-step analysis

The analysis is performed in two steps considering two consecutive evo-epidemiological periods of time: before and after the emergence of the Alpha variant in England (**Fig. 1**). The first step aims to estimate the force of infection in the presence of NPIs. In particular, we quantify $c(t)$, a function measuring the impact of NPIs at time t on the force of infection $\lambda(t)$. This first step takes place temporally before the emergence of the Alpha variant – i.e. before it reaches 5% of the cases tested positive in England – and consists in modeling the epidemiological phase of the previous lineage, which we refer to as the resident strain, disregarding the pre-existing genetic diversity [22]. The second step consists in estimating the differences in contagiousness and in infectious duration in the presence of NPIs during the period when the two strains cohabit – i.e. for each region, from the moment the frequency of the variant reaches 10% of cases tested positive. We combine information from screening and mortality data for the first step (using an epidemiological model), while we focus on the changes in frequency of the variant among positive cases for the second step. See **Table 1** for an overview of this two-step approach.

For both steps, we consider a host population of size N. We note S, E, I and R, respectively, the states (or compartments) of individuals that are Susceptible to the disease, Exposed – i.e. infected but not yet infectious -, Infectious and Recovered. For a given state, for instance S, and current time t (expressed in days), we note $S(t)$ the density of people in that state and $\dot{S}(t)$ its differentiation with respect to time. Let β be the *per capita* transmission rate (direct and horizontal) and γ the *per capita* recovery rate. Control measures implemented by governments such as social distancing, face coverings, lockdowns or travel bans are NPIs that aim to curb the spread of the epidemic by alleviating the force of infection $\lambda(t) = \beta I(t)/N$. Given $c(t)$ the effectiveness of these measures – ranging from 0 (no control)

Table 1: **Overview of the two-step analysis.** This table summarises the main features of the two phases of the analysis. For each one, we recall the aim, dates, circulating SARS-CoV-2 strains that we considered, fixed parameters, data and fitted variables (model) – equation numbers are specified between brackets just after the corresponding variable. For both phases, we also use values of the Stringency Index in the UK. $\mathcal{R}_0$, γ (or γ_w) and β (or β_w) are the basic reproduction number and the *per capita* rates (per day) of recovery and transmission, respectively, of the resident strain w; κ is the *per capita* transition rate (per day) from the exposed to the infectious state (same for both strains); k and a are the parameters linking the Stringency Index to the efficacy of NPIs (same for both strains); $S(t)/N$ is the proportion of susceptible hosts in the population (assumed constant in the second phase); D refers to the cumulative density of COVID-19-related deaths. See **Table S1** and **S3** for a more detailed summary of the parameters involved in phase 1 and 2, respectively.

PHASE 1 – National frequency of the Alpha variant < 5 %				
AIM: Estimating the impact of NPIs (control parameters k and a) on the spread of the virus				
Dates	**Strain(s)**	**Fixed parameters**	**Data**	**Fitted variable(s)**
2020-08-03 – 2020-11-08	Resident strain (WT)	• $\mathcal{R}_0 = 2.5$ • $\gamma = 0.1$ • $\beta = 0.25$ ($\approx \gamma\mathcal{R}_0$) • $\kappa = 0.2$ • $S(t_0^{\text{step 1}})/N = 0.9$	Daily new cases tested negative (UK)	$T^-(t)$ (6)
			Daily new cases tested positive (UK)	$T^+(t)$ (7)
			Daily new fatality cases (UK)	$D(t) - D(t-1)$ (4)
PHASE 2 – Regional frequency of the Alpha variant ≥ 10 %				
AIM: Knowing the impact of NPIs, estimating the phenotypic differences $\Delta\beta$ and $\Delta\gamma$				
Dates	**Strain(s)**	**Fixed parameters**	**Data**	**Fitted variable(s)**
Region-dependant (final week 2021-01-18)	Resident strain (WT) & Alpha variant	• k and a (estimations from phase 1) • $\beta_w = 0.25$ • $\gamma_w = 0.1$ • $\kappa = 0.2$ • $S/N = 0.75$ ($\approx$ final proportion of S at the end of the simulation of phase 1)	Weekly regional logit-frequencies of S Gene Target Failure among cases tested positive (England)	$\text{logit}(\widetilde{f}_m(t))$ (12)

to 1 (total control) –, the expression for the force of infection thus becomes: $\lambda(t) = (1 - c(t))\beta I(t)/N$. Directly estimating the control efficiency $c(t)$ is usually impossible; it results from a multitude of factors that may vary spatially and temporally and is not necessarily proportional to the severity of the measures in place. This is why we choose here to include the *Stringency Index* (which we noted $\psi(t)$), a composite score published by OxCGRT [18]. This index is based on nine component indicators and rescaled to a value between 0 (no control) and 100 (the stringest) in order to reflect the strictness of public health policy. Eight component indicators are related to "*containment and closure*" (school and workplace closing, cancel public events, restrictions on gathering site, close public transport, stay-at-home requirements and restrictions on internal movement and on international travel) and one is related to "*health system*" (public information campaign) [18]. These measures, in contrast with post-symptomatic isolation or contact tracing (not explicitly taken into account in this score), are mainly

limiting the number of contacts *un*conditionally to infection, that is mostly intended to reduce the transmission rate than to shorten the infectious period. We thus assume that NPIs included in the Stringency Index would only affect the transmission rate (and not the infectious period). Although somewhat imperfect, this index has the advantage of integrating many factors into one value, as well as being available per day online since the onset of the pandemic in many countries. We model the link between $c(t)$ and $\psi(t)$ through the following concave or convex relationship:

$$c(t) = k\left(\frac{\psi(t)}{100}\right)^a, \tag{1}$$

with $k \in [0;1]$, the maximum achievable efficiency (when $\psi(t) = 100$), and with $a \in \mathbb{R}_+^*$, a '*shape*' parameter.

4.1.1 Step 1: epidemiological analysis just before the emergence of the Alpha variant

We use a version of the well-known $SEIR$ model (see **Fig. S3**) to estimate the parameters that govern the epidemiological dynamics before the arrival of the Alpha variant. We denote α, the additional *per capita* mortality rate induced by the viral disease (i.e. the virulence) and D, the compartment of (COVID-19-related) deceased individuals. We assume that the (potential) onset of symptoms and the onset of infectiousness occur simultaneously after a latent period of mean duration $1/\kappa$. Within the infectious compartment I, some hosts develop symptoms (I_S) with probability ω while the others remain asymptomatic (I_A) with complementary probability. It is further assumed that individuals I_A systematically recover at a *per capita* rate γ while individuals I_S are divided into two sub-compartments depending on their fate: I_{Sd}, with probability p, for those who will eventually die from the disease (with virulence α), or, alternatively, I_{Sr}, for those who will eventually recover (at the same rate γ as asymptomatic hosts). We model these epidemiological trajectories using the following system of ODEs:

$$\begin{cases} \dot{S}(t) &= -(1-c(t))\beta S(t)\dfrac{I(t)}{N} \\ \dot{E}(t) &= (1-c(t))\beta S(t)\dfrac{I(t)}{N} - \kappa E(t) \\ \dot{I}_A(t) &= (1-\omega)\kappa E(t) - \gamma I_A(t) \\ \dot{I}_{Sr}(t) &= (1-p)\omega\kappa E(t) - \gamma I_{Sr}(t) \\ \dot{I}_{Sd}(t) &= p\omega\kappa E(t) - \alpha I_{Sd}(t) \\ \dot{R}(t) &= \gamma\Big(I_A(t) + I_{Sr}(t)\Big) \\ \dot{D}(t) &= \alpha I_{Sd}(t) \end{cases} \tag{3}$$

Following [10] for the construction of the Next Generation Matrix, the basic reproduction number $\mathcal{R}_0$ – i.e. the expected number of *infectees* from one *infector* in a fully susceptible population – is then given in the absence of NPI by:

$$\mathcal{R}_0 = \beta \left(\frac{1-\omega p}{\gamma} + \frac{\omega p}{\alpha} \right).$$

In the context of COVID-19, the product ωp – i.e. the probability of developing symptoms and dying from the disease – is very low. We then approximate the basic reproduction number of the resident strain of SARS-CoV-2 as $\mathcal{R}_0 \approx \beta/\gamma$.

At each time point (each day), only a small fraction of the population is tested and hosts with symptoms are more likely to be tested than others. In order to take these biases into account, we use the following range of assumptions:

- Individuals S and I_A are tested with the same probability / reporting rate ρ;
- Individuals S and I_A can be tested several times;
- All new individuals I_S (symptomatic) are tested (reporting rate of 1);
- Screening of individuals E and R is neglected (reporting rate of 0);
- All new disease-related deaths are reported (reporting rate of 1).

Furthermore, screening efforts in the UK tended to be strengthened over time during this period (as shown for instance by the increasing number of negative tests in **Fig. S6**). As the reporting rate for individuals without symptoms S and I_A can no longer be considered constant, we also assume a linear increase with time:

- The reporting rate ρ for individuals S and I_A (without symptoms) increases linearly over time: $\rho(t) = \eta\ t + \mu$.

The reporting rate is not identifiable in an SIR model when only a fraction of the compartment I is observed [19]. Thus, we also consider the disease-related deaths in the observation process. The combination of information, that is daily new cases tested negative and tested positive and daily new fatality cases, allow us to identify the reporting rate. Between two consecutive time points $t-1$ and t, the number of new fatality cases is given by:

$$D(t) - D(t-1) = \int_{t-1}^{t} \alpha I_{Sd}(t)\, \mathrm{d}t, \tag{4}$$

and, given $\displaystyle\int_{t-1}^{t} \omega\kappa E(t)\, \mathrm{d}t$, the *incidence* of symptomatic cases (i.e. new incomers in compartment I_S), we decomposed the number of performed tests $T(t)$ as follows:

$$T(t) \quad = \quad T^-(t) + T^+(t) \quad = \quad \underbrace{\rho(t) S(t)}_{T^-(t)} + \underbrace{\rho(t) I_A(t) + \int_{t-1}^{t} \omega\kappa E(t)\, \mathrm{d}t}_{T^+(t)}, \tag{5}$$

with $T^-(t)$ and $T^+(t)$, the number of cases tested negative and tested positive, respectively. Thus:

$$T^{-}(t) = \left(\eta\ t + \mu\right)\ S(t) \tag{6}$$

$$T^{+}(t) = \left(\eta\ t + \mu\right)\ I_A(t) + \int_{t-1}^{t} \omega\kappa E(t)\,\mathrm{d}t \tag{7}$$

4.1.2 Step 2: Evolutionary analysis

We now consider that two distinguishable pathogenic strains compete: the resident (or WT) strain, represented with the subscript w, and the mutant strain (or variant), represented with the subscript m. The total number of exposed hosts $E(t)$, where t is the current time, can therefore be decomposed into: $E(t) = E_m(t) + E_w(t)$. Likewise, for the infectious hosts $I(t)$: $I(t) = I_w(t) + I_m(t)$, and we denote $q_m(t) = I_m(t)/I(t)$, the frequency of the variant in I. We propose that the variant may differ phenotypically from the resident strain in its effective transmission rate $\beta_m = \beta_w + \Delta\beta$ and/or its recovery rate $\gamma_m = \gamma_w + \Delta\gamma$. In contrast, we neglect any difference in terms of latent period ($\kappa_m = \kappa_w = \kappa$), and we neglect the virulence of both strains ($\alpha_m = \alpha_w = 0$). For SARS-CoV-2, frequencies of the Alpha variant did not seem to depend on the age of hosts [5]. Assuming furthermore that over-infections do not occur – including co-infections with both strains – and that (persistent) immunity acquired with either strain protects effectively against both, we start with the simple following $SEIR$ model:

$$\begin{cases} \dot{S}(t) = -(1-c(t))\overline{\beta}(t)S(t)\dfrac{I(t)}{N} \\ \dot{E}(t) = (1-c(t))\overline{\beta}(t)S(t)\dfrac{I(t)}{N} - \kappa E(t) \\ \dot{I}(t) = \kappa E(t) - \overline{\gamma}I(t) \\ \dot{R}(t) = \overline{\gamma}(t)I(t) \end{cases} \tag{8}$$

where the overlines refer to mean values of the phenotypic traits after averaging over the distribution of strain frequencies:

$$\begin{cases} \overline{\beta}(t) &= (1-q_m(t))\beta_w + q_m(t)\beta_m \\ \overline{\gamma}(t) &= (1-q_m(t))\gamma_w + q_m(t)\gamma_m \end{cases}$$

As described in [27, 28], under the assumption of weak selection, the overall frequency of the variant $\widetilde{f}_m(t)$ can be tracked using:

$$\frac{\mathrm{d}\widetilde{f}_m(t)}{\mathrm{d}t} = \underbrace{\widetilde{f}_m(t)(1-\widetilde{f}_m(t))}_{\text{Genetic variance}}\ \underbrace{\mathbf{v}(t)^{\top}\Delta\mathbf{R}(t)\mathbf{f}(t)}_{\text{s(t), selection coefficient}}\ , \tag{9}$$

with $\mathbf{v}(t)$ and $\mathbf{f}(t)$, the vectors of reproductive values and class frequencies, respectively, within the infected states (E and I), and $\Delta\mathbf{R}(t)$, the matrix of differences in transition rates between the mutant

strain and the resident strain (for more details, see **SI Appendix**). An easier way to study $s(t)$ in time series analyses is not to directly work with frequencies but with logit-frequencies instead, that is ln(frequency of the variant / frequency of the resident strain). Indeed, it may easily be shown that:

$$\frac{\mathrm{d}\ \mathrm{logit}(\widetilde{f}_m(t))}{\mathrm{d}t} = s(t) \tag{10}$$

We then focus on the selection coefficient of the variant $s(t)$ (also known as the selection gradient). According to its value (weakly or strongly positive, weakly or strongly negative), this selection coefficient quantifies over time the success or the disadvantage of the variant over the resident strain through natural selection [8, 7, 9]. In the **SI Appendix**, we show that, using quasi-equilibrium approximations for fast variables, the selection coefficient of the variant $s(t)$ may be approximated as:

$$s(t) \approx \frac{2\kappa(1-c(t))\Delta\beta\frac{S(t)}{N} - \left(\kappa - \overline{\gamma}(t) + \sqrt{\left(\kappa-\overline{\gamma}(t)\right)^2 + 4\kappa(1-c(t))\overline{\beta}(t)\frac{S(t)}{N}}\right)\Delta\gamma}{2\sqrt{\left(\kappa-\overline{\gamma}(t)\right)^2 + 4\kappa(1-c(t))\overline{\beta}(t)\frac{S(t)}{N}}} \tag{11}$$

For the SIR model nested in the $SEIR$ model (8), the selection coefficient is merely: $s(t) = (1 - c(t))\Delta\beta S(t)/N - \Delta\gamma$ [8, 7], which shows analytically the importance of the control through $c(t)$ to distinguish the scenario where the selective advantage of the variant stems from a higher transmission rate ($\Delta\beta > 0$; $\Delta\gamma = 0$) from the scenario with a longer duration of infectiousness ($\Delta\beta = 0$; $\Delta\gamma < 0$), or from an intermediate scenario ($\Delta\beta \neq 0$; $\Delta\gamma \neq 0$). In other words, it is particularly the variations in $c(t)$ that might help us to decouple the effects of these two phenotypic traits. Simply adding an exposed state E makes the selection gradient surprisingly much more difficult to express but the importance of the variations in $c(t)$ for this purpose (although less clear-cut) remains nevertheless relevant as suggested by (11).

4.2 Statistical inference

4.2.1 Programming

Numerical simulations and data analyses were carried out using R [37] version 4.1.1 (2021-08-10). ODEs were solved numerically by the function '*ode*' (method '*ode45*') from the package '*deSolve*' [42].

4.2.2 Step 1

We used daily screening data between 2020-08-03 and 2020-11-08 in the UK (a period for which the Alpha variant was below 5% among cases tested positive in England); 7-day rolling average data were used in order to mitigate the effects of variation in testing activity, e.g. during weekends. We also included daily COVID-19-related deaths in the UK ('*Daily deaths with COVID-19 on the death certificate by date of death*') as well as the Stringency Index.

The goal of this part is to compute $c(t)$ from the Stringency Index and thus to focus on the

estimation of the parameters k and a. We used additional information from the literature to fix the value of some parameters of the model (3): we set the mean latent period to 5 days [11] and the mean duration of infectiousness to 10 days [4] – that is, an average infection period of 15 days –; we also set the basic reproduction number $\mathcal{R}_0$ to 2.5 [12, 26] and the initial proportion of susceptible hosts to 0.9. Besides, we approximated the initial states within compartment I. This is summarised with further details in **Table S1**. With these parameters fixed, the model (3) is identifiable (**Fig. S12**, following [39]). The remaining parameters of the first phase were estimated using weigthed least squares (WLS). Let $\theta = \left(k,\ a,\ E(t_0^{\text{step 1}}),\ \alpha,\ \omega,\ p,\ \eta,\ \mu\right)^\top$ be the vector of parameters to estimate, with $t_0^{\text{step 1}}$ the initial time point of the first step, and $\widehat{\theta}$ its estimator such that:

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_i \sum_t \frac{\left(y_i(t) - f_i(\theta, t)\right)^2}{f_i(\theta, t)},$$

where the subscript i refers to our three observation states – i.e. daily new cases tested negative and tested positive and daily new fatality cases –, which were modelled through functions f_i. $f_i(\theta, t)$ corresponds thus to the expected observations while $y_i(t)$ corresponds to the real observations (data). With WLS, squared residuals are weighted by the inverse of the variance of the observations $y_i(t)$; these weights balanced the contrasting intrinsic contributions of each observation – e.g. negative tests and deaths are not on the same order of magnitude. Assuming $y_i(t)$ to be Poisson-distributed – consistent with ODEs where sojourn times are exponentially distributed – then the variance of the observations is $f_i(\theta, t)$. This would correspond to the Pearson χ^2 function in [2]. Non-linear optimizations were tackled with the R function '*optim*', from the basic package '*stats*', using the Nelder-Mead (or downhill simplex) method – maximum number of iterations $maxit = 2000$, absolute and relative convergence tolerance $abstol = reltol = 10^{-6}$. This optimization procedure was iterated for 1500 sets of uniformly drawn initial values (because of the presence of local minima) and was restricted to certain ranges of values through parameter transformations (cf. **Table S2**). Only parameter estimates from the best fit – i.e. successful completion with the lowest WLS value – were kept and we refer to them as the best WLS estimates.

Parameter distributions were then computed using wild bootstrap [29, 24], which allow in particular to take into account any heteroscedasticity in the residuals. To do this, 2000 sets of bootstraped data were generated: residuals were perturbed by an i.i.d. sequence of n random weights $\{W_i\}_{i=1}^n$ following Mammen's 2-points distribution (that is, $(1-\sqrt{5})/2$ with probability $(\sqrt{5}+1)/(2\sqrt{5})$ and $(1+\sqrt{5})/2$ with probability $(\sqrt{5}-1)/(2\sqrt{5})$), which satisfies $\mathbb{E}(W_i) = 0$ and $\mathbb{E}(W_i^2) = 1$ [24]. Non-linear optimizations were then reiterated, but starting only from the best WLS estimates and the corresponding set of initial values.

As a sensitivity analysis, ±10% and ±20% perturbations were applied to the fixed parameters of the first phase separately (β, κ, γ and $S(t_0^{\text{step 1}})/N$) and non-linear optimizations were each time reiterated starting from a set of 500 initial conditions (uniformly drawn, as before).

4.2.3 Step 2

We used weekly regional frequencies of S Gene Target Failure (SGTF) in England from the technical briefing 5 of Public Health England (PHE), which was investigating the new VOC 202012/01 variant between September 2020 and January 2021 [36]. Briefly, qPCR from the ThermoFisher TaqPath kit (designed to target three genes: ORF1ab, N and S) were performed after swab sampling in the the wider population – i.e. outside NHS hospitals and PHE labs. Due to the deletion ΔH69/V70 in the genome of the Alpha variant, a mismatch between one of the three molecular probes and the viral sequence encoding for the glycoprotein Spike (S) resulted in a failure of detection, or SGTF, a genomic signature that was then used as a proxy for this variant [36, 47]. As in the first step, we also included values of the Stringency Index in the UK.

Under the assumption that variations in $S(t)/N$ on short time scales may be neglected for a controlled epidemic ($S(t)/N \approx S/N$) and by neglecting the effect of the rise in frequency of the variant on the average phenotypic trait values – i.e. $\overline{\beta}(t) \approx \beta_w$ and $\overline{\gamma}(t) \approx \gamma_w$ (weak selection approximation) –, we may integrate (11) in accordance with (10) to find an expression for the overall logit-frequency of the variant:

$$\begin{aligned}\operatorname{logit}(\widetilde{f}_m(t)) \approx \operatorname{logit}(\widetilde{f}_m(t_0^{\text{step 2}})) + \kappa \int_{t_0^{\text{step 2}}}^{t} \left(\frac{(1-c(t))}{\sqrt{(\kappa-\gamma_w)^2 + 4\kappa(1-c(t))\beta_w S/N}} \right) \mathrm{d}t \;\; \Delta\beta \frac{S}{N} - \\ \frac{1}{2}\left[(\kappa-\gamma_w) \int_{t_0^{\text{step 2}}}^{t} \left(\frac{1}{\sqrt{(\kappa-\gamma_w)^2 + 4\kappa(1-c(t))\beta_w S/N}} \right) \mathrm{d}t + \Delta t \right] \;\; \Delta\gamma, \end{aligned} \tag{12}$$

where $\Delta t = t - t_0^{\text{step 2}}$ is the period of time between the system at time t and its initial state.

$\Delta\beta$ appears as a product with S/N in (12), which implies that they are likely not to be separately identifiable. At the final time point of the first phase, our best fit ended up with a proportion of susceptible hosts around 0.75. Hence, we consistently set S/N to 0.75 for the second phase. As in the first phase, we also set: $\kappa = 0.2$, $\beta_w = 0.25$ and $\gamma_w = 0.1$. Phenotypic differences relative to the previous lineage ($\Delta\beta$ and $\Delta\gamma$ in (12)) were estimated using a linear mixed-effects model (MEM) to fit weekly logit-frequencies of SGTF among cases tested positive to COVID-19 as a proxy of the Alpha variant in the nine regions of England late 2020 early 2021. We assumed that these frequencies were representative of the infected population and that the regions were independent of each other – i.e. no inter-region flows. In more detail, $\operatorname{logit}(\widetilde{f}_m(t))$ was the response variable, $\Delta\beta$ and $\Delta\gamma$ were treated as fixed effects and the region (nine in total) was treated as a random effect on the intercept of the model. Hence, for the region i at time point t (i and t are now noted as indexes for clarity):

$$\underbrace{\operatorname{logit}(\widetilde{f}_m)_{t,i}}_{\text{Response variable}} = intercept + \Delta\beta\ C_t^{\beta} + \Delta\gamma\ C_t^{\gamma} + Region_i + \varepsilon_{t,i},$$

with:

- *intercept*, the common fixed effect (reference);

- $C_t^{\beta} = \kappa \int_{t_0}^{t} \frac{(1-c(t))}{\sqrt{(\kappa-\gamma_w)^2 + 4\kappa(1-c(t))\beta_w S/N}}\, dt\ \frac{S}{N}$, the covariate associated with $\Delta\beta$ (fixed effect);
- $C_t^{\gamma} = -\frac{1}{2}\left[\left(\kappa-\gamma_w\right)\int_{t_0}^{t} \frac{dt}{\sqrt{(\kappa-\gamma_w)^2 + 4\kappa(1-c(t))\beta_w S/N}} + \Delta t\right]$, the covariate associated with $\Delta\gamma$ (fixed effect);
- $Region_i \sim \mathcal{N}\left(0,\ \nu^2\right)$, the random effect (with variance ν^2) of the region i on the intercept of the model;
- $\varepsilon_{t,i} \sim \mathcal{N}\left(0,\ \sigma^2\right)$, the residual error (with variance σ^2).

This MEM was implemented using the function '*lmer*' from the R package '*lme4*': $\text{logit}(\widetilde{f_m}) \sim \Delta\beta + \Delta\gamma + (1|Region)$, and 95% CIs of parameters $\Delta\beta$ and $\Delta\gamma$ were computed using the function '*confint*' from the package '*stats*'. For each region, the initial date corresponds to the moment the Alpha variant reached 10% of cases tested positive – i.e. above horizontal lines in **Fig. S1-D**. Below this threshold, the dynamics of the variant could not be considered as deterministic. The parameters k and a that govern the link between the Stringency Index the and intensity of control (1) were set according to their best WLS estimates and joint distribution that were previously computed in the first step (cf. **§4.2.2**).

As for the first step, we investigated the robustness of our estimations. First, keeping our best WLS estimates for parameters k and a, linear MEM were reiterated with $\pm 10\%$ and $\pm 20\%$ perturbations in the fixed parameters of the second phase separately (β_w, κ, γ_w and S/N). Secondly, we used estimations of parameters k and a that we obtained after perturbing the fixed parameters of the first phase (cf. **§4.2.2**) to propagate these perturbations to the outcomes of the second step; the value of the fixed parameters of the second phase were each time updated in accordance.

Abbreviations (alphabetical order)

ACE2: Angiotensin-Converting Enzyme 2
CI: Confidence Interval
COVID-19: COronaVIrus Disease-2019
i.i.d.: independent and identically distributed
MEM: Mixed-Effects Model
NHS: National Health Service *(UK)*
NPI: Non-Pharmaceutical Intervention
ODE: Ordinary Differential Equation
ORF1ab: Open Reading Frames 1a and 1b
OxCGRT: Oxford COVID-19 Government Response Tracker
PHE: Public Health England *(now replaced by UKHSA)*
qPCR: quantitative Polymerase Chain Reaction
S: Spike *(viral gene and protein)*
SARS-CoV-2: Severe Acute Respiratory Syndrome-CoronaVirus-2
$\boldsymbol{S(E)IR}$: Susceptible-(Exposed-)Infectious-Recovered
SGTF: S Gene Target Failure
VOC: Variant Of Concern
WHO: World Health Organization
WLS: Weighted Least Squares
WT: Wild Type

Author contributions

Conceptualization, Methodology, Investigation and Writing (original draft and review & editing): WB, SL, RC and SG; Formal analysis: WB; Vizualisation: WB and SG; Supervision: SL, RC and SG.

Data accessibility statement

Data that were used in this study are available in a GitHub repository, along with the scripts for the analyses (in R).

Acknowledgements

We thank Troy Day and François Blanquart for inspiring discussions.

Funding statement

This work was funded by grants ANR-16-CE35-0012 "STEEP" to SL and ANR-17-CE35-0012 "EVOMALWILD" to SG from the Agence Nationale de la Recherche. We also thank the MESRI (French Ministry of Research) and the École Normale Supérieure Paris-Saclay for the PhD scholarship of WB.

References

[1] Abbott, S., Sherratt, K., Gerstung, M., and Funk, S. "Estimation of the test to test distribution as a proxy for generation interval distribution for the Omicron variant in England". *medRxiv* (2022). DOI: 10.1101/2022.01.08.22268920.

[2] Berkson, J. "Minimum Chi-Square, not Maximum Likelihood!" *Ann. Stat.* 8.3 (1980), 457–487. URL: https://www.jstor.org/stable/2240587.

[3] Blanquart, F., Hozé, N., Cowling, B. J., Débarre, F., and Cauchemez, S. "Selection for infectivity profiles in slow and fast epidemics, and the rise of SARS-CoV-2 variants". *eLife* (2022). DOI: 10.7554/eLife.75791.

[4] Byrne, A. W., McEvoy, D., Collins, A. B., Hunt, K., Casey, M., Barber, A., Butler, F., Griffin, J., Lane, E. A., McAloon, C., O'Brien, K., Wall, P., Walsh, K. A., and More, S. J. "Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases". *BMJ Open* 10.8 (2020). DOI: 10.1136/bmjopen-2020-039856.

[5] Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A. B., Russell, T. W., Tully, D. C., Washburne, A. D., Wenseleers, T., Gimma, A., Waites, W., Wong, K. L. M., van Zandvoort, K., Silverman, J. D., CMMID COVID-19 Working Group, COVID-19 Genomics UK (COG-UK) Consortium, Diaz-Ordaz, K., Keogh, R., Eggo, R. M., Funk, S., Jit, M., Atkins, K. E., and Edmunds, W. J. "Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England". *Science* 372.6538 (2021). DOI: 10.1126/science.abg3055.

[6] Day, T. "Virulence evolution and the timing of disease life-history events". *TREE* 18.3 (2003). DOI: 10.1016/S0169-5347(02)00049-6.

[7] Day, T. and Gandon, S. "Applying population-genetic models in theoretical evolutionary epidemiology". *Ecology Letters* 10 (2007), 876–888. DOI: 10.1111/j.1461-0248.2007.01091.x.

[8] Day, T. and Gandon, S. "Insights From Price's Equation into Evolutionary Epidemiology". *Disease Evolution: Models, Concepts, and Data Analysis* (2006), 23–44.

[9] Day, T., Gandon, S., Lion, S., and Otto, S. P. "On the evolutionary epidemiology of SARS-CoV-2". *Curr. Biol.* 30.15 (2020), R841–R870. DOI: 10.1016/j.cub.2020.06.031.

[10] Diekmann, O., Heesterbeek, J. A. P., and Roberts, M. G. "The construction of next-generation matrices for compartmental epidemic models". *J.R. Soc. Interface* 7.47 (2010), 873–885. DOI: 10.1098/rsif.2009.0386.

[11] Ding, Z., Wang, K., Shen, M., Wang, K., Zhao, S., Song, W., Li, R., Li, Z., Wang, L., Feng, G., Hu, Z., Wei, H., Xiao, Y., Bao, C., Hu, J., Zhu, L., Li, Y., Chen, X., Yin, Y., Wang, W., Cai, Y., Peng, Z., and Shen, H. "Estimating the time interval between transmission generations and the presymptomatic period by contact tracing surveillance data from 31 provinces in the mainland of China". *Fundamental Research* 1.2 (2021), 104–110. DOI: 10.1016/j.fmre.2021.02.002.

[12] Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L. C., van Elsland, S., Thompson, H., Verity, R., Volz, E., Wang, H., Wang, Y., Walker, P. G., Walters, C., Winskill, P., Whittaker, C., Donnelly, C. A., Riley, S., Ghani, A. C., and on behalf of the Imperial College COVID-19 Response Team. "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand" (2020). URL: https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf.

[13] Forien, R., Pang, G., and Pardoux, É. "Estimating the state of the COVID-19 epidemic in France using a model with memory". *R. Soc. Open Sci.* 8.3 (2021). DOI: 10.1098/rsos.202327.

[14] Gandon, S. and Lion, S. "Targeted vaccination and the speed of SARS-CoV-2 adaptation". *PNAS* 119.3 (2022). DOI: 10.1073/pnas.2110666119.

[15] Grubaugh, N. D., Hanage, W. P., and Rasmussen, A. L. "Making Sense of Mutation: What D614G Means for the COVID-19 Pandemic Remains Unclear". *Cell* 182 (2020), 794–795. DOI: 10.1016/j.cell.2020.06.040.

[16] Grubaugh, N. D., Petrone, M. E., and Holmes, E. C. "We shouldn't worry when a virus mutates during disease outbreaks". *Nat. Microb.* 5 (2020), 529–530. DOI: 10.1038/s41564-020-0690-4.

[17] Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R. A. "Nextstrain: real-time tracking of pathogen evolution". *Bioinformatics* 34.23 (2018), 4121–4123. DOI: 10.1093/bioinformatics/bty407.

[18] Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., and Tatlow, H. "A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)". *Nat. Hum. Behav* 5 (2021), 529–538. DOI: 10.1038/s41562-021-01079-8.

[19] Hamelin, F., Iggidr, A., Rapaport, A., and Sallet, G. "Observability, Identifiability and Epidemiology A survey". *HAL* (2021). URL: https://hal.archives-ouvertes.fr/hal-02995562/document.

[20] He, X., Lau, E. H. Y., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X., Mo, X., Chen, Y., Liao, B., Chen, W., Hu, F., Zhang, Q., Zhong, M., Wu, Y., Zhao, L., Zhang, F., Cowling, B. J., Li, F., and Leung, G. M. "Temporal dynamics in viral shedding and transmissibility of COVID-19". *Nat. Med.* 26 (2020), 672–675. DOI: 10.1038/s41591-020-0869-5.

[21] Hodcroft, E. B. *CoVariants: SARS-CoV-2 Mutations and Variants of Interest.* 2021. URL: https://covariants.org/.

[22] Hodcroft, E. B., Zuber, M., Nadeau, S., Vaughan, T. G., Crawford, K. H. D., Althaus, C. L., Reichmuth, M. L., Bowen, J. E., Walls, A. C., Corti, D., Bloom, J. D., Veesler, D., Mateo, D., Hernando, A., Comas, I., González-Candelas, F., SeqCOVID-SPAIN consortium, Stadler, T.,

and Neher, R. A. "Spread of a SARS-CoV-2 variant through Europe in the summer of 2020". *Nature* 595 (2021), 707–712. DOI: 10.1038/s41586-021-03677-y.

[23] Hoffmann, M., Kleine-Weber, H., and Pöhlmann, S. "A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells". *Mol. Cell.* 78.4 (2020), 779–784. DOI: 10.1016/j.molcel.2020.04.022.

[24] Kline, P. M. and Santos, A. "A Score Based Approach to Wild Bootstrap Inference". *J. Econom. Methods* 1.1 (2012), 23–41. DOI: 10.1515/2156-6674.1006.

[25] Lai, J., Coleman, K. K., Sheldon Tai, S.-H., German, J., Hong, F., Albert, B., Esparza, Y., Srikakulapu, A. K., Schanz, M., Sierra Maldonado, I., Oertel, M., Fadul, N., Louie Gold, T., Weston, S., Mullins, K., McPhaul, K. M., Frieman, M., and Milton, D. K. "Evolution of SARS-CoV-2 Shedding in Exhaled Breath Aerosols". *medRxiv* (2022). DOI: 10.1101/2022.07.27.22278121.

[26] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S. M., Lau, E. H. Y., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., Chen, C., Jin, L., Yang, R., Wang, Q., Zhou, S., Wang, R., Liu, H., Luo, Y., Liu, Y., Shao, G., Li, H., Tao, Z., Yang, Y., Deng, Z., Liu, B., Ma, Z., Zhang, Y., Shi, G., Lam, T. T. Y., Wu, J. T., Gao, G. F., Cowling, B. J., Yang, B., Leung, G. M., and Feng, Z. "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia". *N. Engl. J. Med.* 382.13 (2020), 1199–1207. DOI: 10.1056/NEJMoa2001316.

[27] Lion, S. "Class structure, demography and selection: reproductive-value weighting in nonequilibrium, polymorphic populations". *Am. Nat.* 191.5 (2018). DOI: 10.1086/696976.

[28] Lion, S. and Gandon, S. "Evolution of class-structured populations in periodic environments". *Evolution* 76.8 (2022), 1674–1688. DOI: 10.1101/2021.03.12.435065.

[29] Liu, R. Y. "Bootstrap Procedures under some Non-I.I.D. Models". *Ann. Stat.* 16.4 (1988), 1696–1708. DOI: 10.1214/aos/1176351062.

[30] Lu, H., Stratton, C. W., and Tang, Y.-W. "Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle". *J. Med. Virol.* 92.4 (2020), 401–402. DOI: 10.1002/jmv.25678.

[31] Obermeyer, F., Jankowiak, M., Barkas, N., Schaffner, S. F., Pyle, J. D., Yurkovetskiy, L., Bosso, M., Park, D. J., and Babadi, M. "Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness". *Science* (2022). DOI: 10.1126/science.abm1208.

[32] Otto, S. P., Day, T., Arino, J., Colijn, C., Dushoff, J., Li, M., Mechai, S., Van Domselaar, G., Wu, J., Earn David, J. D., and Ogden, N. H. "The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic". *Current Biology* 31.14 (2021). DOI: 10.1016/j.cub.2021.06.049.

[33] Park, S. W., Bolker, B. M., Funk, S., Metcalf, C. J. E., Weitz, J. S., Grenfell, B. T., and Dushoff, J. "The importance of the generation interval in investigating dynamics and control of new SARS-CoV-2 variants". *J.R. Soc. Interface* 19.191 (2022). DOI: 10.1098/rsif.2022.0173.

[34] Park, S. W., Champredon, D., Weitz, J. S., and Dushoff, J. "A practical generation-interval-based approach to inferring the strength of epidemics from their speed". *Epidemics* 27 (2019), 12–18. DOI: 10.1016/j.epidem.2018.12.002.

[35] Paton, R. S., Overton, C. E., and Ward, T. "The rapid replacement of the Delta variant by Omicron (B. 1.1. 529) in England". *Sci. Transl. Med.* (2022), eabo5395. DOI: 10.1126/scitranslmed.abo5395.

[36] Public Health England. "Investigation of novel SARS-COV-2 variant 202012/01: technical briefing 5" (2020). URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959426/Variant_of_Concern_VOC_202012_01_Technical_Briefing_5.pdf.

[37] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[38] Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, D. L., Volz, E., and on behalf of COVID-19 Genomics Consortium UK (CoG-UK). "Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations". *Virological* (2020). URL: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.

[39] Raue, A., Kreutz, C., Maiwald, T., Bachman, J., Schilling, M., Klingmüller, U., and Timmer, J. "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood". *Bioinformatics* 25.15 (2009), 1923–1929. DOI: 10.1093/bioinformatics/btp358.

[40] Rausch, J. W., Capoferri, A. A., Katusiime, M. G., Patro, S. C., and Kearney, M. F. "Low genetic diversity may be an Achilles heel of SARS-CoV-2". *PNAS* 117.40 (2020), 24614–24616. DOI: 10.1073/pnas.2017726117.

[41] Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., Zimmer, T., Thiel, V., Janke, C., Guggemos, W., Seilmaier, M., Drosten, C., Vollmar, P., Zwirglmaier, K., Zange, S., Wölfel, R., and Hoelscher, M. "Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany". *N. Engl. J. Med.* 382 (2020), 970–971. DOI: 10.1056/NEJMc2001468.

[42] Soetaert, K., Petzoldt, T., and Setzer, R. W. "Solving Differential Equations in R: Package deSolve". *J. Stat. Softw.* 33.9 (2010), 1–25. DOI: 10.18637/jss.v033.i09.

[43] Sofonea, M. T., Reyné, B., Elie, B., Djidjou-Demasse, R., Selinger, C., Michalakis, Y., and Alizon, S. "Memory is key in capturing COVID-19 epidemiological dynamics". *Epidemics* 35 (2021). DOI: 10.1016/j.epidem.2021.100459.

[44] Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veesler, D., and Bloom, J. D. "Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding". *Cell* 182.5 (2020), 1295–1310. DOI: 10.1016/j.cell.2020.08.012.

[45] van Dorp, L., Richard, D., Tan, C. C. S., Shaw, L. P., Acman, M., and Balloux, F. "No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2". *Nat. Commun.* 11.5986 (2020). DOI: 10.1038/s41467-020-19818-2.

[46] Vasilarou, M., Alachiotis, N., Garefalaki, J., Beloukas, A., and Pavlidis, P. "Population Genomics Insights into the First Wave of COVID-19". *Life* 11.129 (2021). DOI: 10.3390/life11020129.

[47] Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., Hinsley, W. R., Laydon, D. J., Dabrera, G., O'Toole, Á., Amato, R., Ragonnet-Cronin, M., Harrison, I., Jackson, B., Ariani, C. V., Boyd, O., Loman, N. J., McCrone, J. T., Gonçalves, S., Jorgensen, D., Myers, R., Hill, V., Jackson, D. K., Gaythorpe, K., Groves, N., Sillitoe, J., Kwiatkowski, D. P., The COVID-19 Genomics UK (COG-UK) consortium, Flaxman, S., Ratmann, O., Bhatt, S., Hopkins, S., Gandy, A., Andrew, R., and Ferguson, N. M. "Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England". *Nature* 593 (2021), 266–269. DOI: 10.1038/s41586-021-03470-x.

[48] Wallinga, J. and Lipsitch, M. "How generation intervals shape the relationship between growth rates and reproductive numbers". *Proc. R. Soc. B* 274 (2007), 599–604. DOI: 10.1098/rspb.2006.3754.

[49] World Health Organization (WHO). "WHO Situation Report on 11 February 2020". 22 (2020). URL: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200211-sitrep-22-ncov.pdf?sfvrsn=fb6d49b1_2.