Using early detection data to estimate the date of emergence of an epidemic outbreak

Jijón, S.[1], Czuppon, P.[2], Blanquart, F.[3] and Débarre, F.[1]

[1]Institute of ecology and environmental sciences of Paris (iEES-Paris, UMR 7618),

Sorbonne Université, CNRS, UPEC, IRD, INRAE, Paris 75005, France

[2]Institute for Evolution and Biodiversity, University of Münster, Münster 48149, Germany

[3]Center for Interdisciplinary Research in Biology, CNRS, Collège de France,

PSL Research University, Paris 75005, France

Short title: Using data to estimate the date of emergence of an epidemic outbreak

Abstract

While the first infection of an emerging disease is often unknown, information on early cases can be used to date it. In the context of the COVID-19 pandemic, previous studies have estimated dates of emergence (e.g., first human SARS-CoV-2 infection, emergence of the Alpha SARS-CoV-2 variant) using mainly genomic data. Another dating attempt used a stochastic population dynamics approach and the date of the first reported case. Here, we extend this approach to use a larger set of early reported cases to estimate the delay from first infection to the Nth case. We first validate our model using data on Alpha variant infections in the UK, dating the first Alpha infection at (median) August 21, 2020 (95% interquantile range across retained simulations, IqR: July 23 – September 5, 2020. Next, we apply our model to data on COVID-19 cases with symptom onset before mid-January 2020. We date the first SARS-CoV-2 infection in Wuhan at (median) November 28, 2019 (95%IqR: November 2–December 9, 2019). Our results fall within ranges previously estimated by studies relying on genomic data. Our population dynamics-based modelling framework is generic and flexible, and thus can be applied to estimate the starting time of outbreaks in contexts other than COVID-19.

Keywords. Epidemic emergence; Early dynamics; Stochastic simulations; COVID-19; SARS-CoV-2.

Author summary

While the first infection of an emerging disease is often unknown, information on early cases can be used to date it. In the context of the COVID-19 pandemic, previous studies have estimated dates of emergence of epidemic outbreaks (e.g., first human SARS-CoV-2 infection, emergence of the Alpha SARS-CoV-2 variant) using mainly genomic data. Another dating attempt used a population-level stochastic approach and the date of the first reported case. Here, we extend this generic and flexible approach to use a larger set of early reported cases to estimate the time elapsed between the first infection and the Nth case. Our model dates the first Alpha infection at around August 21, 2020, and the first SARS-CoV-2 infection in Wuhan at around November 28, 2019. Our findings fall within ranges previously estimated by studies relying on genomic data.

1 Introduction

Dating the first infection of an emerging infectious disease is a step towards tracing the disease's origin and understanding early epidemic dynamics. Beyond the early transmission of a new pathogen, estimating the date of first infection is also of interest while studying the initiations of local clusters in naïve populations, such as when the pathogen is first introduced to a new location, but also when the pathogen evolves to distinct genotypes such as emerging variants of concern (VOCs).

Various attempts have been made to date the first human infections by SARS-CoV-2 that led to the COVID-19 pandemic (noting that earlier spillovers, leading to dead-ends, may have occurred), using case data and/or viral genomic data. Using a stochastic model for the epidemic spread coupled with genomic data allowing to trace transmission at the individual level, Pekar et al. [1] estimated that the first human infection took place between late October and early December 2019. This estimate was slightly later than a previous one by the same authors, who had previously found an emergence date between mid-October and mid-November 2019 [2]. The revision notably included updating the dates of the first case reported [3], and better cleaning up genomic data to exclude sequencing errors [1]. Another, earlier, modeling study used case data only and dated the first COVID-19 case between early October and mid-November 2019, by adapting a technique used in conservation science for dating extinction using observation events [4]. The analysis had been conducted on outdated case data [5]. Re-running the analysis on updated case data had major effects on both the estimated date of the first infection and 95% confidence interval (see Appendix, Supplementary Figure S3).

Other studies have focused on the introductions of SARS-CoV-2 to different countries. Some have used molecular clock analyses relying on genomic data to determine the time of most recent common ancestor (tMRCA ; providing an upper bound for the time of the first infection, see Figure 1) of lineages introduced in a focal country. For instance, studies have been conducted using genomic data from France [6], the United States [7] and the United Kingdom (UK) [8]. Another study, based on case data only, used a stochastic non-Markovian approach relying on mortality data to estimate the date of SARS-CoV-2 introduction to France [9]; the first COVID-19 wave was estimated to have been initiated mid-January 2020 in France.

Finally, similar frameworks have been used to date the emergence of SARS-CoV-2 variants. One study focusing on the emergence of the 'EU1' SARS-CoV-2 variant (B.1.177) circulating among European countries during the summer of 2020 used genomic data and dated most introductions to June, 2020 [10]. Dating attempts have also been done for the 'Alpha' variant (B.1.1.7), whose date of

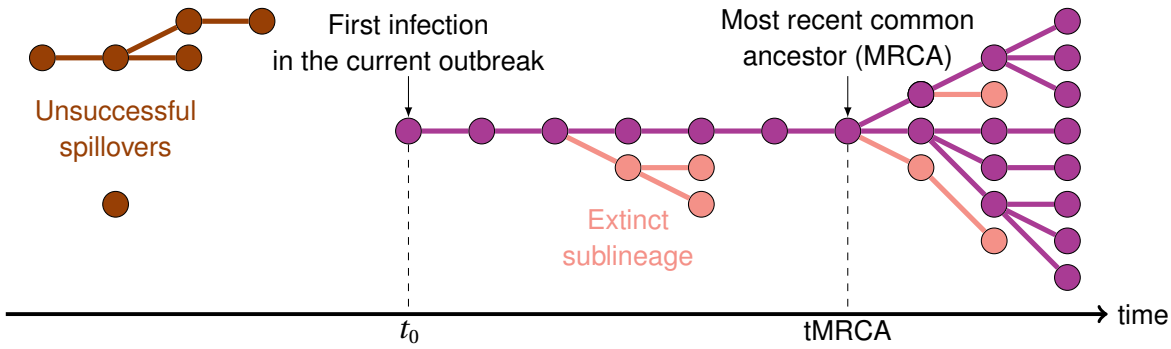


Figure 1: Schematic of early transmission dynamics. Each dot corresponds to an infected individual, and the link between dots represent pathogen transmission. Only a fraction of all infections will be detected as cases. We aim to estimate t_0, the time of the first infection which is at the origin of the ongoing outbreak. Genomic methods can date a most recent common ancestor (MRCA). Adapted from Ref. [2].

emergence was estimated at early August 2020 using a stochastic, non-Markovian approach relying on the date of the first observed case [11], and whose tMRCA was estimated at late August 2020 [12].

Genomic analyses on their own are usually insufficient to estimate the time of the first infections: the tMRCA does not necessarily approximate the emergence date [13], which can have taken place earlier (see Figure 1). Infection times occurring earlier than the tMRCA may be estimated thanks to other mathematical models using population dynamics methodologies. Moreover, modeling studies have helped unveiling other unobserved indicators during the early stages of epidemics, such as the epidemic size at the time of first detection [11] or at the time of the first reported death [14]. In particular, because infection numbers are low, stochastic approaches are key to studying early dynamics and their long-term effects on the epidemic outbreak [15–19]. Hence, methodological developments of stochastic models to study the early stages of infectious diseases remain of great interest in the field of mathematical epidemiology [e.g. 20]. Finally, we think that there is value in approaching an inference question via different methods (especially on potentially contentious topics), in order to obtain methodologically independent confirmations of estimates.

The main objective of our study is the estimation of the date of the first infection leading to a sustained epidemic, using available data on the first N detected and reported cases (hereafter referred to as the date of epidemic/outbreak *emergence*). To this end, we build a stochastic model and design a simulation framework extending our previous work that only used information on the first detected case [11] . Our rationale for extending the analysis to more cases is two-fold. First and foremost, including more cases can make the analysis more robust to changes in the underlying case data. Second, a new methodology using more than just the first case will also give access to important information such as the proportion of detected cases among infections. Our simulations construct a transmission tree, and identify detected cases among the infections. The comparison

of simulation outputs and available case data then allow the estimation of the date of emergence of the outbreak, as well as other key epidemiological pieces of information such as the proportion of the epidemic that remains undetected. We apply the framework to two examples: the emergence of the Alpha variant in 2020 in the UK, and the first human SARS-CoV-2 infection in Wuhan in late 2019.

2 Results

2.1 Modeling the early dynamics of an epidemic outbreak

2.1.1 Model description

We develop a stochastic epidemic model that estimates the time elapsed between the first infection and N reported cases (i.e., infections that were detected and reported), using available data and estimates on key epidemiological parameters. We model infectious disease transmission with a general branching process starting from a single infectious individual. This implies that times from infection to transmission events are not exponentially distributed as assumed by ordinary differential equation models. We then model the detection of infected individuals, which constitutes the modeled time series of cases. Both infection and detection processes follow distributions with known fixed parameters, listed in Table 2. Importantly, we assume fixed parameters as estimated by previous studies (in particular, the detection probability), but we later examine the robustness of our findings to these exact values. We consider the time-series of infections and detections up to the day of occurrence of the Nth case. We calibrate our model to reproduce the observed epidemic, using available data on disease cases. For more details on the model, we refer to the Methods section.

Our model generates a time series of cases, from which we deduce the delay between emergence (first infection) and N^{th} case. By first infection, we refer to successful epidemic outbreaks only; that is, we do not account for the first infections that may have led to epidemics that went extinct. In addition, by keeping track of the whole epidemic, we retrieve the time series of infections and the number of secondary cases produced by each infectious individual. We can thereby compute, for instance, the epidemic size at the time where the Nth case is infected to deduce the proportion of infections that are detected (i.e., cases) within the study period (i.e., up to the day of detection of the Nth case). This proportion is impacted by detection delays and stochasticity and thus, it is not

straightforwardly obtained from the probability of detection considered in the simulations.

We run as many numerical simulations as needed to obtain 5 000 successful epidemics, i.e., epidemics that were sustained after a predetermined period of time and that verified the calibration conditions imposed by the data from a certain epidemiological context (details in the Methods section). We apply our model to two epidemiological contexts: the emergence of the 'Alpha' variant in the UK and the emergence of SARS-CoV-2 in Wuhan. Both applications and the corresponding results are described in more detail below.

2.1.2 Testing the model on simulated data

To check the validity of our framework, we ran our model to generate epidemics and create case data from them. With these simulated data, the actual date of the first infection is known. We then applied our framework to generate estimates of the date of first infection, and compared it to the actual dates. We confirmed that our framework could recover the date of the first infection, the median being within a 2–3 day delay (see Appendix, Figure S6 for details). Median estimates of the date of the first infection were not affected by changes in the number of considered cases N. Confidence intervals, and therefore the range of potential dates, are larger as we include more cases, in particular because our conditions to retain a simulation (see Methods) are less stringent on the first cases, as the total number of cases increases.

2.2 Estimating the date of the first infection with the Alpha variant in the UK

Moving on to real case data, we first ran our model to estimate the date of emergence of the Alpha variant in the UK, which was the main result of the numerical applications presented in [11]. We applied our model to a dataset of $N = 406$ samples carrying the Alpha variant collected and sequenced between September 20 and November 11, 2020 [21]. Table 2 summarizes the parameter values used in our simulations. The 5 000 simulated epidemics that we analyse below result from model calibration (i.e., epidemics arising from a single infectious individual and verifying the calibration constraints; details in the Methods section), and represent ∼36% of all simulations run with the input parameters. The cumulative cases of the accepted epidemics are depicted in Supplementary Figure S4.

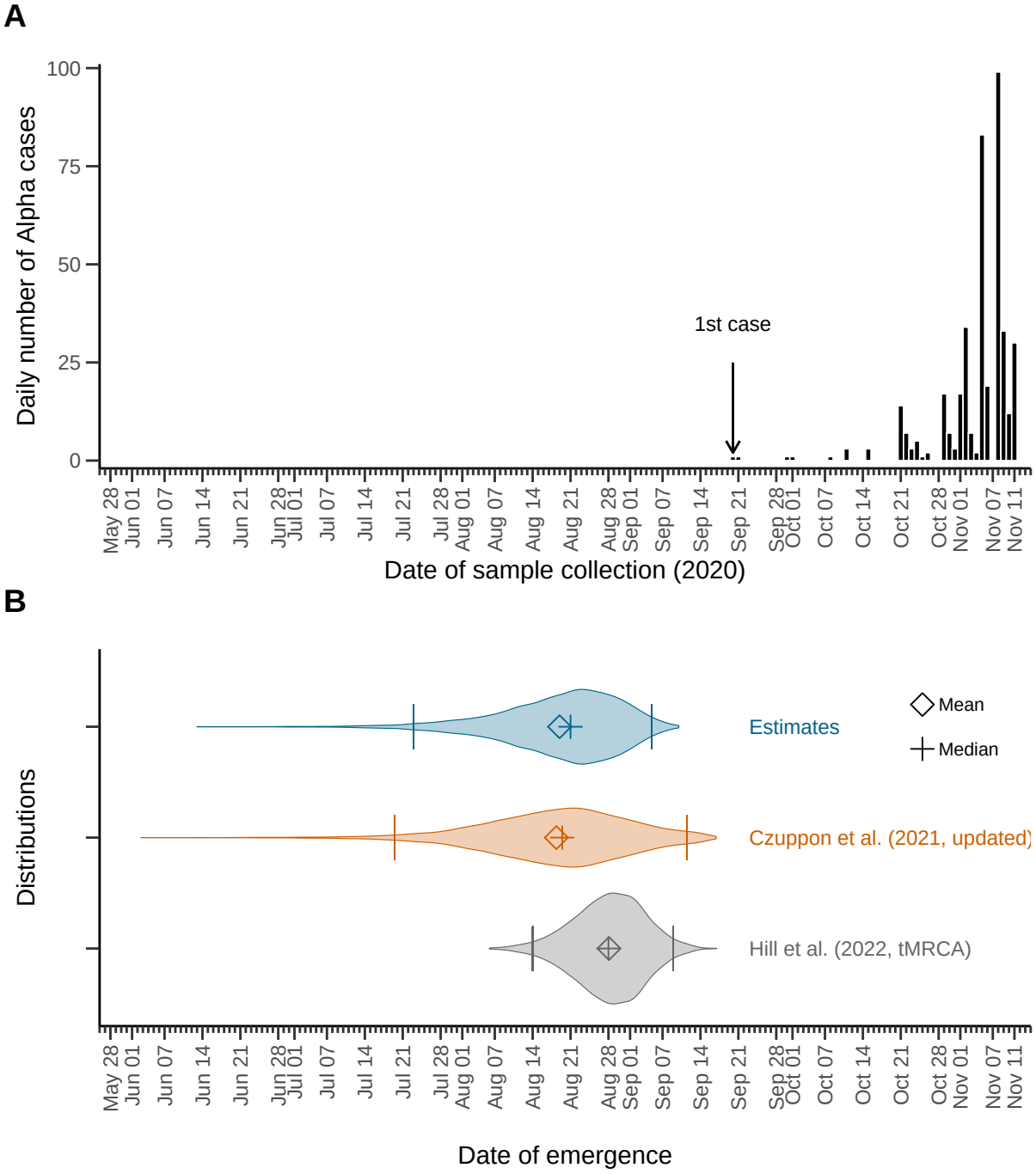
Hereafter, we summarize our results using median values and 95% interquantile ranges (95%IqR; values between the 2.5th and the 97.5th percentiles) from the distributions of the different epidemio-

logical indicators obtained from the 5 000 simulated epidemics, similar to an approximation of the posterior distribution obtained in an Approximate Bayesian Computation framework (see Methods). We also use the minimum values from those distributions to unveil the earliest infection occurrences among our simulated epidemics (that is, the earliest dates possible for the first infection to occur).

We estimated the number of days between the 1st infection and the Nth case at 82 (95%IqR: 67–111), dating the emergence of the Alpha variant in the UK at August 21 (95%IqR: July 23–September 5), 2020, and not earlier than June 13, 2020. Alpha transmissions were ongoing about 30 (95%IqR: 10–60) days before the date at which the first known case was sampled and sequenced. Table S1 in the Appendix provides other calibration metrics like the delay between the 1st and N^{th} sequenced samples in the simulations. Figure 2 depicts our estimates of the date of emergence along with the epidemic curve (i.e., the daily number of sequenced samples; by sampling date), for context, as well as previous estimates, for comparison. In particular, we ran an updated version of the model presented in [11] (the distribution of the number of secondary cases is negative-binomial instead of Poisson previously, and its mean, R, is now equal to 1.9 instead of 1.5 in [11]). We also compare our results to tMRCA estimates [12] (personal communication of the distributions). Our median estimates for the emergence date fall within a very close, slightly narrower range than that found by running an updated version of [11], while falling ~ 1 week earlier than the estimated tMRCA [12]. These comparisons are summarized in Supplementary Table S2.

We further estimated the epidemic size at the date of infection of the Nth case at about 90 800 (95%IqR: 80 500-102 000). The simulated detected cases thus represent a proportion of about 0.48% (95%IqR: 0.43%–0.52%) of the total number of infections. In this section, a case is an infected individual who underwent a PCR test and whose sample was sequenced and contained the Alpha variant, which accounts for the low probability of detection. Our results are summarized in Table 1. Note that when $N = 1$, we recover results similar to [11], cf. Supplementary Figure S7.

Figure 2: Estimates of the date of emergence of the Alpha variant in the United Kingdom (UK). (A) The epidemic curve corresponding to the data used to calibrate the model, for context. A total of $N = 406$ samples carrying the Alpha variant were collected between September 20 and November 11, 2020 [21]. **(B)** Violin plots for the distributions of the date of emergence. We estimated the emergence of the Alpha variant in the UK at August 21 (95%IqR: Jul 23–Sep 5), 2020 (top, blue; upper and lower bound of the 95%IqR depicted by bars). For comparison, we also show the distributions of the estimates from an updated version of the model developed in [11]—which relies on a single observation on September 20—where we set $R = 1.9$ and a negative-binomial distribution for the number of the secondary cases (middle, orange). The distribution for the estimated time of most recent ancestor (tMRCA) [12] is also shown (bottom, gray).



2.3 Estimating the date of emergence of SARS-CoV-2 in Wuhan

Next, we applied our model to the dataset of the early cases of COVID-19 considered in [1] (personal communication). A total of 3 072 confirmed COVID-19 cases were reported to have had symptoms onset between December 10, 2019 and January 19, 2020, the day before the first public declaration of human-to-human transmission, shortly before the first lockdown interventions [22]. We parameterized our model using estimates from the literature; they are listed in Table 2. The 5 000 selected simulations represent ∼14% of all runs (cf. Supplementary Table S1, Appendix) and are depicted in Supplementary Figure S5.

Our simulations yield an estimated median number of days between the 1st SARS-CoV-2 infection to the Nth symptomatic COVID-19 case recorded of 52 (95%IqR: 41–78) days, dating the emergence (i.e., the first sustained human infection) of SARS-CoV-2 to November 28 (95%IqR: November 2–December 9), 2019, and not earlier than September 26, 2019. This also implies that the epidemic remained completely undetected (i.e., no detected infections) for about 9 (95%IqR 3–21) days. These findings are depicted in Figure 3, along with the observed epidemic curve (i.e., COVID-19 cases dataset) as well as previously published estimates of the date of emergence of the COVID-19 pandemic [1] (personal communication of the distributions), for comparison. Our estimates of the date of first infection are consistent with those previously found in [1], but yield a lower bound closer to the date of first case detection (cf. Supplementary Table S2, Appendix). We also compare our results to those obtained through other more computationally expensive methods (namely, Approximate Bayesian Computation approaches) to illustrate the benefit of using our model.

We further estimate the median number of infections on the day of infection of the Nth case at about 63 400 (95%IqR: 57 300–69 900), which results in a median proportion of detected infections of 5.31% (95%IqR: 5.06%–5.57%). Our results are summarized in Table 1.

Figure 3: Estimates of the emergence of SARS-CoV-2 in Wuhan. **(A)** Observed epidemic curve, for context. A total of $N = 3\,072$ COVID-19 cases with symptom onset between December 10, 2019 and January 19, 2020, the day before the first public statement on human-to-human transmission. **(B)** Violin plots for the distributions of the date of emergence. We estimated the median date of the first SARS-CoV-2 infection (i.e., emergence) at November 28 (95%IqR: November 2–December 9), 2019 (top, blue; upper and lower bound of the 95%IqR depicted by bars). For comparison, the distribution for the estimates obtained implementing more classical approximate bayesian computation (middle, green), and estimates from [1] (bottom, purple) are also shown.

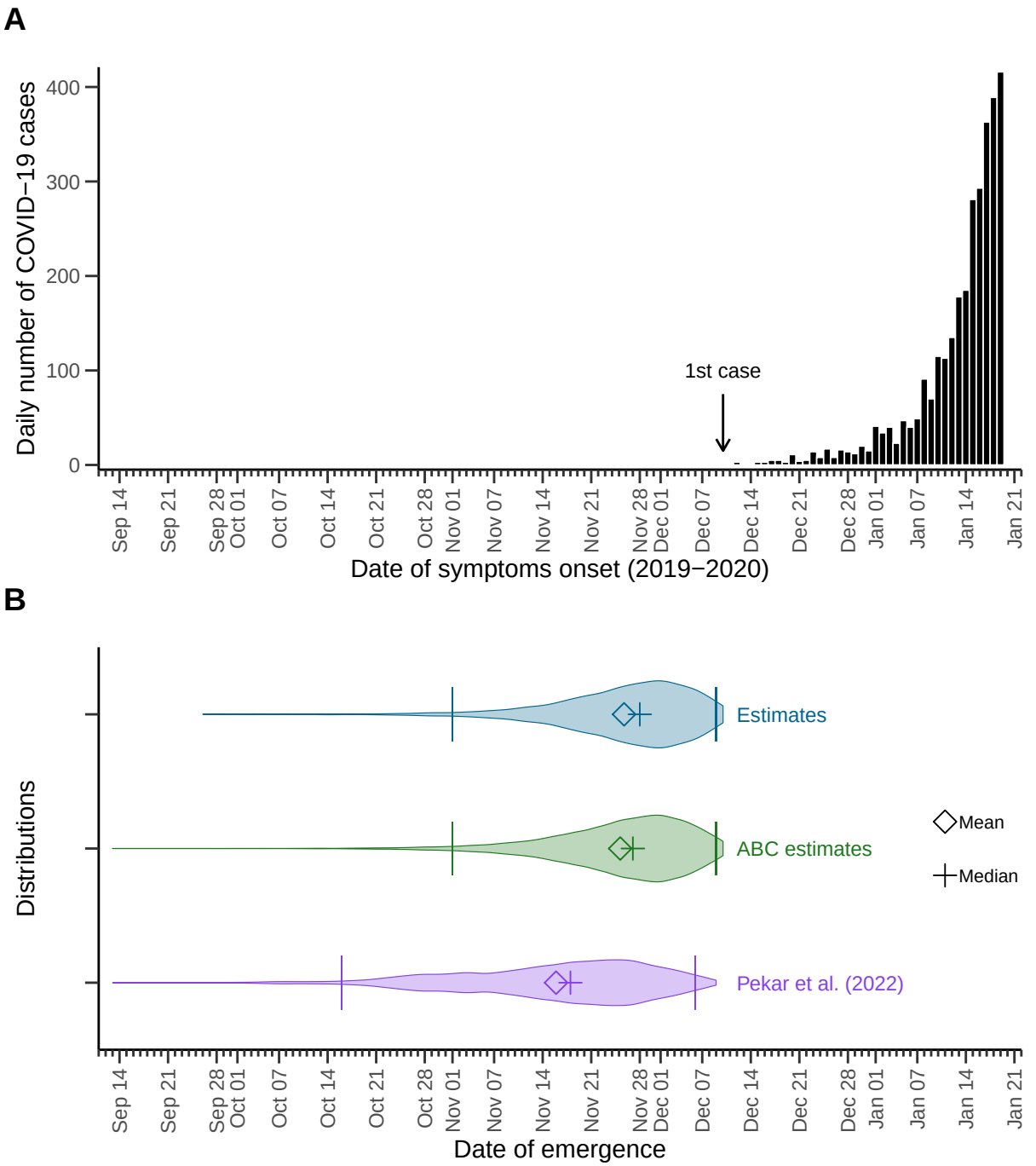


Table 1: Estimates of the date of emergence and other epidemiological indicators resulting from the calibrated model. Here we summarize the estimates obtained from the numerical application of our model to two epidemiological contexts: the Alpha variant infections in UK, and the first COVID-19 cases reported in Wuhan. Median and 95% interquantile ranges across retained simulations are shown, unless stated otherwise. The estimated time elapsed between the first infection to the Nth observed case yields the estimated emergence date. The earliest date corresponds to the lower bound of the emergence dates distribution. In addition, we estimate the epidemic size at the date of detection of the Nth case. The proportion of detected infections and mean secondary cases are retrieved for comparison with the input epidemic parameters.

Epidemiological indicator	**Alpha (UK, 2020)**	**COVID-19 (Wuhan, 2019)**
Number of days from 1st infection to N^{th} case*	82 (67–111)	52 (41–78)
Date of first infection	Aug 21 (Jul 23–Sep 5), 2020	Nov 28 (Nov 2–Dec 9), 2019
Date of earliest infection	Jun 13, 2020	Sep 26, 2019
Epidemic size at day of infection of the Nth case	∼ 90 800 (80 500-102 000)	∼63 400 (57 300–69 900)
Proportion of detected infections	0.48% (0.43%–0.52%)	5.31% (5.06%–5.57%)

* $N = 406$ on November 11, 2020 for the Alpha application, and $N = 3\,072$ on January 19, 2019 for the COVID-19 application.

2.4 Sensitivity analyses

Here, we evaluate the impact of uncertainty around the main model parameters on the results, by running our simulations while varying the input values of the expected number of secondary cases (R), the overdispersion parameter (κ), the probability of detecting an infection (p_{detect}) and the number of days elapsed between infection and detection, via the distribution parameter θ_τ. We find that, as expected, increasing p_{detect}, R or decreasing θ_τ results in a reduction in the number of days between the first infection and the Nth case, meaning that the epidemic emerges later in time; cf. Figure 4 and Table S3. In addition, we find that increasing R (i.e., higher number of per-day transmission events) yields distributions with more pronounced skewness, reflecting the occurrence of larger numbers of transmissions generated by few individuals (*'super-spreaders'*); cf. the Methods section. Median estimates and 95%IqR are summarized in the Supplementary Table S2 in the Appendix.

We further evaluate the impact of the criteria chosen for accepting the simulated epidemics on our results (cf. Methods section for details). We run our simulations while varying the tolerance for the difference between the simulated and the observed daily number of infections, δ^{Y}_{tol}, for both applications. Our estimates are robust with respect to these variations.

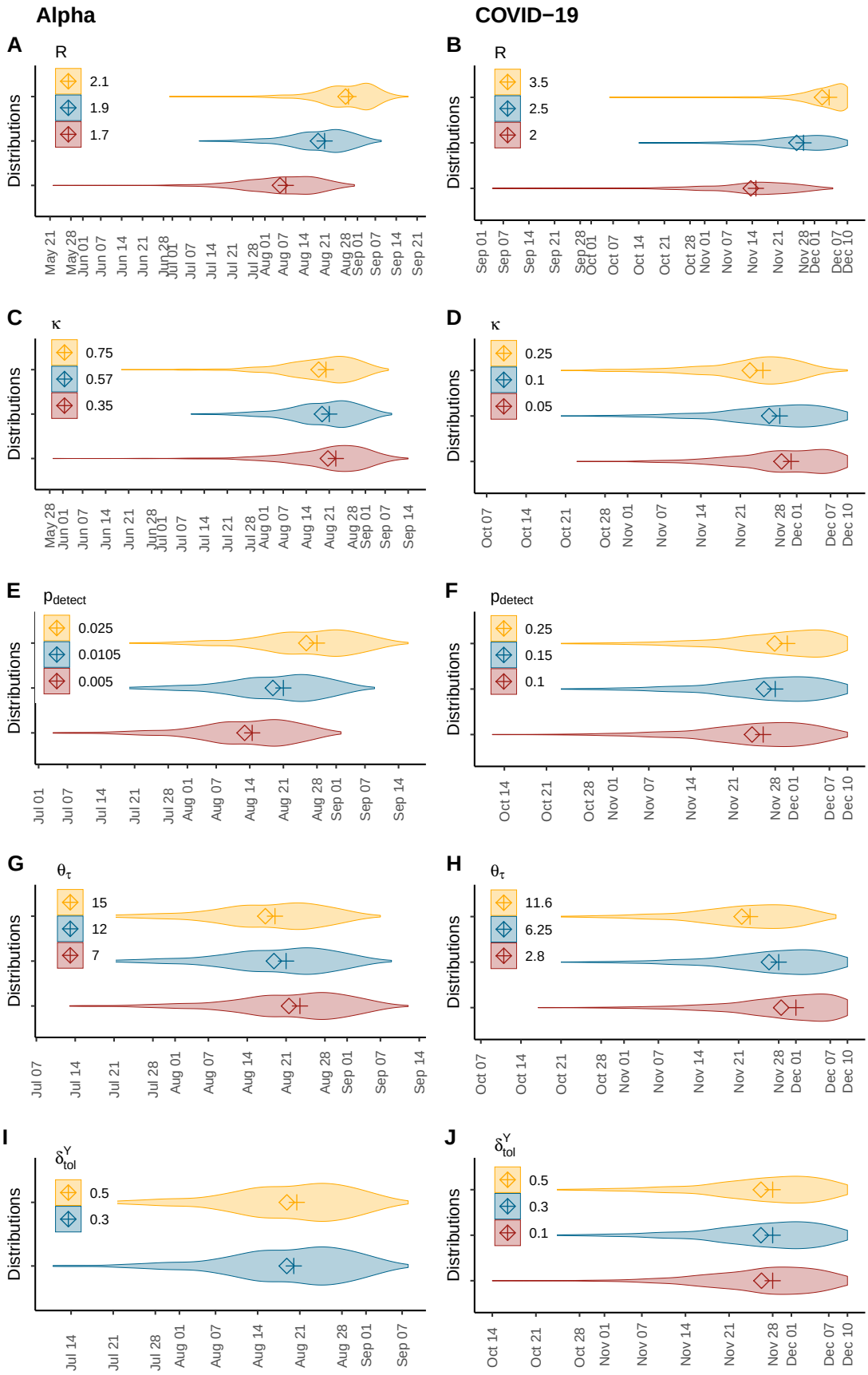


Figure 4: Sensitivity analyses. Distributions for the date of emergence of the Alpha variant in the UK (left) and the COVID-19 epidemic in Wuhan (right) obtained from running our simulations setting different input values for key model parameters: the reproduction number, R (panels A and B), the overdispersion parameter, κ, (panels C and D), the probability of detection, p_{detect} (panels E and F), the time elapsed between infection and detection, by varying θ_τ (panels G and H) and the tolerance for the difference between the simulated and the observed daily number of infections, δ^Y_{tol} (panels I and J). The absence of a violin plot for $\delta^Y_{\text{tol}} = 0.1$ in panel G results from the absence of selected simulations in > 3 million runs. Median values are depicted by crosses and mean values by diamonds. Baseline parameterization of the model is depicted in blue (middle).

Uncertainty around very early case declaration

We ran additional sensitivity analyses for the application to the emergence of the COVID-19 cases in Wuhan, to evaluate the impact of the uncertainty around very early case declaration, by applying our model to different datasets; cf. Section 2.4 of the Appendix. We first used a shorter dataset, with data on COVID-19 cases with symptoms onset only up to December 31, 2019, the day of the first public declaration of a cluster of pneumonia of unknown etiology [22] ($N = 169$). We also used an outdated, later corrected dataset published by the WHO [23] ($N = 202$) where the first case reported symptoms onset was on December 2, 2019. We found that using the shorter dataset dates the epidemic emergence 2–3 days earlier than our main results, and selects simulations with a slightly higher mean number of secondary infections (c.f. Table 1 and Supplementary Table S4), which may reflect a change in the population behavior following the first public announcement. However, using the outdated dataset [23] dates the epidemic emergence about a week prior to our main results, while the mean number of secondary infections remained about the same, which reflects the impact of the date of first detection on the results.

3 Discussion

We estimate the date of emergence of an epidemic outbreak, defined as the first infection leading to a sustained transmission chain, relying on estimates of key epidemiological parameters as well as available data on the first N observed cases. With our population-dynamics approach, we recover estimates very close to those of previous studies [1, 12], which were obtained using information from whole genome sequences. Our results constitute methodologically independent confirmations for the previously published estimates.

Our model was conceived as an extension of the numerical application presented in [11]. This methodology relies on a general branching process to model disease transmission and detection. Our model is informed by available data on the first N observed cases (unlike [11], who used the date of first detection only). We further account for super-spreaders using a negative-binomial distribution for the generation of secondary infections. This assumption may reduce the time elapsed between the first infection and the first detection or increase its variance, in comparison to ignoring super-spreading by considering other distributions (e.g. Poisson) [11] or by ignoring individual heterogeneity in infectiousness [9], as well as by using deterministic approaches [9].

We first study the emergence of the Alpha variant. We ensure that our model recovers the results from [11] when we set $N = 1$. Then we run our simulations on a larger set of data. Our results suggest that the Alpha variant emerged in the UK around August 21, 2020 and not earlier than June 13, 2020. Our results fall indeed within the same ranges as those of the approach presented in [11] when updated to match our parameterization, and fall shortly earlier than previous tMRCA estimates of the Alpha variant [12]. This result makes sense: tMRCA does not necessarily yield the date of first infection, which may have occurred before the most recent common ancestor.

Next, we apply our model to data on the early COVID-19 cases in Wuhan, estimating the date of the first SARS-CoV-2 infection around November 28 (95%IqR: Nov 2–Dec 9) , 2019, and not earlier than September 26, 2019. These ranges also fall remarkably close to—slightly later than—previously published estimates [1]. To the best of our knowledge, the novelty of these findings rests on using exclusively a population-dynamics approach, unlike previous studies aiming to date the emergence of the COVID-19 epidemic. The median estimate in Pekar et al. [1] (which is a first infection, and not a tMRCA) falls ~ 8 days before ours, which can be explained by the fact that our approach ignores genomic information. Namely, with our approach, having two cases with the exact same infecting virus, or two cases with viruses two mutations away, are treated the same way, as our approach only uses case numbers. Using genomic information would however yield an earlier date of emergence

if the infecting viruses are genetically more distant. Our findings are thus also compatible with a previous study that found no evidence of widespread transmission in Wuhan before December 2019, using serological data [24]. Hence, in accordance to previous discussions [1, 2], our results suggest that widespread SARS-CoV-2 circulation (and even more so, international spread) earlier than the end of 2019 is unlikely. Assuming an origin of the pandemic in China [25], claims of large early (i.e., before January 2020) circulation outside of China [e.g. 26, and references therein] would be therefore extraordinary, and require extraordinary evidence, excluding potential false positives by setting appropriate controls. Our study differs from [1, 2] in that we use population-level rather than genomic data, and does not intend to replace nor outperform other approaches. Rather, our results provide further support for these previous estimates relying in different methods.

Quantifying the time that emerging epidemics remain undetected before detecting the first cases is particularly important in the context of emergent pathogens such as SARS-CoV-2, where very early cases may remain unidentified, especially if a high proportion of the infections are asymptomatic [27] (N.B.: most COVID-19 cases with symptoms onset up to December 31, 2019 were declared retrospectively [28]). There is also evidence that SARS-CoV-2 may have been introduced in other countries for some time before the first reported cases [6–8, 29]. Since the first known cases are likely not to be the first human infections, there is a need for methods estimating the time of the first infection. The value of such work is essentially historical, but also helps understand how long an epidemic may have gone undetected, and design early warning systems accordingly. For instance, our approach shows that early SARS-CoV-2 infections of the transmission chain that was first detected at Wuhan's Huanan market [25] remained undetected for between about a week and three weeks before the first case of symptomatic COVID-19 was observed and reported.

The impact of uncertainty on our results is assessed by varying the main transmission and detection parameters, as well as the rejection criteria for the simulated epidemics. These sensitivity analyses unveil the robustness of our approach regarding the choice of the rejection criteria. On the other hand, we find that the main model parameters (expected number of secondary infections, R, and probability of detection, p_{detect}) have a greater impact on the model outcomes: as expected, both higher transmissibility and higher detection shorten the time between emergence and N detected infections. This variation in the results is particularly true for the application on the Alpha variant data, probably due to the notably smaller case dataset we use.

Our study has several limitations. First, our results depend heavily on input data, while access to good quality data on the early stages of an epidemic outbreak may be challenging. Datasets may be scarce, they may face reporting delays, early cases may be detected retrospectively and detec-

tion protocols may change. Early Wuhan COVID-19 cases with—severe—symptoms onset before December 30 (the day of issue of the emergency notice from the Wuhan Municipal Health Commission) [30, 22] were diagnosed clinically before tests for SARS-CoV-2 infection were available [3]. Second, our model requires early estimates of the distributions for key epidemiological indicators such as the mean number of secondary infections, the secondary infection generation time, the probability of case detection and the incubation period, which depend themselves on the quality of early observed data and may not be available for new emerging infectious diseases. In particular, it can be challenging to estimate the probability of case detection (or ascertainment rate) during early stages of an emergent infectious diseases, and it may vary between countries [31]. Third, our methods rely on the hypothesis that by the time an infection is detected, a sustained epidemic is ongoing, and all cases contained in the data set belong to the same transmission tree. That is, our model does not deal with earlier transmission trees that have gone extinct. We model epidemic spread starting from a single infectious individual, thus neglecting scenarios of multiple introductions. This impedes, for instance, the application of our model to contexts such as SARS-CoV-2 importation to France [6]. This may also be a limitation in the context of epidemics emerging from multiple spillover events, such as has been concluded by [1] relying on data on the early SARS-CoV-2 lineages. Fourth, the forward simulations of our model do not allow to consider time-dependent parameters, but a more classic Approximate Bayesian Computation approach will. Hence, we are constrained to use data on relatively short periods of time to ensure that epidemiological parameters remain nearly constant over the study period. This may not reflect early epidemic dynamics, where public outbreak alerts may provoke, on the one hand, an increase in testing efforts and thus, rapid changes in the probability of detection and, on the other hand, changes in individual behaviors that may impact the probability of disease transmission and thus, the expected number of secondary infections.

The numerous, fast and free availability of genomic data for the COVID-19 epidemic is unprecedented. Here, we built our model in a parsimonious, generic and flexible manner intended to be applied to contexts other than COVID-19, provided that key epidemiological parameters are known and transmission chains arise from a single infectious individual. Further developments of our model need to include genomic data on top of case data, as it is likely that sequencing will remain as intensive for other infectious diseases as it has been for COVID-19.

In conclusion, our study contributes to the literature with estimates of the date of first infection of early COVID-19 cases and infections with the SARS-CoV-2 Alpha variant. Our results fall within ranges previously found by studies relying on genomic data, thus offering weight to these estimates using different methodologies.

4 Methods

4.1 Model

We extend the methods presented in the numerical applications of [11], where the time of emergence was estimated from data on the first reported case only, using a stochastic population-dynamics approach. We model the early stages of the epidemic and use dates of the first N observed cases to estimate the date of emergence. We use the term '*case*' to refer to infections that are ascertained and reported: time series of cases may thus correspond to one of the following types of time series: infection detection, sample sequences, symptoms onset declarations, etc. We use the term '*probability of detection*' for the probability of such ascertainment to occur.

Our model is defined by a non-Markovian branching process to model the transmission of an infectious disease, starting from a single infectious individual in a fully susceptible population [9, 11]. Since we study early epidemic dynamics, i.e., for a relatively short period of time, all distribution parameters are assumed to be constant during the modeled time period. At any time t, infected individuals may transmit the disease. We account for super-spreading by assuming that the number of secondary cases follows a negative binomial distribution,

$$NegBinom\left(\text{number of failures} = \kappa, \text{probability of transmission} = \frac{\kappa}{\kappa + R}\right), \tag{1}$$

where R denotes the expected number of secondary infections (i.e., the effective reproduction number). Then, we model the ascertainment of infections by drawing the number of detected cases among the secondary infections from a Binomial distribution

$$Binom\left(\text{number of trials} = I(d_k),\ \text{probability of success} = p_{\text{detect}}\right), \tag{2}$$

where $I(d_k)$ is the number of incident infections at day d_k, with $k = 1, 2, \ldots$, and p_{detect} is the probability of infection detection. The generation time of each new infection, $\{t_i\}_{i=1,2,\ldots}$, as well as the time from infection to detection of a case are drawn from a Gamma distribution

$$Gamma\,(\text{shape} = \omega_x,\ \text{scale} = \theta_x), \tag{3}$$

with $x \in \{t, \tau\}$, respectively, where we use $\left\{\tau_j\right\}_{j=1,2,\ldots}$ to denote the time series of observed cases. N.B. we assume that t and τ are independent; that is, we do not model the generation time depending on the infectious individuals' own time elapsed between infection and detection. Note that the

finite number of secondary infections per individual together with the generation time distribution mean that an individual is able to transmit the pathogen for a limited period of time.

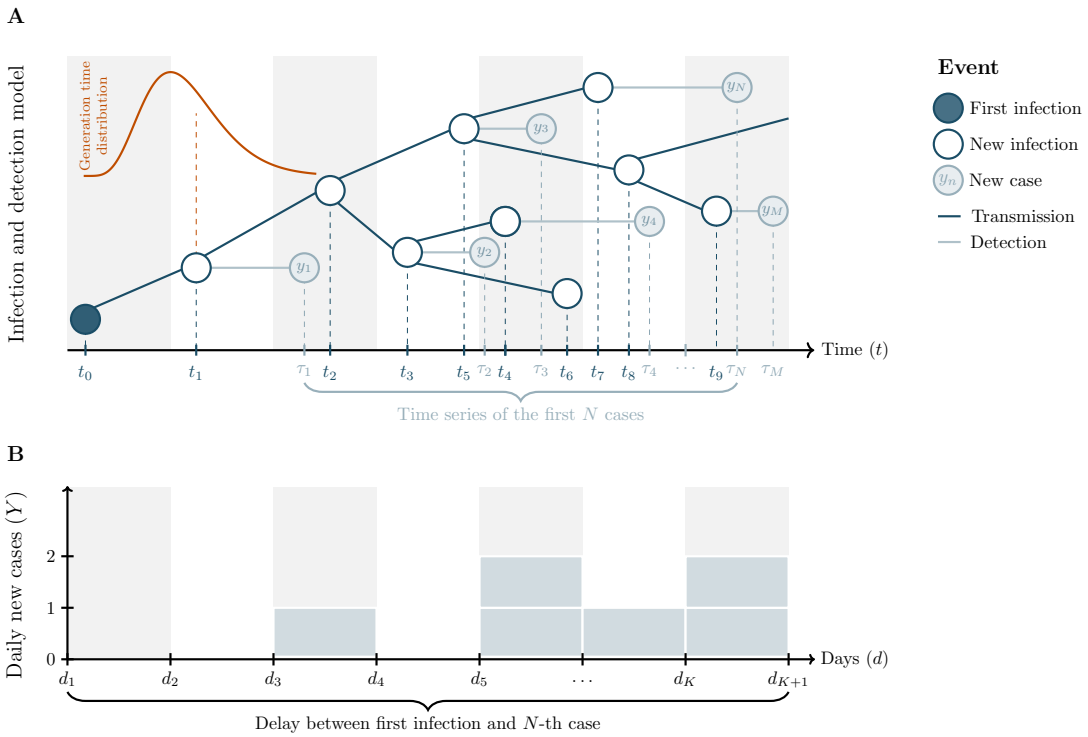
4.2 Estimation of the date of first infection

In our simulations, we discretize time by $\Delta t = 0.1$ days. Note that times to infection and detection are drawn from their corresponding distributions and thus kept on a continuous scale. We then aggregate epidemiological indicators such as the per-day number of new infections and cases (denoted by d_k, where $k = 1, 2, ...$), since this is the time scale at which most data are presented. We run stochastic simulations forward in time, from $t = t_0$, the time of occurrence of the first (undetected) infection (i.e., $I(t_0) = 1$), until the end of the day of detection of the Nth case, d_K, which is determined by $d_K \leq \tau_N < d_{K+1}$. Note that while the Nth case is a stopping criterion of our model, our analyses still deal with all $M \geq N$ cases occurring up to the end of day d_K: the number of infections detected on day k is defined by

$$Y_k \equiv \sum_{j=1}^{M} \mathbb{1}_{d_k \leq \tau_j < d_{k+1}}, \qquad k = 1, 2, \ldots, K. \tag{4}$$

A depiction of our infectious disease transmission and detection model is shown in Figure 5.

Figure 5: Model diagram. (A) We model infectious disease transmission (dark blue elements) starting from a single infectious individual (full dark blue dot), using a general branching process. The generation time of secondary infections, $\{t_{i+1} - t_i\}_{i=0,1,\ldots}$, follows a Gamma distribution (shown in orange, above the first transmission event). In addition, we model the detection of infected individuals (light blue elements), which yields the time series of observed cases, $\{\tau_j\}_{j=1}^{M}$, with $M \geq N$. The number of cases are aggregated daily. Days are denoted by $\{d_k\}_{k=1}^{K}$ and depicted by alternating gray and white bands. Our algorithm stops the day at which the Nth case is observed, d_K, but our analyses deal with the set of all cases detected on day d_K, $\{y_j\}_{j=1}^{M}$, where y_j denotes the jth detected infection (i.e., the jth case). **(B)** Resulting epidemic curve (cases per day). Our model is calibrated so that the simulated epidemic curve, $\{Y_k\}_{k=1}^{K}$, reproduces the observed number of cases per day. The main outcome of our model is the number of days elapsed between the first infection and the Nth observed case, d_K. NB. the time scale in the figure is not representative of our simulations: infection and detection delays in the simulations usually span multiple days.



The goal is to estimate the date of the first infection. To this end, we follow a strategy that is similar to an Approximate Bayesian Computation (ABC) algorithm, but meant to be computationally more efficient. In ABC, one would define a (typically uniform) prior distribution for the date of the first infection (i.e., date of emergence), randomly draw emergence dates from this distribution, simulate stochastic epidemics and compute a distance between the data (the observed time series of cases) and each simulation. One simple algorithm (*rejection* algorithm) consists in accepting the small fraction of trajectories that are closest to the data. The distribution of emergence dates from the accepted simulations then approximates the posterior distribution of the date of first infection. For comparison, we have implemented this computationally intensive strategy for the COVID-19 (Wuhan) data set. We simulated 10 000 established epidemics, i.e. the simulated epidemic reached at least 1 000 infections, for each possible date of first infection, ranging from September 2 till De-

cember 10, 2019 (the end date corresponding to condition (C1) below). Simulations were stopped when the simulation reached January 19, 2020, the date of the 3 072th detected case in the empirical data set. Simulated epidemics were then rejected based on condition (C2) below to obtain the approximate posterior distribution of the first infection dates. Using this approach, we retained 3.5% of all established epidemic simulations, which is considerably less than the 14% retained simulations with our alternative approach, which we outline now.

Our inference method is an alternative to this computationally intensive ABC algorithm and yields an approximate posterior distribution of the date of emergence. In a nutshell, this method makes use of the fact that parameters are constant in time, and that dates are therefore arbitrary; more than absolute dates, what matters in the end is the time *interval* between the first infection and the Nth case. We can therefore run simulations without a reference to absolute time; we define absolute time in a second step only, by aligning the day at which the Nth case is attained (if ever) in the simulation with the day at which the Nth case is attained in the data. In detail, let us assume for simplicity that we want to infer with ABC a single parameter, the date of emergence, defining to this end a distance metric based only on the date d_K when the simulations reach the Nth case. We only accept simulations with identical date (distance 0) to that observed in the data. If the prior distribution for the dates of emergence is uniform, the set of accepted simulations is identical to a set of simulations where simulation time is shifted to absolute date such that the date d_K is indeed identical to that observed in the data. This is possible thanks to our assumption that all parameters are constant in time, implying that the epidemiological dynamics do not depend on absolute date but solely on time elapsed since the first infection. Thus, simulations can be shifted to absolute dates in such a way that they correspond to a sample from the posterior distribution in this simple ABC algorithm. Thanks to this property, we can sample one accepted trajectory after another and have more precise control on the number of accepted simulations that are underlying the approximate posterior distribution. In contrast, with a classical ABC approach there are a lot more of rejected simulations (and wasted computation time). Our algorithm is an order of magnitude faster than classical ABC because all our simulations are constrained to match exactly the observed date of the Nth case.

The conditions to accept or reject a simulation are the following. Simulated epidemics ('sim') are calibrated to reproduce the observed ('obs') epidemic, via two conditions.

First, we require that the first simulated infection predates the first observed case:

$$t_1^{\text{sim}} \leq \tau_1^{\text{obs}}. \tag{C1}$$

Second, we accept simulations in which an epidemic occurs with enough cases, and we require the epidemic curve to resemble the whole time series of case data. More specifically, the daily number of cases of the simulated epidemic is required to resemble (under a certain tolerance δ^Y_{tol}) the observed behavior:

$$\max_k \left| \sum_{j=1}^{k} Y^{\text{obs}}(d_j) - \sum_{j=1}^{k} Y^{\text{sim}}(d_j) \right| \leq \delta^Y_{\text{tol}} N, \qquad k = K, K-1, K-2, \ldots, \tag{C2}$$

where $\sum_{j=1}^{k} Y(d_j)$ denotes the cumulative number of cases at day d_k. The daily case count in the accepted simulations thus depends on N, i.e. on the epidemiological context.

This procedure is repeated until we simulated 5 000 accepted epidemics. To compute the dates of emergence from these simulations, we take the final date of the observed case data and subtract the duration of each simulated epidemic, which produces the posterior distribution of the date of emergence.

For more details on the numerical application of the model, please refer to the pseudo-algorithm in the Appendix. The simulations were run in Julia [32] version 1.8, and the results figures were generated in R [33] version 4.1.2, using the `ggplot2` package [34], version 3.3.6. All data and codes needed for reproducibility of our results and the corresponding figures are available at a public Github repository: https://github.com/sjijon/estimate-emergence-from-data.

4.3 Applications

4.3.1 Alpha variant in the UK

The first application concerns the early stages of the spread of the Alpha SARS-CoV-2 variant in the UK. Note that the case $N = 1$ thus serves to confirm whether our extended version of the model presented in [11] recovers its results.

The first reported sequence of the SARS-CoV-2 Alpha variant of concern, was collected on September 20, 2020 [35]. Here, we define a case as a sequenced sample carrying the Alpha variant, and we define the probability of a case detection as the probability of sampling and sequencing such variant. The parameterization of our model is as in [11], except for the expected number of secondary infections, R, which was updated to match the hypotheses made in [12], to ensure comparability of results; cf. Table 2. The data on early Alpha cases were retrieved from the Global Initiative on Sharing Avian Influenza Data (GISAID) [36], available at `doi.org/10.55876/gis8.230104xg` (Ap-

pendix). We use the data on the sequences submitted to GISAID up to November 30, 2020, and used only the samples collected up to November 11, 2020. This choice was done to overcome reporting delays, and to exclude sequences analysed and added retrospectively once the growth of the variant had been identified; cf. Supplementary Figure S1.

4.3.2 COVID-19 in Wuhan

The second application concerns the early COVID-19 cases reported in Wuhan, China. Here, we define a case as a confirmed COVID-19 infection, in many cases determined retrospectively [28]. We use the dataset considered in [1] (personal communication), comprising 3 072 COVID-19 cases with symptoms onset between December 10, 2019 and January 19, 2020, the day before the first public statement on human-to-human transmission [22]; cf. Supplementary Figure S2 in the Appendix. The input parameters are summarized in Table 2.

Table 2: Input parameters. Dates for the first and N^{th} observed cases correspond to the dates in the data set. A case in the context of the Alpha variant in the UK is defined as a sequenced sample, whereas a case in the context of COVID-19 in Wuhan is defined as a confirmed, symptomatic case. The total numbers of observed cases correspond to the size of the data set used to inform the model. The key epidemiological parameters are obtained from available literature (references in brackets, next to the parameter values).

Parameter	**Symbol**	**Alpha**	**SARS-CoV-2**
Total number of observed cases	N	406	3072
Date of first reported case	d_1^{obs}	Sep 20, 2020	Dec 10, 2019
Date of Nth observed case	$d_{K'}^{\text{obs}}$	Nov 11, 2020	Jan 19, 2020
Expected number of secondary cases	R	1.90 [12]	2.50 [37]
	κ	0.57 [11]	0.10 [38]
Secondary infection generation time	ω_t	0.83 [39]†	0.83 [39]
	θ_t	6.60 [39]†	6.60 [39]
Probability of case detection	p_{detect}	0.01 [40, 41]†	0.15 [37]
Time from infection to detection	ω_τ	0.58 [11]	1.04 [42]
	θ_τ	12.0 [11]	6.25 [42]
Tolerance on the daily number of cases*	δ_{tol}^Y	0.3	0.3

*Used to select the simulations that resemble the observed data; cf. the Methods section.

†We use the parameters from the studies cited in [11].

Acknowledgements

SJ's postdoctoral fellowship was funded by a grant from the MODCOV19 platform of the National Institute of Mathematical Sciences and their Interactions (Insmi, CNRS) to FD. FD was funded by ANR-19-CE45-0009 (TheoGeneDrive). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We also thank Verity Hill [12] and Jonathan Pekar [1] for sharing their data and results for comparison with ours. We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. We thank four anonymous reviewers for comments on the manuscript.

References

[1] Jonathan E. Pekar, Andrew Magee, Edyth Parker, Niema Moshiri, Katherine Izhikevich, Jennifer L. Havens, Karthik Gangavarapu, Lorena Mariana Malpica Serrano, Alexander Crits-Christoph, Nathaniel L. Matteson, Mark Zeller, Joshua I. Levy, Jade C. Wang, Scott Hughes, Jungmin Lee, Heedo Park, Man-Seong Park, Katherine Zi Yan Ching, Raymond Tzer Pin Lin, Mohd Noor Mat Isa, Yusuf Muhammad Noor, Tetyana I. Vasylyeva, Robert F. Garry, Edward C. Holmes, Andrew Rambaut, Marc A. Suchard, Kristian G. Andersen, Michael Worobey, and Joel O. Wertheim. The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. *Science*, page eabp8337, July 2022. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abp8337.

[2] Jonathan Pekar, Michael Worobey, Niema Moshiri, Konrad Scheffler, and Joel O. Wertheim. Timing the SARS-CoV-2 index case in Hubei province. *Science*, 372(6540):412–417, April 2021. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abf8003.

[3] Michael Worobey. Dissecting the early COVID-19 cases in Wuhan. *Science*, 374(6572):1202–1204, December 2021. doi: 10.1126/science.abm4454.

[4] David L. Roberts, Jeremy S. Rossman, and Ivan Jarić. Dating first cases of COVID-19. *PLOS Pathogens*, 17(6):e1009620, June 2021. ISSN 1553-7374. doi: 10.1371/journal.ppat.1009620.

[5] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, 395(10223):497–506, 2020. doi: https://doi.org/10.1016/S0140-6736(20)30183-5. URL `https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30183-5/fulltext`.

[6] Fabiana Gámbaro, Sylvie Behillil, Artem Baidaliuk, Flora Donati, Mélanie Albert, Andreea Alexandru, Maud Vanpeene, Méline Bizard, Angela Brisebarre, Marion Barbet, Fawzi Derrar, Sylvie van der Werf, Vincent Enouf, and Etienne Simon-Loriere. Introductions and early spread of SARS-CoV-2 in France, 24 January to 23 March 2020. *Eurosurveillance*, 25(26), July 2020. ISSN 1560-7917. doi: 10.2807/1560-7917.ES.2020.25.26.2001200.

[7] Michael Worobey, Jonathan Pekar, Brendan B. Larsen, Martha I. Nelson, Verity Hill, Jeffrey B. Joy, Andrew Rambaut, Marc A. Suchard, Joel O. Wertheim, and Philippe Lemey. The emergence of SARS-CoV-2 in Europe and North America. *Science*, 370(6516):564–570, October 2020. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abc8169.

[8] Louis du Plessis, John T. McCrone, Alexander E. Zarebski, Verity Hill, Christopher Ruis, Bernardo Gutierrez, Jayna Raghwani, Jordan Ashworth, Rachel Colquhoun, Thomas R. Connor, Nuno R. Faria, Ben Jackson, Nicholas J. Loman, Áine O'Toole, Samuel M. Nicholls, Kris V. Parag, Emily Scher, Tetyana I. Vasylyeva, Erik M. Volz, Alexander Watts, Isaac I. Bogoch, Kamran Khan, COVID-19 Genomics UK (COG-UK)

Consortium, David M. Aanensen, Moritz U. G. Kraemer, Andrew Rambaut, and Oliver G. Pybus. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science*, 371(6530):708–712, February 2021. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abf2946.

[9] Thomas Beneteau, Baptiste Elie, Mircea T. Sofonea, and Samuel Alizon. Estimating dates of origin and end of COVID-19 epidemics. *Peer Community Journal*, 1:e70, December 2021. ISSN 2804-3871. doi: 10.24072/pcjournal.63.

[10] Philippe Lemey, Nick Ruktanonchai, Samuel L. Hong, Vittoria Colizza, Chiara Poletto, Frederik Van den Broeck, Mandev S. Gill, Xiang Ji, Anthony Levasseur, Bas B. Oude Munnink, Marion Koopmans, Adam Sadilek, Shengjie Lai, Andrew J. Tatem, Guy Baele, Marc A. Suchard, and Simon Dellicour. Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature*, 595(7869):713–717, July 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03754-2.

[11] Peter Czuppon, Emmanuel Schertzer, François Blanquart, and Florence Débarre. The stochastic dynamics of early epidemics: probability of establishment, initial growth rate, and infection cluster size at first detection. *Journal of The Royal Society Interface*, 18(184):20210575, November 2021. ISSN 1742-5662. doi: 10.1098/rsif.2021.0575.

[12] Verity Hill, Louis Du Plessis, Thomas P Peacock, Dinesh Aggarwal, Rachel Colquhoun, Alesandro M Carabelli, Nicholas Ellaby, Eileen Gallagher, Natalie Groves, Ben Jackson, J T McCrone, Áine O'Toole, Anna Price, Theo Sanderson, Emily Scher, Joel Southgate, Erik Volz, Wendy S Barclay, Jeffrey C Barrett, Meera Chand, Thomas Connor, Ian Goodfellow, Ravindra K Gupta, Ewan M Harrison, Nicholas Loman, Richard Myers, David L Robertson, Oliver G Pybus, and Andrew Rambaut. The Origins and Molecular Evolution of SARS-CoV-2 Lineage B.1.1.7 in the UK. *Virus Evolution*, page veac080, August 2022. ISSN 2057-1577. doi: 10.1093/ve/veac080.

[13] James B. Pettengill. The Time to Most Recent Common Ancestor Does Not (Usually) Approximate the Date of Divergence. *PLOS ONE*, 10(8):e0128407, August 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0128407.

[14] Thibaut Jombart, Kevin van Zandvoort, Timothy W. Russell, Christopher I. Jarvis, Amy Gimma, Sam Abbott, Sam Clifford, Sebastian Funk, Hamish Gibbs, Yang Liu, Carl A. B. Pearson, Nikos I. Bosse, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Rosalind M. Eggo, Adam J. Kucharski, and W. John Edmunds. Inferring the number of COVID-19 cases from recently reported deaths. *Wellcome Open Research*, 5:78, April 2020. ISSN 2398-502X. doi: 10.12688/wellcomeopenres.15786.1.

[15] Onno A. van Herwaarden and Johan Grasman. Stochastic epidemics: major outbreaks and the duration of the endemic period. *Journal of Mathematical Biology*, 33(6):581–601, 1995. doi: 10.1007/bf00298644.

[16] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359, 2005. doi: 10.1038/nature04153.

[17] Tom Britton. Stochastic epidemic models: A survey. *Mathematical Biosciences*, 225(1):24–35, 2010. doi: 10.1016/j.mbs.2010.01.006.

[18] Linda J.S. Allen. A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling*, 2(2):128–142, 2017. doi: 10.1016/j.idm.2017.03.001.

[19] Nicolas P. Rebuli, N.G. Bean, and J.V. Ross. Estimating the basic reproductive number during the early stages of an emerging epidemic. *Theoretical Population Biology*, 119:26–36, 2018. doi: 10.1016/j.tpb.2017.10.004.

[20] Tom Britton and Etienne Pardoux. *Stochastic Epidemic Models with Inference*. Springer International Publishing, 2019. doi: 10.1007/978-3-030-30900-8.

[21] Andrew Rambaut, Edward C. Holmes, Áine O'Toole, Verity Hill, John T. McCrone, Christopher Ruis, Louis du Plessis, and Oliver G. Pybus. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, 5(11):1403–1407, November 2020. ISSN 2058-5276. doi: 10.1038/s41564-020-0770-5. Number: 11 Publisher: Nature Publishing Group.

[22] WHO. Listings of WHO's response to COVID-19, June 2020. URL `https://www.who.int/news/item/29-06-2020-covidtimeline`. [Accessed: Nov 25, 2022].

[23] The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19) – china. *China CDC Weekly*, 2:113, 2020. ISSN 2096-7071. doi: 10.46234/ccdcw2020.032.

[24] Le Chang, Lei Zhao, Yan Xiao, Tingting Xu, Lan Chen, Yan Cai, Xiaojing Dong, Conghui Wang, Xia Xiao, Lili Ren, and Lunan Wang. Serosurvey for SARS-CoV-2 among blood donors in Wuhan, China from September to December 2019. *Protein & Cell*, page pwac013, May 2022. ISSN 1674-800X, 1674-8018. doi: 10.1093/procel/pwac013.

[25] Michael Worobey, Joshua I. Levy, Lorena Malpica Serrano, Alexander Crits-Christoph, Jonathan E. Pekar, Stephen A. Goldstein, Angela L. Rasmussen, Moritz U. G. Kraemer, Chris Newman, Marion P. G. Koopmans, Marc A. Suchard, Joel O. Wertheim, Philippe Lemey, David L. Robertson, Robert F. Garry, Edward C. Holmes, Andrew Rambaut, and Kristian G. Andersen. The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic. *Science*, page abp8715, July 2022. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abp8715.

[26] Antonella Amendola, Marta Canuti, Silvia Bianchi, Sudhir Kumar, Clara Fappani, Maria Gori, Daniela Colzani, Sergei L. Kosakovsky Pond, Sayaka Miura, Melissa Baggieri, Antonella Marchi, Elisa Borghi, Gianvincenzo Zuccotti, Mario C. Raviglione, Fabio Magurano, and Elisabetta Tanzi. Molecular evidence for SARS-CoV-2 in samples collected from patients with morbilliform eruptions since late 2019 in Lombardy, northern Italy. *Environmental Research*, 215:113979, December 2022.

[27] Mercedes Yanes-Lane, Nicholas Winters, Federica Fregonese, Mayara Bastos, Sara Perlman-Arrow, Jonathon R. Campbell, and Dick Menzies. Proportion of asymptomatic infection among COVID-19 positive persons and their transmission potential: A systematic review and meta-analysis. *PLOS ONE*, 15 (11):e0241536, November 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0241536.

[28] WHO. WHO-convened global study of origins of SARS-CoV-2: China Part. Technical report, WHO, 2021. URL `https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part`.

[29] Francesca Di Giallonardo, Sebastian Duchene, Ilaria Puglia, Valentina Curini, Francesca Profeta, Cesare Cammà, Maurilia Marcacci, Paolo Calistri, Edward Holmes, and Alessio Lorusso. Genomic Epidemiology of the First Wave of SARS-CoV-2 in Italy. *Viruses*, 12(12):1438, December 2020. ISSN 1999-4915. doi: 10.3390/v12121438.

[30] Congressional Research Service. COVID-19 and China: A Chronology of Events (December 2019-January 2020), 2020. URL `https://crsreports.congress.gov/product/pdf/r/r46354`. [Accessed on: December 15, 2022].

[31] Timothy W. Russell, Nick Golding, Joel Hellewell, Sam Abbott, Lawrence Wright, Carl A. B. Pearson, Kevin van Zandvoort, Christopher I. Jarvis, Hamish Gibbs, Yang Liu, Rosalind M. Eggo, W. John Edmunds, Adam J. Kucharski, CMMID COVID-19 working group, Arminder K. Deol, C. Julian Villabona-Arenas, Thibaut Jombart, Kathleen O'Reilly, James D. Munday, Sophie R. Meakin, Rachel Lowe, Amy Gimma, Akira Endo, Emily S. Nightingale, Graham Medley, Anna M. Foss, Gwenan M. Knight, Kiesha Prem, Stéphane Hué, Charlie Diamond, James W. Rudge, Katherine E. Atkins, Megan Auzenbergs, Stefan Flasche, Rein M. G. J. Houben, Billy J. Quilty, Petra Klepac, Matthew Quaife, Sebastian Funk, Quentin J. Leclerc, Jon C. Emery, Mark Jit, David Simons, Nikos I. Bosse, Simon R. Procter, Fiona Yueqian Sun, Samuel Clifford, Katharine Sherratt, Alicia Rosello, Nicholas G. Davies, Oliver Brady, Damien C. Tully, and Georgia R. Gore-Langton. Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. *BMC Medicine*, 18(1):332, December 2020.

[32] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1):65–98, January 2017. ISSN 0036-1445, 1095-7200. doi: 10.1137/141000671.

[33] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

[34] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

[35] Andrew Rambaut, Nick Loman, Oliver Pybus, Wendy Barclay, Jeff Barrett, Alesandro Carabelli, Tom Connor, Tom Peacock, David L Robertson, Erik Volz, 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 - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology, December 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. [Accessed: Nov 25, 2022].

[36] Shruti Khare, Céline Gurry, Lucas Freitas, Mark B Schultz, Gunter Bach, Amadou Diallo, Nancy Akite, Joses Ho, Raphael TC Lee, Winston Yeo, Gisaid Core Curation Team, Sebastian Maurer-Stroh, GISAID Global Data Science Initiative (GISAID), Munich, Germany, Bioinformatics Institute, Agency for Science Technology and Research, Singapore, Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil, Institut Pasteur de Dakar, Dakar, Senegal, National Institutes of Biotechnology Malaysia, Selangor, Malaysia, Smorodintsev Research Institute of Influenza, St. Petersburg, Russia, Genome Institute of Singapore, Agency for Science Technology and Research, Singapore, China National GeneBank, Shenzhen, China, A*STAR Infectious Disease Labs (ID Labs), Singapore, National Public Health Laboratory, National Centre for Infectious Diseases, Ministry of Health, Singapore, and Department of Biological Sciences, National University of Singapore, Singapore. GISAID's Role in Pandemic Response. *China CDC Weekly*, 3 (49):1049–1051, 2021. ISSN 2096-7071. doi: 10.46234/ccdcw2021.255.

[37] Xingjie Hao, Shanshan Cheng, Degang Wu, Tangchun Wu, Xihong Lin, and Chaolong Wang. Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature*, 584(7821):420–424, August 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2554-8.

[38] Akira Endo, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Sam Abbott, Adam J. Kucharski, and Sebastian Funk. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Research*, 5:67, July 2020. ISSN 2398-502X. doi: 10.12688/wellcomeopenres.15842.3.

[39] Robert Hinch, William J. M. Probert, Anel Nurtay, Michelle Kendall, Chris Wymant, Matthew Hall, Katrina Lythgoe, Ana Bulas Cruz, Lele Zhao, Andrea Stewart, Luca Ferretti, Daniel Montero, James Warren, Nicole Mather, Matthew Abueg, Neo Wu, Olivier Legat, Katie Bentley, Thomas Mead, Kelvin Van-Vuuren, Dylan Feldner-Busztin, Tommaso Ristori, Anthony Finkelstein, David G. Bonsall, Lucie Abeler-Dörner, and Christophe Fraser. Openabm-covid19—an agent-based model for non-pharmaceutical interventions against covid-19 including contact tracing. *PLOS Computational Biology*, 17(7):1–26, 07 2021. doi: 10.1371/journal.pcbi.1009146. URL https://doi.org/10.1371/journal.pcbi.1009146.

[40] COVID-19 Genomics Consortium UK. How do we collect and sequence SARS-CoV-2 samples?, 2020. URL https://www.cogconsortium.uk/news_item/how-do-we-collect-and-sequence-sars-cov-2-samples/. [Accessed on: Feb 9, 2020].

[41] Ewan Colman, Gavrila A. Puspitarani, Jessica Enright, and Rowland R. Kao. Ascertainment rate of SARS-CoV-2 infections from healthcare and community testing in the UK. *Journal of Theoretical Biology*, 558:111333, 2023. ISSN 0022-5193. doi: https://doi.org/10.1016/j.jtbi.2022.111333. URL https://www.sciencedirect.com/science/article/pii/S0022519322003241.

[42] Jantien A Backer, Don Klinkenberg, and Jacco Wallinga. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Eurosurveillance*, 25 (5), February 2020. ISSN 1560-7917. doi: 10.2807/1560-7917.ES.2020.25.5.2000062.

Supporting information

S1 Text. Appendix. Case datasets, supplementary results and pseudo-algorithm