Epidemics 44 (2023) 100713

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Bayesian reconstruction of SARS-CoV-2 transmissions highlights substantial proportion of negative serial intervals

Cyril Geismar [a,b,*], Vincent Nguyen [b], Ellen Fragaszy [c,d], Madhumita Shrotri [b], Annalan M.D. Navaratnam [b], Sarah Beale [b,c], Thomas E. Byrne [b], Wing Lam Erica Fong [b], Alexei Yavlinsky [b], Jana Kovar [c], Susan Hoskins [b], Isobel Braithwaite [b], Robert W. Aldridge [b], Andrew C. Hayward [c], Peter J. White [a], Thibaut Jombart [a], Anne Cori [a]

[a] MRC Centre for Global Infectious Disease Analysis and NIHR Health Protection Research Unit in Modelling and Health Economics, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK
[b] Centre for Public Health Data Science, Institute of Health Informatics, University College London, London, UK
[c] Institute of Epidemiology and Health Care, University College London, London, UK
[d] Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK

ARTICLE INFO

Keywords:
Sars-cov-2
Variants of concern
Serial interval
Bayesian modelling
Infectious Disease Epidemiology

ABSTRACT

Background: The serial interval is a key epidemiological measure that quantifies the time between the onset of symptoms in an infector-infectee pair. It indicates how quickly new generations of cases appear, thus informing on the speed of an epidemic. Estimating the serial interval requires to identify pairs of infectors and infectees. Yet, most studies fail to assess the direction of transmission between cases and assume that the order of infections - and thus transmissions - strictly follows the order of symptom onsets, thereby imposing serial intervals to be positive. Because of the long and highly variable incubation period of SARS-CoV-2, this may not always be true (i.e an infectee may show symptoms before their infector) and negative serial intervals may occur. This study aims to estimate the serial interval of different SARS-CoV-2 variants whilst accounting for negative serial intervals.

Methods: This analysis included 5 842 symptomatic individuals with confirmed SARS-CoV-2 infection amongst 2 579 households from September 2020 to August 2022 across England & Wales. We used a Bayesian framework to infer who infected whom by exploring all transmission trees compatible with the observed dates of symptoms, based on a wide range of incubation period and generation time distributions compatible with estimates reported in the literature. Serial intervals were derived from the reconstructed transmission pairs, stratified by variants.

Results: We estimated that 22% (95% credible interval (CrI) 8–32%) of serial interval values are negative across all VOC. The mean serial interval was shortest for Omicron BA5 (2.02 days, 1.26–2.84) and longest for Alpha (3.37 days, 2.52–4.04).

Conclusions: This study highlights the large proportion of negative serial intervals across SARS-CoV-2 variants. Because the serial interval is widely used to estimate transmissibility and forecast cases, these results may have critical implications for epidemic control.

1. Introduction

The serial interval is a key epidemiological measure, defined as the time between an infectee's onset of symptoms and its infector's onset of symptoms. Characterising its distribution helps investigate epidemiological links between cases and is usually needed for estimating transmissibility (Vink et al., 2014; Cori et al., 2013; Wallinga and Teunis, 2004). The serial interval is informative of how fast an epidemic is spreading and a key component to link the reproduction number and the epidemic growth rate (Cori et al., 2013; Wallinga and Teunis, 2004). The emergence and rapid spread of SARS-CoV-2 variants of concern (VOC) throughout the COVID-19 pandemic highlighted how small genetic

* Corresponding author at: MRC Centre for Global Infectious Disease Analysis and NIHR Health Protection Research Unit in Modelling and Health Economics, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK.
E-mail address: c.geismar21@imperial.ac.uk (C. Geismar).

https://doi.org/10.1016/j.epidem.2023.100713
Received 15 December 2022; Received in revised form 25 July 2023; Accepted 31 July 2023
Available online 7 August 2023

changes can have a strong impact on transmission dynamics (Davies et al., 2021). While changes in growth rates among VOC are often believed to reflect changes in the reproduction number, it remains unclear whether such changes might be the result of different serial interval distributions.

Throughout the pandemic, commonly used models aimed at inferring the SARS-CoV-2 reproduction number did not account for the differences in the serial intervals of variants nor for negative values of the serial interval, leading to potential biases in their results (Bhatia et al., 2021; Anderson et al., 2020). Since the generation time (the time between the infection of a primary case and its secondary cases) is seldom observable, only a few estimates have been published (Anderson et al., 2020; Griffin et al., 2020). In practice, the serial interval is used as a proxy for the generation time as they typically have similar means although they have different variances (Svensson, 2007). Most of the published serial interval estimates were reported early in the pandemic from studies conducted in Asia, predominantly in China, based on a relatively small number of infector-infectee pairs, often well below 100 pairs (Anderson et al., 2020; Griffin et al., 2020; Alene et al., 2021). Several studies remained unclear how the infector-infectee pairs were identified (Kwok et al., 2020; Bao et al., 2021; Nishiura et al., 2020; Zhang et al., 2021; Du et al., 2020). Although negative serial intervals are possible, in practice many studies fail to assess the direction of transmission between pairs of related cases and assume a minimum serial interval value of zero days (Griffin et al., 2020; Nishiura et al., 2020; UKHSA, 2022a; Lavezzo et al., 2020).

By accounting for the uncertainty in who infected whom using a Bayesian approach, this study is able to infer negative serial intervals. We estimated and compared the serial intervals of all major SARS-CoV-2 VOC using data from thousands of confirmed cases reported in households across England and Wales between September 2020 and August 2022.

2. Methods

2.1. Data

Virus Watch is a household community cohort study following up entire households in England and Wales since mid-June 2020 (Hayward et al., 2021). Eligible households were 1–6 persons in size, with internet access and at least one household member able to complete surveys in English. By February 2022, 58 566 individuals in 28 495 households had registered to take part in the study. Participants completed a weekly online survey reporting the date of any respiratory, constitutional, gastrointestinal, ocular, or skin symptoms experienced and, the date and result of any testing for SARS-CoV-2 by Lateral Flow Test (LFT) or Polymerase Chain Reaction (PCR).

2.2. Data processing

Symptom data were extracted and grouped into illness episodes. The start date of an illness episode was defined as the first day any symptoms were reported, and the end date was the final day of reported symptoms. A seven-day washout period where no symptoms were reported was used to identify separate illness episodes. Swab results were matched to illnesses that were within seven days of the illness start date. In addition to SARS-CoV-2 self-reported test results, test results from the UK Second-Generation Surveillance System (SGSS) dataset were linked to the Virus Watch dataset (appendix A1).

Households with a single case were not included in the analysis. We removed all households where the index case's symptom onset date was within two weeks of our most recent survey date to allow for a minimum fourteen-day follow-up. Testing (by LFT or PCR) detects the presence of SARS-CoV-2 but does not identify the variant. Therefore, using national surveillance data (UKHSA, 2022b), we designated a variant to a household if that variant was making up at least 75% of all regional sequenced genomes at the time of the index case's symptom onset. Illnesses in regions and weeks that did not have a dominant variant reaching at least 75% of all sequenced genomes were excluded. Variants are defined as per the UK Health Secretary Agency (UKHSA) definition and wild-type refers to all SARS-CoV-2 variants circulating before the Alpha variant (UKHSA, 2022b).

2.3. Model

We used the R package *outbreaker2* to reconstruct, independently for each household, plausible within-household transmission chains from the members' symptom onset dates. *Outbreaker2* can integrate various data sources - epidemiological (e.g. symptom onset dates, collection dates), contact and genetic data - in a Bayesian framework to reconstruct transmission trees (Campbell et al., 2018). Swab tests from positive cases in the Virus Watch cohort were not sequenced at the time of the analysis. Transmission pairs were therefore inferred based on the *epidemiological likelihood:*

$$\begin{aligned} \mathrm{p}\left(t_i \middle| T_i^{inf},\ \alpha_i, T_{\alpha_i}^{inf}\right) &= \mathrm{p}\left(t_i \middle| T_i^{inf}\right)\ *\ \mathrm{p}\left(T_i^{inf} \middle| \alpha_i, T_{\alpha_i}^{inf}\right) \\ &= f\left(t_i - T_i^{\mathrm{inf}}\right) * w\left(T_i^{\mathrm{inf}} - T_{\alpha_i}^{\mathrm{inf}}\right) \end{aligned} \tag{1}$$

Where:

- i is the index of cases ($i = 1, \ldots, n$),
- n is the number of SARS-CoV-2 positive symptomatic cases in the household,
- t_i is the observed date of symptom onset of case i,
- T_i^{inf} is the unobserved date of infection of case i,
- α_i is the unobserved infector of case i.

This likelihood describes the probability that individual i has symptoms at time t_i conditional on being infected at time T_i^{inf} by individual α_i, where α_i was infected at time $T_{\alpha_i}^{\mathrm{inf}}$. It is calculated as the probability of the implied incubation period distribution f and generation time distribution w. Although, *outbreaker2* can infer unobserved intermediate cases in a transmission chain, in our context of household outbreaks, we assumed that all cases were observed. *Outbreaker2* uses a Markov Chain Monte Carlo (MCMC) algorithm to derive samples from the posterior distributions (Campbell et al., 2018). The MCMC explored the parameter space of the dates of infection and who infected whom, conditional on an incubation period and a generation time distribution which enabled us to quantify the uncertainty in the serial interval distribution.

To address the uncertainty in the generation times and incubation periods reported in the literature, we employed the Latin Hypercube Sampling (LHS) method outlined in Fig. 1 and appendix A2. Our approach involved constructing 100 pairs of incubation period and generation time distributions by sampling from a 2-dimensional parameter space of mean and coefficient of variation. Values for the mean and coefficient of variation were respectively sampled from the quantile function of the normal and truncated normal distribution, parameterised with values reported in the literature. Truncation was implemented to restrict the minimum mean incubation period and generation time to 1 day. The LHS approach thereby allowed us to explore 100 pairs of incubation period and generation time distributions ("natural histories") compatible with findings from the literature on SARS-CoV-2. For every household, the MCMC ran for 10 000 iterations, including a burnin of 500 iterations, and one in 50 iterations being recorded thereafter totalling to 190 $\left(\frac{10\,000-500}{50}\right)$ samples. This process was repeated 100 times with all natural histories, and results were aggregated to obtain a single sample from the joint posterior distribution across all natural histories (Fig. 1). Altogether, for every household, we obtained a posterior sample size of 19 000 i.e. 190 (posterior sample size

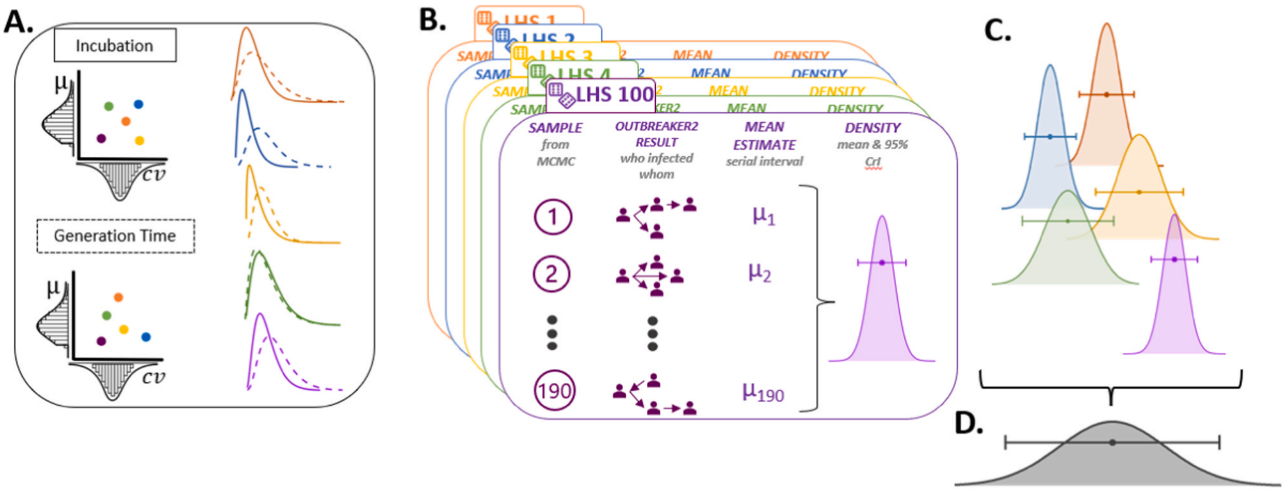

Fig. 1. Diagram describing the method to compute the mean serial interval posterior distribution, mean estimate and its 95% credible interval. A) Latin Hypercube Sampling (LHS). We generated 100 pairs of incubation period and generation time distributions compatible with findings from the literature on SARS-CoV-2. Dashed lines refer to the generation time distribution, solid lines refer to the incubation period distribution. Colours refer to a unique LHS sample. B) The outbreaker2 model inferred transmission chains using an MCMC algorithm. For each posterior sample, the mean serial interval was calculated from the inferred transmission pairs. Finally, we calculated the kernel density of the mean serial interval, the overall mean estimate and its 95% credible interval from the 2.5 and 97.5 percentiles. C) We repeated the process described in B) considering the 100 pairs of natural histories drawn in A) as input for the outbreaker2 model. D) The mean serial interval estimates were aggregated across the 100 natural histories considered.

for each natural history) × 100 (number of natural histories). We assessed the convergence of our model by manual inspection of the MCMC traces. To obtain the serial intervals, we computed, for each transmission tree in the posterior sample, the time difference (in days) between each infectee's and their infector's symptom onset date. A household with n cases will have $n - 1$ transmission pairs, resulting in a posterior sample of the serial interval of size (n-1)× 19 000.

There was a large proportion of households where all cases reported the same symptom onset date, leading to an unusual peak at 0 in the serial interval histograms (appendix A3, A4). The frequency of observations at 0 days was at least 2.5 times higher than any other day, suggesting 0 to be an outlier. A discretised shifted gamma distribution was thus fitted to non-zero observations, using a maximum likelihood method. The shift was used to account for potential negative serial intervals. Log densities were rescaled to account for the fact that we only fitted non-zero data points (Eq. (2)).

$$P(si = x) = \frac{F_\Gamma(x + s + 0.5) - F_\Gamma(x + s - 0.5)}{1 - (F_\Gamma(0.5) - F_\Gamma(-0.5))} \quad \text{where} \quad P(si = 0) = 0 \tag{2}$$

In the following, adjusted mean refers to the mean of the shifted discretised gamma distribution fitted to the posterior sample (Eq. (2)). This fitting procedure was performed at every step of the MCMC and for every natural history to explore the range of valid parameters (Fig. 1). We also considered alternative shifted Normal and Log-normal distributions that were also fitted to our data (appendix A5). Mean estimates and 95% credible intervals were obtained by computing the mean, the 2.5th and 97.5th percentiles from the posterior densities (Fig. 1).

We set bounds of −10 to +20 days for the serial interval values as most studies report extremum values well within these bounds (Griffin et al., 2020; Nishiura et al., 2020). Any infector-infectee pair with values outside these bounds were discarded and considered re-introduction from infections contracted outside the household (appendix A6).

2.4. Comparisons with alternative approaches

To assess the added value of the *outbreaker2* model, we performed two comparisons with alternative approaches. First, we compared our results to a "*pairwise*" model, in which all pairs of individuals in a household were considered as equally likely transmission pairs. Thus, the *pairwise* model derives the serial interval distribution strictly from our data. Second, we compared our results to a "*theoretical*" model, where the serial interval distribution is strictly inferred from the input incubation period and generation time distributions. The serial interval could theoretically be derived from the two aforementioned parameters. Following the equation (2.1c) of the generation time described by Lehtinen et al (Lehtinen et al., 2021)., the serial interval can be derived as Eq. (3) below:

$$S_{ij} = G_{ij} + I_j - I_i \tag{3}$$

S represents the serial interval of the pair and is equal to the generation time of the pair, plus the incubation of the infectee j (I_j) minus the incubation of the infector i (I_i), assuming independence between the generation time and incubation period distributions.

For every pair of natural histories considered, 10 000 independent random numerical draws from the generation time distribution G and the incubation period distribution I were computed and assigned to G_{ij}, I_j and I_i respectively. The theoretical serial interval distribution was then derived from Eq. (3) above.

Outbreaker2 R package and description of the model is available online (Campbell et al., 2018; Jombart et al., 2014).

2.5. Simulation study & estimates reliability

We conducted a simulation study to evaluate the reliability of the *outbreaker2* estimates when the natural histories' moments are misspecified. To simulate the "true" outbreaks, we used the *simulacr* R package (https://github.com/CyGei/simulacr) which utilises the generation time and incubation period distributions within branching processes. In conjunction with our LHS method, we used *outbreaker2* to reconstruct transmission chains and estimate the mean serial interval (Fig. 1). Subsequently, we compared the reconstructed (*outbreaker2*) serial interval with the true (*simulacr*) serial interval (Appendix B). Additionally, to investigate how different assumptions about the incubation period and generation time may affect our estimates, we conducted a multivariate linear regression to explore the relationship between the natural histories' moments and the adjusted mean serial interval (Appendix B).

3. Results

There were 81 892 illness episodes reported by 29 766 individuals from 17 794 households between 22 June 2020 and 14 August 2022, which were evenly distributed across sex. 49 367 (60%) of illness episodes were tested for SARS-CoV-2, with weekly proportions ranging from 13.00% to 90.50% (appendix A7). The proportion of test results that were positive was 22% (10,663/49,367), with weekly proportions ranging from 0% to 54.20% (appendix A8). Table 1 displays the summary characteristics of 5 842 symptomatic SARS-CoV-2 positive illnesses reported by 5 799 cases from 2 579 households included in the analysis. Over half of households were composed of only 2 members. Households with 3 or less cases constituted 89.5% of our data. Amongst the 10 regions represented in our data, the most common were the East of England, the South-East and London in that order**.**

About 64% of illnesses occurred during the Omicron BA1 and BA2 wave (Fig. 2, appendix A9). About 86% of illnesses were occurring amongst individuals above the age of 15 years.

Fig. 3 displays the kernel posterior densities of the adjusted serial interval mean and standard deviation by VOC. The adjusted mean serial interval was shortest for Omicron BA5 (2.02 days, 95%CrI:1.26–2.84) and longest for Alpha (3.37, 2.52–4.04). The adjusted mean serial interval was 2.29 days (1.39–2.94) for wild-type, 3.11 (2.28–3.90) for Delta, 2.72 (2.01–3.47) for Omicron BA1 and 2.67 (1.90–3.46) for Omicron BA2. The estimated gamma parameters are reported in appendix A10. When running our model with only a single pair of natural histories (selected to be closest to the median estimate from our Latin Hypercube sample, appendix A2), our estimates remained consistent with our findings, but the credible intervals significantly narrowed by an average factor of 2.34 (appendix A11, A12). Without accounting for the uncertainty in the natural histories, we find that there is a statistically significant difference between the adjusted mean serial interval of Omicron BA5 and Alpha, Delta, Omicron BA1, Omicron BA2 (appendix A11, A12).

Despite the overlapping credible intervals, we note a consistent trend where the adjusted mean serial interval shortened since the appearance of the Delta variant late 2020. We aggregated the adjusted mean serial interval across the entire posterior sample for each natural histories and computed the pairwise differences by VOC (Fig. 4). Across all 100 pairs of natural histories, the adjusted mean serial interval of Omicron BA5 was consistently shorter than any other variant except for wild type. In comparison with Omicron BA1 & BA2, the adjusted mean serial interval of Omicron BA5 was, on average, shorter by 0.75 day, for all natural histories. Compared to Delta and Alpha, the Omicron BA5 adjusted mean serial interval was respectively 1 and 1.25 day shorter.

Fig. 5 displays the posterior cumulative distribution function of the adjusted serial interval by VOC. On average, 22% (95% CrI: 8–32%) of serial interval values were negative across all VOC, with proportion varying from 19.8 (6.68 – 29.9) for Delta to 24.4 (7.48 – 36.7) for wild type (appendix A13). Most negative values lied between −4 and −1 day. The median serial interval was either 2 or 3 days depending on the VOC (appendix A13).

Appendix A4 displays the distribution of the serial interval by VOC for *outbreaker2* and the *pairwise* model. The *pairwise* model, solely driven by data, highlights how the observed peaks at 0 are the result of our data and is explained by the large proportion (14%) of households reporting the same symptom onset date. Fig. 6 illustrates the distribution of the SARS-CoV-2 serial interval of *outbreaker2, the pairwise* and the *theoretical*

Table 1
Summary characteristics of individual SARS-CoV-2 positive symptomatic illnesses included in the analysis.

Characteristic	Overall, N = 5842[a]	Wild Type, N = 139[a]	Alpha, N = 242[a]	Delta, N = 1052[a]	Omicron BA1, N = 1357[a]	Omicron BA2, N = 2380[a]	Omicron BA5, N = 672[a]
Age							
0–15	821 (14%)	14 (10%)	28 (12%)	300 (29%)	290 (21%)	151 (6.3%)	38 (5.7%)
16–44	1186 (20%)	47 (34%)	94 (39%)	284 (27%)	333 (25%)	352 (15%)	76 (11%)
45–64	2143 (37%)	46 (33%)	88 (36%)	342 (33%)	438 (32%)	954 (40%)	275 (41%)
65+	1692 (29%)	32 (23%)	32 (13%)	126 (12%)	296 (22%)	923 (39%)	283 (42%)
Sex							
Female	2972 (52%)	68 (51%)	127 (55%)	568 (55%)	674 (52%)	1208 (52%)	327 (50%)
Male	2725 (48%)	65 (49%)	105 (45%)	470 (45%)	633 (48%)	1127 (48%)	325 (50%)
(Missing)	145	6	10	14	50	45	20
Number of household members							
2	3210 (55%)	74 (53%)	95 (39%)	344 (33%)	596 (44%)	1634 (69%)	467 (69%)
3	933 (16%)	26 (19%)	74 (31%)	210 (20%)	234 (17%)	295 (12%)	94 (14%)
4	1226 (21%)	21 (15%)	51 (21%)	363 (35%)	362 (27%)	349 (15%)	80 (12%)
5	370 (6.3%)	12 (8.6%)	16 (6.6%)	109 (10%)	128 (9.4%)	78 (3.3%)	27 (4.0%)
6	103 (1.8%)	6 (4.3%)	6 (2.5%)	26 (2.5%)	37 (2.7%)	24 (1.0%)	4 (0.6%)
Number of cases per household							
2	4243 (73%)	100 (72%)	162 (67%)	626 (60%)	878 (65%)	1939 (81%)	538 (80%)
3	985 (17%)	22 (16%)	64 (26%)	263 (25%)	264 (19%)	289 (12%)	83 (12%)
4	527 (9.0%)	11 (7.9%)	15 (6.2%)	150 (14%)	160 (12%)	142 (6.0%)	49 (7.3%)
5	75 (1.3%)	6 (4.3%)	1 (0.4%)	13 (1.2%)	43 (3.2%)	10 (0.4%)	2 (0.3%)
6	12 (0.2%)	0 (0%)	0 (0%)	0 (0%)	12 (0.9%)	0 (0%)	0 (0%)
Region							
East Midlands	608 (10%)	25 (18%)	12 (5.0%)	118 (11%)	146 (11%)	252 (11%)	55 (8.2%)
East of England	1116 (19%)	10 (7.2%)	58 (24%)	189 (18%)	265 (20%)	467 (20%)	127 (19%)
London	721 (12%)	17 (12%)	69 (29%)	113 (11%)	203 (15%)	246 (10%)	73 (11%)
North East	303 (5.2%)	8 (5.8%)	7 (2.9%)	62 (5.9%)	59 (4.4%)	130 (5.5%)	37 (5.5%)
North West	628 (11%)	31 (22%)	14 (5.8%)	103 (9.8%)	159 (12%)	245 (10%)	76 (11%)
South East	1146 (20%)	6 (4.3%)	59 (24%)	228 (22%)	247 (18%)	480 (20%)	126 (19%)
South West	459 (7.9%)	10 (7.2%)	6 (2.5%)	100 (9.5%)	91 (6.8%)	189 (8.0%)	63 (9.4%)
Wales	115 (2.0%)	0 (0%)	0 (0%)	0 (0%)	29 (2.2%)	71 (3.0%)	15 (2.2%)
West Midlands	374 (6.4%)	8 (5.8%)	13 (5.4%)	71 (6.7%)	79 (5.9%)	159 (6.7%)	44 (6.5%)
Yorkshire and The Humber	337 (5.8%)	24 (17%)	4 (1.7%)	68 (6.5%)	60 (4.5%)	125 (5.3%)	56 (8.3%)
(Missing)	35	0	0	0	19	16	0
DATE OF ONSET	2020–09–01 to 2022–08–10	2020–09–01 to 2020–12–08	2020–12–10 to 2021–04–26	2021–06–03 to 2021–12–11	2021–12–14 to 2022–02–06	2022–02–27 to 2022–05–27	2022–07–04 to 2022–08–10

[a] n (%); Range

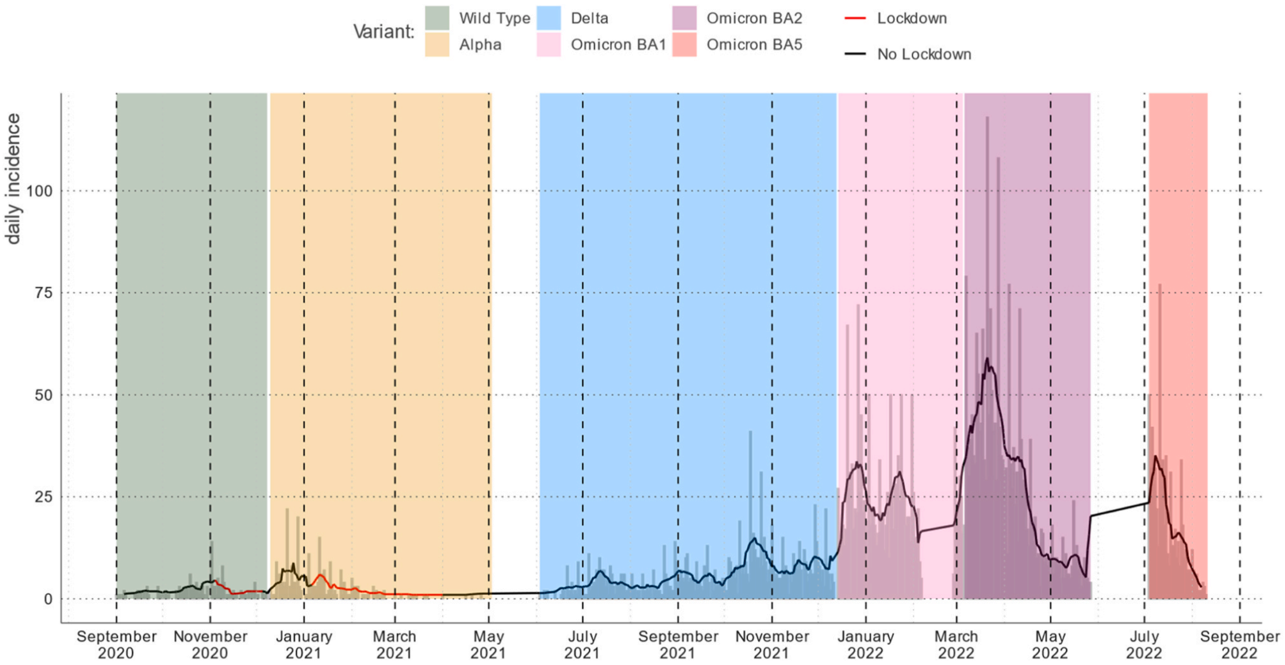

Fig. 2. Daily incidence of SARS-CoV-2 positive symptomatic illnesses amongst the participants selected for the analysis. Grey bars represent daily confirmed illnesses, the black line represents the 7-day rolling average. Periods with no variant designation were not recorded for this analysis.

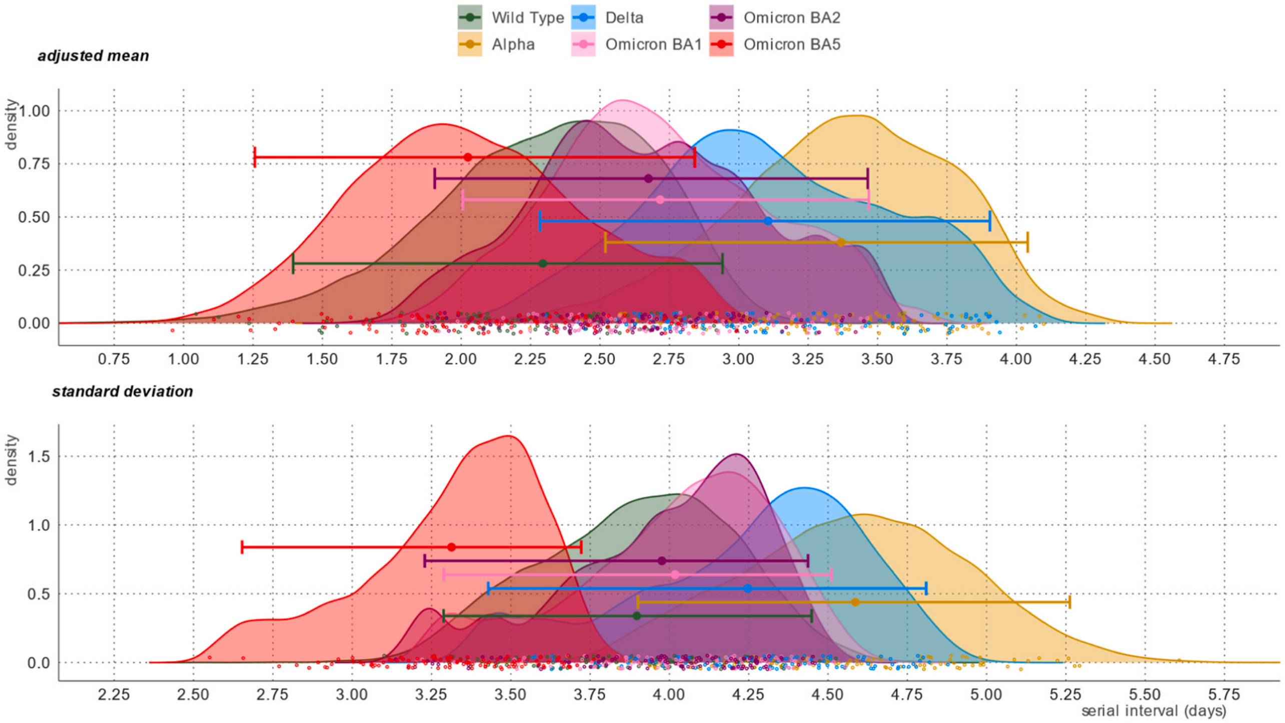

Fig. 3. Posterior distributions of the adjusted serial interval mean and standard deviation. Posterior densities have been computed from the posterior sample for all natural histories. Point represents the overall mean estimate. Error bar represents the 95% credible interval.

models across all VOC. In contrast to the *theoretical* model, the *outbreaker2* serial interval exhibits a higher proportion of negative serial intervals, driven by the data. However, the occurrence of negative serial intervals in the *outbreaker2* distribution remains lower than that of the *pairwise* model, which can be attributed to the input generation time and incubation period distributions. The discrepancy between the *theoretical* and observed (*outbreaker2*) serial interval is primarily attributable to the fact that the *theoretical* distribution fails to account for the correlation between the generation time and incubation period (Lehtinen et al., 2021). However, this correlation would be reflected in the actual transmission pairs, and therefore in the *outbreaker2* serial interval. Additionally, the theoretical serial interval does not account for the rapid decline of susceptibles in a household (i.e. saturation) which leads to shortened serial intervals (Svensson, 2007; Kenah et al., 2008).

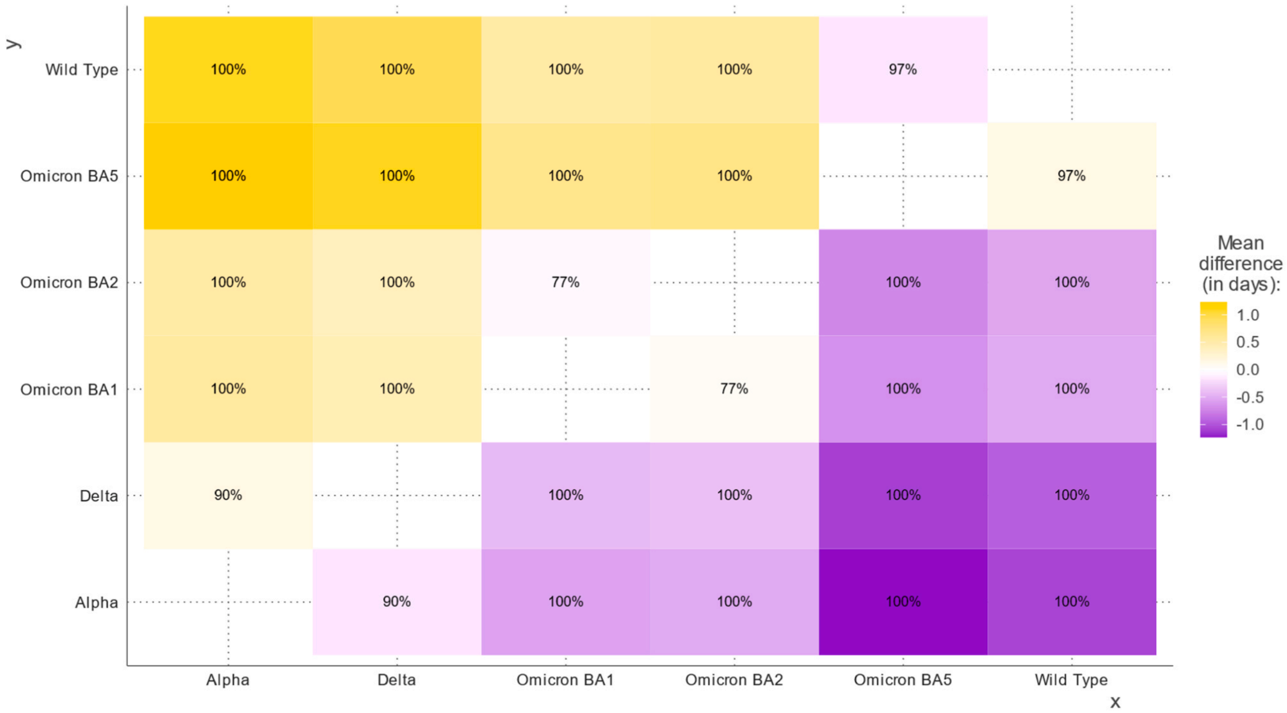

Fig. 4. Pairwise difference of the adjusted mean serial interval by VOC. The adjusted mean serial interval has been computed from the posterior sample for every pair of natural history, resulting in 100 estimates for every variant. Pairwise differences between each variant can be read from x to y (i.e. x - y) or y to x (i.e. y - x). The colour informs on the mean serial interval difference (in days), the text displays the frequency of that difference (e.g. Compared to Alpha, the adjusted mean serial interval of Delta was shorter by 0.5 day for 90 out of the 100 pairs of natural histories considered).

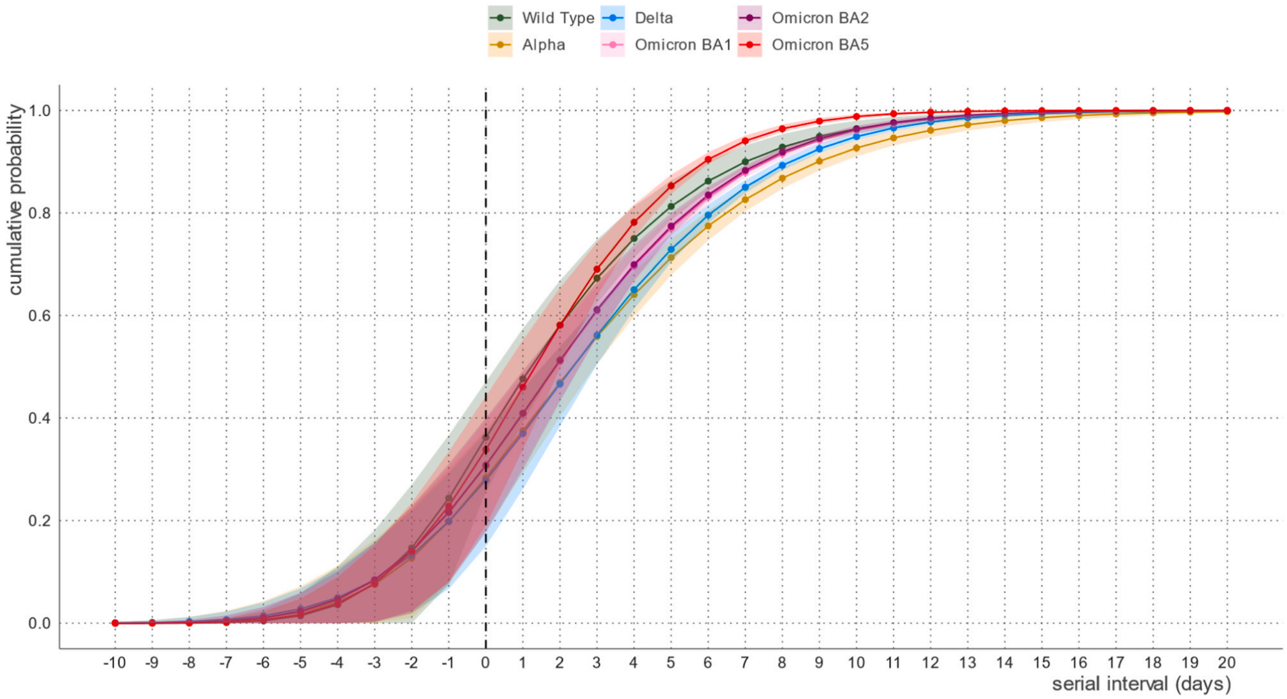

Fig. 5. Cumulative distribution function of the SARS-CoV-2 adjusted serial interval derived from the posterior sample by VOC.

4. Discussion

This study characterised the household serial interval by SARS-CoV-2 VOC using a Bayesian framework to reconstruct plausible transmission chains. The adjusted mean serial interval differed by variant with Omicron BA5 (2.02 days, 95%CrI: 1.26–2.84) being the shortest and Alpha (3.37 days, 95%CrI: 2.52–4.04) the longest.

Several factors could explain the observed changes in the serial interval over time. A change in the virus biology, through increased viral shedding or improved receptor binding, are likely to lead to quicker transmissions and shorter incubation periods (Kidd et al., 2021; Wu et al., 2022). Non-pharmaceutical interventions (NPIs) may also modify

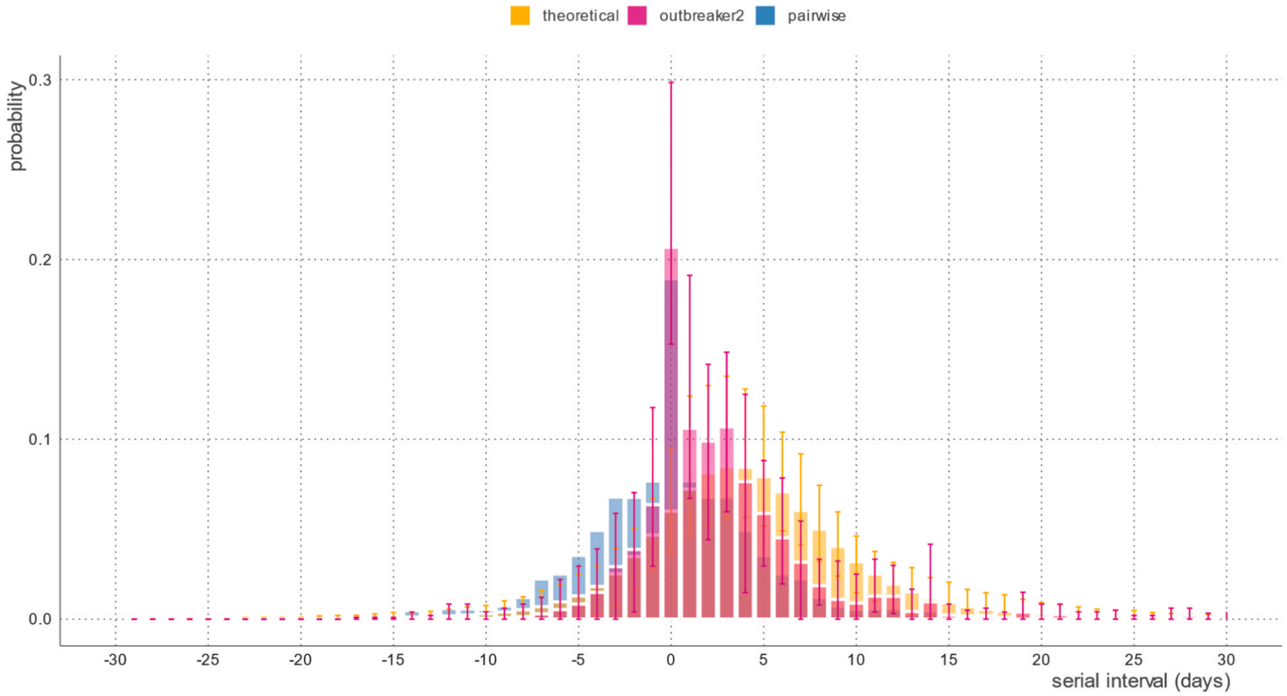

Fig. 6. SARS-CoV-2 serial interval distributions derived from the *theoretical* (yellow), *pairwise* (blue) & *outbreaker2* (red) models. Error bars display de 95% credible interval. The *theoretical* serial interval is solely derived from the 100 natural histories' probability mass functions (pmf) used as input for the *outbreaker2* model. The *pairwise* model is solely derived from the data and considers all infector infectee pairs with the same probability. The *outbreaker2* model considers both the natural histories pmf and the data to derive the serial interval.

the contact patterns within the household. Periods of lockdowns and movement restrictions are likely to increase the duration and frequency of household contacts, increasing the exposure amongst susceptible household members (Liu et al., 2021). Changing levels of immunity in the population during the pandemic through natural infections (Fig. 2), vaccination (appendix A14) and natural and vaccine-induced immunity waning may also affect the generation time and the serial interval by affecting the proportion of susceptible. Svensson's analytical demonstrations suggest that the generation time will significantly shorten as the proportion of susceptible decreases (Svensson, 2007). This "generation interval contraction" effect is particularly relevant for small household outbreaks (Kenah et al., 2008). Vaccinated individuals generally have milder symptoms (Bergwerk et al., 2021; Chen et al., 2022) and may take longer to report illness. Further research is required to find out the underlying factors related to the observed changes in the mean serial interval across the differing VOC periods.

Studies on the SARS-CoV-2 serial interval often rely on a small number of cases (generally in the hundreds) and do not clearly explain the approach used to determine the direction of the transmission (Kwok et al., 2020; Bao et al., 2021; Nishiura et al., 2020; Lavezzo et al., 2020; Wang and Teunis, 2020). Several studies, including our previous work (Geismar et al., 2021), were not able to infer negative serial interval values (Griffin et al., 2020; Nishiura et al., 2020; UKHSA, 2022a), thus reporting serial interval distributions with positive values only. Yet this study estimates that 22% of serial interval values are negative across all VOC (Fig. 5). This means that in one in five transmission events, the first person to present symptoms is not the first infected with SARS-CoV-2. This highlights the value of making SARS-CoV-2 tests available for the whole household as soon as a symptomatic case is declared as it might not be the first infected case.

Most methods and studies estimating the SARS-CoV-2 reproduction number (*R*) from incidence time series assume that the serial interval is strictly positive (Cori et al., 2013; Anderson et al., 2020; You et al., 2020). Our results suggest that this assumption may misrepresent how infectiousness evolves over time, which may cause biases in the estimates of *R*, and may also invalidate short-term forecasts derived from branching process models. Further research is needed to assess the potential impact of negative serial intervals on these methods.

Given that most studies have only examined positive serial intervals derived from contact tracing data, our findings exhibit lower values compared to previously published reports. Our results are lower compared to the UKHSA report on the serial interval of Omicron BA1 and Delta (UKHSA, 2022a) and compared to pooled estimates from meta-analyses of data from international studies published early in the pandemic prior to the emergence of VOC (mean range = 3.03 – 7.6 days) (mean = 5.2 days, 95% Confidence Interval (CI): 4.9–5.5) (Griffin et al., 2020; Alene et al., 2021). However, our estimates were similar to those of Ganyani et.al for wild-type circulating in Tianjin (China) up to March 2020, when the authors considered all possible negative serial intervals (Ganyani et al., 2020). Our study produced estimates for Delta and Omicron BA.1 that align with those reported by Kremer et al., who also adjusted for negative serial intervals (Kremer et al., 2022). Our estimates for Omicron BA.1 and BA.2 align with those of Manica et al. who used a Bayesian inference model to reconstruct household transmission pairs in Italy during January 2022 (Manica et al., 2022). The close and frequent nature of contacts amongst household cases and rapid saturation might explain shorter serial intervals in contrast with studies estimating the serial interval from contact tracing data (Svensson, 2007; Alene et al., 2021). Additionally, differences in populations, social contact, timeframes, NPI, immunity levels and variants may explain the range of estimates reported (UKHSA, 2022a; Ali et al., 2020).

Strengths of the study include the large number of transmission pairs, the weekly reporting of symptoms and swab test results in a large household cohort over 24 months, allowing to compare all major VOC in the UK. Our method accounts for the uncertainty in the direction of transmissions and the uncertainty of the underlying generation time and incubation period distributions. The code for this analysis is available online (https://github.com/CyGei/serial_interval). This should allow the rapid assessment of the serial interval in different contexts in future epidemics.

Limitations of our analysis include the reliance on self-reported samples meaning that we would miss cases that did not test. However, most reported illness episodes (60%) were tested for SARS-CoV-2 across the study period (appendix A7). Removing our 20-day serial interval limit would lead to longer serial intervals, however most studies report values below this threshold (Griffin et al., 2020; Alene et al., 2021). In the absence of pathogen genetic sequence data, our model uses solely information on the symptom onset dates to infer who infected whom, with the assumption of within household transmission being supported by previous studies suggesting that households have the highest transmission rates among indoor settings (Jombart et al., 2014). Sequence data would have been beneficial to inform on transmissions and to determine whether positive cases are the result of within-household transmissions or are imported cases. In fourteen percent of households, all cases reported the same symptom onset date. This observation could potentially be attributed to a mixture of “co-primaries” i.e., cases that were exposed and infected outside the household at the same event and reported symptoms at the same time, and secondary transmissions. Although participants reported weekly, dates of symptom onset may be subject to inaccurate recall where some households might have recorded a single symptom onset date for cases that occurred close together but on different days. We are not aware of other studies that reported serial interval distributions with substantial density at zero. Thus, we report the SARS-CoV-2 serial interval estimates from gamma distributions fitted to non-zero observations to account for the aforementioned limitation. This approach had negligible impact on the serial interval estimate (appendix A15).

Our results show that uncertainty in the incubation period and generation time can have a significant impact on the estimation of the serial interval distribution (appendix B). Further improvements of this work could focus on using variant-specific natural histories, which may impact further estimations of the serial interval. Nonetheless, our simulation study in Appendix B suggests that unless the mean generation time or the standard deviation of the incubation period vary greatly between variants (over a factor of 2) - which seems unlikely based on the current literature -, using the same natural histories for all variants should not significantly bias our estimates.

In conclusion, our analysis highlights some difference in the mean serial interval between variants and the large proportion of negative serial intervals. Further research is needed to explain those differences and assess the extent to which negative serial intervals impact estimates of *R* and short-term forecasts of case incidence.

Ethics

The Virus Watch study was approved by the Hampstead NHS Health Research Authority Ethics Committee. Ethics approval number - 20/HRA/2320. All members of participating households provided informed consent for themselves and, where relevant, for children that they were responsible for. This was electronically collected during registration. All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived.

Funding

CG is supported by a PhD studentship at Imperial College London funded by the National Institute for Health Research (NIHR) Health Protection Research Unit (HPRU) in Modelling and Health Economics, which is a partnership between the UK Health Security Agency (UKHSA), Imperial College London, and the London School of Hygiene & Tropical Medicine (grant code NIHR200908). AC, PJW are supported by the HPRU in Modelling and Health Economics. AC, PJW are also supported by the MRC Centre for Global Infectious Disease Analysis (grant number MR/R015600/1); this award is jointly funded by the MRC and Foreign, Commonwealth and Development Office (FCDO) under the MRC/FCDO Concordat agreement, and is also part of the European and Developing Countries Clinical Trials Partnership programme supported by the EU. The Virus Watch study is supported by the MRC Grant Ref: MC_PC 19070 awarded to University College London on 30 March 2020 and MRC Grant Ref: MR/V028375/1 awarded on 17 August 2020. The study also received $15,000 of Facebook advertising credit to support a pilot social media recruitment campaign on 18th August 2020. This study was also supported by the Wellcome Trust through a Wellcome Clinical Research Career Development Fellowship to RWA [206602]. The funders had no role in study design, data collection, analysis and interpretation, in the writing of this report, or in the decision to submit the paper for publication. The views expressed are those of the authors and not necessarily those of the UK Department of Health and Social Care (DHSC), Foreign, Commonwealth and Development Office (FCDO), European Union (EU), UK Medical Research Council (MRC), National Institute for Health and Care Research (NIHR), UK Health Security Agency (UKHSA), or Wellcome Trust. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

CRediT authorship contribution statement

Conceptualization, Ideas; formulation or evolution of overarching research goals and aims, **CG, AC, TJ, PJW**. Methodology, Development or design of methodology; creation of models, **CG, AC, TJ, PJW**. Software, Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components, **CG, AC, TJ**. Validation, Verification, whether as a part of the activity or separate, of the overall replication/ reproducibility of results/experiments and other research outputs, **AC, TJ, PJW, CG**. Formal analysis, Application of statistical, mathematical, computational, or other formal techniques to analyse or synthesize study data, **CG**. Resources, Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools, Virus Watch Team: **RWA, ACH, JK, CG, VN, EF, MS, AMDN, SB, TEB, WLEF, AY, IB, SH**. Data Curation, Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later reuse, **CG, VN, EF, MS, AMDN**. Writing - Original Draft, Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation), **CG**. Writing - Review & Editing, Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre-or postpublication stages, **CG, TJ, PJW, AC**. Visualization, Preparation, creation and/or presentation of the published work, specifically visualization/ data presentation, **CG**. Supervision, Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team, **AC, TJ, PJW**. Project administration, Management and coordination responsibility for the research activity planning and execution, **JK, ACH, RWA**. Funding acquisition, Acquisition of the financial support for the project leading to this publication, **RWA, ACH** (funding for running the Virus Watch study) **AC, TJ, PJW** (funding for the data analysis).

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Cyril Geismar: National Institute for Health Research (NIHR) Health Protection Research Unit (HPRU) in Modelling and Health Economics PhD funding, grant code: NIHR200908. Anne Cori: NIHR Grant paid to institution (HPRU in Modelling and Health Economics); Sergei Brin foundation Grant paid to institution (Mapping the Risk of International Infectious Disease Spread II); USAID Grant paid to institution (Mapping the Risk of International Infectious Disease Spread); Academy of Medical Sciences Grant paid to institution (Developing reliable epidemic forecasting using branching processes: Ebola as a case study); Pfizer Money paid to institution. Lecturing on a course on mathematical modelling of infectious disease transmission and vaccination. Peter J White: Medical Research CouncilFunding for the MRC Centre for Global Infectious Disease Analysis; Public Health England Salary; National Institute for Health Research, Funding for the Health Protection Research Unit in Modelling and Health Economics (2020–2025), Funding for the Health Protection, Research Unit in Modelling Methodology (2014–2020), Funding for work on tuberculosis; Pfizer Payment for teaching of mathematical modelling of infectious disease transmission and vaccination. Robert Aldridge: The research costs for the study have been supported by the MRC Grant Ref: MC_PC 19070 awarded to UCL on 30 March 2020 and MRC Grant Ref: MR/V028375/1 awarded on 17 August 2020. The study also received $15,000 of Facebook advertising credit to support a pilot social media recruitment campaign on 18th August 2020. This study was supported by the Wellcome Trust through a Wellcome Clinical Research Career Development Fellowship to RWA [206602]. Thibaut Jombart: SPI-M I have been a member of SPI-M in 2021 as an infectious disease modeller advising the response to COVID-19 in the UK. No funding received; RECON President of the R Epidemics Consortium, a non-profit organization developing free tools for outbreak analytics.

All other authors declare no conflict of interests.

Data availability

We aim to share aggregate data from this project on our website www.ucl-virus-watch.net and via a "Findings so far" section on our website. We are sharing individual record-level data on the Office of National Statistics (ONS) Secure Research Service MDX Browser > Virus Watch - England and Wales @89201 (metadata.works). The data are available under restricted access as they contain sensitive health data. Access can be obtained by ONS Secure Research Service. Code for the analysis is available on Github: https://github.com/CyGei/serial_interval. For a fully reproducible example of our methods, we provided the code of our simulation study available at https://github.com/CyGei/SI_simulation.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.epidem.2023.100713.

References

Alene, M., Yismaw, L., Assemie, M.A., Ketema, D.B., Gietaneh, W., Birhan, T.Y., 2021. Serial interval and incubation period of COVID-19: a systematic review and meta-analysis. BMC Infect. Dis. 21 (1), 257.

Ali, S.T., Wang, L., Lau, E.H., Xu, X.-K., Du, Z., Wu, Y., et al., 2020. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. Science 369 (6507), 1106–1109.

Anderson R., Donnelly C., Hollingsworth D., Keeling M., Vegvari C., Baggaley R., et al., Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK: methods of estimation, data sources, causes of heterogeneity, and use as a guide in policy formulation. The Royal Society. 2020;2020.

Bao, C., Pan, E., Ai, J., Dai, Q., Xu, K., Shi, N., et al., 2021. COVID-19 outbreak following a single patient exposure at an entertainment site: an epidemiological study. Transbound. Emerg. Dis. 68 (2), 773–781.

Bergwerk, M., Gonen, T., Lustig, Y., Amit, S., Lipsitch, M., Cohen, C., et al., 2021. Covid-19 Breakthrough Infections in Vaccinated Health Care Workers. N. Engl. J. Med. 385 (16), 1474–1484.

Bhatia S., Wardle J., Nash R.K., Nouvellet P., Cori A. A generic method and software to estimate the transmission advantage of pathogen variants in real-time: SARS-CoV-2 as a case-study. medRxiv. 2021:2021.11.26.21266899.

Campbell, F., Didelot, X., Fitzjohn, R., Ferguson, N., Cori, A., Jombart, T., 2018. outbreaker2: a modular platform for outbreak reconstruction. BMC Bioinforma. 19 (S11).

Chen, J., Wang, R., Gilby, N.B., Wei, G.-W., 2022. Omicron variant (B.1.1.529): infectivity, vaccine breakthrough, and antibody resistance. J. Chem. Inf. Model. 62 (2), 412–422.

Cori, A., Ferguson, N.M., Fraser, C., Cauchemez, S., 2013. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am. J. Epidemiol. 178 (9), 1505–1512.

Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J.D., et al., 2021. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science 372 (6538), eabg3055.

Du, Z., Xu, X., Wu, Y., Wang, L., Cowling, B.J., Meyers, L.A., 2020. COVID-19 serial interval estimates based on confirmed cases in public reports from 86 Chinese cities. medRxiv.

Ganyani, T., Kremer, C., Chen, D., Torneri, A., Faes, C., Wallinga, J., et al., 2020. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. Eurosurveillance 25 (17), 2000257.

Geismar, C., Fragaszy, E., Nguyen, V., Fong, W., Shrotri, M., Beale, S., et al., 2021. Household serial interval of COVID-19 and the effect of Variant B.1.1.7: analyses from prospective community cohort study (Virus Watch) [version 2; peer review: 2 approved]. Wellcome Open Res. 6, 224.

Griffin, J., Casey, M., Collins, Á., Hunt, K., McEvoy, D., Byrne, A., et al., 2020. Rapid review of available evidence on the serial interval and generation time of COVID-19. BMJ Open 10 (11), e040263.

Hayward, A., Fragaszy, E., Kovar, J., Nguyen, V., Beale, S., Byrne, T., et al., 2021. Risk factors, symptom reporting, healthcare-seeking behaviour and adherence to public health guidance: protocol for Virus Watch, a prospective community cohort study. BMJ Open 11 (6), e048042.

Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., Ferguson, N., 2014. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLOS Comput. Biol. 10 (1), e1003457.

Kenah, E., Lipsitch, M., Robins, J.M., 2008. Generation interval contraction and epidemic data analysis. Math. Biosci. 213 (1), 71–79.

Kidd, M., Richter, A., Best, A., Cumley, N., Mirza, J., Percival, B., et al., 2021. S-variant SARS-CoV-2 lineage B1.1.7 is associated with significantly higher viral load in samples tested by taqpath polymerase chain reaction. J. Infect. Dis. 223 (10), 1666–1670.

Kremer, C., Braeye, T., Proesmans, K., André, E., Torneri, A., Hens, N., 2022. Serial intervals for SARS-CoV-2 omicron and delta variants, Belgium, November 19-December 31, 2021. Emerg. Infect. Dis. 28 (8), 1699–1702.

Kwok, K.O., Wong, V.W.Y., Wei, W.I., Wong, S.Y.S., Tang, J.W.-T., 2020. Epidemiological characteristics of the first 53 laboratory-confirmed cases of COVID-19 epidemic in Hong Kong, 13 February 2020. Eurosurveillance 25 (16), 2000155.

Lavezzo, E., Franchin, E., Ciavarella, C., Cuomo-Dannenburg, G., Barzon, L., Del Vecchio, C., et al., 2020. Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'. Nature 584 (7821), 425–429.

Lehtinen, S., Ashcroft, P., Bonhoeffer, S., 2021. On the relationship between serial interval, infectiousness profile and generation time. J. R. Soc. Interface 18 (174), 20200756.

Liu, C.Y., Berlin, J., Kiti, M.C., Del Fava, E., Grow, A., Zagheni, E., et al., 2021. Rapid review of social contact patterns during the COVID-19 pandemic. Epidemiology 32 (6), 781–791.

Manica, M., De Bellis, A., Guzzetta, G., Mancuso, P., Vicentini, M., Venturelli, F., et al., 2022. Intrinsic generation time of the SARS-CoV-2 Omicron variant: an observational study of household transmission. Lancet Reg. Health - Eur. 19, 100446.

Nishiura, H., Linton, N.M., Akhmetzhanov, A.R., 2020. Serial interval of novel coronavirus (COVID-19) infections. Int J. Infect. Dis. 93, 284–286.

Svensson, Å., 2007. A note on generation times in epidemic models. Math. Biosci. 208 (1), 300–311.

UKHSA. Omicron and Delta serial interval distributions from UK contact tracing data. 2022a 14/01/2022.

UKHSA. Investigation of SARS-CoV-2 variants: technical briefings. 2022b.

Vink, M.A., Bootsma, M.C.J., Wallinga, J., 2014. Serial intervals of respiratory infectious diseases: a systematic review and analysis. Am. J. Epidemiol. 180 (9), 865–875.

Wallinga, J., Teunis, P., 2004. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am. J. Epidemiol. 160 (6), 509–516.

Wang, Y., Teunis, P., 2020. Strongly heterogeneous transmission of COVID-19 in mainland China: local and regional variation. Front. Med. 7, 329.

Wu, Y., Kang, L., Guo, Z., Liu, J., Liu, M., Liang, W., 2022. Incubation period of COVID-19 caused by unique SARS-CoV-2 strains. JAMA Netw. Open 5 (8), e2228008.

You, C., Deng, Y., Hu, W., Sun, J., Lin, Q., Zhou, F., et al., 2020. Estimation of the time-varying reproduction number of COVID-19 outbreak in China. Int. J. Hyg. Environ. Health 228, 113555.

Zhang, M., Xiao, J., Deng, A., Zhang, Y., Zhuang, Y., Hu, T., et al., 2021. Transmission dynamics of an outbreak of the COVID-19 delta variant B.1.617.2 — Guangdong Province, China, May–June 2021. China CDC Wkly. 3 (27), 584–586.