**Report 47: A generic method and software to estimate the transmission advantage of pathogen variants in real-time : SARS-CoV-2 as a case-study**

Sangeeta Bhatia[1,*], Jack Wardle[1,*], Rebecca K Nash[1], Pierre Nouvellet[1, 2], Anne Cori[1,**]

**1** MRC Centre for Global Infectious Disease Analysis, Jameel Institute, School of Public Health, Imperial College London.
**2** School of Life Sciences, University of Sussex, Brighton, UK.

* equal contribution
** corresponding author (a.cori@imperial.ac.uk)

**Summary**

Recent months have demonstrated that emerging variants may set back the global COVID-19 response. The ability to rapidly assess the threat of new variants in real-time is critical for timely optimisation of control strategies.
We extend the EpiEstim R package, designed to estimate the time-varying reproduction number ($R_t$), to estimate in real-time the effective transmission advantage of a new variant compared to a reference variant. Our method can combine information across multiple locations and over time and was validated using an extensive simulation study, designed to mimic a variety of real-time epidemic contexts.
We estimate that the SARS-CoV-2 Alpha variant is 1.46 (95% Credible Interval 1.44-1.47) and 1.29, (95% CrI 1.29-1.30) times more transmissible than the wild type, using data from England and France respectively. We further estimate that Beta and Gamma combined are 1.25 (95% CrI 1.24-1.27) times more transmissible than the wildtype (France data). All results are in line with previous estimates from literature, but could have been obtained earlier and more easily with our off-the-shelf open-source tool.
Our tool can be used as an important first step towards quantifying the threat of new variants in real-time. Given the popularity of EpiEstim, this extension will likely be used widely to monitor the co-circulation and/or emergence of multiple variants of infectious pathogens.

doi: https://doi.org/10.25561/92689 

**Significance Statement**

Early assessment of the transmissibility of new variants of an infectious pathogen is critical for anticipating their impact and designing appropriate interventions. However, this often requires complex and bespoke analyses relying on multiple data streams, including genomic data. Here we present a novel method and software to rapidly quantify the transmission advantage of new variants. Our method is fast and requires only routinely collected disease surveillance data, making it easy to use in real-time. The ongoing high level of SARS-CoV-2 circulation in a number of countries makes the emergence of new variants highly likely. Our work offers a powerful tool to help public health bodies monitor such emerging variants and rapidly detect those with increased transmissibility.

**Introduction**

The SARS-CoV-2 pandemic has highlighted the potentially dramatic influence that emerging novel pathogen variants can have on transmission dynamics and on the control measures needed to mitigate the epidemic burden. The emergence of the Alpha variant of SARS-CoV-2 in September 2020, and of the Delta variant in December 2020 drastically altered the trajectory of the COVID-19 epidemic in several countries leading to renewed imposition of public health measures such as lockdowns [1, 2]. The continued high level of transmission of SARS-CoV-2 globally makes the emergence of new variants very likely. As of December 2021, the World Health Organization has classified four variants of SARS-CoV-2 as "variants of concern" (i.e. Alpha, Beta, Gamma and Delta), because of their increased transmissibility, severity, and/or immune escape properties compared to the circulating SARS-CoV-2 variants [3].

Rapidly quantifying characteristics of such emerging variants is critical to anticipate their potential impact and adjust interventions accordingly. Shortly after the emergence of the Alpha variant in England in September 2020 [4], a number of studies aimed to estimate its transmission potential, compared to the previously circulating non-VOC lineages [4–7]. More recently, several papers have evaluated the transmissibility of VOCs compared to non-VOC lineages [8–16]. All of these studies have developed new approaches to estimate the transmission advantages of new VOCs, often synthesising evidence from multiple data sources including genomic data. The time and expertise required to design and implement such approaches, with methods tailored to the specificity of each dataset and context, greatly limit their widescale and real-time use.

In this study, we present a new Bayesian inference method, MV-EpiEstim (for Multi-Variant EpiEstim), to estimate in real-time the transmission advantage of a new variant of a pathogen compared to a reference variant, using simple data consisting of the time series of incidence of cases of each variant in one or more locations. In the rest of the manuscript, we refer to different "variants" but the method can be equally applied to different strains. We present the method for one reference and one new variant, but the method naturally extends to more than one new variant. Our work build on a previously published methodology [17, 18] to estimate the instantaneous reproduction number $R_t$ (defined as the average number of secondary cases that an individual infected at time $t$ would generate if conditions remained the same as at time $t$).

We assume that locally, the transmissibility of all variants follows the same temporal pattern, i.e. the reproduction number of the new variant is the same as that of the reference variant, albeit with a multiplicative factor. We refer to this multiplicative factor as the "effective transmission advantage" of the new variant, compared to the reference variant. We further assume that the effective transmission advantage remains constant over time and across all locations under consideration.

We provide an open source implementation of our method in the R package EpiEstim [19]. The ap-

proach, which we validate on an extensive simulation study, is computationally efficient as it takes advantage of an analytical formulation of marginal posterior densities of the instantaneous reproduction number for the reference variant on the one hand, and the transmission advantage of the new variant on the other hand.
We retrospectively estimate the effective transmission advantage of the Alpha SARS-CoV-2 variant compared to the other non-VOC lineages circulating at the time using data from England and France. We show that the estimates from our method are consistent with those from several bespoke studies, and could have been obtained earlier and with improved accuracy. Our inference framework and open source software should allow rapid quantification of the effective transmission advantage of future new variants in real-time.

**Results**

## Transmission advantage of SARS-CoV-2 variants

We used MV-EpiEstim to retrospectively estimate the transmission advantage of SARS-CoV-2 variants using data from England and France. The Alpha variant originated in late summer to early Autumn 2020 in England (before vaccination was initiated), where it became dominant in early 2021 (Fig. 1A). England never experienced substantial transmission of the Beta and Gamma variants, first detected in South Africa and Brazil respectively [3].
In France, the Alpha variant emerged in early 2021, rapidly dominating cases in metropolitan France and the French West Indies [20, 21]. The Beta and Gamma variants were also circulating from January 2021 in most regions, and accounted for the majority of cases in French Guyana and la Réunion from spring 2021 [22] (Fig. S2).
We considered daily variant-specific incidence data from 7 National Health Service (NHS) regions in England between 1st September 2020 and 14th March 2021 (Fig. S1), and from 18 ADM2 regions in France between 18th February and 30th May 2021 (Figs. S2 and S3).
For simplicity, we refer to all lineages of SARS-CoV-2 other than the VOCs that were circulating at the time as 'wildtype'. $R_t$ estimates obtained independently for the wildtype and for Alpha indicated that Alpha was more transmissible (Fig. 1B). However, the magnitude of the transmission advantage (naively estimated as the ratio between the two $R_t$s, see Suppl Sec. 3 for details) varied over time and across regions. Pooling these naïve estimates over time and regions yielded a highly uncertain and non-significant transmission advantage of 1.41 (95% Credible Interval (CrI) 0.86-2.01) for Alpha compared to the wildtype in England.
By explicitly assuming that the effective transmission advantage, which we denote as $\epsilon$, remains constant over time and across regions, MV-EpiEstim reduces the uncertainty in the estimates. Using MV-EpiEstim with data from all NHS regions, we found strong evidence that Alpha was more transmissible than the wildtype ($\epsilon$ = 1.46, 95% CrI 1.44-1.47. See also Suppl Tab. S1).
To mimic real-time use, we examined how $\epsilon$ estimates varied as more data became available. The central estimate steadily increased from around 1 to approximately 1.5 by early December 2020, with uncertainty decreasing in that period; estimates then remained relatively stable (Fig. 1C). As a comparison, Volz et al. first estimated the multiplicative transmission advantage of Alpha to be 1.74 (95% CrI 1.40-1.80), in a report dated December 31st 2020.
Alpha cases accounted for less than 10% of all cases between mid-September and early-December (Fig. 1C). The variability in $\epsilon$ estimates in this period suggests that accurate estimation of the transmission advantage can only be achieved once enough cases of the new variant have been observed.
We also used MV-EpiEstim to estimate $\epsilon$ separately for each NHS region, highlighting minor regional

differences with $\epsilon$ ranging from 1.36 (95% CrI 1.33-1.39) in the South-East to 1.54 (95% CrI 1.50-1.58) in the Midlands (Fig. 1D, Suppl Tab. S1).
We estimated a similar, albeit slightly lower, effective transmission advantage for Alpha using data from the 18 ADM2 regions in France ($\epsilon$ = 1.29, 95% CrI 1.29-1.30, see Fig. S3). In region-specific analyses (excluding regions where Beta/Gamma were dominant), $\epsilon$ varied from 1.21 (95% CrI 1.20-1.23) in Île-de-France to 1.41 (95% CrI 1.37-1.46) in Bourgogne-Franche-Comté (Fig. S3 and Suppl Tab. S2).
Following the same approach, and using data from France, we demonstrated that the Beta and Gamma variants (combined) are also more transmissible than the wildtype ($\epsilon$ = 1.25, 95% CrI 1.24-1.27, Fig. S4 and Suppl Tab. S3).

## Method validation

We assessed the validity of our method using simulations under several scenarios with different values for the transmissibility of each variant, allowing for superspreading and under-reporting as well as differences in natural history between variants (Suppl Sec. 5). The method performed well across all scenarios considered, with a small bias (defined as the difference between the mean posterior estimate and the true $\epsilon$ value, Fig. 2). MV-EpiEstim was able to accurately estimate the transmission advantage when variants were known to differ in their natural history (characterised by the serial interval distribution, i.e. the delay between onset of symptoms in a case and their infector, Fig. 2c and e). We also explored a scenario typical of real-time outbreak analysis where the natural history of the new variant is different, but in the absence of information, is assumed to be the same as that of the reference. Misspecifying the mean serial interval led to substantial biases, especially when the transmission advantage was moderate (more than 1.5) and the mean serial interval of the new variant was much shorter than (less than half) that of the reference (Fig. 2d). Misspecifying the coefficient of variation of the serial interval had little impact on the quality of the estimates, unless the transmission advantage was very high (more than 2, Fig. 2f).
Even in the presence of substantial superspreading (equivalent to that of SARS-CoV-1, Fig. 2b) or poor case-reporting (up to 80% cases not reported, Fig. S18), neither of which is explicitly accounted for by MV-EpiEstim, the transmission advantage remained unbiased.
In all scenarios, using more days of data reduced both the bias and the uncertainty in the estimated effective transmission advantage (Suppl Secs. 5.3 to 5.7).
We used the full posterior distribution of $\epsilon$ to classify the variant as more or less transmissible than the reference (see Methods). Crucially, in many scenarios including some where the bias was large, MV-EpiEstim was able to correctly characterise a variant as being more transmissible than the reference. For instance, when the mean serial interval of the new variant was shorter but misspecified, the variant was still correctly classified as more transmissible since $\epsilon$ was over-estimated (Fig. S10, scenario type low). Conversely, when the mean serial interval of the new variant was longer but was misspecified, correct classification was only feasible with sufficient days of data and a large transmission advantage (Fig. S10, scenario type high).
More results using fewer days of data, two locations, time-varying $R_t$ and accounting for under-reporting are shown in Suppl Sec. 5.

## Discussion

In this study we present a novel method, MV-EpiEstim, to estimate the transmission advantage of a new variant of a pathogen over a reference variant. MV-EpiEstim builds on the EpiEstim method [17], which was found to perform better than other approaches for estimating the instantaneous reproduc-

tion number [23]. As such, MV-EpiEstim offers the same functionalities as EpiEstim, including explicitly accounting for imported cases [18]. Because MV-EpiEstim is based on analytical formulations of the marginal posterior densities, the run time of a typical analysis is very short (a few minutes at most on a standard laptop for all analyses presented here). MV-EpiEstim is implemented as a new function ("estimate_advantage") in the R package EpiEstim [24].

We show that MV-EpiEstim could have precisely estimated the effective transmission advantage of the SARS-CoV-2 Alpha variant a few weeks before the earliest published estimate. Importantly, our method only requires as inputs time-series of incident cases and serial interval distributions for each variant. If specific bio-markers are sufficient to distinguish variants (e.g. S-gene), no Whole-Genome-Sequence data is required. Therefore, MV-EpiEstim could be used in near real-time, relying only on routinely collected incidence data and not necessarily suffering from potential delays in the sequencing pipeline.

Our method works well across a range of simulated scenarios, designed to mimic a variety of real-time epidemic contexts, including in the presence of superspreading and when the natural history of the new variant is imperfectly characterised. In the absence of precise information on this natural history, the fast run time offers the possibility of exploring various assumptions and in turn estimate a range of plausible transmission advantages. Our method is robust to under-reporting and temporal changes in reporting if these affect both the reference and the variant equally.

Importantly, we show that our method can accurately characterise a variant as being 'more' or 'less' transmissible than a reference variant. This simple but robust characterisation could be as important as estimating the exact value of the transmission advantage, especially in informing public health response during the early emergence of a new variant.

We emphasise that our method estimates the *effective* transmission advantage, which will often reflect a combination of several factors such as a true increase in underlying transmissibility and the ability of a new variant to escape immunity. Disentangling these effects is particularly challenging in the context of changing population immunity e.g. due to vaccination roll-out, and may require additional data [25]. However, regardless of its drivers, early identification of a transmission advantage is a critical first step to a timely response.

MV-EpiEstim allows combining information across time and locations, assuming that the effective transmission advantage is constant across these. This allows reducing the uncertainty in the estimates. Temporal or spatial heterogeneity in the transmission advantage (e.g. reflecting heterogeneity in population immunity) can also be characterised by applying the method separately by location or time period, which is easy to do in our software.

Our estimated transmission advantage of the SARS-CoV-2 Alpha variant (over the wildtype) is consistent with those from bespoke analyses using multiple data streams including whole-genome-sequence data [26–30].

The incidence of SARS-CoV-2 in much of the world is still high with nearly 3 million cases reported every week in November 2021 [31]. Given the continued high levels of SARS-CoV-2 transmission and low vaccination coverage globally [32], new variants are likely to continue emerging. Our tool can be used to monitor their transmissibility and rapidly identify variants of concern.

Applications of our work are not limited to SARS-CoV-2; our generic method could easily be used to monitor other pathogens with multiple co-circulating strains such as influenza or *streptococcus pneumoniae.*

**Figures**

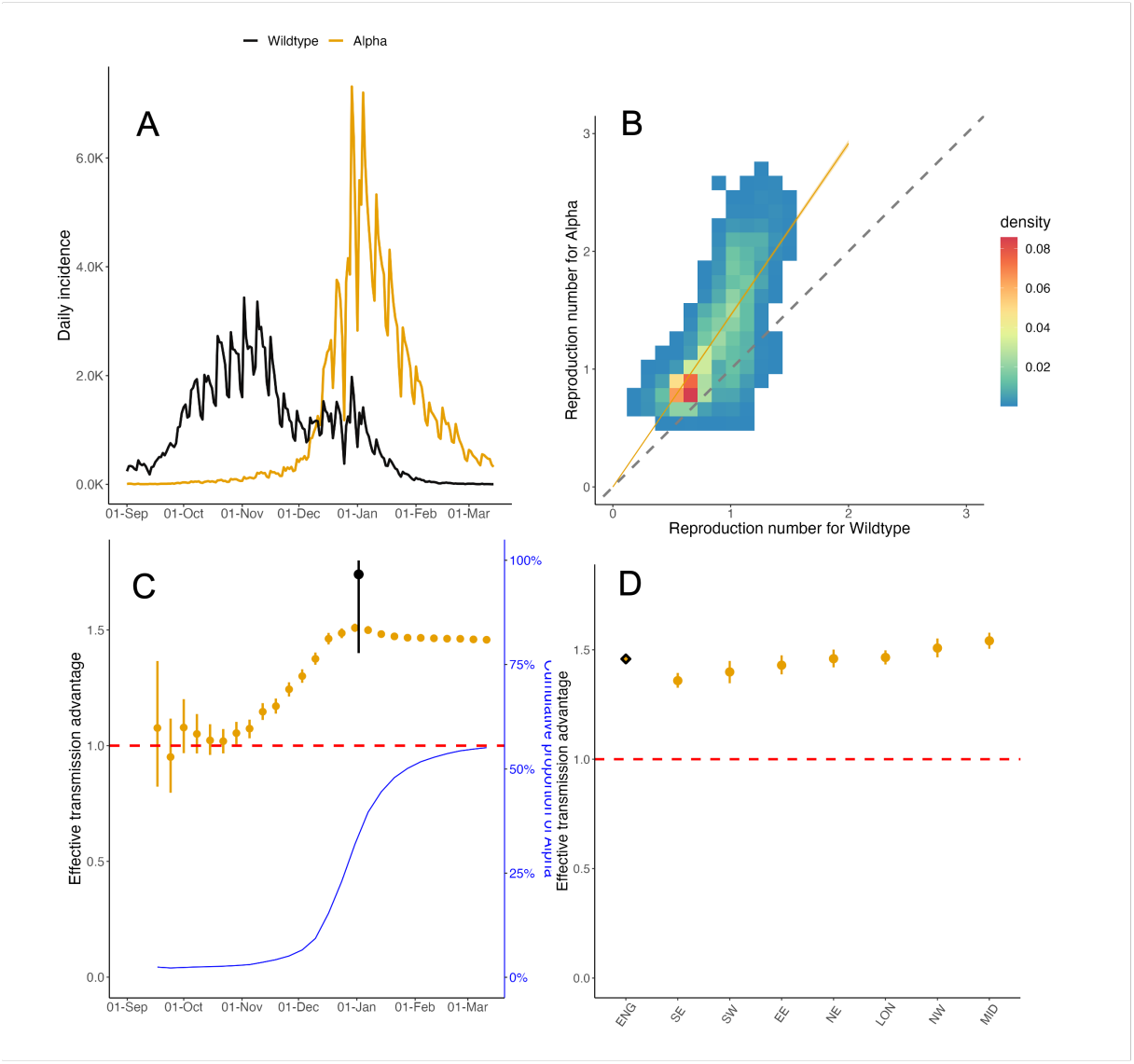


Figure 1: **Effective transmission advantage of the Alpha SARS-CoV-2 variant over the wildtype in England** (A) The daily reported incidence of cases of the wildtype (black) and Alpha (orange) in England from September 2020 to March 2021. (B) The effective reproduction number $R_t$ estimated independently for the wildtype (x-axis) and Alpha (y-axis) on sliding weekly windows. The colour of the cells indicates the density of the draws from the respective posterior distributions of $R_t$. The dashed diagonal line indicates the $x = y$ threshold. Coloured cells lying above the diagonal line suggest that Alpha is more transmissible. The orange line denotes the median effective transmission advantage estimated using MV-EpiEstim. 95% CrI were so narrow that they could not be distinguished from the line. (C) Effective transmission advantage estimated using MV-EpiEstim using data available up to the date specified on the x-axis. The dark blue line denotes the proportion of cumulative incidence of Alpha (right y-axis) counted from 1st September 2020. The black estimate corresponds to the multiplicative transmission advantage of Alpha estimated by Volz et al [33] in a report published on 31st December 2020. (D) Effective transmission advantage estimated using MV-EpiEstim for all NHS England regions together (diamond) and separately (solid circles), using data from 1st September 2020 to 14th March 2021. The NHS England regions are - East of England (EE), London (LON), Midlands (MID), North-East (NE), North-West (NW), South-East (SE), South-West (SW). In panels (C) and (D), the solid circles denote the median estimate, the vertical lines indicate the 95% CrI, and the red dashed line denotes the $\epsilon = 1$ threshold.

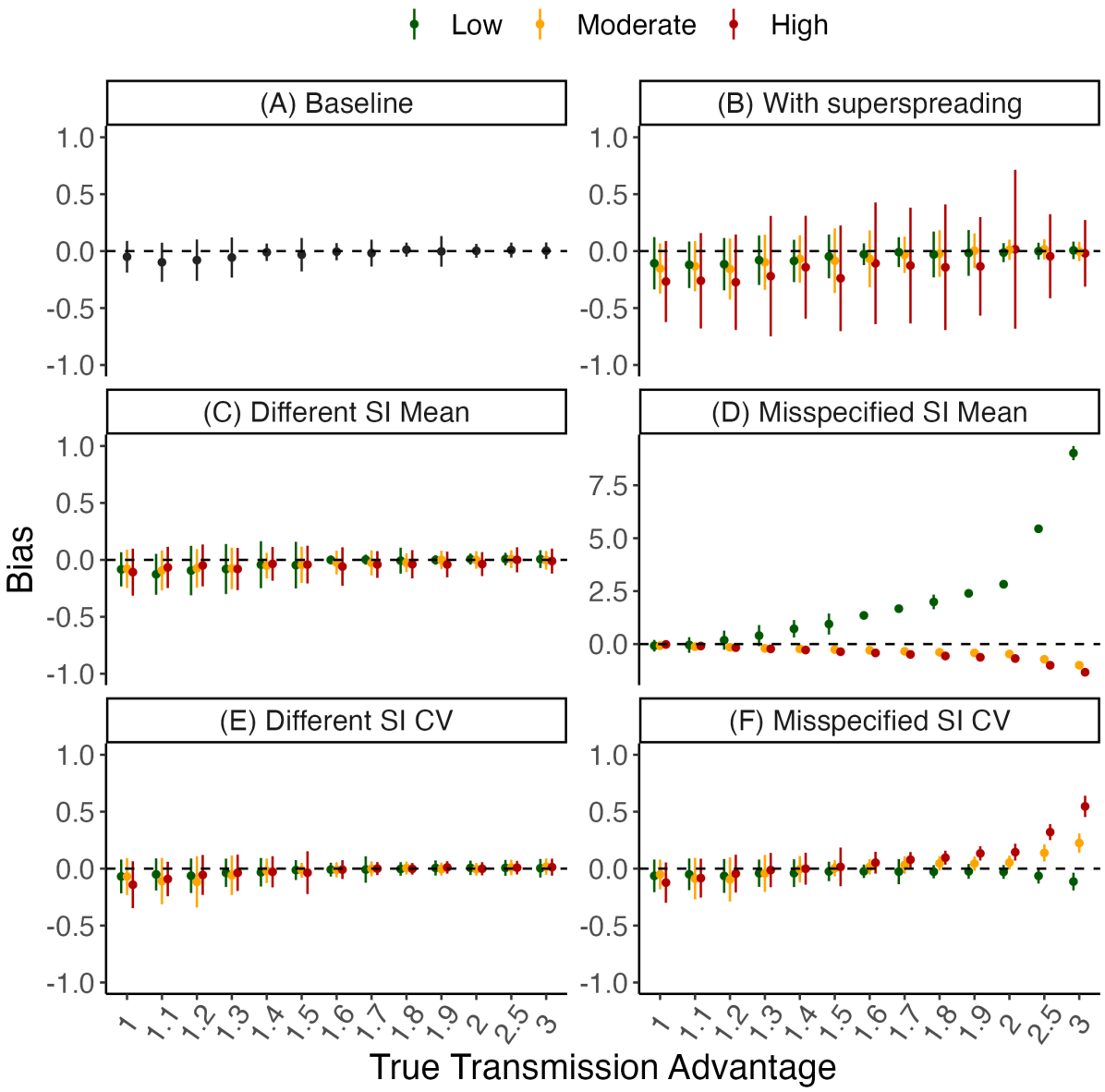


Figure 2: **Method performance on simulated data.** We assessed the performance of MV-EpiEstim on a range of scenarios. In each panel, the x-axis shows the true value of the effective transmission advantage, $\epsilon$ (on categorical scale). The y-axis shows the bias i.e., the difference between the posterior mean estimate of the transmission advantage and the true value. The solid dots represent the mean bias (across 100 simulations) and the vertical bars show the standard deviation (SD) of the bias. Each panel corresponds to a different simulation scenario. In all scenarios, the $R_t$ for the reference variant was 1.1 and the $R_t$ for the new variant was $\epsilon$ times the reference $R_t$ (see Suppl Sec. 5 for details). (A) In the baseline scenario, we assumed no superspreading and the same natural history for both variants. (B) As (A), but with low (overdispersion parameter $\kappa$ = 1), moderate ($\kappa$ = 0.5) and high ($\kappa$ = 0.1) levels of superspreading. (C) As (A), but the mean serial interval of the new variant is 0.5 (low), 1.5 (moderate) or 2 (high) times that of the reference and is correctly specified during estimation. (D) As (C), but the mean serial interval of the variant is assumed to be the same as that of the reference during estimation. (E) As (A), but the coefficient of variation (CV, ratio of standard deviation to mean) of the serial interval of the new variant is 0.5 (low), 1.5 (moderate) or 2 (high) times that of the reference and is correctly specified during estimation. (F) As (E), but the CV of the serial interval of the new variant is assumed to be the same as that of the reference during estimation. Note that the y-axis range is different for panel D. Results using $R_t = 1.6$ for the reference variant and using fewer days of data are presented in Suppl Secs. 5.3 to 5.9. Results using time-varying reference $R_t$ in one or two locations are shown in Suppl Sec. 5.10 and Suppl Sec. 5.11.

**References**


[1] *Prime Minister's statement on coronavirus (COVID-19): 19 December 2020*. Dec. 2020. URL: https://www.gov.uk/government/speeches/prime-ministers-statement-on-coronavirus-covid-19-19-december-2020 (visited on 11/08/2021).

[2] *PM statement at coronavirus press conference: 14 June 2021*. June 2021. URL: https://www.gov.uk/government/speeches/pm-statement-at-coronavirus-press-conference-14-june-2021 (visited on 10/25/2021).

[3] *Tracking SARS-CoV-2 Variants*. 2021.

[4] Voloch, Carolina M and da Silva Francisco Jr, Ronaldo and de Almeida, Luiz GP and Cardoso, Cynthia C and Brustolini, Otavio J and Gerber, Alexandra L and Guimarães, Ana Paula de C and Mariani, Diana and da Costa, Raissa Mirella and Ferreira Jr, Orlando C and others. "Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil". In: *Journal of Virology* 95.10 (2021), e00119–21.

[5] E. Volz et al. "Assessing transmissibility of SARS-CoV-2 lineage B. 1.1. 7 in England". In: *Nature* 593.7858 (2021), pp. 266–269. DOI: 10.1038/s41586-021-03470-x.

[6] N. G. Davies et al. "Increased mortality in community-tested cases of SARS-CoV-2 lineage B. 1.1. 7". In: *Nature* 593.7858 (2021), pp. 270–274.

[7] K. Leung et al. "Early Transmissibility Assessment of the N501Y Mutant Strains of SARS-CoV-2 in the United Kingdom, October to November 2020". In: *Eurosurveillance* 26.1 (Jan. 2021). ISSN: 1560-7917. DOI: 10.2807/1560-7917.ES.2020.26.1.2002106.

[8] *SARS-CoV-2 variants of concern and variants under investigation in England: Technical Briefing 15*. 2021. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/993879/Variants_of_Concern_VOC_Technical_Briefing_15.pdf (visited on 09/01/2021).

[9] R. Sonabend et al. "Non-Pharmaceutical Interventions, Vaccination, and the SARS-CoV-2 Delta Variant in England: A Mathematical Modelling Study". In: *The Lancet* (Oct. 2021), S0140673621022765. ISSN: 01406736. DOI: 10.1016/S0140-6736(21)02276-5.

[10] N. M. Ferguson. "B.1.617.2 Transmission in England: Risk Factors and Transmission Advantage". In: (), p. 14.

[11] N. R. Faria et al. "Genomics and Epidemiology of the P.1 SARS-CoV-2 Lineage in Manaus, Brazil". In: *Science* 372.6544 (May 2021), pp. 815–821. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.abh2644.

[12] S. Alizon et al. "Rapid Spread of the SARS-CoV-2 Delta Variant in Some French Regions, June 2021". In: *Eurosurveillance* 26.28 (July 2021). ISSN: 1560-7917. DOI: 10.2807/1560-7917.ES.2021.26.28.2100573.

[13] W. Yang et al. *Epidemiological Characteristics of the B.1.526 SARS-CoV-2 Variant*. Preprint. Public and Global Health, Aug. 2021. DOI: 10.1101/2021.08.04.21261596.

[14] W. Yang and J. Shaman. "Development of a Model-Inference System for Estimating Epidemiological Characteristics of SARS-CoV-2 Variants of Concern". In: *Nature Communications* 12.1 (Dec. 2021), p. 5573. ISSN: 2041-1723. DOI: 10.1038/s41467-021-25913-9.

[15] R. M. Coutinho et al. "Model-Based Estimation of Transmissibility and Reinfection of SARS-CoV-2 P.1 Variant". In: *Communications Medicine* 1.1 (Dec. 2021), p. 48. ISSN: 2730-664X. DOI: 10.1038/s43856-021-00048-6.

[16] F. Campbell et al. “Increased Transmissibility and Global Spread of SARS-CoV-2 Variants of Concern as at June 2021”. In: *Eurosurveillance* 26.24 (June 2021). ISSN: 1560-7917. DOI: 10.2807/1560-7917.ES.2021.26.24.2100509.

[17] A. Cori et al. “A New Framework and Software to Estimate Time-Varying Reproduction Numbers during Epidemics”. In: *American Journal of Epidemiology* 178.9 (2013), pp. 1505–1512. DOI: 10.1093/aje/kwt133.

[18] R. Thompson et al. “Improved inference of time-varying reproduction numbers during infectious disease outbreaks”. In: *Epidemics* 29 (2019), p. 100356.

[19] A. Cori. *EpiEstim*. Dec. 2021. URL: https://github.com/mrc-ide/EpiEstim (visited on 11/16/2021).

[20] *COVID-19 Point épidémiologique hebdomadaire du 28 janvier 2021*. URL: https://www.santepubliquefrance.fr/content/download/315275/2903017 (visited on 10/25/2021).

[21] *COVID-19 Point épidémiologique hebdomadaire du 4 mars 2021*. URL: https://www.santepubliquefrance.fr/content/download/324805/2944195 (visited on 10/25/2021).

[22] *Tableau synthétique des résultats par vague d’enquêtes*. URL: https://www.santepubliquefrance.fr/content/download/368440/3132368 (visited on 10/25/2021).

[23] K. M. Gostic et al. “Practical Considerations for Measuring the Effective Reproductive Number, Rt”. In: *PLOS Computational Biology* 16.12 (Dec. 2020). Ed. by V. E. Pitzer, e1008409. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008409.

[24] *EpiEstim package - RDocumentation*. 2021. URL: https://www.rdocumentation.org/packages/EpiEstim/versions/2.2-4 (visited on 09/19/2021).

[25] P. Mlcochova et al. “SARS-CoV-2 B.1.617.2 Delta Variant Replication and Immune Evasion”. In: *Nature* (Sept. 2021). ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-021-03944-y.

[26] The COVID-19 Genomics UK (COG-UK) consortium et al. “Assessing Transmissibility of SARS-CoV-2 Lineage B.1.1.7 in England”. In: *Nature* 593.7858 (May 2021), pp. 266–269. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-021-03470-x.

[27] M. Chand et al. *Investigation of novel SARS-COV-2 variant: Variant of Concern 202012/01*. 2021. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959438/Technical_Briefing_VOC_SH_NJL2_SH2.pdf (visited on 09/01/2021).

[28] N. G. Davies et al. “Estimated Transmissibility and Impact of SARS-CoV-2 Lineage B.1.1.7 in England”. In: *Science* 372.6538 (Apr. 2021). ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.abg3055.

[29] M. S. Graham et al. “Changes in Symptomatology, Reinfection, and Transmissibility Associated with the SARS-CoV-2 Variant B.1.1.7: An Ecological Study”. In: *The Lancet Public Health* 6.5 (May 2021), e335–e345. ISSN: 24682667. DOI: 10.1016/S2468-2667(21)00055-4.

[30] C. Piantham et al. *Estimating the Elevated Transmissibility of the B.1.1.7 Strain over Previously Circulating Strains in England Using GISAID Sequence Frequencies*. Preprint. Epidemiology, Mar. 2021. DOI: 10.1101/2021.03.17.21253775.

[31] *WHO Coronavirus Disease (COVID-19) Dashboard*. 2021. URL: https://covid19.who.int (visited on 09/19/2021).

[32] E. Mathieu et al. “A Global Database of COVID-19 Vaccinations”. In: *Nature Human Behaviour* 5.7 (July 2021), pp. 947–953. ISSN: 2397-3374. DOI: 10.1038/s41562-021-01122-8.

[33] E. Volz et al. *Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from Linking Epidemiological and Genetic Data*. Preprint. Infectious Diseases (except HIV/AIDS), Jan. 2021. DOI: 10.1101/2020.12.30.20249034.

[34] A. Gelman and D. B. Rubin. "Inference from Iterative Simulation Using Multiple Sequences". In: *Statistical Science* 7.4 (Nov. 1992). ISSN: 0883-4237. DOI: 10.1214/ss/1177011136.

**Methods**

We extend the methodology from Cori et al. [17] and Thompson et al. [18] to develop an inference framework for jointly estimating the transmissibility (instantaneous reproduction number $R_t$) of a reference variant and the effective transmission advantage of novel variants, compared to the reference. For simplicity, we present the method for two variants only (a reference and a new variant). The method is applicable to, and has been implemented for, estimating the transmission advantages of multiple variants over a single reference.

**Assumptions.** Our method relies on daily incidence data of the reference and the variant. Where data from more than one location are used, we assume that the epidemic in each location are independent and closed. That is, we do not account for spatial interaction between various locations and assume that all new cases in any location arise from previously infected cases in that location unless identified as imported cases in the dataset. The effective reproduction number is defined as the ratio of locally infected cases to the total infectiousness (due to local or imported cases) in a location. For more details, see [18].

**Notations.** We use the following notations:

- Indexes $t$ for time, $l$ for location and $v$ for variant, with $v$ = 0 denoting the reference variant and $v$ = 1 the new variant,
- $n_l$ the number of locations considered
- $T$ the number of days of observation
- $I_t^{\text{local},l,v}$ denotes the number of locally infected incident cases of variant $v$ at time $t$ in location $l$,
- $I_t^{\text{imported},l,v}$ denotes the number of imported infected incident cases of variant $v$ at time $t$ in location $l$ (in the absence of information on imported cases, $I_t^{\text{imported},l,v} = 0$ except on the first day of observation, where all cases are assumed to be imported),
- $I_t^{l,v}$ denotes the total number of incident cases of variant $v$ at time $t$ in location $l$, with $I_t^{l,v} = I_t^{\text{local},l,v} + I_t^{\text{imported},l,v}$,
- $R_t^{l,v}$ denotes the instantaneous reproduction number for variant $v$ at time $t$ in location $l$. For simplicity we use $R_t^l$ to denote the instantaneous reproduction number for the reference variant in location $l$ i.e. $R_t^l = R_t^{l,0}$.
- $w^v$ is the probability mass function of the discrete serial interval for variant $v$, assumed the same across all locations, but potentially different between variants ($w_s^v$ is the probability that the serial interval lasts s days, $s = 1, \ldots, SI_{max}$; and we assume $w_0^v = 0$).
- $\Lambda_t^{l,v} = \sum_{s=1}^{t} I_{t-s}^{l,v} w_s^v$ is the overall infectiousness for variant $v$ at time $t$ and in location $l$ due to past incident cases of that variant in that location (both imported and locally infected cases).
- For simplicity we introduce the generic notation $\mathbf{X}_t^{L,V} = \left\{ X_t^{l,v} \right\}_{l=1,\ldots,n_l; v=0,1}$ for the variable $X$ at time $t$ across all locations and both variants.

We assume that $R_t^{l,1} = \epsilon R_t^l$, i.e. the reproduction number of the new variant is proportional to that for the reference variant at all times and in all locations; the proportional factor $\epsilon$ is the effective transmission advantage (if $\epsilon > 1$, or disadvantage if $\epsilon < 1$) of the new variant compared to the reference variant, assumed constant over time and across locations. We explored values of $\epsilon > 1$ in all simulation scenarios as values of $\epsilon < 1$ correspond to swapping the reference and new variant.
We assume the number of secondary infections generated by each case is Poisson distributed. Under these assumptions, the likelihood of the time series of incident cases of the reference and the new variants can be written as

$$\begin{aligned}
\mathcal{L}_t &= P\left(\mathbf{I}_t^{\text{local},L,V} | \mathbf{I}_0^{\text{local},L,V}, \ldots, \mathbf{I}_{t-1}^{\text{local},L,V}, \mathbf{I}_0^{\text{imported},L,V}, \ldots, \mathbf{I}_{t-1}^{\text{imported},L,V}, \mathbf{w}^V, \mathbf{R}_t^{L,V}, \epsilon\right) \\
&= \prod_{l=1}^{n_l} \prod_{v=0}^{1} \frac{\left(R_t^{l,v}\right)^{I_t^{\text{local},l,v}} e^{-R_t^{l,v}\Lambda_t^{l,v}}}{I_t^{\text{local},l,v}!} \\
&\propto \prod_{l=1}^{n_l} \left( \left(R_t^l\right)^{\sum_{v=0}^{1} I_t^{\text{local},l,v}} \epsilon^{I_t^{\text{local},l,1}} e^{-R_t^l\left(\Lambda_t^{l,0} + \epsilon\Lambda_t^{l,1}\right)} \right) \\
&\propto \epsilon^{\sum_{l=1}^{n_l} I_t^{\text{local},l,1}} e^{-\sum_{l=1}^{n_l} R_t^l\left(\Lambda_t^{l,0} + \epsilon\Lambda_t^{l,1}\right)} \prod_{l=1}^{n_l} \left(R_t^l\right)^{\sum_{v=0}^{1} I_t^{\text{local},l,v}}
\end{aligned}$$

We assume Gamma priors for each $R_t^l$, with same shape $a$ and scale $b$ across times and locations, and for $\epsilon$, with shape $c$ and scale $d$. The joint posterior distribution of parameters given the observations is (assuming the serial interval distributions for both variants $\mathbf{w}^V$ are known):

$$\begin{aligned}
&P\left(\epsilon, \mathbf{R}_1^L \ldots, \mathbf{R}_T^L | \mathbf{I}_0^{\text{local},L,V}, \ldots, \mathbf{I}_T^{\text{local},L,V}, \mathbf{I}_0^{\text{imported},L,V}, \ldots, \mathbf{I}_T^{\text{imported},L,V}, \mathbf{w}^V\right) \\
&\propto \prod_{t=1}^{T} \left( \mathcal{L}_t \prod_{l=1}^{n_l} \left(R_t^l\right)^{a-1} e^{-\frac{R_t^l}{b}} \right) \epsilon^{c-1} e^{-\frac{\epsilon}{d}} \\
&\propto \left( \prod_{t=1}^{T} \prod_{l=1}^{n_l} \left(R_t^l\right)^{a-1+\sum_{v=0}^{1} I_t^{\text{local},l,v}} \right) \epsilon^{c-1+\sum_{t=1}^{T}\sum_{l=1}^{n_l} I_t^{\text{local},l,1}} \\
&\quad \times e^{-\frac{\epsilon}{d} - \sum_{t=1}^{T}\sum_{l=1}^{n_l} R_t^l\left(\frac{1}{b} + \Lambda_t^{l,0} + \epsilon\Lambda_t^{l,1}\right)}
\end{aligned}$$

The marginal posterior distribution for $\epsilon$ given the data (i.e. the incidence for all variants, at all locations and for all time steps) and given the reproduction number for the reference variant in all locations and at all time steps is given by:

$$\begin{aligned}
&P\left(\epsilon | \mathbf{I}_0^{\text{local},L,V}, \ldots, \mathbf{I}_T^{\text{local},L,V}, \mathbf{I}_0^{\text{imported},L,V}, \ldots, \mathbf{I}_T^{\text{imported},L,V}, \mathbf{R}_1^L \ldots, \mathbf{R}_T^L, \mathbf{w}^V\right) \\
&\propto \epsilon^{c-1+\sum_{t=1}^{T}\sum_{l=1}^{n_l} I_t^{\text{local},l,1}} e^{-\epsilon\left(\frac{1}{d} + \sum_{t=1}^{T}\sum_{l=1}^{n_l} R_t^l \Lambda_t^{l,1}\right)}
\end{aligned}$$

Therefore, the marginal posterior distribution of $\epsilon$ given the data and other parameters is a Gamma distribution with shape $c + \sum_{t=1}^{T}\sum_{l=1}^{n_l} I_t^{\text{local},l,1}$ and scale $\frac{1}{\frac{1}{d} + \sum_{t=1}^{T}\sum_{l=1}^{n_l} R_t^l \Lambda_t^{l,1}}$.
Similarly, the marginal posterior distribution for $R_t^l$ at time step $t$ and in location $l$ given the data, $\epsilon$, and the reproduction number at other locations and time steps, is given by:

$$P(R_t^l|\mathbf{I}_0^{\text{local},L,V},\ldots,\mathbf{I}_T^{\text{local},L,V},\mathbf{I}_0^{\text{imported},L,V},\ldots,\mathbf{I}_T^{\text{imported},L,V},\epsilon,$$
$$\mathbf{R}_1^L\ldots,\mathbf{R}_{t-1}^L,\mathbf{R}_{t+1}^L,\ldots,\mathbf{R}_T^L,\mathbf{R}_t^1\ldots,\mathbf{R}_t^{l-1},\mathbf{R}_t^{l+1}\ldots,\mathbf{R}_t^{n_l},\mathbf{w}^V)$$
$$\propto \left(R_t^l\right)^{a-1+\sum_{v=0}^{1} I_t^{\text{local},l,v}} e^{-R_t^l\left(\frac{1}{b}+\Lambda_t^{l,0}+\epsilon\Lambda_t^{l,1}\right)}$$

Therefore, the marginal posterior distribution of $R_t^l$ given the data and other parameters is a Gamma distribution with shape $a+\sum_{v=0}^{1} I_t^{\text{local},l,v}$ and scale $\frac{1}{\frac{1}{b}+\Lambda_t^{l,0}+\epsilon\Lambda_t^{l,1}}$.

**Monte Carlo Markov Chain (MCMC) inference.** The analytical formulation of the marginal posterior distributions for $R_t^l$ and $\epsilon$ allow us to use a multi-stage Gibbs sampler for the MCMC inference.

To initialise $R_t^l$, we use EpiEstim to estimate a single reproduction number for the reference variant over the entire time period of observations, and using incidence aggregated across all locations. The posterior mean is then used as the initial value for $R_t^l$. We independently use the same approach to estimate a single reproduction number for the new variant; $\epsilon$ is then initialised to the median of the ratio of the reproduction numbers for the new variant and the reference.

We first sample from the marginal distribution of $R_t^l$, conditional on $\epsilon$, and then we sample from the marginal distribution of $\epsilon$, conditional on the newly sampled value of $R_t^l$. We repeat this procedure for a fixed number of iterations or until convergence is achieved. Convergence is assessed using Gelman-Rubin convergence diagnostic [34] using 1.1 as a cut-off value.

**Choosing a time-period for estimation of $\epsilon$.** Users can set the time period over which estimation will be carried out. We recommend that the estimation is started after at least one generation of cases has been observed. The default starting point in the software is set to the first day of non-zero incidence across all locations plus the 95$^{\text{th}}$ percentile of the serial interval distribution.

**Classification of a variant.** We used the posterior distribution of the effective transmission advantage to classify a new variant (in relation to the reference variant) as:

- 'More transmissible' if the 2.5$^{\text{th}}$ quantile of the posterior distribution was greater than 1;
- 'Less transmissible' if the 97.5$^{\text{th}}$ quantile of the posterior distribution was less than 1; and,
- 'Unclear' if the 95% CrI contained 1.

**Implementation.** The inference method is implemented in a new function "estimate_advantage" of the development version of the R package EpiEstim available at `https://github.com/mrc-ide/EpiEstim`.

**Acknowledgements**

The use of pillar-2 PCR testing data was made possible thanks to PHE colleagues, and we extend our thanks to N Gent for facilitation and insights into these data. We also thank Edward S Knock for his inputs on the data for England. This study is partially funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Modelling and Health Economics, a partnership between Public Health England, Imperial College London and LSHTM (grant code NIHR200908); the authors acknowledge funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), which is jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union. JW acknowledges research funding from the Wellcome Trust (grant 102169/Z/13/Z). SB acknowledges funding from the Wellcome

Trust (grant 219415). RKN acknowledges funding from the Medical Research Council Doctoral Training Partnership. Disclaimer: The views expressed are those of the author(s) and not necessarily those of the NIHR, Public Health England or the Department of Health and Social Care.