Supplementary Material: Extending EpiEstim to estimate the transmission advantage of pathogen variants in real-time: SARS-CoV-2 as a case-study

Contents

1 Overview

In this document, we present details of the SARS-CoV-2 variant-specific incidence data used in the analysis (Sec. 2) and describe the method used for obtaining estimates of the transmission advantage for SARS-CoV-2, both using a non-parametric approach and using MV-EpiEstim (Sec. 3). Sec. 4 shows additional results for the estimation of the transmission advantage of SARS-CoV-2 variants of concern (VOCs), including more detailed results on the estimated transmission advantage of Alpha over the wildtype, but also estimates of the transmission advantages of Beta and Gamma (combined) over the wildtype, and of Delta over Alpha. Sec. 5 presents an overview of the simulation study used to assess the validity of our method. We describe

the methodology used for the simulation and present the range of scenarios we explored, as well as more comprehensive results from our simulation study.

2 SARS-CoV-2 variant-specific incidence data

2.1 Incidence data from England

We used the daily number of positive tests from England's community SARS-CoV-2 testing system (also called Pillar 2) from 1st September 2020 to 20th June 2021, stratified by NHS region (Fig S1). The Pillar 2 testing data were shared with Imperial College London by Public Health England. Up to 14th March 2021, we interpreted the number of samples with no S-gene target failure (SGTF) in this data as incidence of the wildtype, and from 14th March 2021 onwards as incidence of the Delta variant. Samples with S-gene failure were considered to be of the Alpha variant throughout. The weekly proportion of reported cases for which the SGTF status was known varied from 57% to 74% over the study period (Fig S4). We assumed that the daily proportion of the variants (Alpha and wildtype, or Alpha and Delta) was identical in the cases for which SGTF was known (S-gene positive and S-gene negative) and in those for which SGTF was not known. We adjusted the daily incidence of the variants by distributing the cases with unknown S-gene status using this daily proportion.

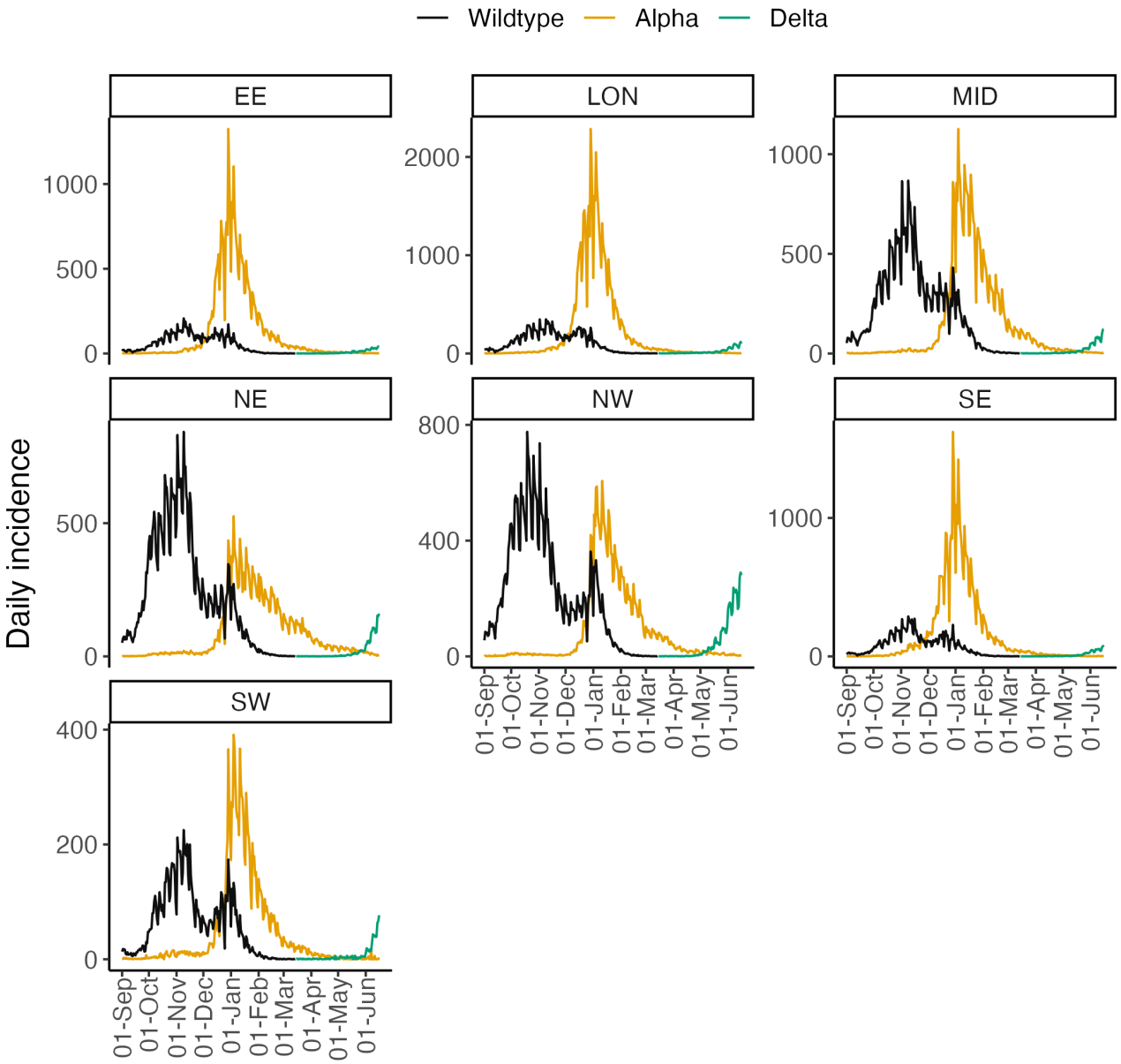


Figure S1: Daily incidence of SARS-CoV-2 wildtype (black), Alpha (yellow) and Delta (green) variants in the 7 NHS regions in England. Note that the y-axis in each panel is different. The NHS England regions are - East of England (EE), London (LON), Midlands (MID), North-East (NE), North-West (NW), South-East (SE), South-West (SW).

2.2 Incidence data from France

Santé Publique France reports the age-disaggregated number of PCR tests with Alpha, Beta, and Gamma variants of SARS-CoV-2 at a sub-national level in France [1] with the incidence of Beta and Gamma variants reported as an aggregate. The absence of labelling with a specific VOC was interpreted as an infection with the wildtype. The 18 ADM2 units for which data were reported include metropolitan France and overseas regions. We aggregated the data across all age groups to obtain a daily incidence time series for each variant from 28th February to 30th May 2021 (Fig S2). The proportion of PCR tests screened for the presence of variants among positive PCRs varied from 66.1% in the week starting 28th February to 26.9% in the week ending 30th May 2021. The variant was know for all cases that were used for estimation.

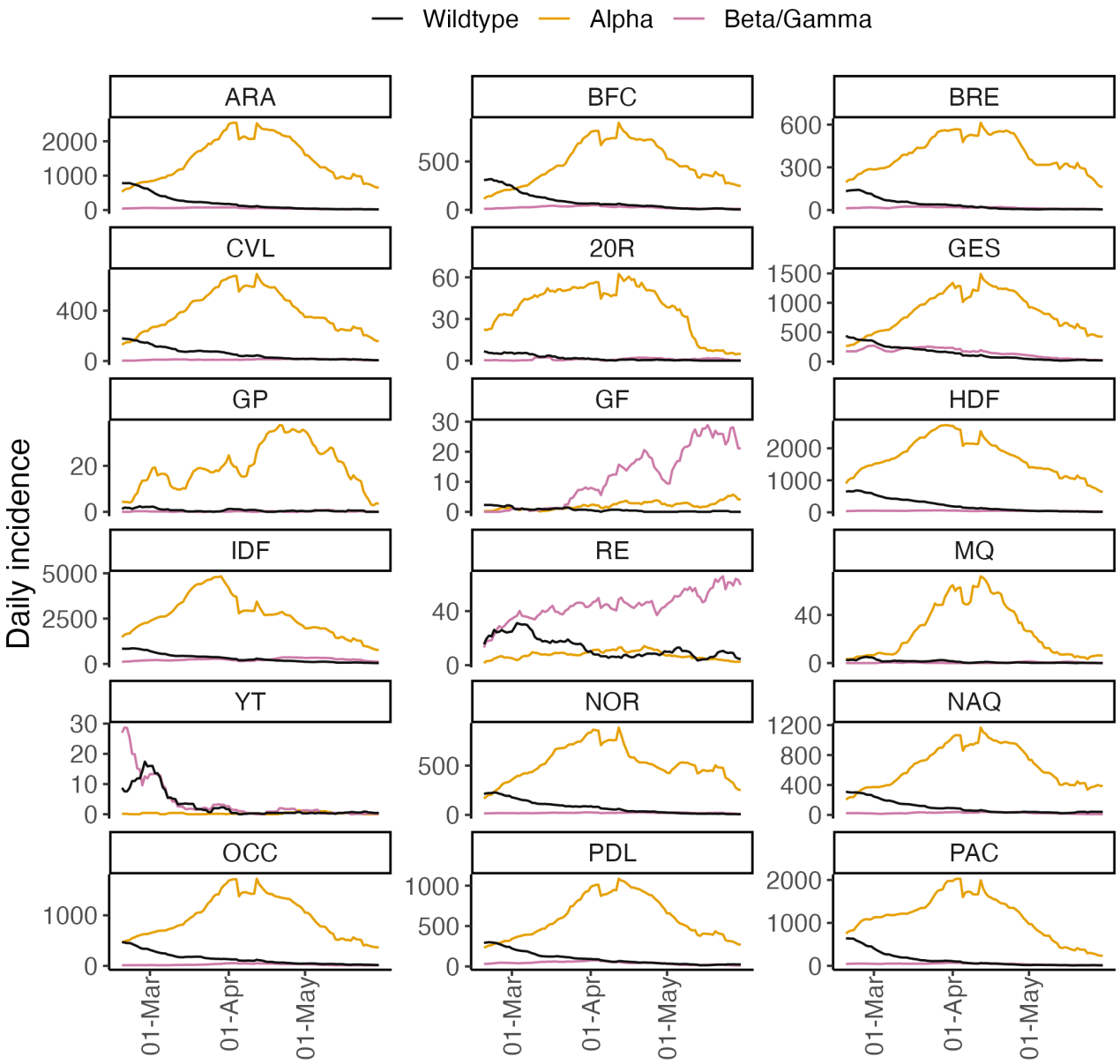


Figure S2: Daily incidence of SARS-CoV-2 wildtype (black), Alpha (yellow) and Beta/Gamma (blue) variants for the 18 ADM2 regions in France. Note that the y-axis in each panel is different. The ADM2 regions are - ARA : Auvergne-Rhône-Alpes, BFC : Bourgogne-Franche-Comté, BRE : Bretagne, CVL : Centre-Val de Loire, 20R : Corse, GES : Grand Est, GP : Guadeloupe, GF : Guyane, HDF : Hauts-de-France, IDF : Île-de-France, RE : La Réunion, MQ : Martinique, YT : Mayotte, NOR : Normandie, NAQ : Nouvelle-Aquitaine, OCC : Occitanie, PDL : Pays de la Loire, and PAC : Provence-Alpes-Côte d'Azur.

3 Estimating the effective transmission advantage

3.1 Serial interval distribution

Both the non-parametric (see next paragraph) and the MV-EpiEstim approaches use the discrete distribution of the serial interval (time between symptom onset in a case and their infector) as an input. We assumed a discrete gamma distributed serial interval for SARS-CoV-2 with mean 5.4 days and standard deviation of 1.5 days following [2]. We used the same serial interval distribution across all variants.

3.2 Non-parametric approach

To obtain a non-parametric estimate of the effective transmission advantage of a VOC over the reference SARS-CoV-2 variant, we first estimated the daily effective reproduction number independently for each variant (wildtype or VOC) for 18 ADM2 units in France and 7 NHS regions in England using the R package EpiEstim [3]. We used a sliding weekly window, and set the prior R_t to have a mean and a standard deviation of 1.

To exclude region-weeks where the R_t estimates were highly uncertain, we only used estimates from region-weeks where the width of 95% CrI of R_t was less than 0.5. We started estimation of R_t on the week starting on the 11^{th} day. The threshold of 11 days was chosen as it is the 99^{th} percentile of the serial interval distribution. That is, 99% of the cases that were infected by an index cases from day 1 in our analysis are expected to have been observed by day 11. Note that because of these exclusion criteria, some of the non-parametric estimates are missing in the tables shown in section Sec. 4, when no weeks could be included for a particular region or time period.

For each region-week included in the analysis, we drew a sample of 100 values from the posterior distribution of R_t for each variant. Non-parametric estimates of a variant's transmission advantage over a reference variant were obtained by dividing the sampled values from their respective posterior R_t distributions (with random pairing). To account for sub-national variation in R_t profile, estimates at the national level were obtained by pooling the sub-national estimates (thereby giving the same weight to each week of data from any region). To gain insight into the potential temporal heterogeneity of the effective transmission advantage, we divided the incidence time series into four non-overlapping periods of equal duration and estimated the transmission advantage in each period.

3.3 Estimation using MV-EpiEstim

We set the priors for both R_t and ϵ to have mean and standard deviation 1. We ran the multi-stage Gibbs sampler for 20,000 iterations. The first 5,000 iterations were discarded as burn-in and thinning was set to keep 1 in 10 iterations, leading to a final posterior sample of size 1,500.

Posterior samples of the transmission advantage were obtained for (i) each region independently and (ii) nationally but using regional data, by assuming a single underlying transmission advantage and region-specific R_t profiles. Independent estimates of the transmission advantage were obtained for the same non-overlapping time period as for the non-parametric estimates.

To mimic real-time epidemic context and examine how estimates changed as more data became available, we estimated the effective transmission advantage using data available up to successive weeks.

4 Estimates of the effective transmission advantages of SARS-CoV-2 variants

4.1 Alpha over wildtype

4.1.1 England

Main results for the estimated transmission advantage of Alpha over the wildtype using data from England are shown in the main text Figure 1. Unlike in Figure 1 (where panel C shows transmission advantages estimated using data from only the week ending on the date specified on the x-axis), Fig S3C shows additional results where we estimate ϵ using the entire time series from the start of the study period up to the time-point shown on the x-axis (implicitly assuming a transmission advantage that is constant over time) (Fig S3). This mimics real-time analyses using all the data accumulated up to the present time. Since this approach uses more information, there is less uncertainty in the estimates. However, compared to estimates using only most recent week of data (Fig 1C), the estimates shown in Fig S3C, obtained using progressively more data, smooth out any underlying temporal trend.

To explore the robustness of estimates to under-reporting, we estimated the transmission advantage of Alpha using only 50% of reported cases of Alpha and wildtype. As expected, the estimates with under-reporting are very similar to that obtained using the full data set, albeit with slightly more uncertainty (Fig S4).

Region/Time Period	Non-parametric	MV-EpiEstim
All	1.41 (0.86, 2.01)	1.46 (1.44, 1.47)
East of England	1.38 (1.02, 2.13)	1.43 (1.39, 1.47)
London	1.40 (0.93, 1.82)	1.46 (1.43, 1.50)
Midlands	1.44 (0.92, 2.04)	1.54 (1.50, 1.58)
North East and Yorkshire	1.41 (0.86, 2.05)	1.46 (1.42, 1.50)
North West	1.50 (0.71, 2.08)	1.51 (1.47, 1.55)
South East	1.35 (1.00, 1.87)	1.36 (1.33, 1.39)
South West	1.41 (0.78, 1.82)	1.40 (1.35, 1.45)
Quarter 1	0.89 (0.62, 1.17)	1.03 (0.99, 1.08)
Quarter 2	1.36 (0.82, 1.95)	1.48 (1.45, 1.51)
Quarter 3	1.45 (1.07, 2.02)	1.50 (1.48, 1.52)
Quarter 4	1.39 (0.93, 2.12)	1.28 (1.23, 1.33)

Table S1: Estimates of the effective transmission advantage of SARS-CoV-2 Alpha variant over the wildtype using the non-parametric approach and MV-EpiEstim for 7 NHS regions in England and 4 non-overlapping time periods. Estimates shown are the posterior median with 95% CrI in parenthesis. Quarters correspond to - Quarter 1: 11$^{\text{th}}$ September - 27$^{\text{th}}$ October 2020; Quarter 2: 27$^{\text{th}}$ October - 12$^{\text{th}}$ December 2020; Quarter 3: 12$^{\text{th}}$ December 2020 - 27$^{\text{th}}$ January 2021; Quarter 4: 27$^{\text{th}}$ January - 14$^{\text{th}}$ March 2021.

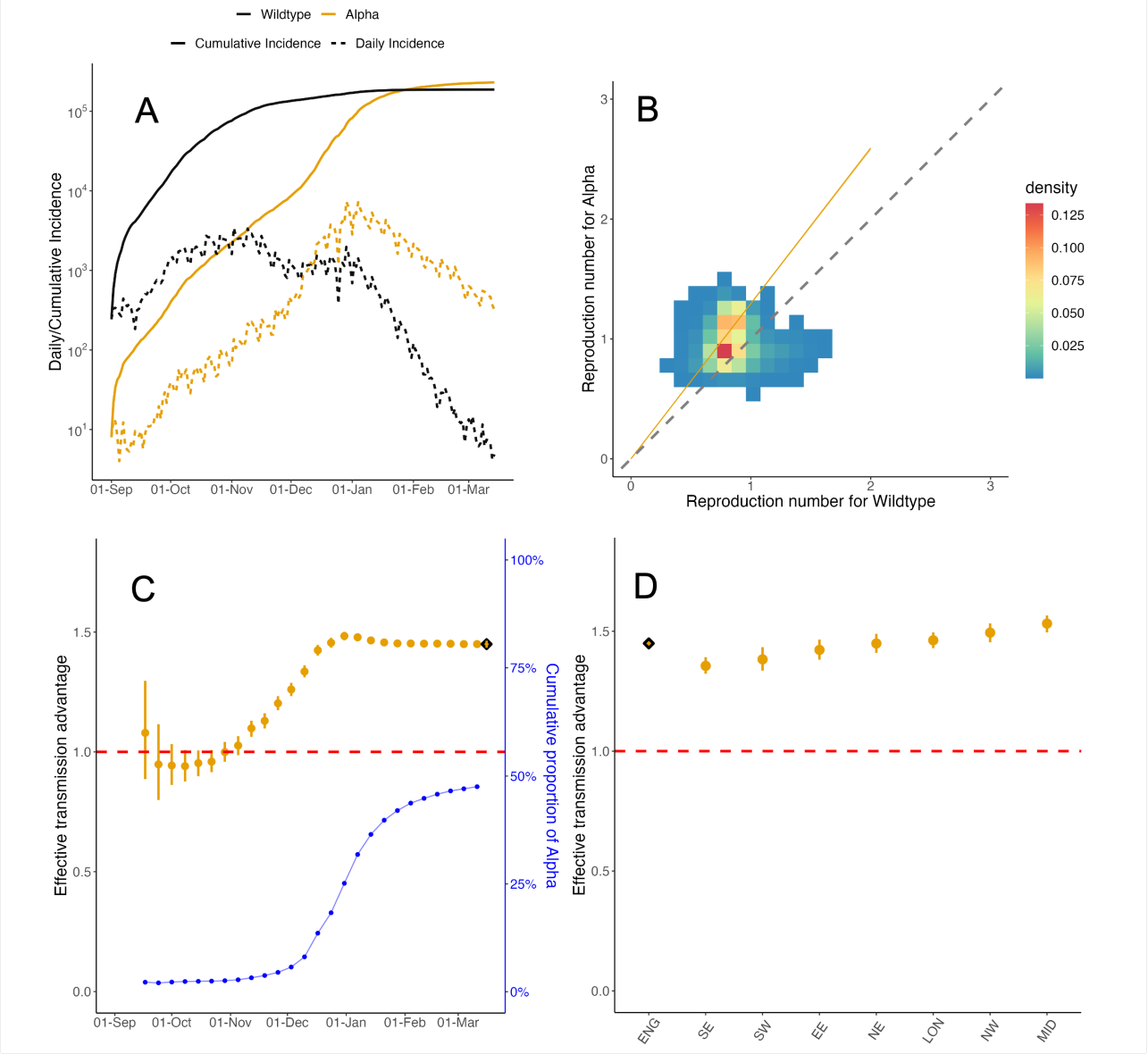


Figure S3: **Effective transmission advantage of Alpha over wildtype in England.** ***(Note this figure duplicates Fig 1 from the main text with the exception of panel C. Further explanation is provided in Sec. 4.1.1).*** (A) The daily reported incidence of cases of the wildtype (black) and Alpha (yellow) 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 yellow line denotes the median effective transmission advantage estimated using MV-EpiEstim, assuming no temporal or spatial heterogeneity. 95% CrI were so narrow that they could not be distinguished from the line. (C) Effective transmission advantage estimated using MV-EpiEstim using data from the start of the time series up to the date specified on the x-axis (yellow). The dark blue circles and the horizontal bars denote respectively the mean and 95% binomial confidence interval of the cumulative proportion of incidence of Alpha (right y-axis). Because of the high incidence of both wildtype and Alpha in the study period, the 95% CI are small and hence difficult to distinguish. (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 yellow circles denote the median estimate, the vertical lines indicate the 95% CrI, and the red dashed line denotes the $\epsilon = 1$ threshold.

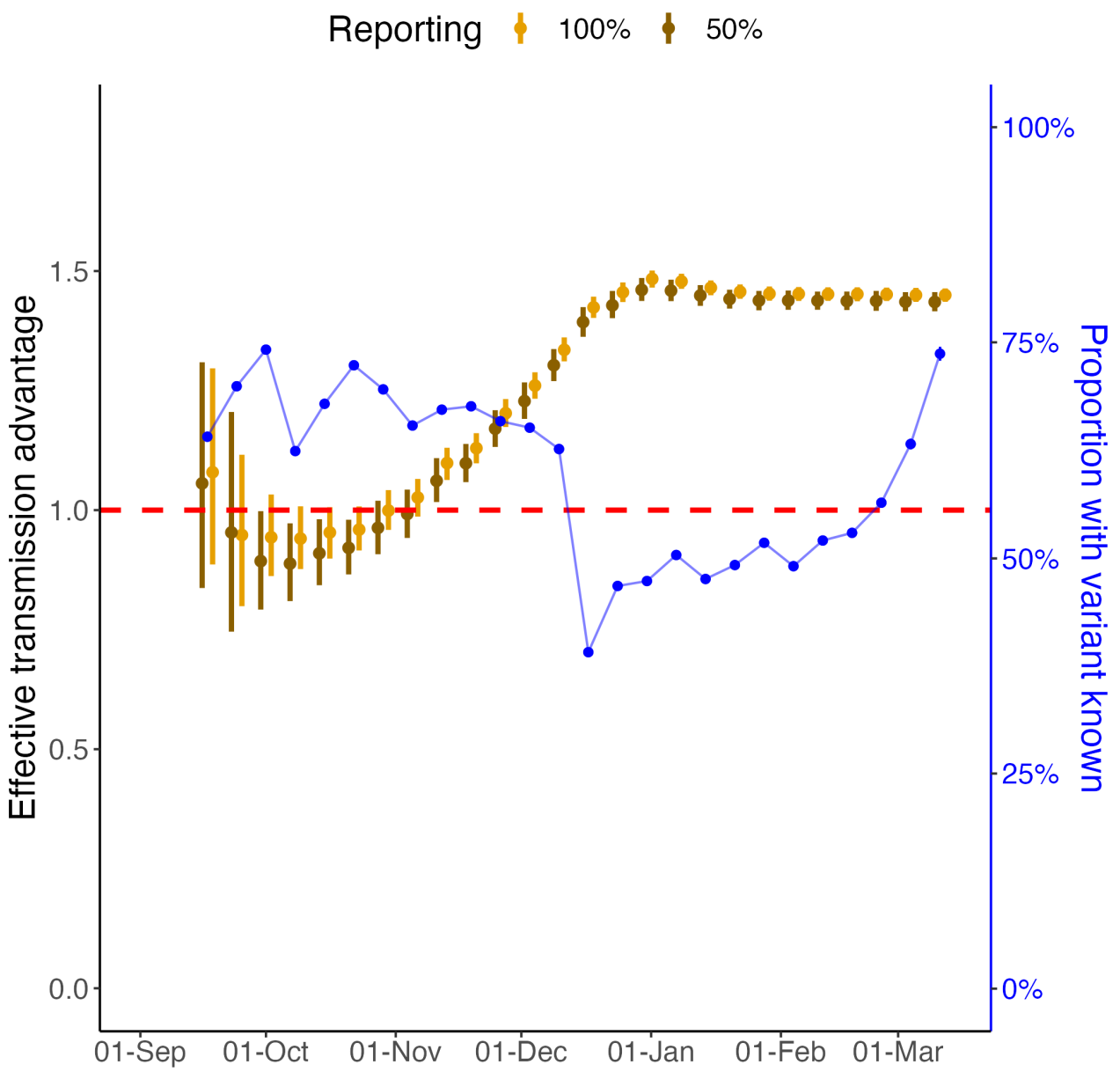


Figure S4: **Effective transmission advantage of Alpha over wildtype in England from 1st September 2020 to 14th March 2021.** Effective transmission advantage estimated using MV-EpiEstim using all data or 50% of the reported cases in the week ending on the date specified on the x-axis. Yellow circles and vertical lines represent the median and 95% CrI of the estimated transmission advantage using all data. The yellow solid circles and lines are estimates based on use of all reported cases, while the dark orange circles and vertical lines are estimates using only 50% of reported cases of Alpha and wildtype. The red dashed line denotes the $\epsilon = 1$ threshold. The blue circles and the vertical bars denote the mean proportion and the 95% binomial CI of incident cases where the variant was known (right y-axis) in the week of estimation. The estimates and CI were obtained using only cases with known SGTF status. Due to the high incidence of the two variants, the 95% CIs are very narrow and difficult to distinguish.

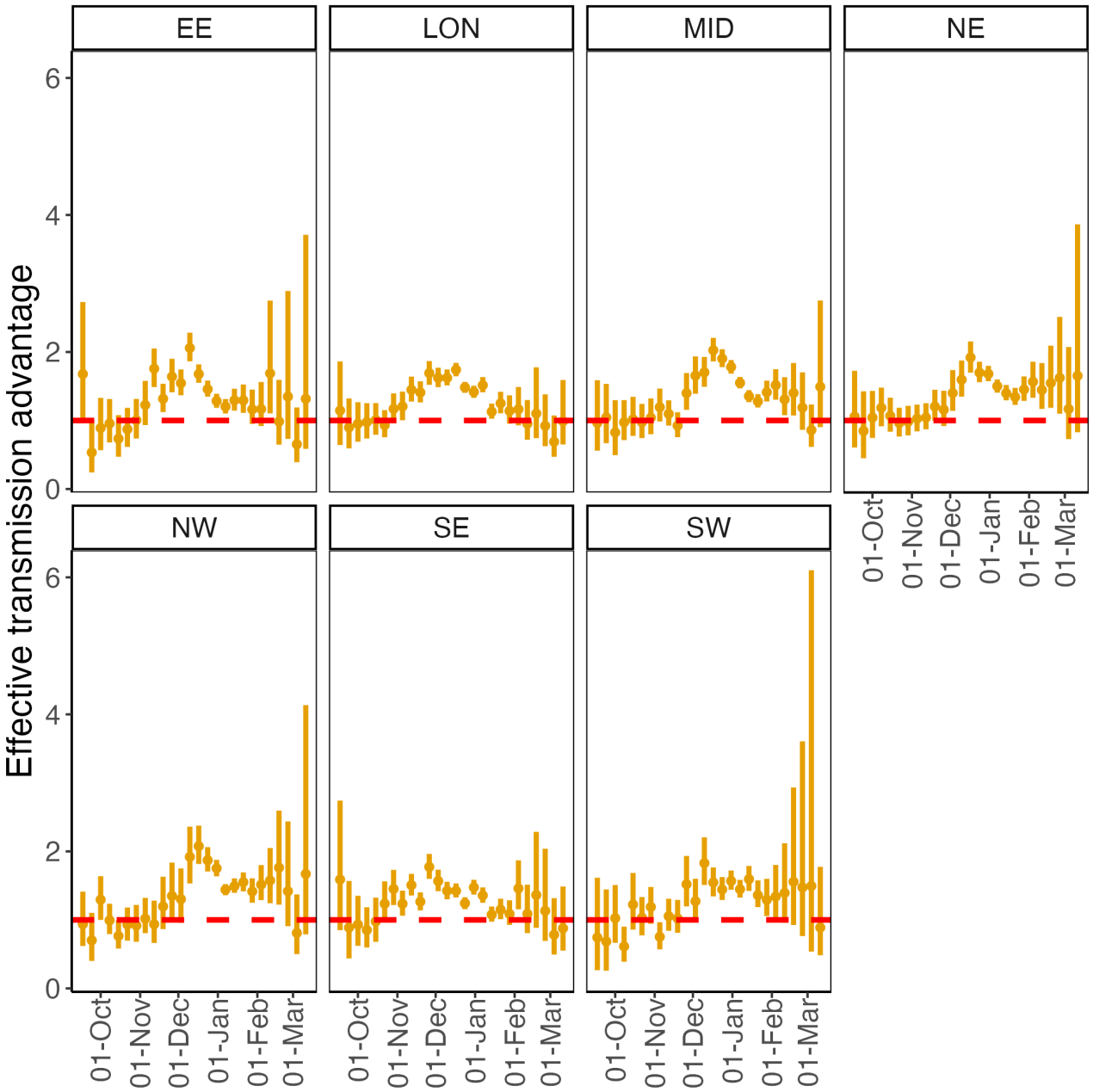


Figure S5: **Effective transmission advantage of Alpha over wildtype for all NHS England regions from 1st September 2020 to 14th March 2021.** The NHS England regions are - East of England (EE), London (LON), Midlands (MID), North-East (NE), North-West (NW), South-East (SE), and South-West (SW). Each point represents the median estimate from using MV-EpiEstim using data in the week ending on the date specified on the x-axis. The vertical yellow lines indicate the 95% CrI, and the red dashed line denotes the $\epsilon = 1$ threshold.

4.1.2 France

Region/Time Period	Non-parametric	MV-EpiEstim
All	1.21 (0.75, 1.65)	1.29 (1.29, 1.30)
Auvergne-Rhône-Alpes	1.25 (0.89, 1.67)	1.40 (1.38, 1.43)
Bourgogne-Franche-Comté	1.27 (0.90, 1.78)	1.41 (1.37, 1.46)
Bretagne	1.34 (0.89, 1.78)	1.35 (1.29, 1.40)
Centre-Val de Loire	1.18 (0.79, 1.61)	1.32 (1.27, 1.36)
Corse	-	1.15 (1.00, 1.31)
Grand Est	1.25 (0.70, 1.48)	1.31 (1.28, 1.34)
Guadeloupe	-	1.12 (0.95, 1.35)
Guyane	-	1.33 (1.05, 1.66)
Hauts-de-France	1.19 (0.97, 1.41)	1.28 (1.26, 1.30)
Île-de-France	1.08 (0.87, 1.47)	1.21 (1.20, 1.23)
La Réunion	1.14 (0.67, 2.08)	1.07 (0.98, 1.18)
Martinique	-	1.27 (1.09, 1.49)
Mayotte	-	1.13 (0.73, 1.69)
Normandie	1.21 (0.91, 1.59)	1.31 (1.28, 1.35)
Nouvelle-Aquitaine	1.22 (0.70, 1.62)	1.28 (1.25, 1.31)
Occitanie	1.18 (0.81, 1.50)	1.27 (1.25, 1.30)
Pays de la Loire	1.20 (0.66, 1.55)	1.29 (1.26, 1.32)
Provence-Alpes-Côte d'Azur	1.22 (0.67, 1.58)	1.37 (1.34, 1.40)
Quarter 1	1.45 (1.26, 1.71)	1.42 (1.41, 1.44)
Quarter 2	1.28 (0.99, 1.62)	1.24 (1.22, 1.25)
Quarter 3	1.15 (0.83, 1.51)	1.11 (1.09, 1.13)
Quarter 4	1.00 (0.67, 1.52)	0.97 (0.95, 0.99)

Table S2: Estimates of the effective transmission advantage of SARS-CoV-2 Alpha variant over the wildtype using the non-parametric approach and MV-EpiEstim for 18 ADM2 regions in France and 4 non-overlapping time periods. Estimates shown are the posterior median with 95% CrI in parenthesis. Quarters correspond to - Quarter 1: 28$^{\text{th}}$ February - 23$^{\text{rd}}$ March 2021; Quarter 2: 23$^{\text{rd}}$ March - 14$^{\text{th}}$ April 2021; Quarter 3: 14$^{\text{th}}$ April - 7$^{\text{th}}$ May 2021; Quarter 4: 7$^{\text{th}}$ May - 20$^{\text{th}}$ May 2021.

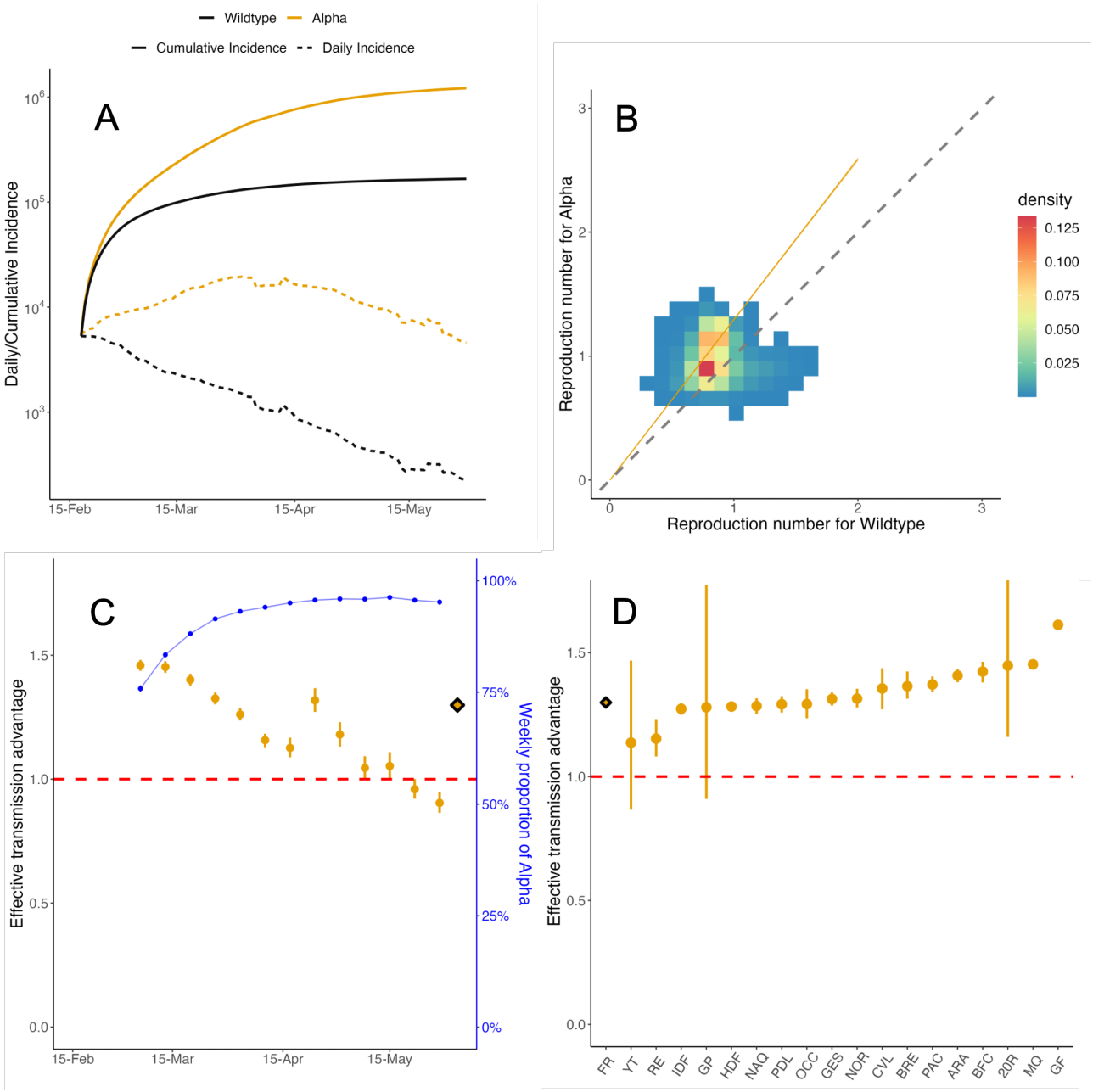


Figure S6: **Effective transmission advantage of Alpha over wildtype in France** (A) The daily reported incidence of cases of the wildtype (black) and Alpha (yellow) in France from 18$^{\text{th}}$ February to 30$^{\text{th}}$ May 2021. (B) The effective reproduction number R_t estimated independently 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 yellow line denotes the median effective transmission advantage estimated using MV-EpiEstim, assuming no temporal or spatial heterogeneity. 95% CrI were so narrow that they could not be distinguished from the line. (C) Effective transmission advantage estimated from MV-EpiEstim using data in the week ending on the date specified on the x-axis (yellow circles) and the entire time series (diamond). The dark blue circles and the vertical bars denote respectively the mean and 95% binomial confidence interval of the proportion of incidence of Alpha (right y-axis) in the week of estimation. Because of the high incidence of both wildtype and Alpha in the study period, the 95% CI are small and hence difficult to distinguish. (D) Effective transmission advantage estimated using MV-EpiEstim for all ADM2 region in France together (diamond) and separately (solid circle) using data from 18$^{\text{th}}$ February to 30$^{\text{th}}$ May 2021. The ADM2 regions are - ARA : Auvergne-Rhône-Alpes, BFC : Bourgogne-Franche-Comté, BRE : Bretagne, CVL : Centre-Val de Loire, 20R : Corse, GES : Grand Est, GP : Guadeloupe, GF : Guyane, HDF : Hauts-de-France, IDF : Île-de-France, RE : La Réunion, MQ : Martinique, YT : Mayotte, NOR : Normandie, NAQ : Nouvelle-Aquitaine, OCC : Occitanie, PDL : Pays de la Loire, and PAC : Provence-Alpes-Côte d'Azur. In panels (C) and (D), the solid yellow circles denote the median estimate, the vertical lines indicate the 95% CrI, and the red dashed line denotes the $\epsilon = 1$ threshold.

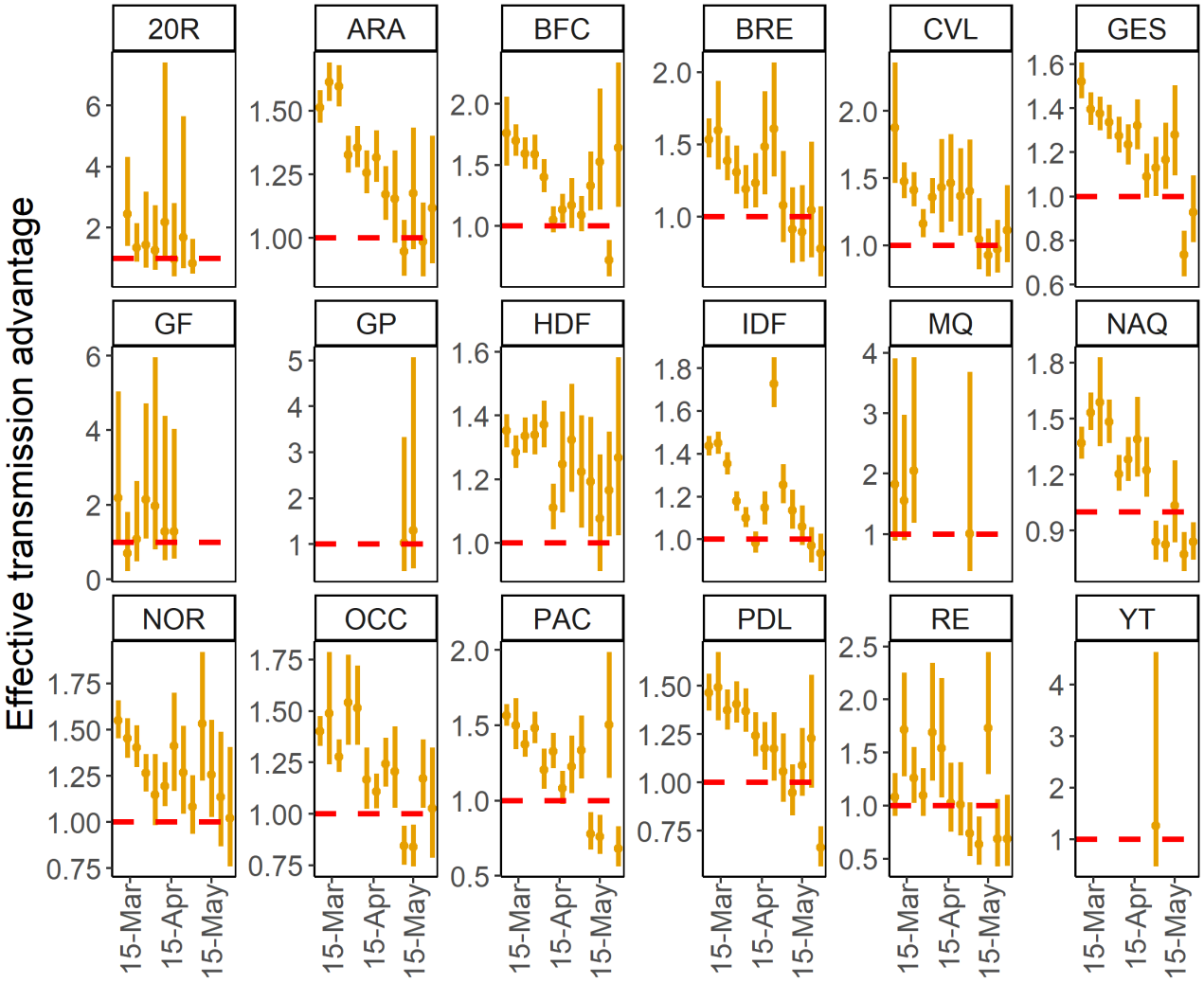


Figure S7: **Effective transmission advantage of Alpha over wildtype for all ADM2 regions in France from 18th February to 30th May 2021.** The ADM2 regions are - 20R : Corse, ARA : Auvergne-Rhône-Alpes, BFC : Bourgogne-Franche-Comté, BRE : Bretagne, CVL : Centre-Val de Loire, GES : Grand Est, GF : Guyane, GP : Guadeloupe, HDF : Hauts-de-France, IDF : Île-de-France, MQ : Martinique, NAQ : Nouvelle-Aquitaine, NOR : Normandie, OCC : Occitanie, PAC : Provence-Alpes-Côte d'Azur, PDL : Pays de la Loire, RE : La Réunion, and YT : Mayotte. Each point represents the median estimate made using MV-EpiEstim using data in the week ending on the date specified on the x-axis. The vertical yellow lines indicate the 95% CrI, and the red dashed line denotes the ϵ =1 threshold.

4.2 Beta and Gamma over wildtype (France)

Region/Time-period	Non-parametric	MV-EpiEstim
All	1.17 (0.69, 1.74)	1.25 (1.24, 1.27)
Auvergne-Rhône-Alpes	1.15 (0.88, 1.50)	1.29 (1.25, 1.34)
Bourgogne-Franche-Comté	1.22 (0.77, 1.90)	1.32 (1.25, 1.38)
Bretagne	1.26 (0.82, 1.89)	1.28 (1.20, 1.37)
Centre-Val de Loire	1.15 (0.68, 1.73)	1.35 (1.27, 1.44)
Corse	-	1.30 (1.03, 1.62)
Grand Est	1.08 (0.61, 1.31)	1.15 (1.12, 1.17)
Guadeloupe	-	1.07 (0.61, 1.67)
Guyane	-	1.40 (1.14, 1.71)
Hauts-de-France	1.21 (0.96, 1.46)	1.26 (1.22, 1.31)
Île-de-France	1.16 (0.85, 1.72)	1.27 (1.25, 1.29)
La Réunion	1.32 (0.89, 2.10)	1.15 (1.07, 1.22)
Martinique	-	1.30 (0.90, 1.84)
Mayotte	0.95 (0.47, 1.54)	0.83 (0.69, 1.00)
Normandie	1.20 (0.88, 1.70)	1.28 (1.22, 1.35)
Nouvelle-Aquitaine	1.16 (0.53, 1.67)	1.23 (1.17, 1.29)
Occitanie	1.16 (0.71, 1.77)	1.29 (1.24, 1.35)
Pays de la Loire	1.12 (0.64, 1.65)	1.19 (1.14, 1.24)
Provence-Alpes-Côte d'Azur	1.21 (0.65, 1.63)	1.33 (1.28, 1.38)
Quarter 1	1.31 (0.85, 1.78)	1.28 (1.26, 1.30)
Quarter 2	1.17 (0.87, 1.69)	1.15 (1.13, 1.17)
Quarter 3	1.17 (0.81, 1.73)	1.20 (1.17, 1.23)
Quarter 4	1.01 (0.58, 1.75)	0.97 (0.94, 0.99)

Table S3: Estimates of the combined effective transmission advantage of SARS-CoV-2 Beta and Gamma variants over the wildtype in France using the non-parametric approach and MV-EpiEstim for 18 ADM2 regions in France and 4 non-overlapping time periods. Estimates shown are the posterior median with 95% CrI in parenthesis. Quarters correspond to - Quarter 1: 28$^{\text{th}}$ February - 23$^{\text{rd}}$ March 2021; Quarter 2: 23$^{\text{rd}}$ March - 14$^{\text{th}}$ April 2021; Quarter 3: 14$^{\text{th}}$ April - 7$^{\text{th}}$ May 2021; Quarter 4: 7$^{\text{th}}$ May - 20$^{\text{th}}$ May 2021.

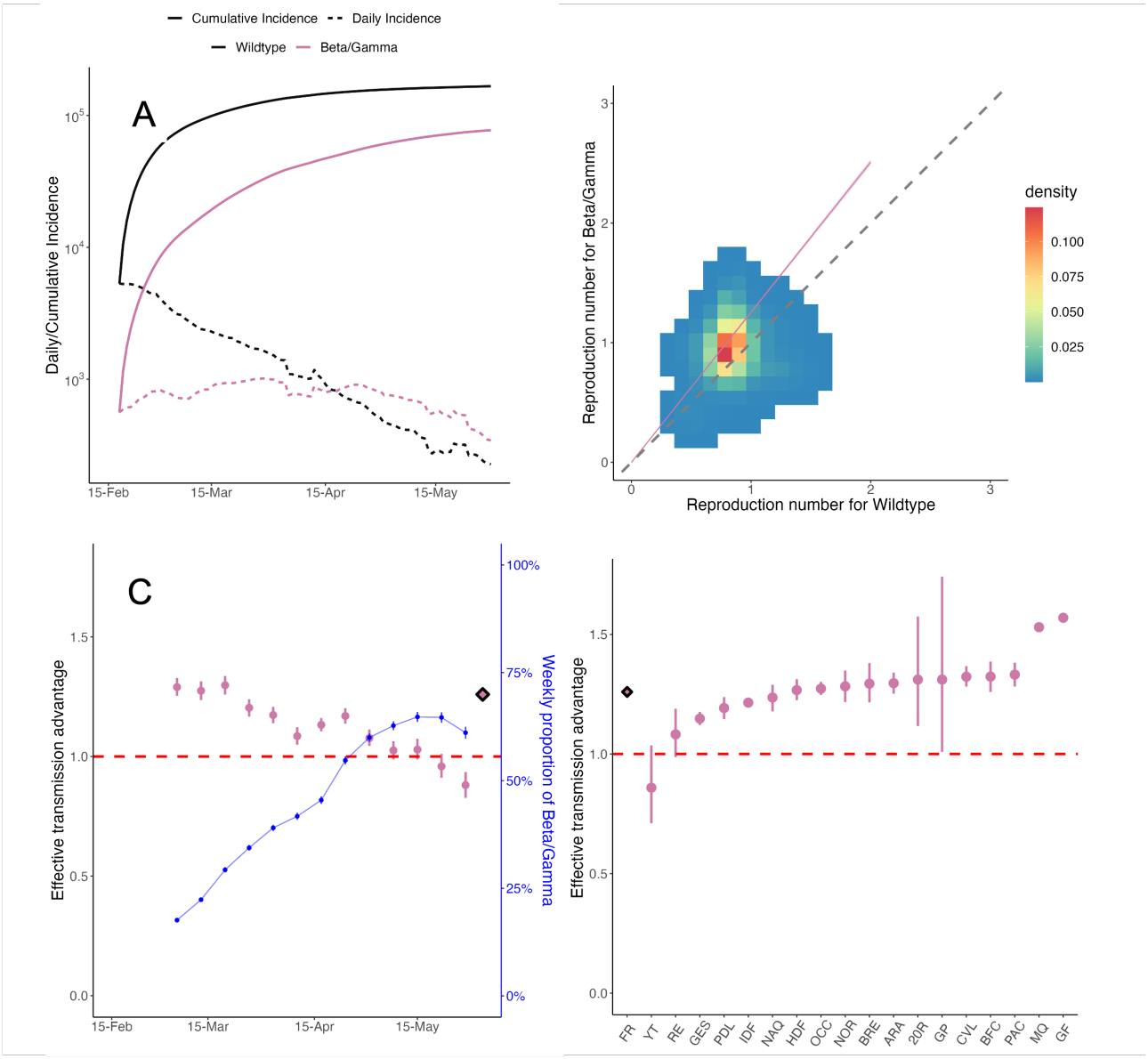


Figure S8: **Effective transmission advantage of Beta and Gamma (combined) over wildtype in France** (A) The daily reported incidence of cases of the wildtype (black) and Beta/Gamma (blue) in France 18$^{\text{th}}$ February to 30$^{\text{th}}$ May 2021. (B) The effective reproduction number R_t estimated independently for the wildtype (x-axis) and Beta/Gamma (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 Beta/Gamma is more transmissible. The pink line denotes the median effective transmission advantage estimated using MV-EpiEstim, assuming no temporal or spatial heterogeneity. 95% CrI were so narrow that they could not be distinguished from the line. (C) Effective transmission advantage estimated from MV-EpiEstim using data in the week ending on the date specified on the x-axis (pink circles) and the entire time series (diamond). The dark blue circles and the vertical bars denote respectively the mean and 95% binomial confidence interval of the proportion of incidence of Beta/Gamma (right y-axis) in the week of estimation. (D) Effective transmission advantage estimated using MV-EpiEstim for all ADM2 regions in France together (diamond) and separately (solid circles) using data from 18$^{\text{th}}$ February to 30$^{\text{th}}$ May 2021. The ADM2 regions are - ARA : Auvergne-Rhône-Alpes, BFC : Bourgogne-Franche-Comté, BRE : Bretagne, CVL : Centre-Val de Loire, 20R : Corse, GES : Grand Est, GP : Guadeloupe, GF : Guyane, HDF : Hauts-de-France, IDF : Île-de-France, RE : La Réunion, MQ : Martinique, YT : Mayotte, NOR : Normandie, NAQ : Nouvelle-Aquitaine, OCC : Occitanie, PDL : Pays de la Loire, and PAC : Provence-Alpes-Côte d'Azur. In panels (C) and (D), the solid pink circles denote the median estimate, the vertical lines indicate the 95% CrI, and the red dashed line denotes the $\epsilon = 1$ threshold.

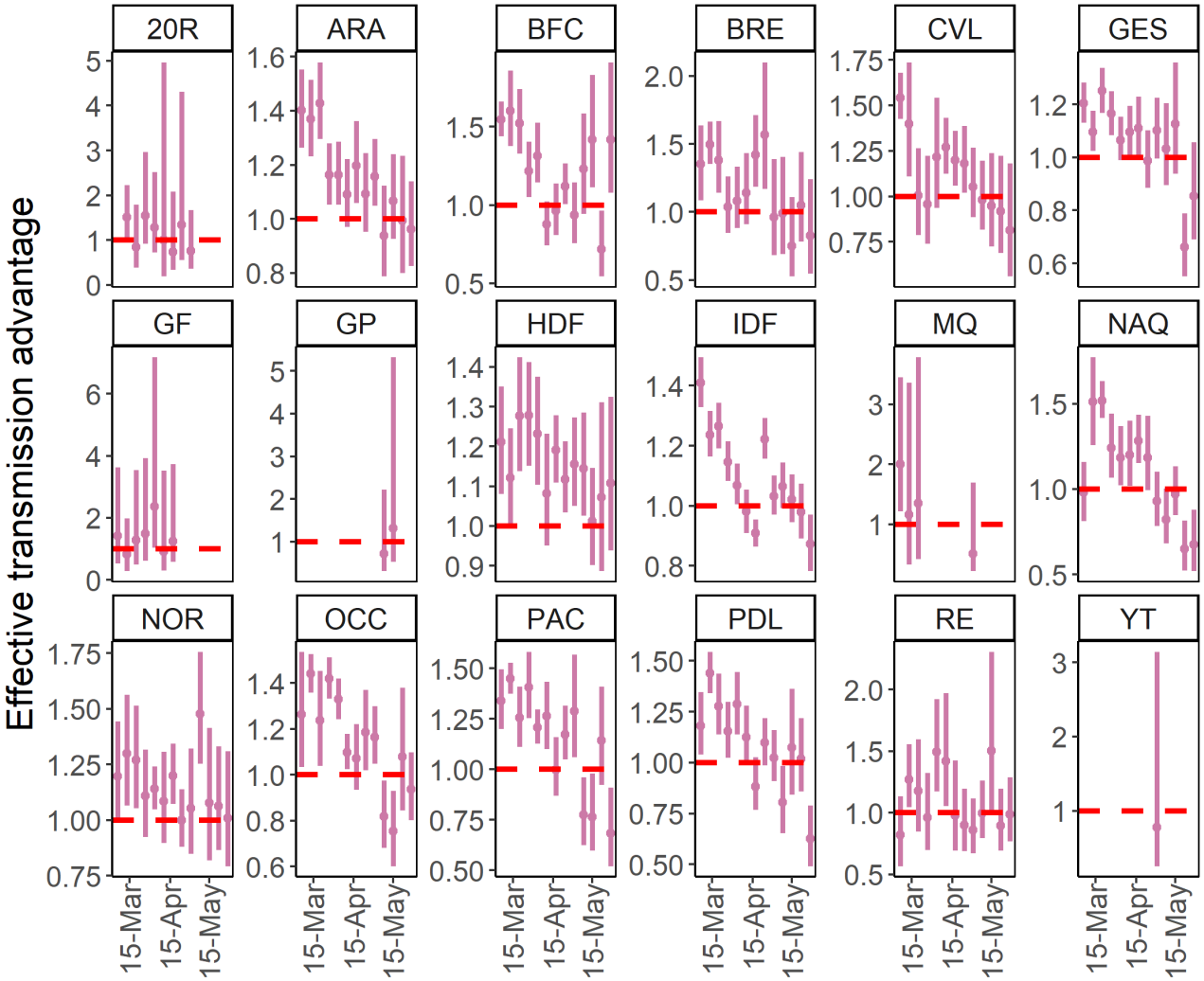


Figure S9: **Effective transmission advantage of Beta and Gamma (combined) over wildtype for all ADM2 regions in France from 18th February to 30th May 2021.** The ADM2 regions are - 20R : Corse, ARA : Auvergne-Rhône-Alpes, BFC : Bourgogne-Franche-Comté, BRE : Bretagne, CVL : Centre-Val de Loire, GES : Grand Est, GF : Guyane, GP : Guadeloupe, HDF : Hauts-de-France, IDF : Île-de-France, MQ : Martinique, NAQ : Nouvelle-Aquitaine, NOR : Normandie, OCC : Occitanie, PAC : Provence-Alpes-Côte d'Azur, PDL : Pays de la Loire, RE : La Réunion, and YT : Mayotte. Each point represents the median estimate made using MV-EpiEstim using data in the week ending on the date specified on the x-axis. The vertical pink lines indicate the 95% CrI, and the red dashed line denotes the $\epsilon = 1$ threshold.

4.3 Delta over Alpha (England)

Region/Time-period	Non-parametric	MV-EpiEstim
All	2.05 (1.09, 6.71)	1.77 (1.69, 1.85)
East of England	-	1.59 (1.37, 1.85)
London	1.23 (0.82, 2.78)	1.51 (1.36, 1.67)
Midlands	1.91 (1.42, 2.77)	1.75 (1.56, 1.95)
North East and Yorkshire	2.55 (1.71, 3.41)	2.12 (1.92, 2.35)
North West	1.84 (1.34, 2.86)	1.87 (1.69, 2.07)
South East	-	1.63 (1.41, 1.86)
South West	7.67 (4.10, 19.21)	1.76 (1.46, 2.13)
Quarter 1	-	1.42 (1.21, 1.66)
Quarter 2	-	1.70 (1.52, 1.90)
Quarter 3	1.75 (0.91, 2.46)	1.71 (1.58, 1.84)
Quarter 4	2.26 (1.42, 7.96)	1.95 (1.81, 2.09)

Table S4: Estimates of the effective transmission advantage of SARS-CoV-2 Delta variant over Alpha using the non-parametric approach and MV-EpiEstim for 7 NHS regions in England and 4 non-overlapping time periods. Estimates shown are the posterior median with 95% CrI in parenthesis. Quarters correspond to - Quarter 1: 25$^{\text{th}}$ March - 15$^{\text{rd}}$ April 2021; Quarter 2: 15$^{\text{th}}$ April - 5$^{\text{th}}$ May 2021; Quarter 3: 5$^{\text{th}}$ May - 26$^{\text{th}}$ May 2021; Quarter 4: 26$^{\text{th}}$ May - 16$^{\text{th}}$ June 2021.

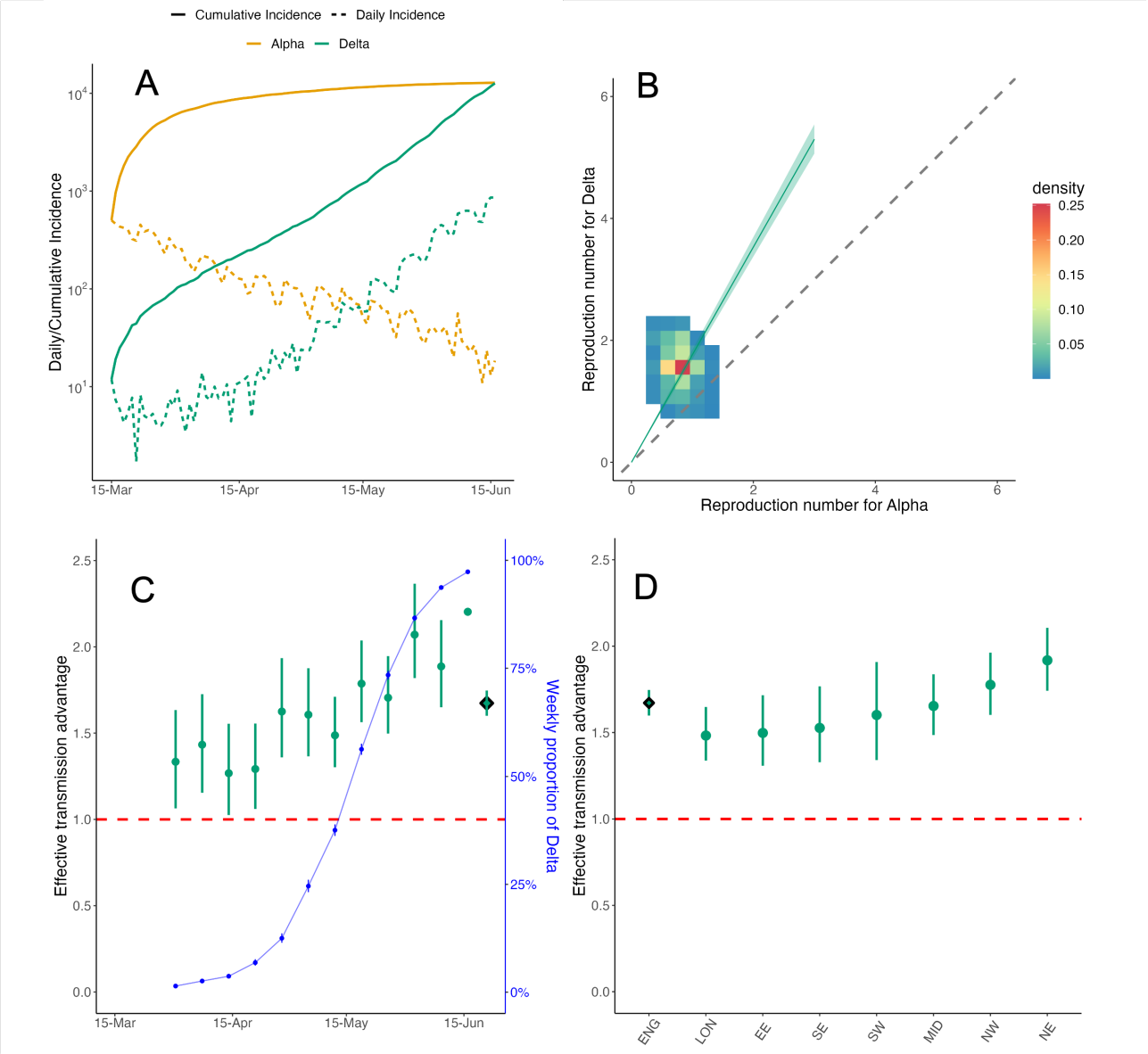


Figure S10: **Effective transmission advantage of Delta over Alpha in England** (A) The daily reported incidence of cases of Delta (green) and Alpha (yellow) in England from 15$^{\text{th}}$ March to 16$^{\text{th}}$ June 2021. (B) The effective reproduction number R_t estimated independently for Alpha (x-axis) and Delta (y-axis) on weekly sliding 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 Delta is more transmissible. The green line and the ribbon denote the median and 95% CrI of the effective transmission advantage estimated using MV-EpiEstim, assuming no temporal or spatial heterogeneity. (C) Effective transmission advantage estimated using MV-EpiEstim using data in the week ending on the date specified on the x-axis (green circles) and the entire time series (diamond). The dark blue circles and the vertical bars denote respectively the mean and 95% binomial confidence interval of the proportion of incidence of Delta (right y-axis) in the week of estimation. (D) Effective transmission advantage estimated using MV-EpiEstim for all NHS England regions together (diamond) and separately (solid circles), using data from 15$^{\text{th}}$ March to 16$^{\text{th}}$ June 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 green circles denote the median estimate, the vertical lines indicate the 95% CrI, and the red dashed line denotes the $\epsilon = 1$ threshold.

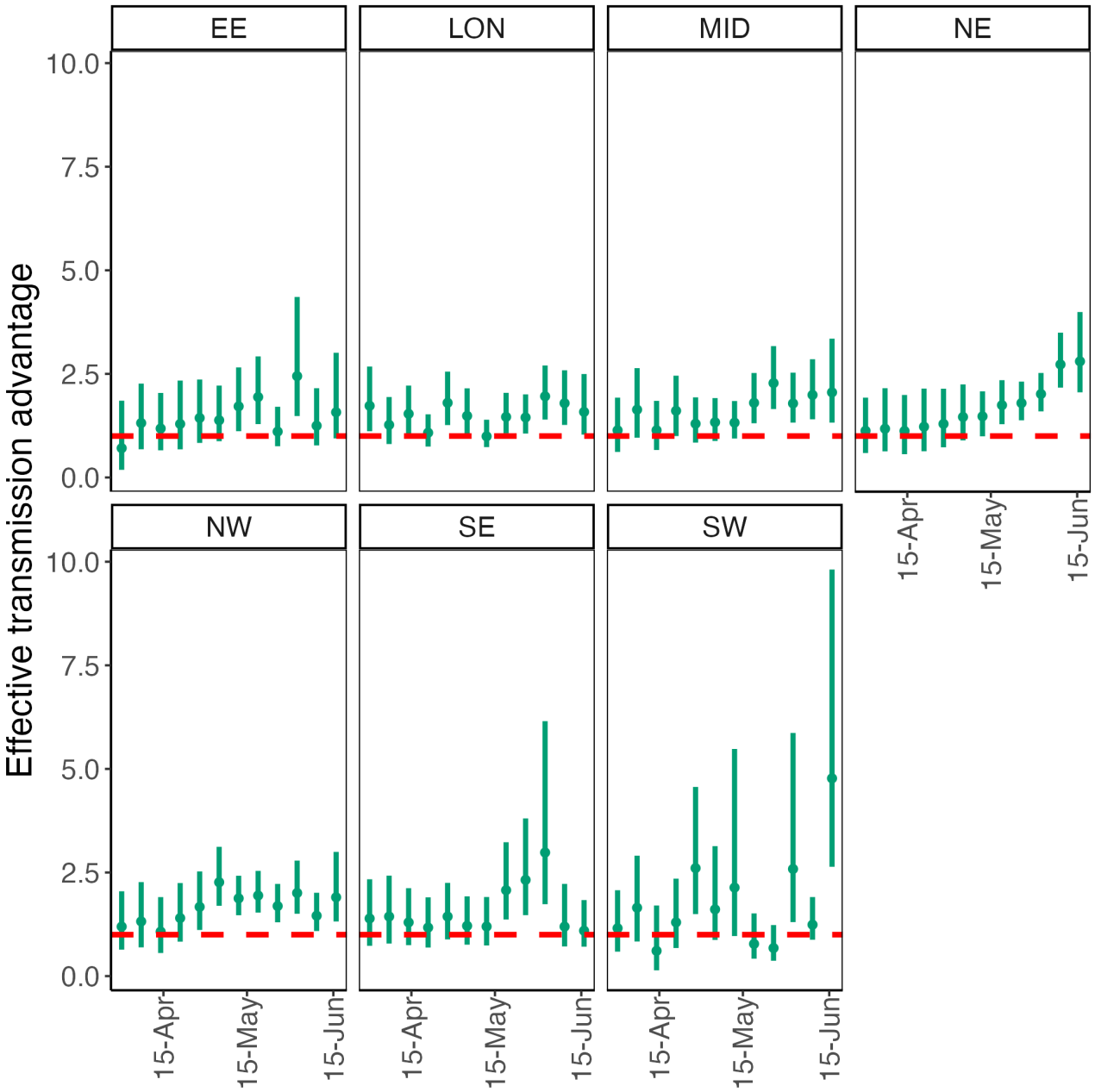


Figure S11: **Effective transmission advantage of Delta over Alpha for all NHS England regions from 15th March to 16th June 2021.** The NHS England regions are - East of England (EE), London (LON), Midlands (MID), North-East (NE), North-West (NW), South-East (SE), and South-West (SW). Each point represents the median estimate obtained using MV-EpiEstim using data in the week ending on the date specified on the x-axis. The vertical green lines indicate the 95% CrI, and the red dashed line denotes the $\epsilon = 1$ threshold.

5 Method performance using simulated data

5.1 Simulation approach

We simulated SARS-CoV-2-like incidence data using a branching process where daily incidence is assumed to follow a Poisson distribution:

$$I_t \sim Poisson(R_t \sum_{s=1}^{t-1} I_s \omega_{t-s}), \quad (1)$$

where ω_s is the probability mass function of the discrete serial interval.

We assume that the effective reproduction number for the variant is $\epsilon \times R_t$, where R_t is the effective reproduction number for the reference variant and ϵ is the effective transmission advantage of the new variant over the reference. We explored values of $\epsilon > 1$ in all simulation scenarios as $\epsilon < 1$ corresponds to swapping the reference and new variant.

We seeded the epidemic with 20 cases of the reference variant for 10 successive days and 1 case of the new variant on the 10^{th} day. We then simulated forward for an additional 100 days, generating 100 stochastic epidemic trajectories for each simulation scenario and each combination of parameters considered for that scenario (Tab S5). We then estimated the effective transmission advantage using 10, 20, 30, or 50 days of data counted from the 11^{th} day (see Sec. 3.2).

In each simulation scenario, we assessed the performance of the method using the following metrics:

- Bias, defined as difference between the mean posterior estimate of the effective transmission advantage and its true value. Bias should be as small as possible and will be zero for a perfect model.
- Uncertainty, defined as the posterior standard deviation (SD). Models with lower uncertainty are preferable as long as they are unbiased.
- Coverage probability, defined as the proportion of simulations where a given credible interval of the transmission advantage contained the true value. The 95% coverage probability for a well-calibrated model should be 95%, i.e. the true value will be contained in the 95% CrI in 95% of the simulations (analogous criterion is applicable for 50% coverage probability) [4, 5];
- Classification. We used the posterior distribution of ϵ to classify the new variant as "more transmissible", "less transmissible" than the reference or "unclear" (see methods in main text). To assess the classification performance, we consider the proportion of simulations where the variant is classified correctly (when the true value of ϵ is 1, we consider the correct classification to be 'unclear', see results in Tab S6). A perfect model would always classify the variant correctly. In practice, the threshold posterior quantile used for classifying a variant as more or less transmissible (see Methods section in the main text) determines the sensitivity of the classification and involves a well-known trade-off with its specificity.

Parameter	Values
Reference R_t	1.1, 1.6
Mean of reference serial interval	5.4 days
Standard deviation of reference serial interval	1.5 days
Mean of variant serial interval	Reference serial interval mean × (0.5, 1, 1.5, 2)
CV of variant serial interval	Reference serial interval CV × (0.5, 1, 1.5, 2)
Overdispersion	0.1, 0.5, 1

Table S5: Parameter values used in the simulations. For each simulation scenario, we considered all (relevant) combinations of parameter values shown in this table; and for each parameter combination, we simulated 100 data sets. CV: coefficient of variation

We first considered a baseline scenario (Sec. 5.3), where we assumed that the natural history of the reference and the new variants are same. We relaxed this assumption in other scenarios, assuming either that the SI distribution of the variant has a different mean (Sec. 5.4) or CV (Sec. 5.6), but that the SI distribution of both the new and reference variants are correctly specified in MV-EpiEstim. We then explored a scenario typical of real-time outbreak analysis where the SI distribution (mean or CV) of the variant is different from that of the reference but in the absence of more information, is assumed to be the same as that of the reference (Secs. 5.5 and 5.7). We also explored the performance of our method in the presence of superspreading (Sec. 5.8), extending the simulation to use a negative binomial offspring distribution, i.e.:

$$I_t \sim NegBin(R_t \sum_{s=1}^{t-1} I_s \omega_{t-s}, \kappa \sum_{s=1}^{t-1} I_s \omega_{t-s}), \tag{2}$$

where κ is the overdispersion parameter (lower values of κ denoting higher levels of superspreading) and $NegBin(\mu, k)$ denotes a negative binomial distribution with mean μ and variance $\mu + \frac{\mu^2}{k}$. In this formulation, we implicitly assume that the number of secondary infections per index case is distributed as $NegBin(R_t, \kappa)$. The sum of n independant identically distributed $NegBin(\mu, k)$ variables is $NegBin(n\mu, nk)$, yielding Eq. (2) when $n = \sum_{s=1}^{t-1} I_s \omega_{t-s}$, $\mu = R_t$ and $k = \kappa$. We use the R package projections [6] to simulate data, which implements superspeading as just described.

Finally, we assessed the sensitivity of MV-EpiEstim to under-reporting (Sec. 5.9), assuming a constant reporting rate for the reference and the new variant.

The scenarios outlined above used a constant R_t over the period of the simulation. We also explored the effect of time-varying R_t profiles on method performance (Sec. 5.10). We also considered simulations with two locations with time-varying R_t (Sec. 5.11), simulating independent epidemics in each location as described above. Finally, we assessed the performance of our method when estimating the transmission advantage in a scenario where the advantage varied over time (Sec. 5.12).

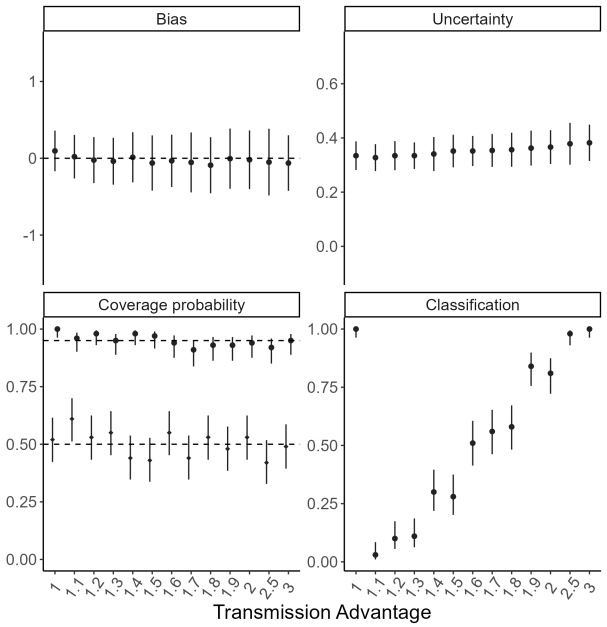
5.2 Description of figures

The figures that follow in the remainder of Sec. 5 are composed of four panels. In each figure panel, we present a performance indicator summarised across 100 simulations. In each figure, the top-left panel shows the mean ± SD of the bias in the estimate of the effective transmission advantage. The dashed horizontal line denotes the threshold bias of 0. The top-right panel shows the mean ± SD of the uncertainty in estimates. The bottom-left panel shows the 95% (circles) and 50% (diamonds) coverage probabilities (mean and 95% binomial confidence interval (CI)). The dashed horizontal lines denote the threshold values of 0.95 and 0.50. The bottom-right panel shows classification performance (mean and 95% binomial CI). For definition of each performance indicator, see Sec. 5.1.

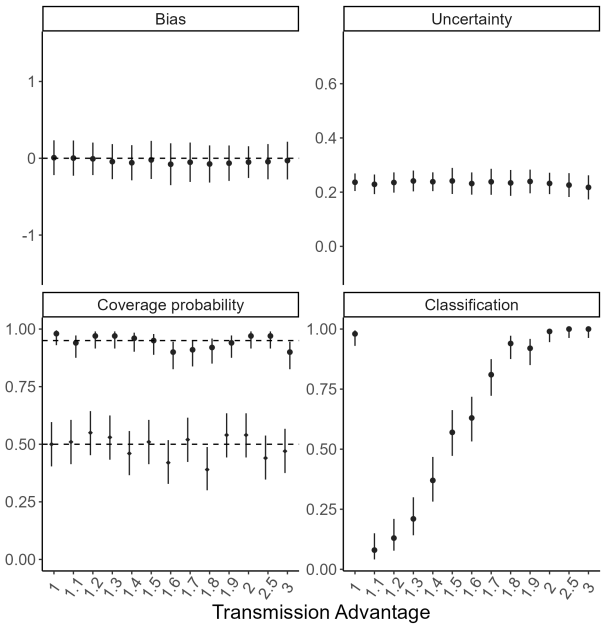
5.3 Baseline scenario

In this section, we present the results for the baseline scenario. That is, we assume no superspreading, and that both the reference and the new variant have the same natural history. Results are shown for estimates obtained using 10, 20, 30 and 50 days of data in MV-EpiEstim.

(A) $R_t = 1.1$ and 10 days of data



(B) $R_t = 1.6$ and 10 days of data



(C) $R_t = 1.1$ and 20 days of data

(D) $R_t = 1.6$ and 20 days of data

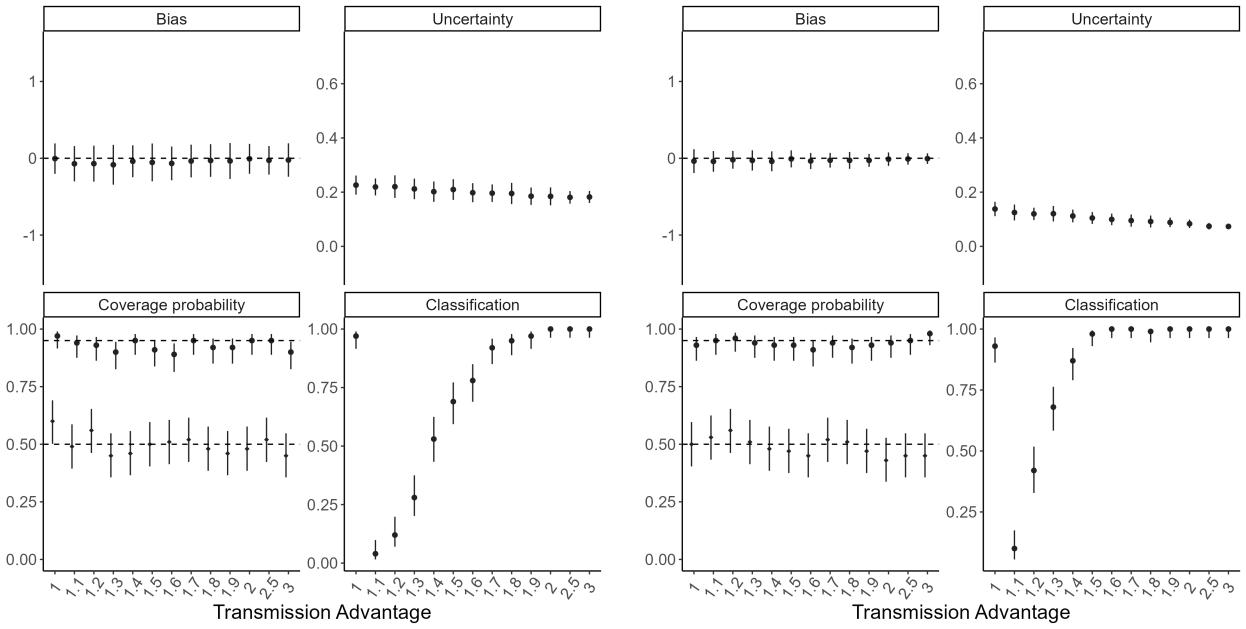


Figure S12: Method performance using simulated data assuming the same natural history for the reference and the variant (using 10 or 20 days of incidence data). In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

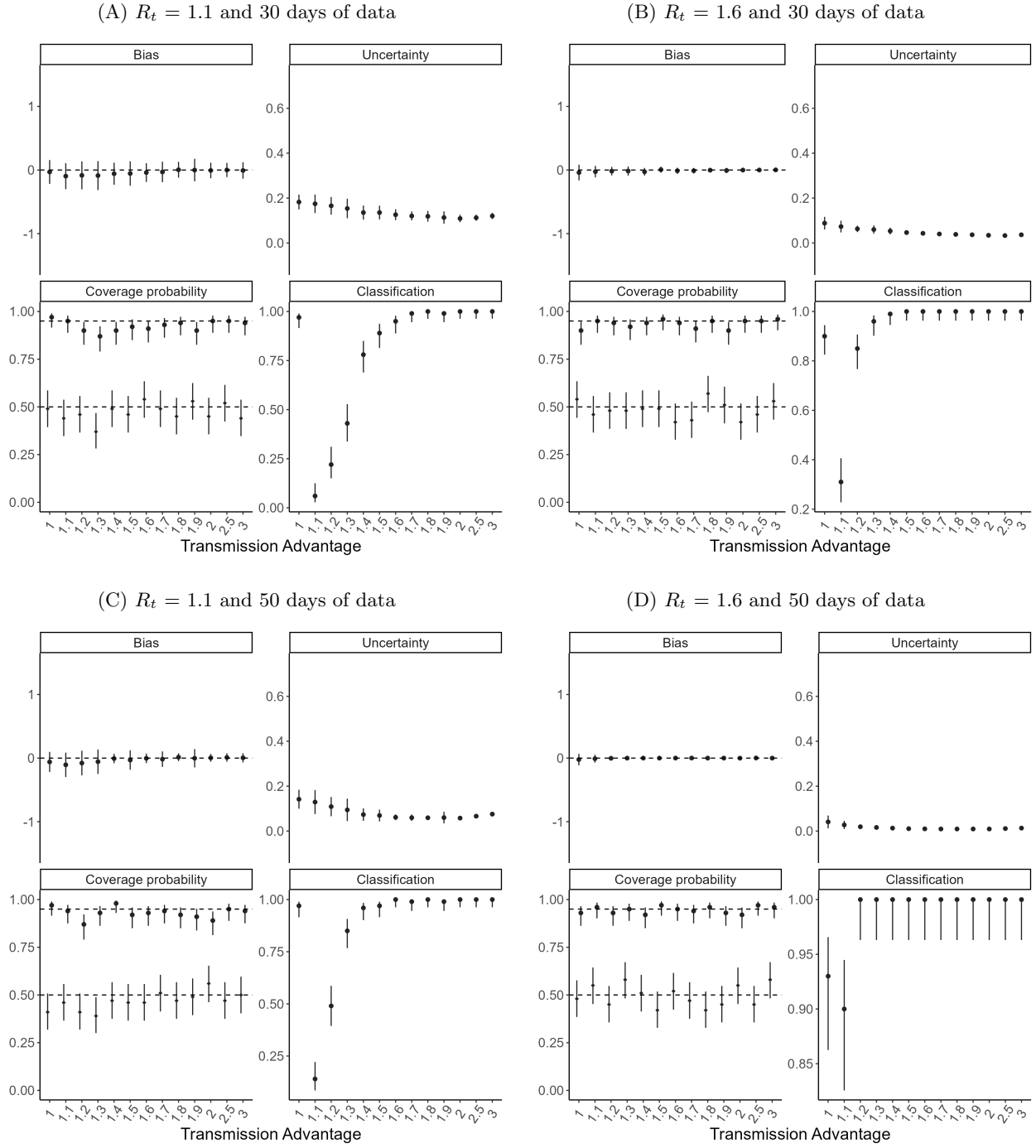


Figure S13: Method performance using simulated data assuming the same natural history for the reference and the variant (using 30 or 50 days of incidence data). In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

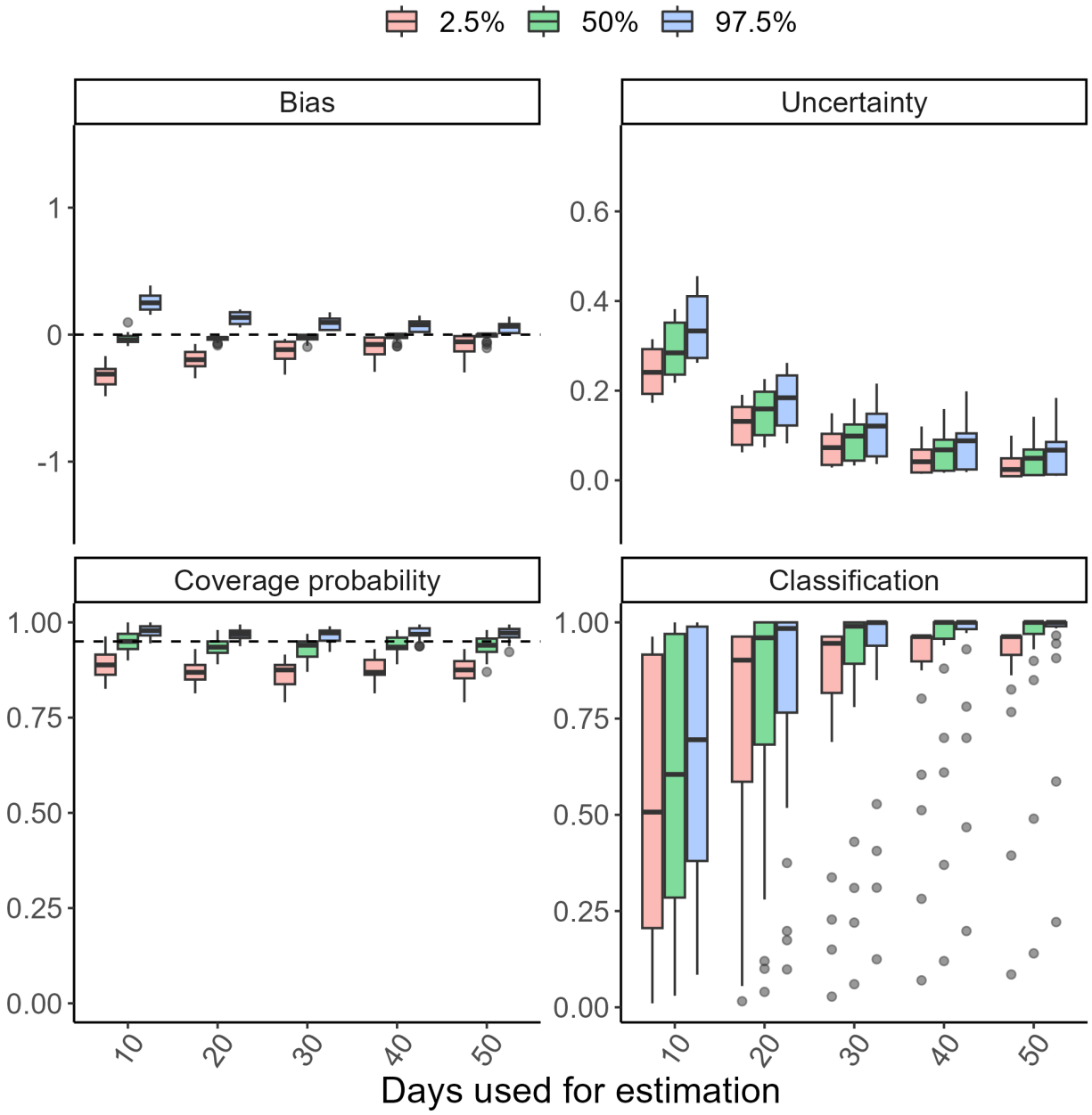
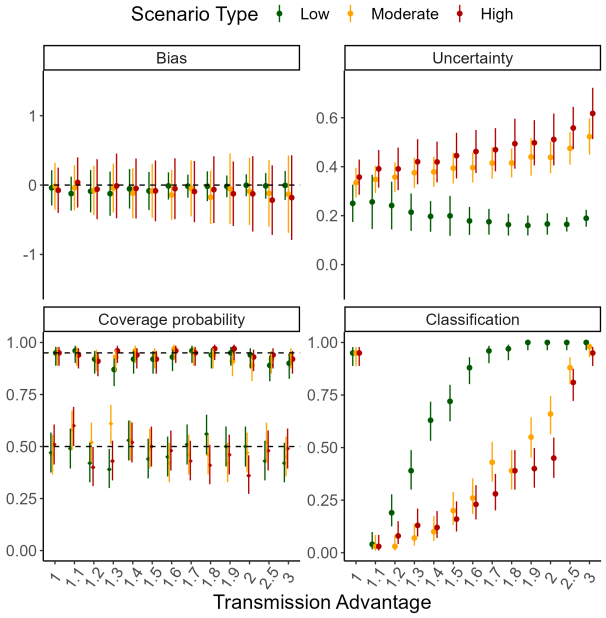


Figure S14: Method performance using simulated data assuming the same natural history for the reference and the variant. In each panel, the boxplots depict the change in the distribution of the 2.5$^{\text{th}}$, 50$^{\text{th}}$ and 97.5$^{\text{th}}$ quantiles of the corresponding metric as progressively more data are used for estimation. For each metric, each quantile is summarised across all simulations for the values of reference R_t (1.1 and 1.6) and the range of ϵ (1 to 3) in the baseline scenario.

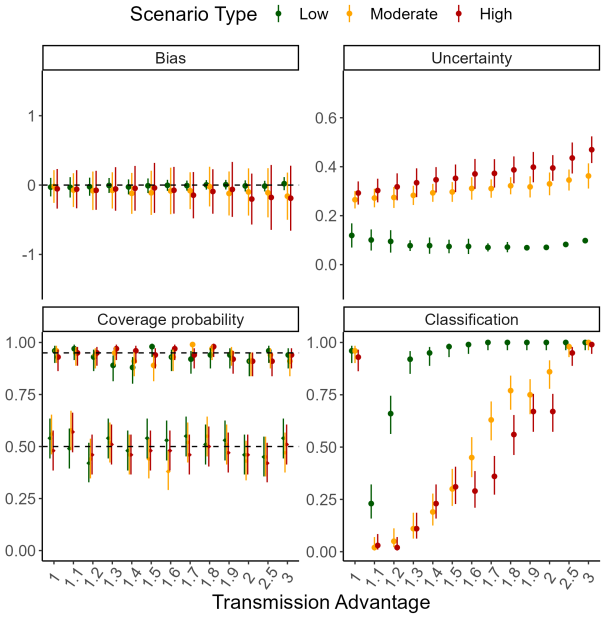
5.4 Sensitivity to serial interval mean

In this section, we present results for the scenario where data were simulated assuming different natural history parameters for the reference and the variant. We assumed that the mean serial interval of the variant is $0.5, 1.5$, or 2 times that of the reference. Further, we assumed that the parameters of both the reference and the variant are correctly specified during estimations. Results are shown using $10, 20, 30$, and 50 days of incidence data.

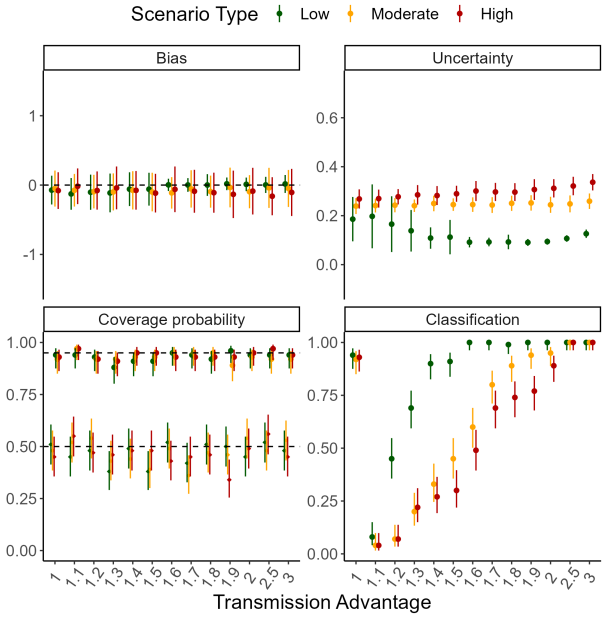
(A) $R_t = 1.1$ and 10 days of data



(B) $R_t = 1.6$ and 10 days of data



(C) $R_t = 1.1$ and 20 days of data



(D) $R_t = 1.6$ and 20 days of data

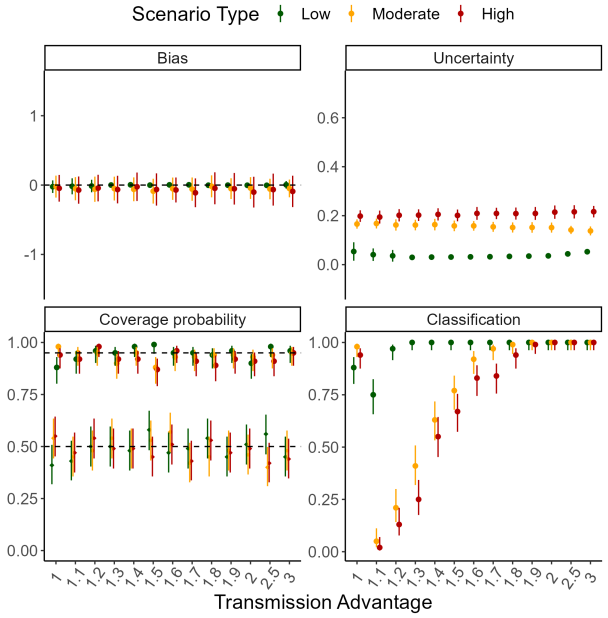
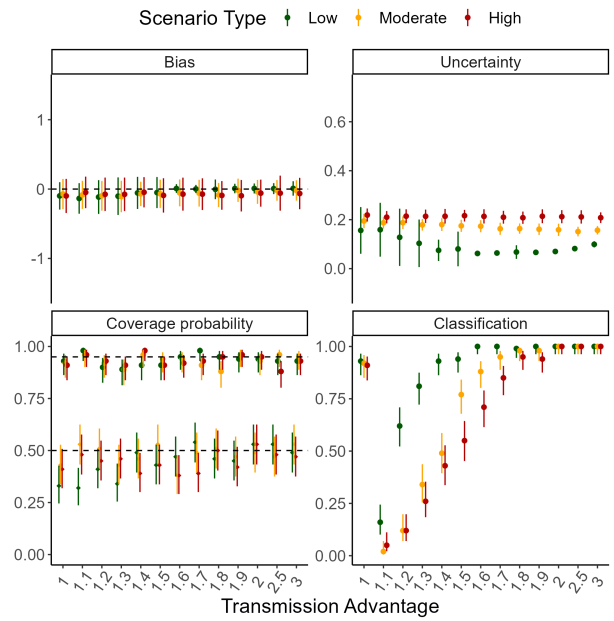
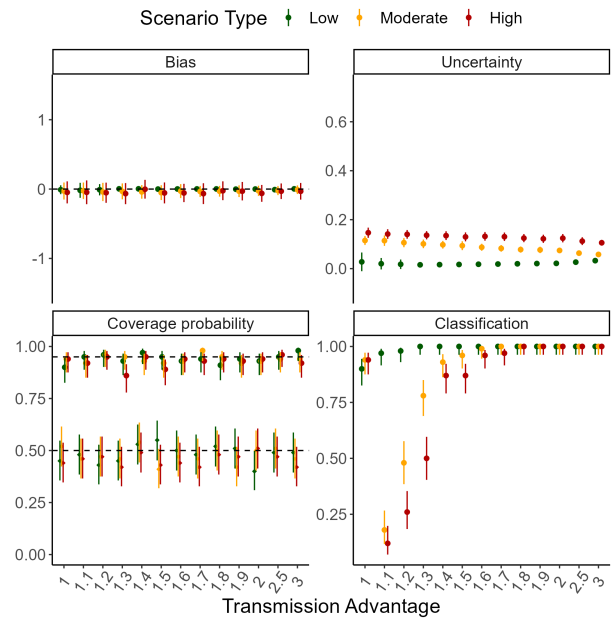


Figure S15: Method performance using simulated data assuming different SI mean for the reference and the variant (using 10 or 20 days of incidence data). The mean serial interval of the variant is assumed to be 0.5 (low), 1.5 (moderate) or 2 (high) times that of the reference. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

(A) $R_t = 1.1$ and 30 days of data



(B) $R_t = 1.6$ and 30 days of data



(C) $R_t = 1.1$ and 50 days of data

(D) $R_t = 1.6$ and 50 days of data

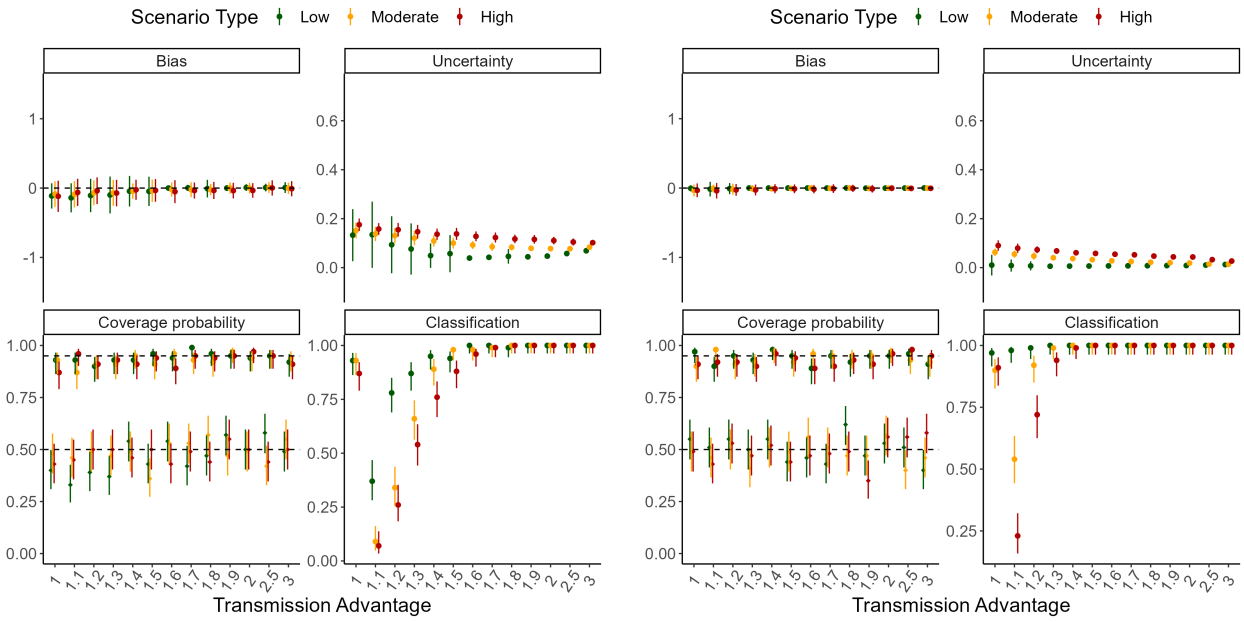
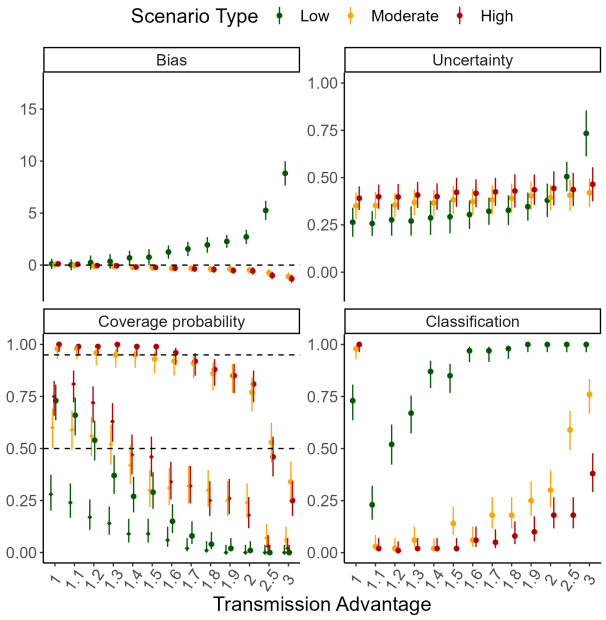


Figure S16: Method performance using simulated data assuming different SI mean for the reference and the variant (using 30 or 50 days of incidence data). The mean serial interval of the variant is assumed to be 0.5 (low), 1.5 (moderate) or 2 (high) times that of the reference. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

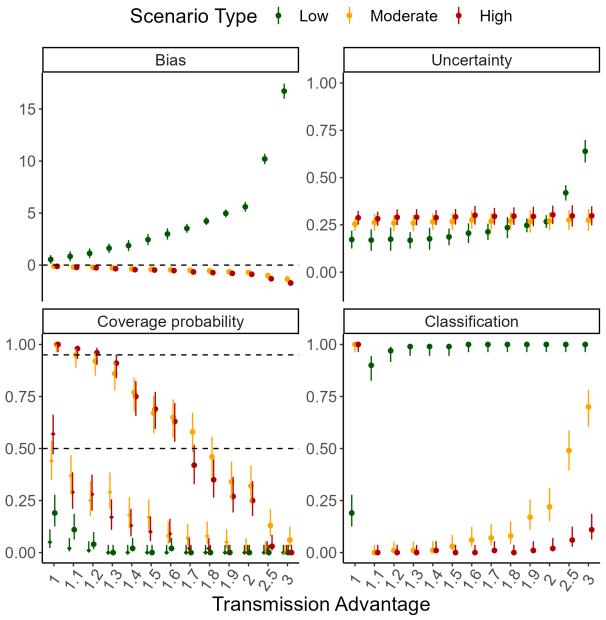
5.5 Misspecification of serial interval mean

In this section, we present results for the scenario where data were simulated assuming different natural history parameters for the reference and the variant. We assumed that the mean serial interval of the variant is 0.5, 1.5, or 2 times that of the reference. However, we assumed that the parameters of both the reference and the variant are assumed to be the same during estimation. Results are shown using 10, 20, 30, and 50 days of incidence data.

(A) $R_t = 1.1$ and 10 days of data



(B) $R_t = 1.6$ and 10 days of data



(C) $R_t = 1.1$ and 20 days of data

(D) $R_t = 1.6$ and 20 days of data

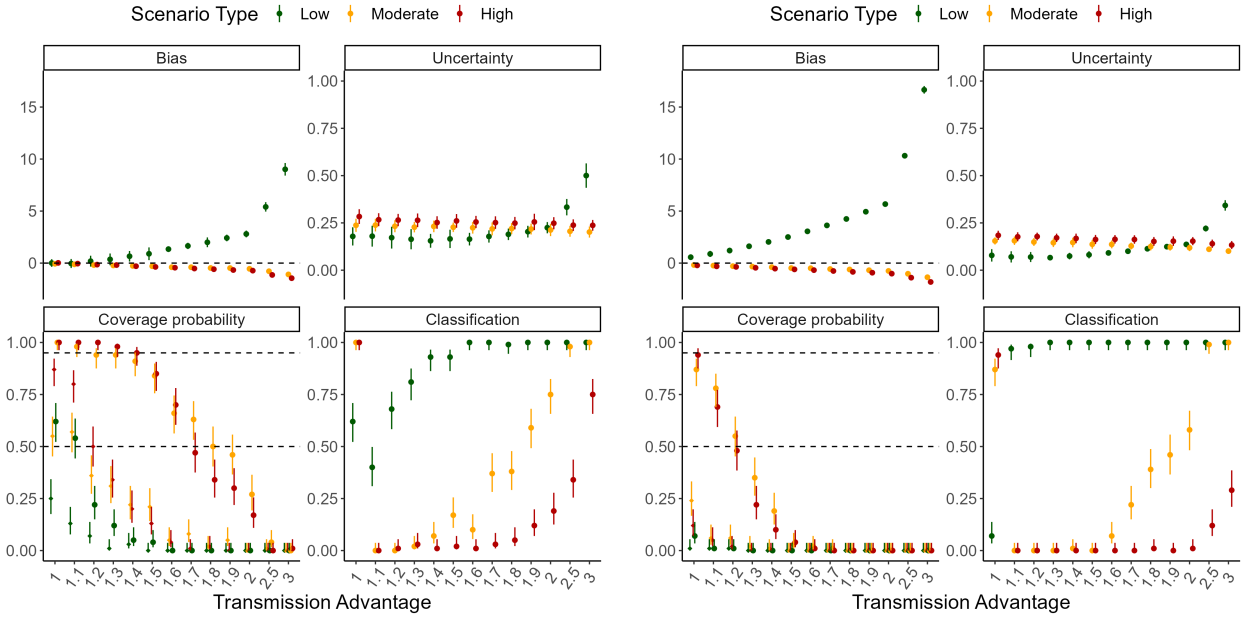
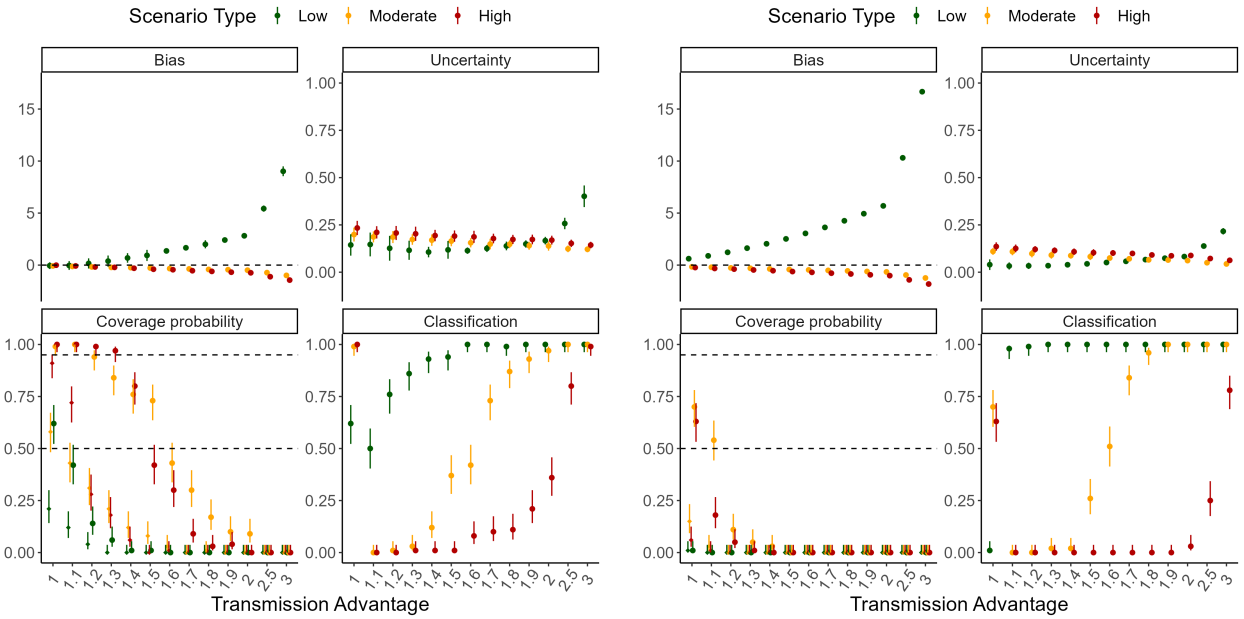


Figure S17: Method performance using simulated incidence data when the mean SI of the variant is different and is misspecified during estimation (using 10 or 20 days of incidence data). The mean serial interval of the variant is assumed to be 0.5 (low), 1.5 (moderate) or 2 (high) times that of the reference. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

(A) $R_t = 1.1$ and 30 days of data

(B) $R_t = 1.6$ and 30 days of data



(C) $R_t = 1.1$ and 50 days of data

(D) $R_t = 1.6$ and 50 days of data

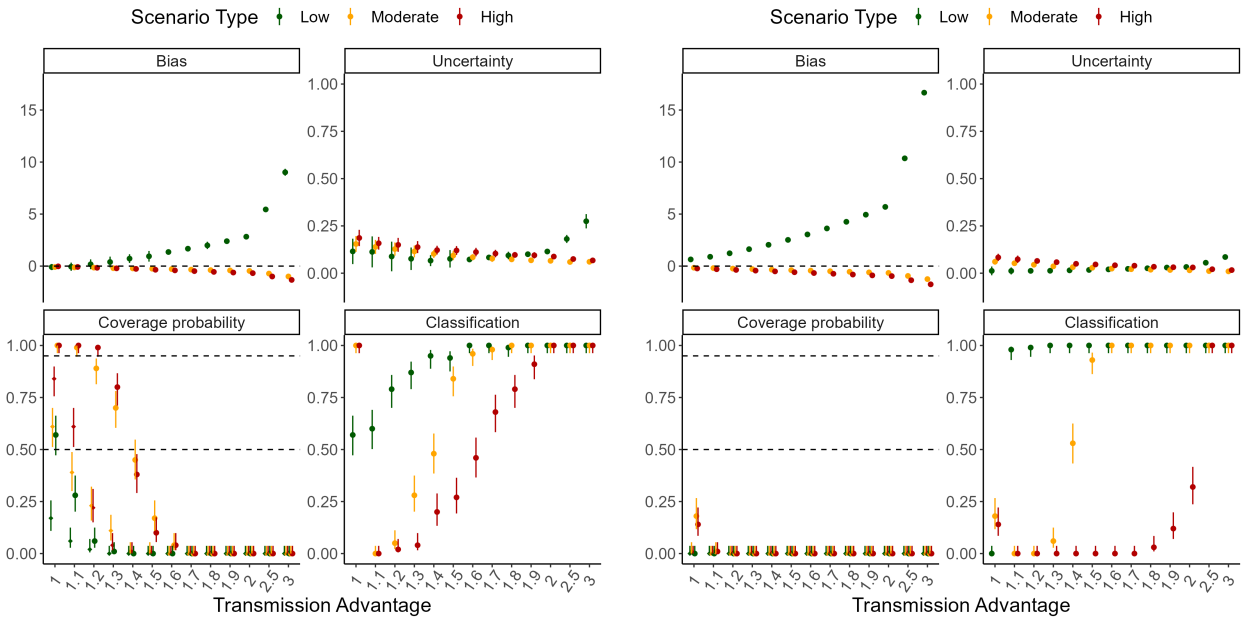
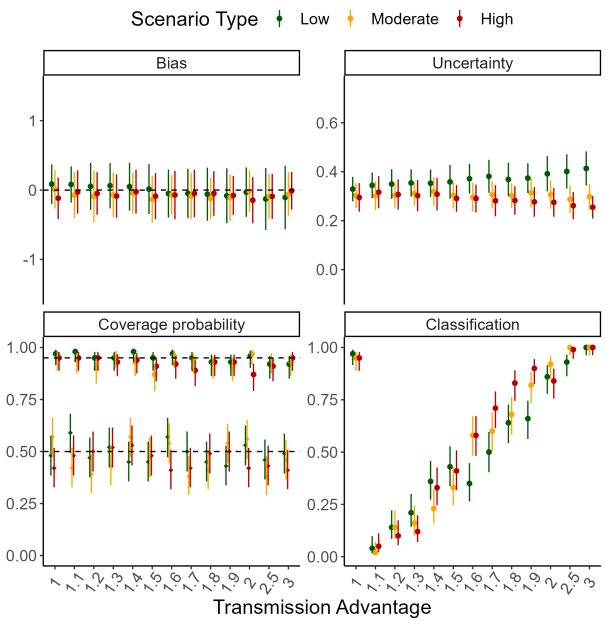


Figure S18: Method performance using simulated incidence data when the mean SI of the variant is different and is misspecified during estimation (using 30 or 50 days of incidence data). The mean serial interval of the variant is assumed to be 0.5 (low), 1.5 (moderate) or 2 (high) times that of the reference. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

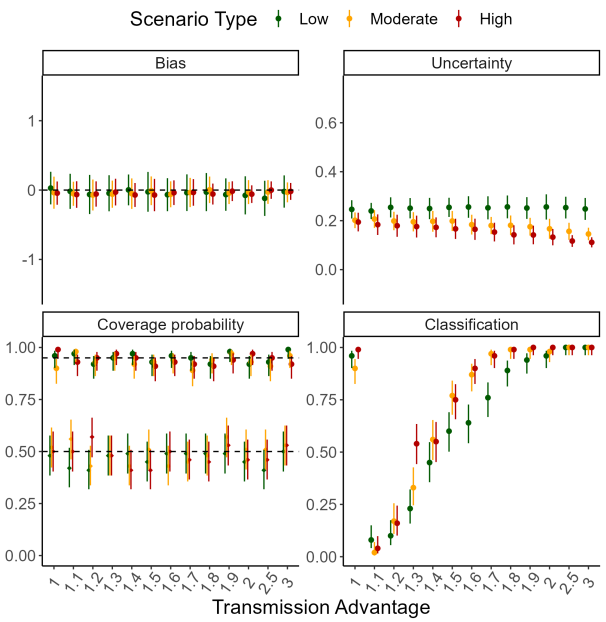
5.6 Sensitivity to serial interval CV

In this section, we present results for the scenario where data were simulated assuming different natural history parameters for the reference and the variant. We assumed that the CV of the serial interval distribution of the variant is $0.5, 1.5$, or 2 times that of the reference. Further, we assumed that the parameters of both the reference and the variant are correctly specified during estimation. Results are shown using $10, 20, 30$, and 50 days of data.

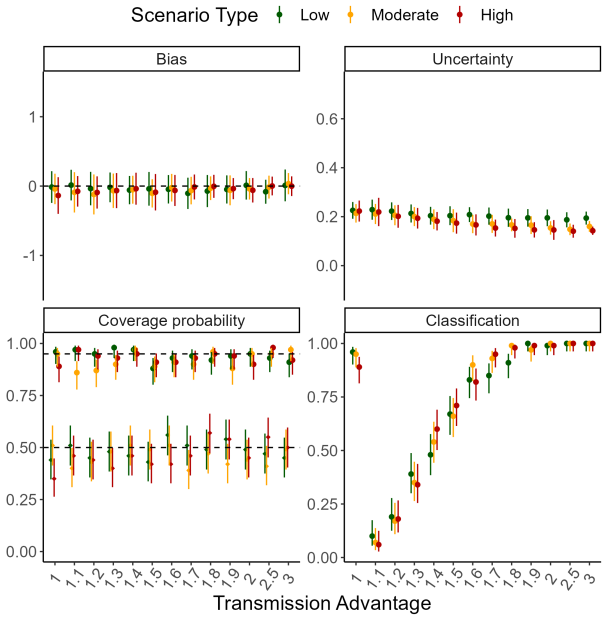
(A) $R_t = 1.1$ and 10 days of data



(B) $R_t = 1.6$ and 10 days of data



(C) $R_t = 1.1$ and 20 days of data



(D) $R_t = 1.6$ and 20 days of data

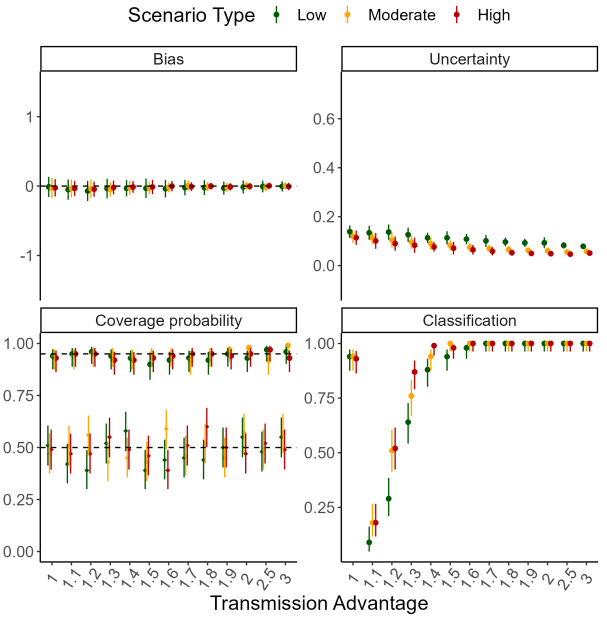
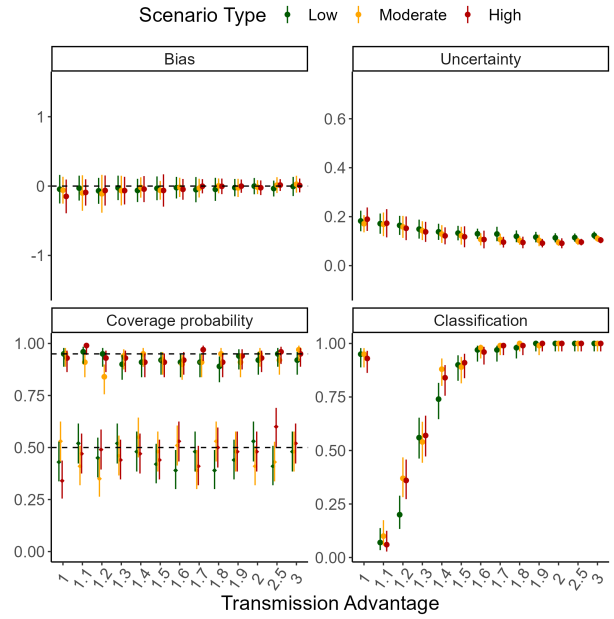
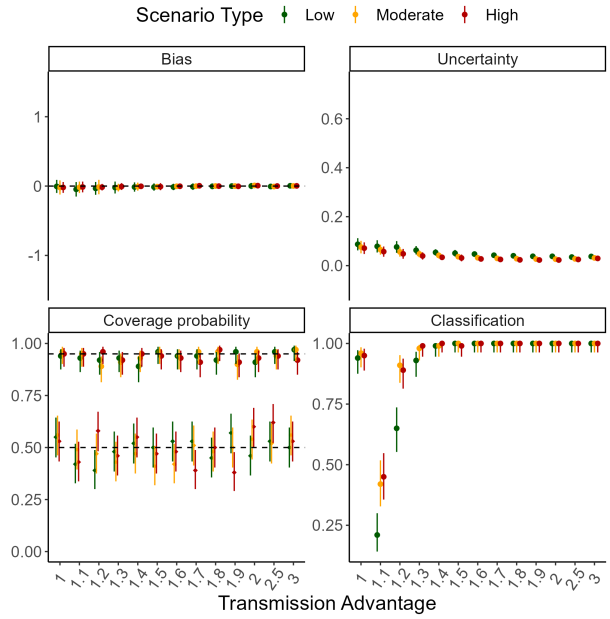


Figure S19: Method performance using simulated incidence data when the CV of the SI distribution of the variant is different and is correctly specified during estimation (using 10 or 20 days of incidence data). The CV of the serial interval distribution of the variant is assumed to be 0.5 (low), 1.5 (moderate) or 2 (high) times that of the reference. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

(A) $R_t = 1.1$ and 30 days of data



(B) $R_t = 1.6$ and 30 days of data



(C) $R_t = 1.1$ and 50 days of data

(D) $R_t = 1.6$ and 50 days of data

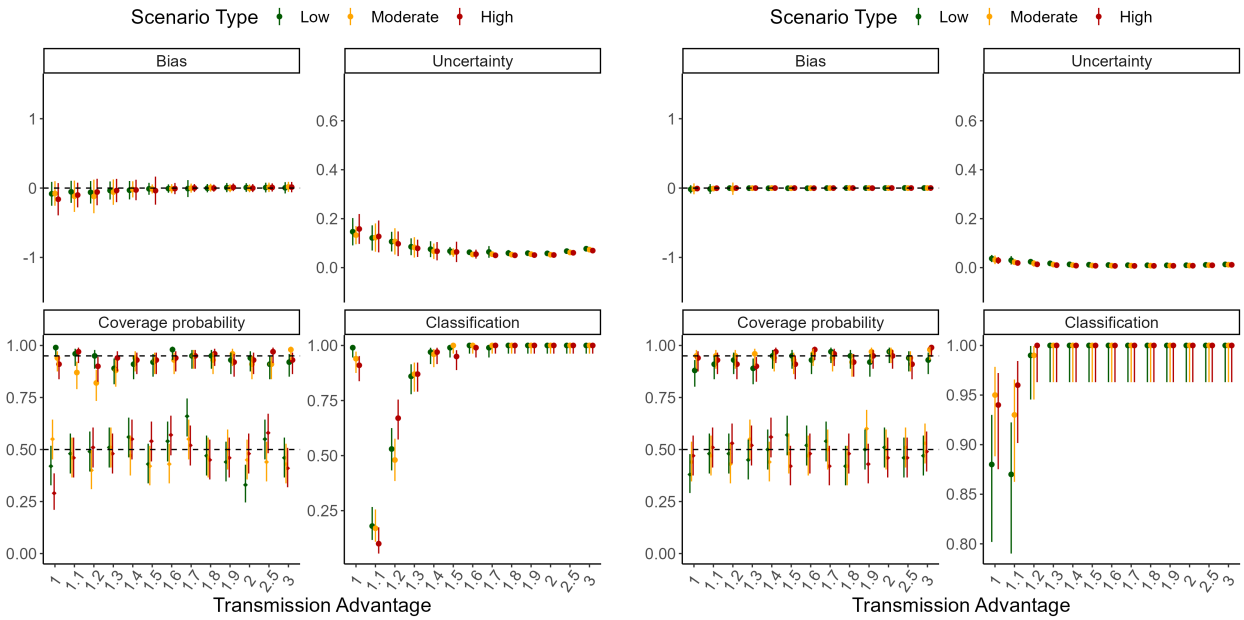
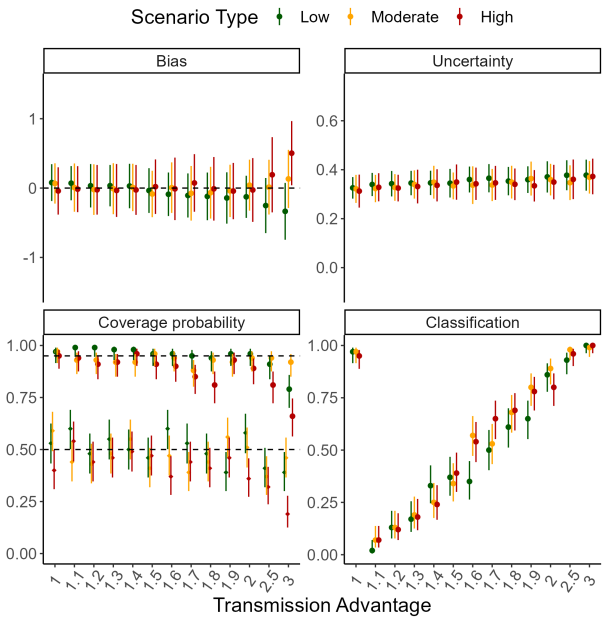


Figure S20: Method performance using simulated incidence data when the CV of the SI distribution of the variant is different and is correctly specified during estimation (using 30 or 50 days of incidence data). The CV of the serial interval distribution of the variant is assumed to be 0.5 (low), 1.5 (moderate) or 2 (high) times that of the reference. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

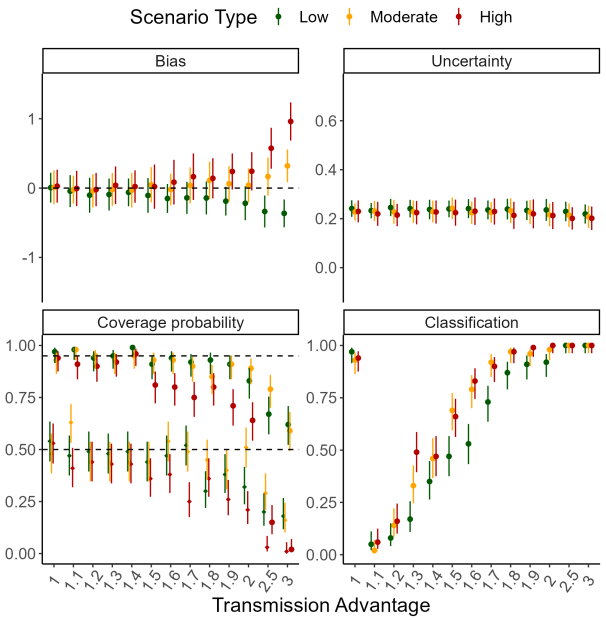
5.7 Misspecification of serial interval CV

In this section, we present results for the scenario where data were simulated assuming different natural history parameters for the reference and the variant. We assumed that the CV of the serial interval distribution of the variant is $0.5, 1.5$, or 2 times that of the reference. However, we assumed that the parameters of both the reference and the variant are assumed to be the same during estimation. Results are shown using $10, 20, 30$, and 50 days of incidence data.

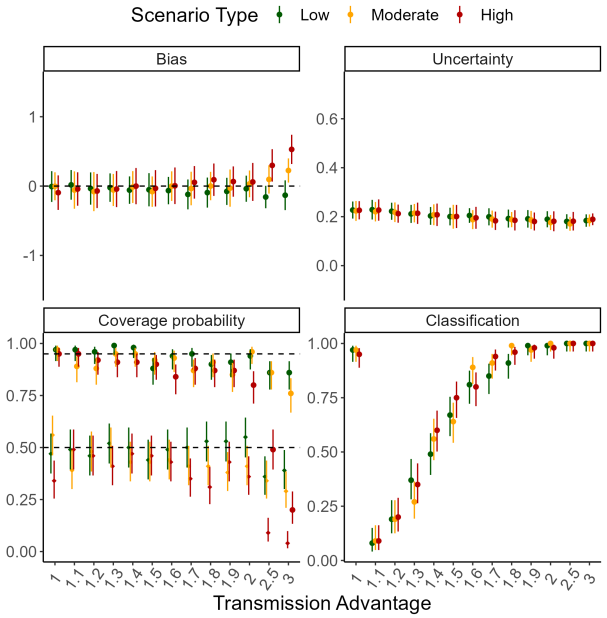
(A) $R_t = 1.1$ and 10 days of data



(B) $R_t = 1.6$ and 10 days of data



(C) $R_t = 1.1$ and 20 days of data



(D) $R_t = 1.6$ and 20 days of data

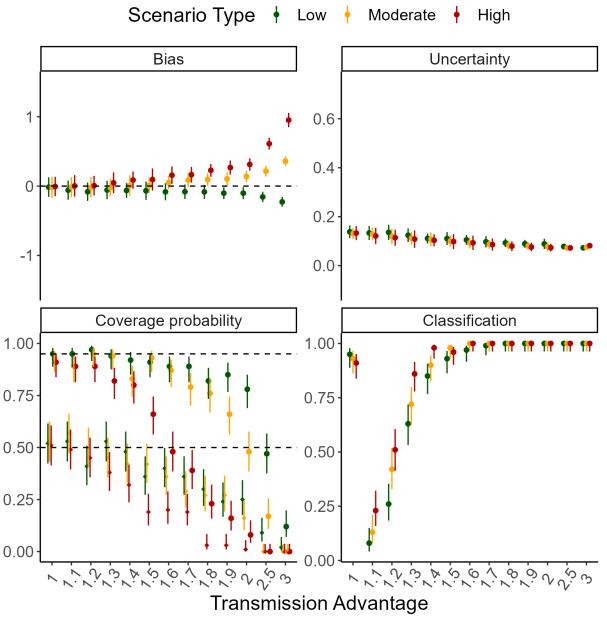
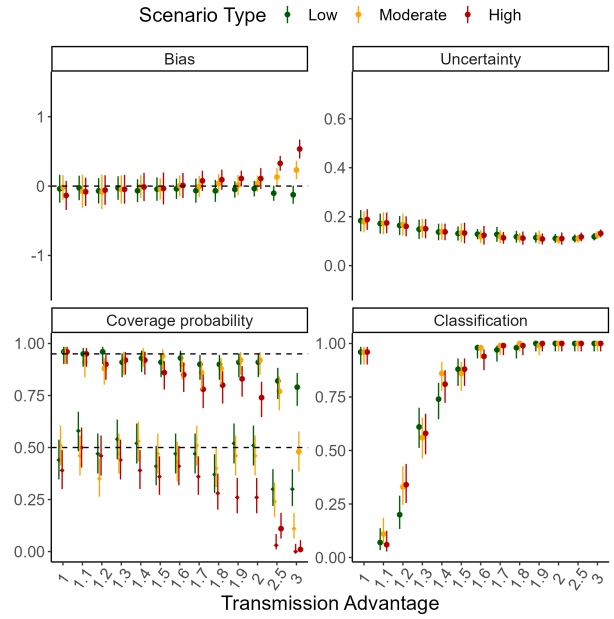
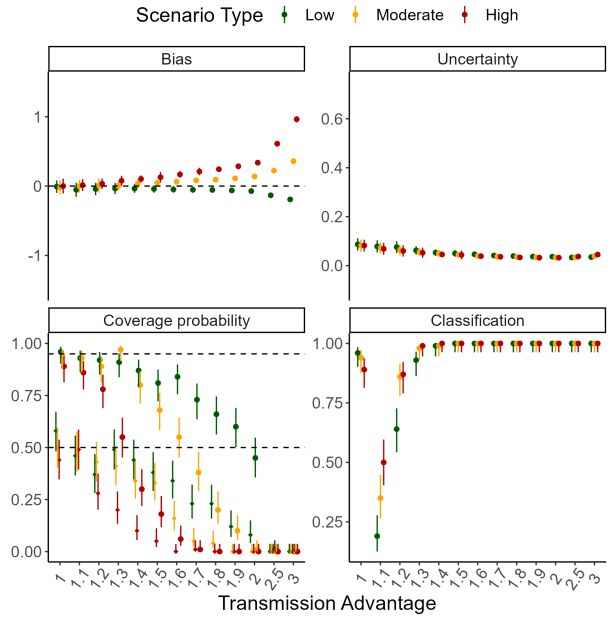


Figure S21: Method performance using simulated incidence data using simulated incidence data when the CV of the SI distribution of the variant is different and is misspecified during estimation (using 10 or 20 days of incidence data). The CV of the serial interval distribution of the variant is assumed to be 0.5 (low), 1.5 (moderate) or 2 (high) times that of the reference. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

(A) $R_t = 1.1$ and 30 days of data



(B) $R_t = 1.6$ and 30 days of data



(C) $R_t = 1.1$ and 50 days of data

(D) $R_t = 1.6$ and 50 days of data

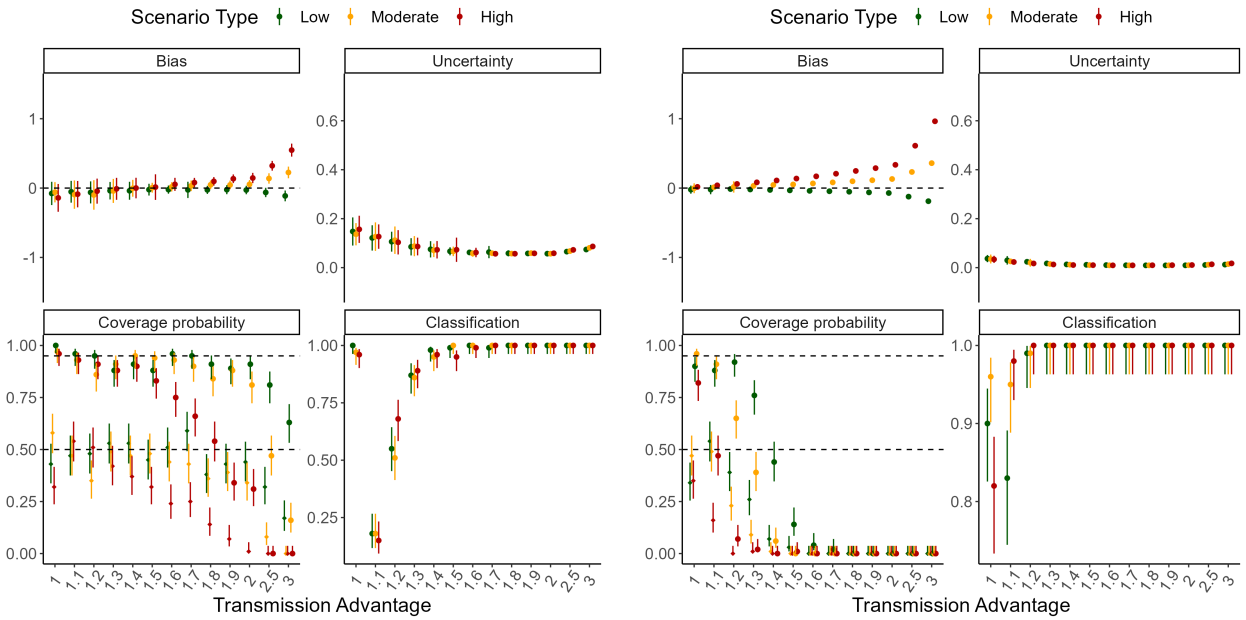
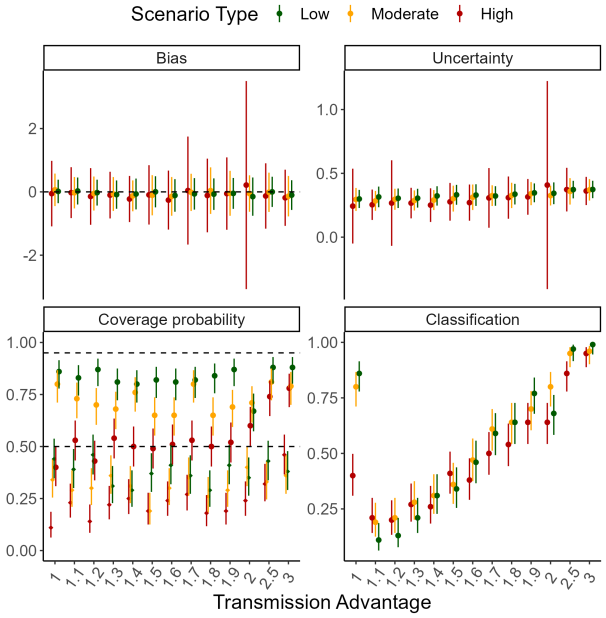


Figure S22: Method performance using simulated incidence data using simulated incidence data when the CV of the SI distribution of the variant is different and is misspecified during estimation (using 30 or 50 days of incidence data). The CV of the serial interval distribution of the variant is assumed to be 0.5 (low), 1.5 (moderate) or 2 (high) times that of the reference. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

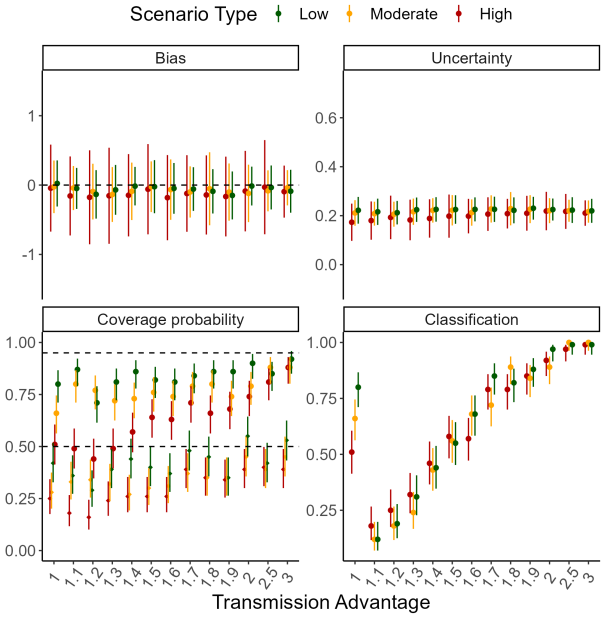
5.8 Sensitivity to superspreading

To explore sensitivity of MV-EpiEstim to superspreading (which is not explicitly accounted for in MV-EpiEstim), we used a Negative Binomial distribution as the offspring distribution as in Eq. (2). We simulated data using low (overdispersion parameter $\kappa = 1$), moderate ($\kappa = 0.5$), and high ($\kappa = 0.1$) levels of superspreading. This section presents the performance metrics when 10, 20, 30, and 50 days of incidence data were used for estimation.

(A) $R_t = 1.1$ and 10 days of data



(B) $R_t = 1.6$ and 10 days of data



(C) $R_t = 1.1$ and 20 days of data

(D) $R_t = 1.6$ and 20 days of data

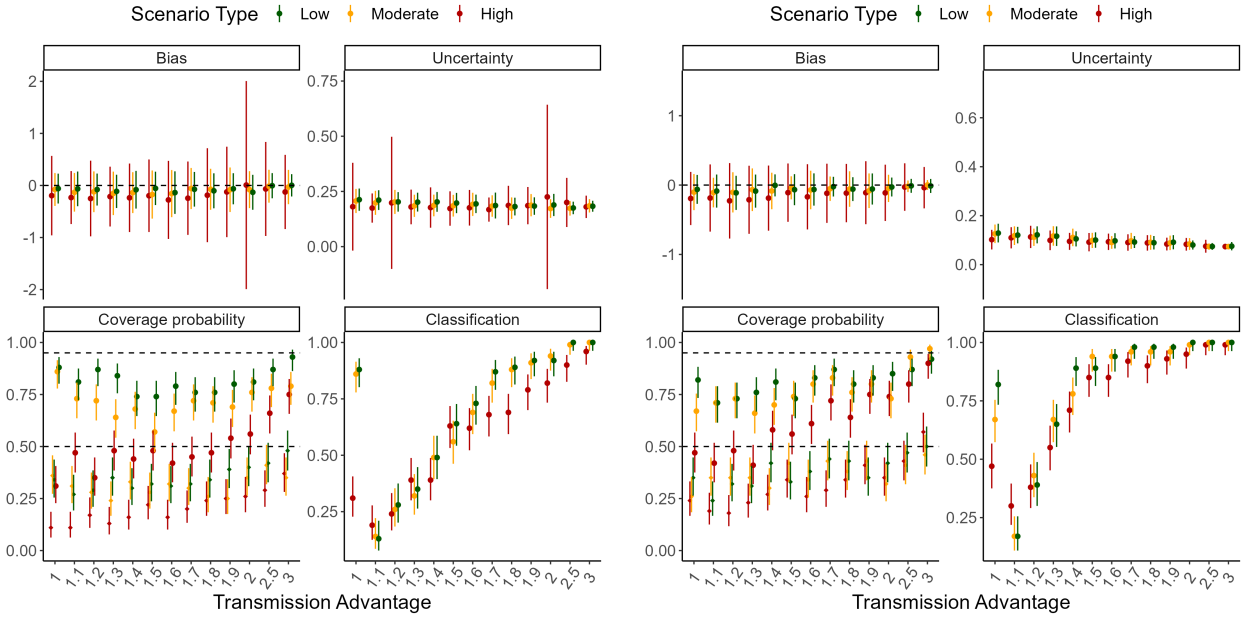
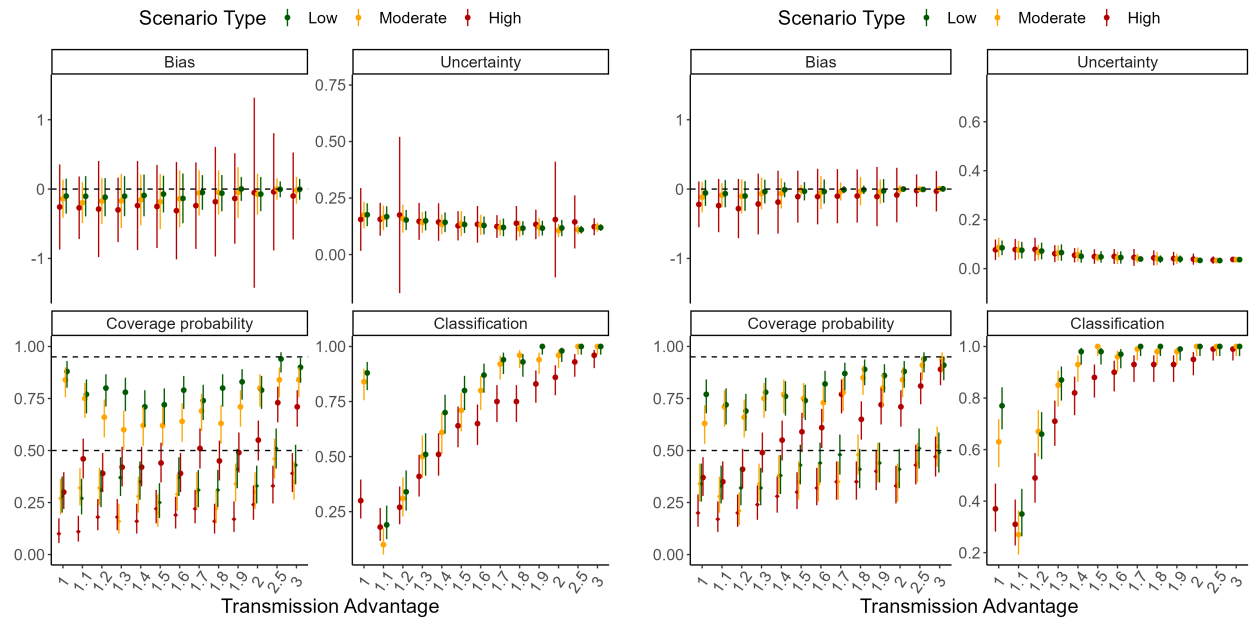


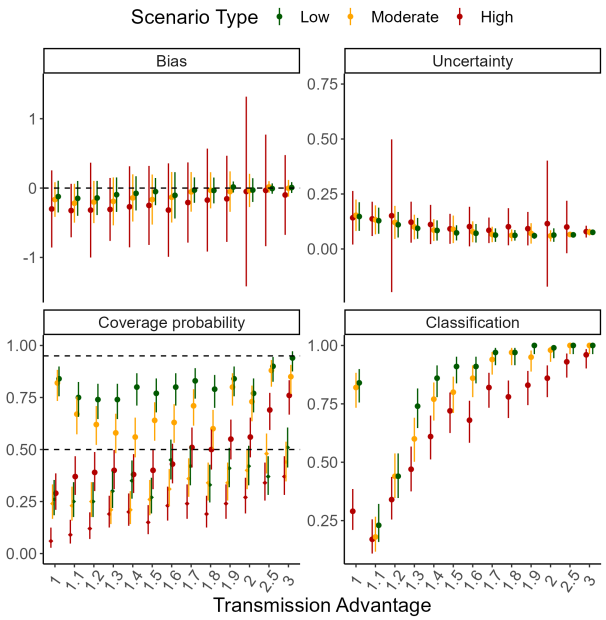
Figure S23: Method performance using incidence data simulated with superspreading (using 10 or 20 days of incidence data). We simulated data with low (overdispersion parameter $\kappa = 1$), moderate ($\kappa = 0.5$) and high ($\kappa = 0.1$) levels of superspreading. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

(A) $R_t = 1.1$ and 30 days of data

(B) $R_t = 1.6$ and 30 days of data



(C) $R_t = 1.1$ and 50 days of data



(D) $R_t = 1.6$ and 50 days of data

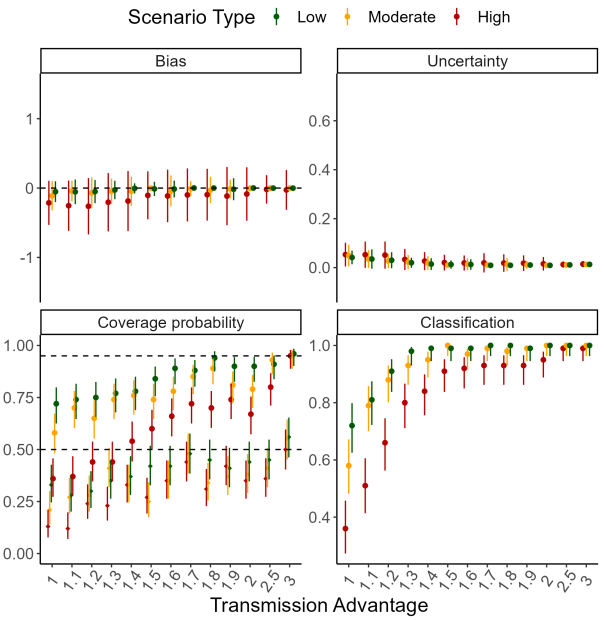
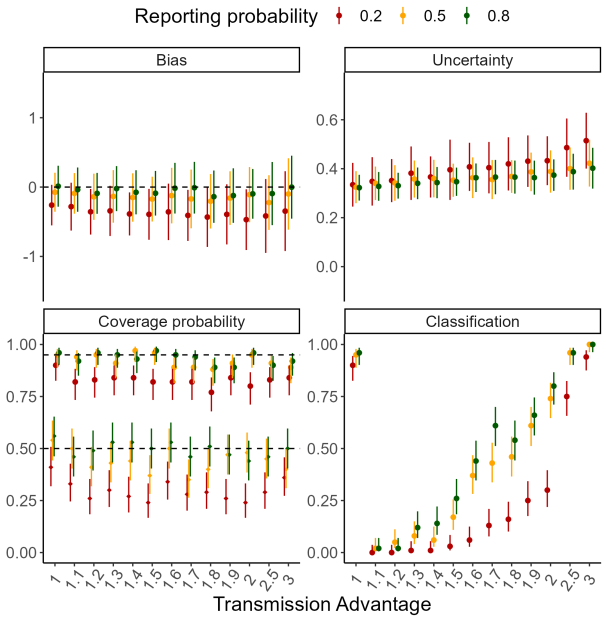


Figure S24: Method performance using incidence data simulated with superspreading (using 30 or 50 days of incidence data). We simulated data with low (overdispersion parameter $\kappa = 1$), moderate ($\kappa = 0.5$) and high ($\kappa = 0.1$) levels of superspreading. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

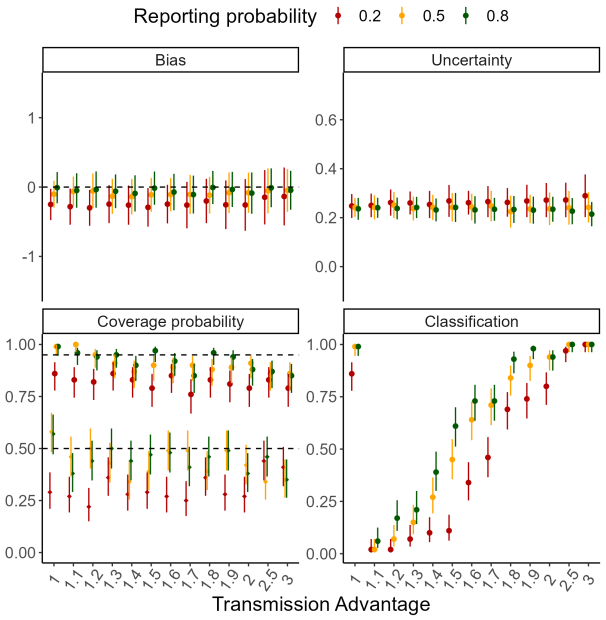
5.9 Sensitivity to under-reporting

To explore the sensitivity of our method to under-reporting, we first simulated data as in the baseline scenario (Sec. 5.3). We then assumed a constant reporting probability for both the reference and the variant and estimated the effective transmission advantage using only the reported cases. We set the reporting probability to 0.2, 0.5, or 0.8. This section presents the performance metrics when 10, 20, 30, and 50 days of incidence data were used for estimation.

(A) $R_t = 1.1$ and 10 days of data



(B) $R_t = 1.6$ and 10 days of data



(C) $R_t = 1.1$ and 20 days of data

(D) $R_t = 1.6$ and 20 days of data

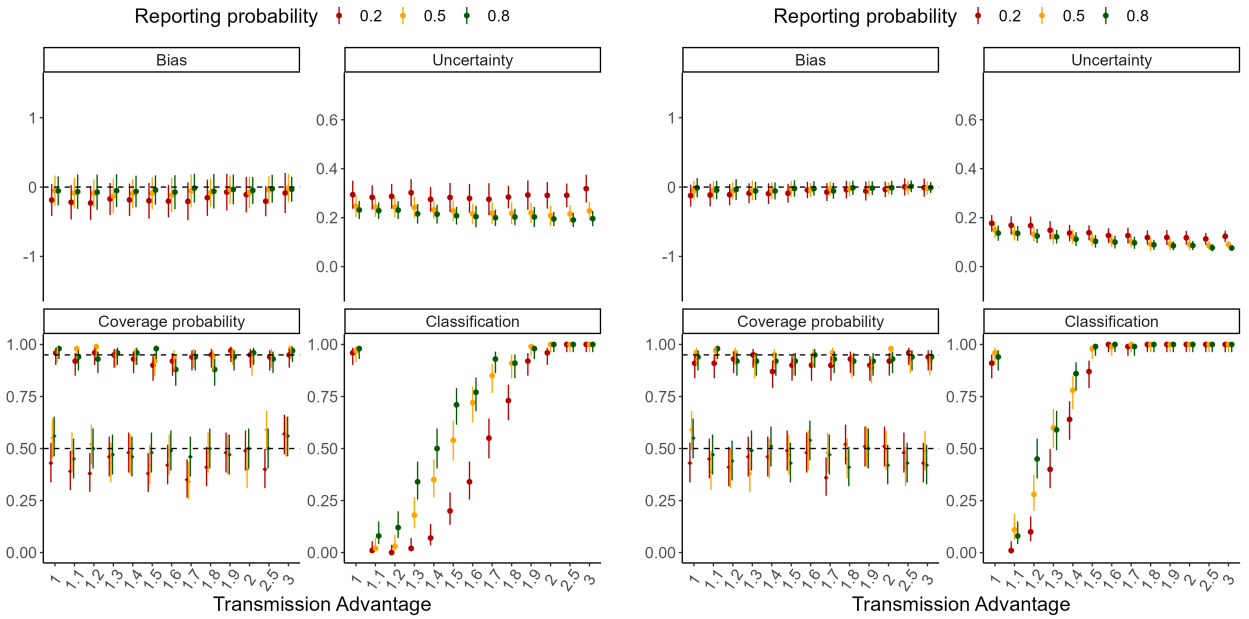
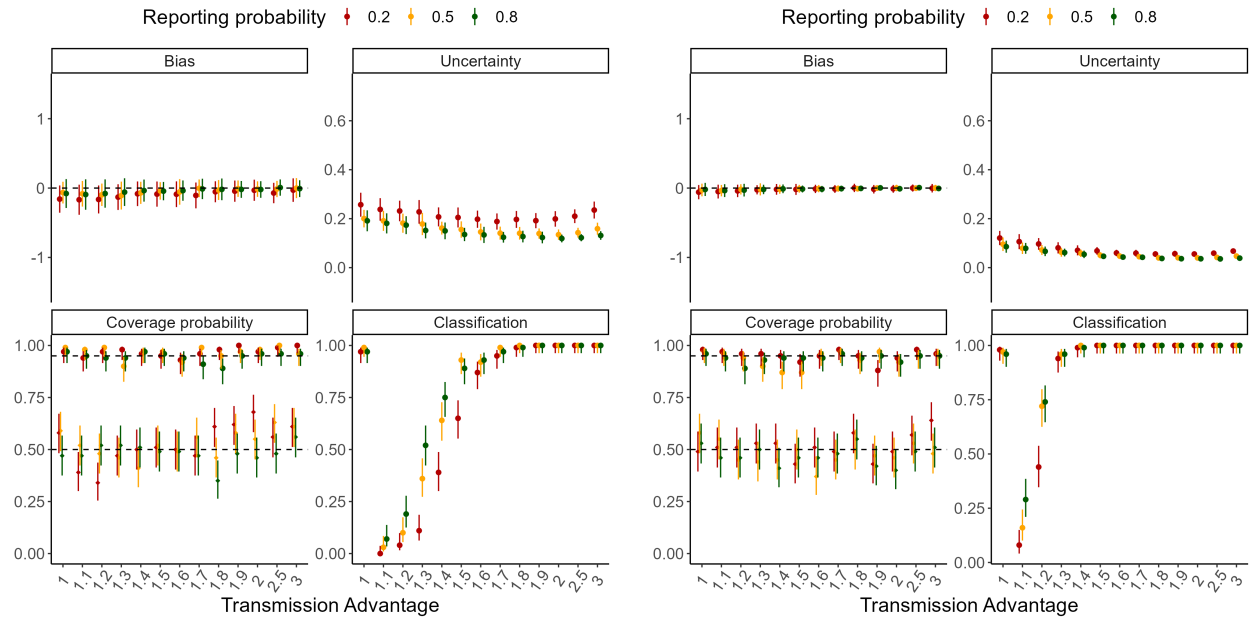


Figure S25: Method performance using using simulated incidence data with under-reporting (using 10 or 20 days of incidence data). We assume that the reporting probability is 0.2, 0.5, or 0.8. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

(A) $R_t = 1.1$ and 30 days of data

(B) $R_t = 1.6$ and 30 days of data



(C) $R_t = 1.1$ and 50 days of data

(D) $R_t = 1.6$ and 50 days of data

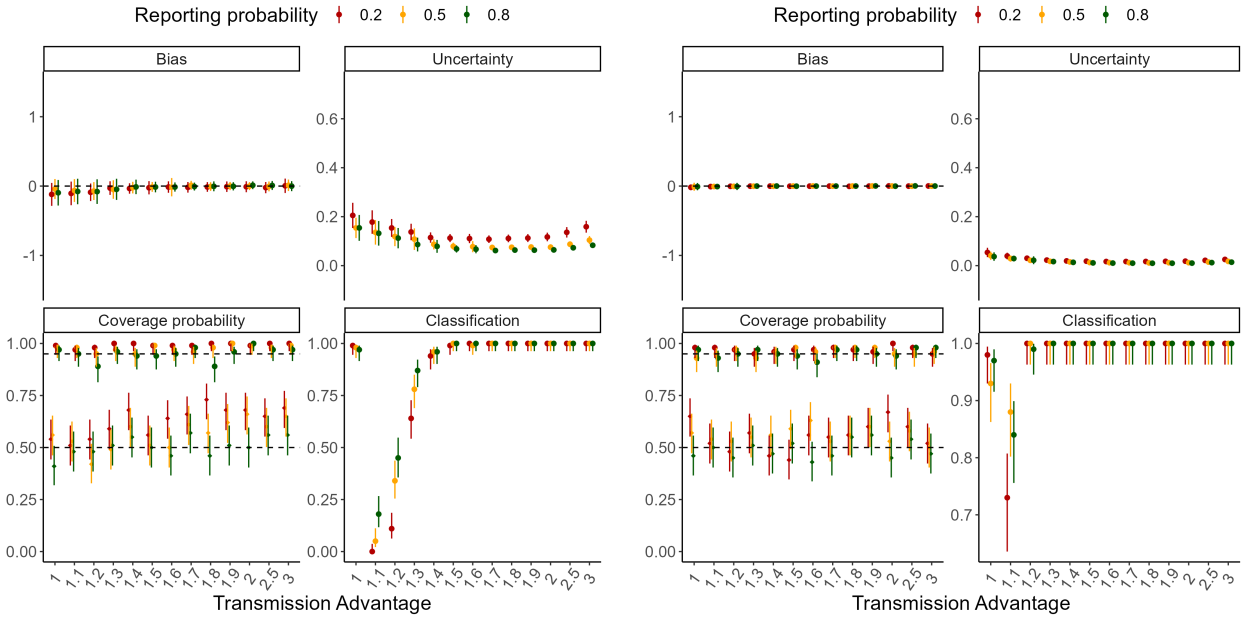


Figure S26: Method performance using using simulated incidence data with under-reporting (using 30 or 50 days of incidence data). We assume that the reporting probability is 0.2, 0.5, or 0.8. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

Scenario	Unclear	Variant less transmissible	Variant more transmissible	Simulations
Baseline	0.96	0.03	0.02	1000
Different SI Mean	0.93	0.05	0.02	3000
Misspecified SI Mean	0.66	0.15	0.20	3000
Different SI CV	0.94	0.04	0.02	3000
Misspecified SI CV	0.94	0.03	0.03	3000
Superspreading	0.64	0.26	0.10	3000
Underreporting	0.96	0.03	0.01	3000

Table S6: Classification of the new variant (as unclear, 'less' or 'more' transmissible) when the true transmission advantage is 1. Note that in this case, 'unclear' is considered as the correct classification (see Sec. 5.1 for more details). For each scenario, we show the proportion of simulations with each classification in 100 simulations using for 10, 20, 30, 40 or 50 days of data each and across all combinations of relevant for the scenario type (reference R_t 1.1 or 1.6, ϵ varying from 1 to 3, and other relevant parameter). See Tab S5 for a full list of parameters for each scenario.

5.10 Time-varying R_t

To explore the effect of changing transmission dynamics on the estimates, we simulated data as in the baseline scenario (Sec. 5.3) but with the reference R_t changing after 30 days from either 1.4 to 1.1, or 1.6 to 1.2. This section presents the performance metrics when 50 days of incidence data were used for estimation (i.e. covering the period both before and after the step-change in R_t).

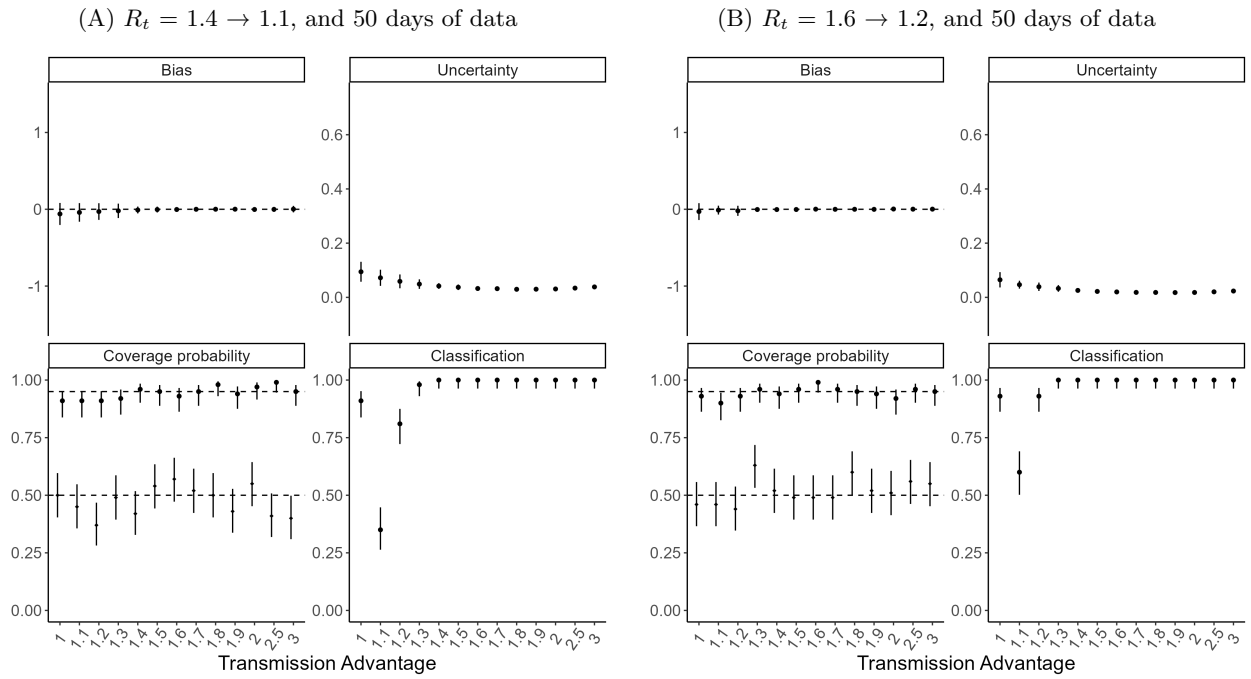


Figure S27: Method performance using incidence data simulated with time-varying R_t. The reference R_t changes after 30 days of the simulation. Here we use 50 days of incidence data for estimation. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

5.11 Two locations with time-varying R_t

We explored the performance of our method in the presence of changing transmission dynamics when data from two locations are used for estimation, instead of a single location. We simulated independent epidemics in two locations as in the baseline scenario (Sec. 5.3) but with the R_t profile changing over time. The reference R_t decreased from (i) 1.4 to 1.1, or (ii) 1.6 to 1.2, after 20 days in the first location and after 40 days in the second location. We also explored a further scenario where the decrease in reference R_t is different in the two locations and occurs at different times (from 1.4 to 1.1 after 40 days in the first location, and from 1.6 to 1.2 after 20 days in the second location). Tab S7 presents a summary of the R_t profiles used for simulations. This section

presents the performance metrics when 50 days of incidence data were used for estimation (i.e. covering the period both before and after the step-changes in R_t).

Location 1			Location 2		
Initial R_t	Final R_t	Time of R_t change	Initial R_t	Final R_t	Time of R_t change
1.4	1.1	20 days	1.4	1.1	40 days
1.6	1.2	20 days	1.6	1.2	40 days
1.4	1.1	40 days	1.6	1.2	20 days

Table S7: Reference R_t values used to simulate incidence data in scenarios with time-varying R_t profiles. The method performance results using incidence data generated by the R_t profiles in rows 1, 2 and 3 are shown in Fig S28A, Fig S28B and Fig S29 respectively.

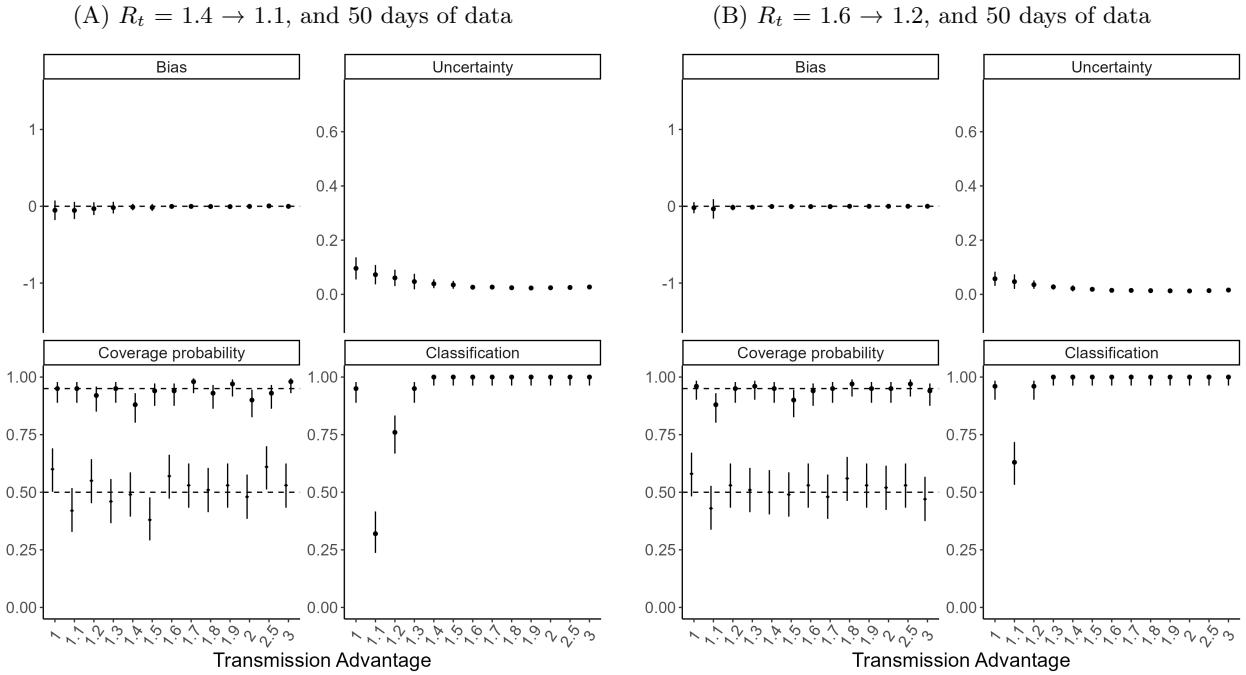


Figure S28: Method performance using incidence data simulated with time-varying R_t in two locations. In both locations the reference R_t decreases in the same way, but the change occurs at day 20 of the simulation for the first location location and day 40 for the second location. Here we use 50 days of incidence data for estimation. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

(A) Location 1: $R_t = 1.4 \to 1.1$, Location 2: $R_t = 1.6 \to 1.2$, and 50 days of data

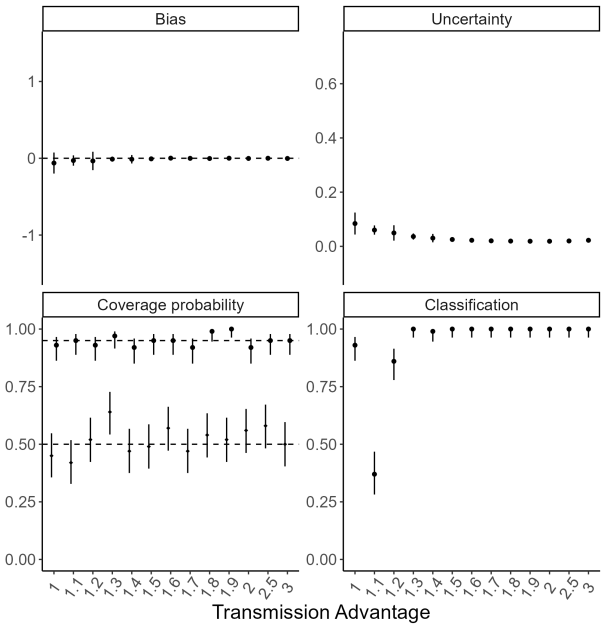


Figure S29: Method performance using incidence data simulated with time-varying R_t in two locations. The reference R_t decreases at day 20 in the first location and day 40 in the second location. Here we use 50 days of incidence data for estimation. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

5.12 Time-varying transmission advantage

To explore how well the transmission advantage could be estimated in a scenario where ϵ varied over time, we simulated data as in the baseline scenario (Sec. 5.3) but with a fixed reference R_t of 1.1. Incidence data were simulated with an initial transmission advantage of 1.1 which then increased linearly to 1.5 over a period of 30 days and then remained constant. This section presents the performance metrics over the duration of the simulated incidence. We estimated ϵ every 10 days using the latest 7 days of data, latest 10 days of data, or all available data.

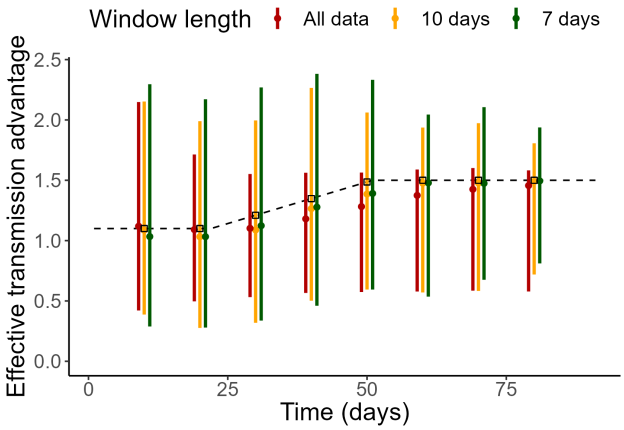


Figure S30: Effective transmission advantage estimated using MV-EpiEstim on simulated data in which the value of the transmission advantage changes over time. The 'true' value of the transmission advantage over time used to simulate the incidence data is shown by the hollow black squares and black dashed line. We show the median (circles) and 95% CrI (lines) of the posterior distribution for ϵ aggregated across all simulations. We estimated ϵ every 10 days into the simulation using the latest 7 days of data (green), latest 10 days of data (yellow), or all the data available up to that point (red). Using the entire time series to estimate ϵ corresponds to ignoring any temporal variability.

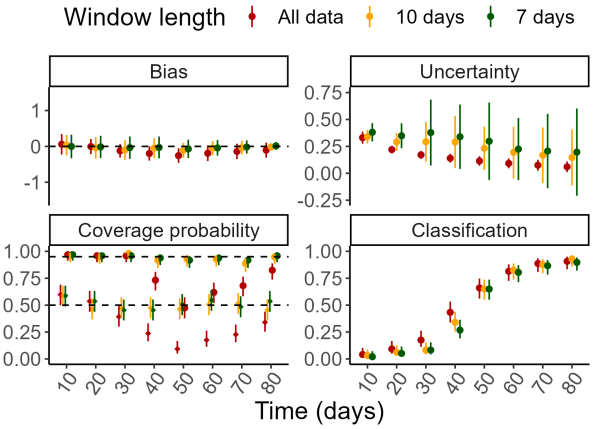


Figure S31: Method performance using incidence data simulated with time-varying ϵ in one location. The reference R_t remains fixed at 1.1 throughout, but the value of ϵ changes over time. It has a value of 1.1 until day 30 of the simulation, at which point it increases linearly to 1.5 over a period of 30 days, then remains constant at 1.5 for the remainder of the simulation. We estimated ϵ every 10 days into the simulation using the latest 7 days of data (green), latest 10 days of data (yellow), or all the data available up to that point (red). Using the entire time series to estimate ϵ corresponds to ignoring any temporal variability. In each panel, the dots and vertical bars represent the central estimate and uncertainty respectively. See Sec. 5.2 for details.

6 Literature review

In total, 66 studies were identified that aimed to compare the transmissibility of SARS-CoV-2 variants. Of those, 53 explicitly provided one or more estimates of the transmission advantage of one variant over the other (see Supplementary Database). 21 studies provided code, of which only 5 studies provided packaged software, which we list in Tab S8. None of these 5 studies assessed the method performance against simulated data.

105 transmission advantage estimates were found across the 53 studies. These were divided into six different advantage types, including the effective transmission advantage in: 1) the reproduction number R, 2) the growth rate, 3) the transmission rate, 4) the secondary attack rate, and two further categories which aimed to disentangle the intrinsic transmission advantage in R versus either: 5) immune escape, or 6) the generation time.

The majority of estimates were, as our method, based solely on incidence data (46/105, 44%), which included raw incidence of each variant or total incidence adjusted by the estimated proportion of each variant based on

Github repository	Number of studies
https://github.com/mrc-ide/reactidd	2
https://github.com/BDI-pathogens/VariantREstimate	1
https://mrc-ide.github.io/sircovid/	1
https://github.com/haschka/SIER_multivariant_epidemic/	1

Table S8: Github repositories for packaged code used in 5 of the studies identified in the literature review.

available sequencing data. For those estimates, the underlying models used were generally renewal equation-based models or exponential growth models.

Many estimates were based on using dynamic transmission model-based inference systems (27/105, 26%), such as compartmental models, which typically require additional data (e.g. on hospitalisations, deaths, population size, interventions) and assumptions (e.g. on disease progression and severity, immunity, etc). Phylodynamic models fitted to a combination of incidence and genomic data were used for a couple of estimates (2/105, 1.9%). A number of estimates (17/105, 16.2%) used household surveys to estimate secondary attack rates for each variant.

Overall, there was only 1 study with broad applicability (in that it only uses incidence data) that was also packaged in a ready-to-use tool [7]. We note that this study did explicitly account for overdispersion. Three other packages were identified through the review (Tab S8), but they required a wealth of additional data for fitting and none of those assessed the method performance against simulated data, including a range of transmission advantage scenarios, and a systematic exploration of the impact of mis-specifying the natural history of the new variant, or of the presence of overdispersion or underreporting on the performance of their method.

A number of other studies explored the trade-off between a change in the generation time and a change in transmission [8, 9], including one with a simulation study assessing the statistical framework performance [8].

See the Supplementary Database for all extracted estimates of transmission advantages and hyperlinks to the available code and R packages.

7 Code and Data availability

All data and code used in this analysis are available at `https://github.com/mrc-ide/epiestims`. MV-EpiEstim is available in the development version of EpiEstim at `https://github.com/mrc-ide/EpiEstim`.

References

[1] *Données de Laboratoires Pour Le Dépistage : Indicateurs Sur Les Variants (SI-DEP) - Data.Gouv.Fr.* URL: `https://www.data.gouv.fr/fr/datasets/donnees-de-laboratoires-pour-le-depistage-indicateurs-sur-les-variants/` (visited on 04/05/2023).

[2] B. Rai et al. "Estimates of Serial Interval for COVID-19: A Systematic Review and Meta-Analysis". In: *Clinical Epidemiology and Global Health* 9 (Jan. 2021), pp. 157–161. DOI: 10.1016/j.cegh.2020.08.007.

[3] 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.

[4] S. Funk et al. "Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area region of Sierra Leone, 2014-15". In: *PLoS Computational Biology* 15.2 (2019), e1006785.

[5] N. I. Bosse et al. "Evaluating Forecasts with scoringutils in R". In: *arXiv preprint arXiv:2205.07090* (2022).

[6] *Projections package - : Project Future Case Incidence*. 2021. URL: `https://cran.r-project.org/web/packages/projections/index.html` (visited on 04/05/2023).

[7] R. Hinch et al. "Estimating SARS-CoV-2 variant fitness and the impact of interventions in England using statistical and geo-spatial agent-based models". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 380.2233 (Aug. 2022). Publisher: Royal Society, p. 20210304. DOI: 10.1098/rsta.2021.0304. URL: `https://royalsocietypublishing.org/doi/10.1098/rsta.2021.0304` (visited on 04/05/2023).

[8] F. Blanquart et al. “Selection for infectivity profiles in slow and fast epidemics, and the rise of SARS-CoV-2 variants”. In: *eLife* 11 (May 2022). Ed. by B. S. Cooper and M. P. Davenport. Publisher: eLife Sciences Publications, Ltd, e75791. ISSN: 2050-084X. DOI: 10.7554/eLife.75791. URL: https://doi.org/10.7554/eLife.75791 (visited on 04/05/2023).

[9] K. Ito et al. “Estimating relative generation times and reproduction numbers of Omicron BA.1 and BA.2 with respect to Delta variant in Denmark”. en. In: *Mathematical Biosciences and Engineering* 19.9 (2022). Cc_license_type: cc_by Number: mbe-19-09-418 Primary_atype: Mathematical Biosciences and Engineering Subject_term: Research article Subject_term_id: Research article, pp. 9005–9017. ISSN: 1551-0018. DOI: 10.3934/mbe.2022418. URL: http://www.aimspress.com/rticle/doi/10.3934/mbe.2022418 (visited on 04/05/2023).