

University of Exeter  
EPSRC Centre for Predictive Modelling in Healthcare

# Addressing local and national priorities for health data science research during the COVID-19 pandemic


Robert Challen

Submitted by Robert Challen, to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Mathematics, January, 2022.

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

Signed:

A handwritten signature in black ink, appearing to read 'R Challen', with a long horizontal flourish extending to the right.

# Abstract

The research in this thesis was conducted in collaboration with Somerset NHS Foundation Trust, and the University of Exeter, during the COVID-19 pandemic. Prior to COVID-19 this research was directed at determining whether predictive models of chronic disease, developed from routinely collected clinical data, could be safely deployed into the clinical workflow. As part of this we published a novel safety framework for data driven clinical decision support.

When the COVID-19 pandemic emerged, the priorities of the NHS changed, shifting towards the acute care of unwell patients. Adjusting for this, we re-oriented the research to answer the many questions about the impact of COVID-19 on the NHS. Our study on the Alpha variant was one of the earliest to show increased severity. Our identification of outbreaks of the Delta variant, and its rapid growth, was at the forefront of decision making in local NHS trusts, NHS England, Public Health England, the UK Health Security Agency, and ultimately the UK Government.

As we move forward from the acute phase of the pandemic, understanding the longer term impact on chronic health management, and as a co-morbidity for chronic disease will become more of a priority for the NHS. A continued collaboration between the University of Exeter, Somerset NHS Foundation Trust and the UK Health Security Agency is well placed to address this need.

# Acknowledgements

I would like to sincerely thank my supervisor Professor Krasimira Tsaneva-Atanasova for her kindness and guidance through the numerous strands that made up this research, many of which could not be represented here. Also at the University of Exeter, I thank Martin Pitt who contributed a valuable perspective.

This thesis would not have happened without the support of Somerset NHS Foundation trust, and the Engineering and Physical Sciences Research Council who jointly funded it. During my time at Somerset NHS foundation trust, the encouragement and support of Luke Gompels, Tom Edwards, and Mark Dayer was very much appreciated, as was the support of the wider digital and research team.

In the response to the pandemic, I am eternally grateful for the guidance of Leon Danon, and the support of other members of the Joint Universities Pandemic Emergency Response (JUNIPER) consortium, Public Health England, and the Scientific Pandemic Influenza: Modelling subgroup (SPI-M).

Finally, I would like to thank my family for their patience and understanding during a challenging time.

# Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgements</b>	<b>ii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Published paper: Artificial intelligence, bias and clinical safety . . . .	4
1.2 Published paper: Factors influencing digital review of pathology test results in an inpatient setting: a cross-sectional study . . . . .	6
1.3 Literature review: Critical early COVID research . . . . .	8
1.4 Submitted paper: Algorithmic hospital catchment area estimation using label propagation . . . . .	12
1.5 Published paper: Meta-analysis of the SARS-CoV-2 serial interval and the impact of parameter uncertainty on the COVID-19 reproduction number . . . . .	15
1.6 Published paper: Estimates of regional infectivity of COVID-19 in the United Kingdom following imposition of social distancing measures	18
1.7 Published paper: Risk of mortality in patients infected with SARS- CoV-2 variant of concern 202012/1: matched cohort study . . . . .	23

1.8	Submitted paper: Early epidemiological signatures of novel SARS-CoV-2 variants: establishment of B.1.617.2 in England . . . . .	27
1.9	Methodology for Estimating the Reproduction Number and Growth Rate . . . . .	31
1.10	Conclusion . . . . .	33
<b>2</b>	<b>Artificial intelligence, bias and clinical safety</b>	<b>37</b>
<b>3</b>	<b>Factors influencing digital review of pathology test results in an inpatient setting: a cross-sectional study</b>	<b>45</b>
<b>4</b>	<b>Algorithmic hospital catchment area estimation using label propagation</b>	<b>56</b>
<b>5</b>	<b>Meta-analysis of the SARS-CoV-2 serial interval and the impact of parameter uncertainty on the COVID-19 reproduction number</b>	<b>80</b>
<b>6</b>	<b>Estimates of regional infectivity of COVID-19 in the United Kingdom following imposition of social distancing measures</b>	<b>115</b>
<b>7</b>	<b>Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study</b>	<b>124</b>
<b>8</b>	<b>Early epidemiological signatures of novel SARS-CoV-2 variants: establishment of B.1.617.2 in England</b>	<b>143</b>
<b>A</b>	<b>Reproduction number validation methodology</b>	<b>200</b>
A.1	Introduction . . . . .	200
A.2	Validation methodology . . . . .	203

A.3	Quantifying estimate delay . . . . .	206
A.4	Quantification of accuracy, bias and calibration . . . . .	206
A.5	Summary . . . . .	216
<b>B</b>	<b>Renewal equation reproduction number estimation and Jepidemic implementation validation</b>	<b>218</b>
B.1	Review of the renewal equation method for estimating the effective reproduction number. . . . .	219
B.2	Limitations . . . . .	225
B.3	Implementation considerations . . . . .	226
B.3.1	Prior selection . . . . .	227
B.3.2	Windowing strategy and posterior estimate selection . . . . .	228
B.3.3	Combining posterior estimates . . . . .	230
B.4	Validation and comparison . . . . .	232
B.4.1	Informed prior selection . . . . .	233
B.4.2	Posterior selection . . . . .	235
B.4.3	Combining posteriors . . . . .	238
B.5	Overall combination . . . . .	239
B.6	Summary . . . . .	242
<b>C</b>	<b>Incidence and growth rate estimation</b>	<b>244</b>
C.1	Introduction . . . . .	244

C.2	Estimating incidence and growth rate. . . . .	245
C.3	Estimating $R_t$ from the growth rate. . . . .	247
C.4	Implementation . . . . .	249
C.5	Growth rate validation . . . . .	250
C.6	$R_t$ validation . . . . .	254
C.7	Summary . . . . .	256
<b>D</b>	<b>Bayesian growth rate and reproduction number estimation</b>	<b>258</b>
D.1	Introduction . . . . .	258
D.2	Estimation of the incidence of infection . . . . .	258
D.3	Estimation of the growth rate . . . . .	260
D.4	Reproduction number estimates from Poisson rate . . . . .	263
D.5	Implementation considerations . . . . .	266
D.6	Growth rate validation . . . . .	268
D.7	$R_t$ validation . . . . .	270
D.8	Summary . . . . .	273

# List of Figures

- 1.1 An entity relationship diagram generated from the database of the laboratory test viewing application demonstrates the complexity of the typical electronic patient record system, with hundreds of tables supporting the internal functions of the application. A very small number of these tables contain clinically relevant data and extracting that data is a significant challenge. . . . . 8
  
- 1.2 Lower tier local authority (LTLA) distribution of case incidence per 1M capita per day (A) and growth rates (B) on March 22, 2021. A nationwide lockdown was introduced the next day. Panel C shows a phase diagram with the growth rate on the X axis and the incidence per 1M capita on the Y axis. Each trace shows the trajectory of a LTLA over the 14 days prior to and including the 22nd March 2021, shown as the point on the head of the path. The 6 highlighted areas show the regions which were experiencing the most new cases per 1M capita each day on the 22nd. Case numbers were under-representative at this time as testing capacity was limited, and testing was being restricted to symptomatic cases presenting to hospital. . . . . 14



1.3 The incidence rate ratio by Lower Tier Local Authority throughout the period of April to late July 2020 shows the evolution of regions with high incidence compared to the UK baseline from areas where high levels of importation and social mixing lead to the initial outbreak to areas where infections persisted, possibly due to societal factors as high levels of key or factory workers, dense housing, close social networks, or high proportions of large and multi-generational households. 22

1.4 A 2 weekly rolling average proportion of hospital admissions by age category in the period from 1st December 2020, to early October 2021. As vaccination rates in older age groups increased over the course of the programme, which was rolled out on an age stratified basis, the proportion of hospitalisations in older age groups dropped. Gradually as younger age groups became eligible for vaccination the benefit inferred to the older age groups was lessened. This coupled with waning immunity in the older age groups results in a reversion to pre-vaccination age distributions. . . . . 23

1.5 S-gene positivity by LTLA at 12 time points between late October 2020 and early January 2021. The data plotted are the 2 week rolling averages of the lower limit of the 95% binomial confidence interval, and represents the minimum proportion estimated at a given time point accounting for areas with small case numbers. . . . . 25

1.6	The reproduction number of S-gene positive and S-gene negative cases over the late autumn and winter 2020-2021. The earliest sign of an advantage for S-gene negative cases was seen in late October in the South East of England, where the growth of S-gene negative cases outpaced the others, leading to a separation of the S-gene negative and S-gene positive reproduction number curves. This pattern is repeated at different times in different regions as B.1.1.7 spread across the country, and once established, the advantage persists as the reproduction number varies. . . . .	26
1.7	The growth rate of reported COVID cases versus incidence per 1 Million population over time in Brazil, India, South Africa, the UK, and the US between February to July 2021. Data are from John Hopkins University Coronavirus tracker. This phase diagram shows case incidence per capita on the y-axis, and growth on the x-axis, and the evolution of growth and incidence over the previous 28 days in each country is shown as a path, with the head on the date of each panel. In a new wave of infection increasing case numbers and growth rates are seen as the path moving up and to the right. As the wave peaks the path trajectory moves anti-clockwise, before crossing the y-axis where growth is zero at the peak. Subsequent resolution of a wave of infection is seen as the trajectory descending on the left hand side of the figure. . . . .	28
1.8	Epidemic curves of 5 limited geographic areas which were identified as having clusters of high numbers of S-gene positive cases and some Delta variant cases (B.1.617.2) identified through delayed genomic sequencing. . . . .	29

A.1	Panel A shows a case count for the Flu2009 outbreak data set, Panel B shows the discrete infectivity profile, and panel C shows the $R_t$ time series estimates for a range of different estimation windows. Vertical error bars show the confidence in $R_t$ whereas horizontal error bars show the date range for which the assumption of constant $R_t$ for each estimate is applied . . . . .	203
A.2	Panel A shows a case count, for a synthetic data sets generated from a spline function for growth rate, with low (100) initial case counts, and a 10% weekly variability. Expected incidence curves are shown in red and black points represent one set of data samples (out of 10 generated). Panel B shows the estimated reproduction number using the renewal equation method with a gamma distributed infectivity profile (mean 5 days, sd 4), and a fixed window of 14 days. The red line shows the theoretical $R_t$ value. . . . .	205
A.3	Analysis of $R_t$ estimate time delays. In panel A as estimate of $R_t$ based on a periodic growth rate clearly shows estimate lag in all methods of estimating $R_t$ . In panel B the root mean squared error for a range of values of lag are calculated. The minimum RMSE is depends on the estimation model employed, however there is not an obvious linear relationship. . . . .	207
A.4	Estimate quality metric summaries for multiple estimation methods. In this instance we compare the performance of the renewal equation method with 14, 7, and 4 days as the windowing. Panel A show summary statistics for the bias of $R_t$ estimates, panel B shows the calibration. Panel C shows the quantile density, panel D the CRPS and panel E the critical threshold calibration. . . . .	211

A.5	Estimate quality metric intermediate detail and model comparison. Panel A-D show summary statistics for the bias of $R_t$ estimates broken down by the simulation smoothness, weekly variability, initial incidence, and time series boundary status. Panels E-F shows the calibration for the same subdivisions. Panels I-L shows the quantile density, panel M-P the CRPS, and panel Q-T the critical thresholds for the same subdivisions. . . . .	212
A.6	Time series of estimate quality metrics for simulations with no weekly variability show which parts of the time-series pose the greatest problem for the estimation method. Panel A shows the modelled case count. Panel B shows estimate of the reproduction number. Panel C shows the bias of each individual $R_t$ estimate. Panel D shows the calibration of individual estimates. Panel E shows the quantile density over time as 10%, 30% 50% 70% and 90% quantiles (rolling 28 day window). In panel F the instantaneous CRPS for individual estimates. In Panel G the critical threshold measure over time. . . . .	215
B.1	A graphical representation of the information involved in estimates of $R_t$ using the renewal equation method with different length windows (y-axis). The highlighted estimates (red dots) all use estimation windows that span the 10th day and during each of these windows there is an assumption of constant $R_t$ . In situations where the true $R_t$ is dynamic, combining these estimates may reflect the overall uncertainty. . . . .	230
B.2	Qualitative estimates of $R_t$ (black) against simulated (red) comparing 3 methods that vary in their prior selection process . . . . .	234
B.3	Time delay analysis of $R_t$ (black) against simulated (red) comparing 3 methods that vary in their prior selection process . . . . .	234
B.4	Quantitative analysis of $R_t$ (black) against simulated (red) comparing 3 methods that vary in their prior selection process . . . . .	235

B.5	Qualitative estimates of $R_t$ (black) against simulated (red) comparing 3 methods that vary in their posterior selection process . . . . .	236
B.6	Time delay analysis of $R_t$ (black) against simulated (red) comparing 3 methods that vary in their posterior selection process . . . . .	237
B.7	Quantitative analysis of $R_t$ (black) against simulated (red) comparing 3 methods that vary in their posterior selection process . . . . .	238
B.8	Qualitative estimates of $R_t$ (black) against simulated (red) comparing 3 methods that vary in their posterior combination process . . . . .	239
B.9	Qualitative estimates of $R_t$ (black) against simulated (red) comparing the reference implementation to a proposed improved method . . . . .	240
B.10	Time delay analysis of $R_t$ (black) against simulated (red) comparing the reference implementation to a proposed improved method . . . . .	241
B.11	Quantitative analysis of $R_t$ (black) against simulated (red) comparing the reference implementation to a proposed improved method . . . . .	241
B.12	Detailed breakdown of quantitative analysis of $R_t$ (black) against simulated (red) comparing the reference implementation to a proposed improved method . . . . .	242
C.1	Qualitative estimates of growth rate (black) against simulated (red) comparing 4 configurations of the locally fitted polynomial method that vary in the parameterisation of the polynomial degree and the bandwidth . . . . .	251
C.2	Time delay analysis of growth rate (black) against simulated (red) comparing 4 configurations of the locally fitted polynomial method that vary in the parameterisation of the polynomial degree and the bandwidth . . . . .	251

C.3	Estimate quality metric summaries for multiple estimation methods. In this instance we compare the performance of the growth rate estimators that vary in the parameterisation of the polynomial degree and the bandwidth. Panel A show summary statistics for the bias of $R_t$ estimates, panel B shows the calibration. Panel C shows the quantile density, panel D the CRPS and panel E the critical threshold calibration. . . . .	252
C.4	Estimate quality metric intermediate detail and model comparison. Panel A-D show summary statistics for the bias of growth rate estimates broken down by the simulation smoothness, weekly variability, initial incidence, and time series boundary status. Panels E-F shows the calibration for the same subdivisions. Panels I-L shows the quantile density, panel M-P the CRPS, and panel Q-T the critical thresholds for the same subdivisions. . . . .	253
C.5	Qualitative estimates of $R_t$ (black) against simulated (red) comparing 2 configurations of the growth rate based $R_t$ estimators using methods presented here and varying by time window (14 or 21 days) with the best performing renewal equation method from Appendix B (labelled “Improved renewal eqn”). . . . .	254
C.6	Time delay analysis of $R_t$ (black) against simulated (red) comparing 2 configurations of the growth rate based $R_t$ estimators using methods presented here and varying by time window (14 or 21 days) with the best performing renewal equation method from Appendix B (labelled “Improved renewal eqn”). . . . .	255

C.7	Estimate quality metric summaries for multiple estimation methods. In this instance we compare the performance of the growth rate estimators that vary in the parameterisation of the polynomial degree and the bandwidth. Panel A show summary statistics for the bias of $R_t$ estimates, panel B shows the calibration. Panel C shows the quantile density, panel D the CRPS and panel E the critical threshold calibration. . . . .	256
D.1	A test of an approximation of a sum of Gamma distributed quantities (panel A) as a single Gamma distribution based on matching moments (panel B). In this particular example the sum of Gamma distributions is well approximated by another single Gamma distribution. . . . .	265
D.2	Qualitative estimates of growth rate (black) against simulated (red) comparing the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C . . . . .	268
D.3	Time delay analysis of growth rate (black) against simulated (red) comparing the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C . . . . .	269
D.4	Estimate quality metric summaries for multiple estimation methods. In this instance we compare the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C. Panel A show summary statistics for the bias of $R_t$ estimates, panel B shows the calibration. Panel C shows the quantile density, panel D the CRPS and panel E the critical threshold calibration.	270
D.5	Qualitative estimates of $R_t$ (black) against simulated (red) comparing the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C and the renewal equation method from Appendix B . . . . .	271

D.6	Time delay analysis of $R_t$ (black) against simulated (red) comparing the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C and the renewal equation method from Appendix B . . . . .	271
D.7	Estimate quality metric summaries for multiple estimation methods. In this instance we compare the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C and the renewal equation method from Appendix B. Panel A show summary statistics for the bias of $R_t$ estimates, panel B shows the calibration. Panel C shows the quantile density, panel D the CRPS and panel E the critical threshold calibration. . . . .	272



# List of Tables

A.1	Estimate quality metric summaries for the three estimation methods compared here. . . . .	212
B.1	Summary of quantitative analysis of $R_t$ (black) against simulated (red) comparing 3 methods that vary in their prior selection process . . . .	235
B.2	Summary of quantitative analysis of $R_t$ (black) against simulated (red) comparing 3 methods that vary in their posterior selection process . .	238
B.3	Summary of quantitative analysis of $R_t$ (black) against simulated (red) comparing 3 methods that vary in their posterior combination process	239
B.4	Summary of quantitative analysis of $R_t$ (black) against simulated (red) comparing the reference implementation to a proposed improved method	242
C.1	Summary of quantitative analysis of growth rate (black) against simulated (red) comparing 4 configurations of the locally fitted polynomial method that vary in the parameterisation of the polynomial degree and the bandwidth . . . . .	252
C.2	Summary of quantitative analysis of $R_t$ (black) against simulated (red) comparing 3 methods that vary in their prior selection process . . . .	255

D.1	Summary of quantitative analysis of $R_t$ (black) against simulated (red) comparing the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C . . . . .	269
D.2	Summary of quantitative analysis of $R_t$ (black) against simulated (red) comparing the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C and the renewal equation method from Appendix B . . . . .	272

# 1. Introduction

This thesis summarises a body of research conducted between 2017 and 2021. During this time the world experienced an unprecedented change resulting from the emergence and rapid spread of SARS-CoV-2 and the subsequent COVID-19 pandemic. During the pandemic the priorities of healthcare organisations around the globe have been necessarily redirected to the management of acute respiratory illness, fundamentally changing the management of patients suffering from all diseases. Local health service capacity issues, local infection control measures and national guidance on the reduction of physical interactions meant care pathways needed to be redesigned and adoption of telemedicine accelerated. This change in service provision was coupled with a seismic shift in health care seeking behaviour, due to fear of COVID-19 and prolonged periods of social isolation. This has delayed the presentation of chronic diseases such as cancers, and many non-essential operations. It will take years for the impact of the pandemic on the health of the nation to be fully felt and longer still for it to recover. To compound this challenging situation, the pandemic continues to evolve, as novel variants of SARS-CoV-2 emerge, vaccination programmes are rolled out, and the longer term health impacts of both COVID-19 disease and vaccination become apparent.

The role of health data science has also changed over the course of this research, and can be expected to continue to change into the future. This in part reflects the shift in the priorities of the health service, and change in the burden of disease. The pandemic has also driven fundamental changes in demand for clinical testing, the

priorities for data sharing for analysis, the demands on information technology to conduct analyses, and the dissemination of those analyses. These changes, although vastly accelerated, are taking time to be realised. As we move into the future, we can expect this to continue to evolve.

The purpose of this extended introduction is to place this body of health data science research, conducted over a challenging period, within the context of the questions that were important at the time, framed by the emerging evidence and the sometimes limited, and potentially biased, data available. Over this time some of the analyses were conducted rapidly to inform immediate local and national policy decisions based on the best information available. Not all of these have been mature enough to contribute to the scientific literature, and some of these findings have dated already.

The overall aim of the research that is described here was to use the best available data to identify timely evidence to guide local and national decision making, and provide interpretation of the evidence in as robust a way as possible. At the same time accounting for the uncertainty resulting from the analysis and understanding limitations and biases present in the data.

The specific objectives changed over the course of the research with changing local and national priorities. Prior to the COVID-19 pandemic we focused on the local needs of our industrial partner, Somerset NHS Foundation Trust, whose priority was in the safe implementation of machine learning based decision support systems in association with Google DeepMind, for which a framework for the evaluation of machine learning based decision support systems was needed. Associated with this was the planned implementation of a mobile device based laboratory test review platform, and evidence was needed to guide the implementation of clinical workflow changes. Shortly afterwards the SARS-CoV-2 pandemic arrived and both local and national priorities changed radically. The pressing question at the outset was around how best to plan local and national clinical capacity to meet the needs of the epidemic,

and to this end we developed evidence on how capacity and demand can be balanced. The demand for hospital beds was driven by the speed of the outbreak, and it became a national priority to provide evidence that the overall growth of the epidemic was in control, to detect regional variations, and to understand the age structure of infections, particularly with regards to school reopening. Over the winter of 2020 the emergence of the Alpha variant of SARS-CoV-2 in Kent changed the epidemic dynamics and raised the need for new evidence on the severity of this emerging variant, an issue of national and international importance. Finally in the Spring of 2021, as case numbers in India rose rapidly and genomic sequencing identified the Delta variant, the national priority once again changed to rapid detection and monitoring, with evidence needed for signs of the Delta variant becoming established in the UK and an understanding of how it would spread in England given the degree of vaccination and immunity in the population at that stage.

With these evolving objectives our research findings have contributed in diverse ways. Prior to the pandemic the principal achievement was the development of an AI safety framework that has been widely adopted and influenced policy documents on the implementation of machine learning systems in healthcare worldwide. Our findings that the speed of laboratory test result review depends on the clinical workflow and factors such as the time of day helped Somerset NHS trust plan their deployment of the mobile device based pathology test viewer, but this was placed on hold due to the pandemic. Our work on hospital demand and catchment areas provided novel methods for linking COVID demand to hospital capacity and allowed us to develop operational analytic models that link localised predictions in infections to expected intensive care capacity for local NHS trusts. In tracking the overall progress, regional variation and age structure of the epidemic we identified numerous transient patterns, which fed into the national Scientific Pandemic Influenza Modelling subgroup and henceforth to the government, including the weekly estimate of the SARS-CoV-2 effective reproduction number. Our analysis on the severity of the Alpha variant determined that there was an elevated risk of death with infections caused by the

Alpha variant and by association hospital demand, which influenced national and international control measures and hospital capacity planning. Finally our operational monitoring for the emergence of the Delta variant, provided critical early warning to the UK Government of the outbreak in Bolton at a time when they were considering relaxing social distancing measures, this shaped local public health responses to the outbreak in Bolton, and influenced the Government to delay relaxing restrictions until further vaccination was completed.

The remainder of this introduction focuses on each of the published and submitted research articles in turn. Within each section we aim to describe in more detail the context, methods used and developed, work done, limitations and impact for each publication. Within each section we will reference our contributions to the scientific literature, both published and under review, and which are included within the body of this thesis. As the main body of this thesis is related to the epidemic, a brief literature review of key foundational work relevant to the SARS-CoV-2 outbreak and characterisation of early epidemiological parameters is also included.

## **1.1 Published paper: Artificial intelligence, bias and clinical safety**

Somerset NHS Foundation trust (SFT) provides care to a population of over 340,000. It also provides some specialist services for the whole of Somerset, making the catchment population around 544,000. The hospital has over 700 beds, 30 wards, 15 operating theatres, a fully equipped diagnostic imaging department and a purpose-built cancer treatment centre (Somerset NHS Foundation Trust, 2022). SFT is an internationally recognised leader in the use of digital technology, and is part of the NHS Global Digital Exemplar (GDE) programme (NHS England, 2018).

In 2017 SFT forged a partnership with Google DeepMind (Deepmind, 2017; Powles and Hodson, 2017) to develop the opportunity of using observational clinical data to develop clinical decision support algorithms to detect disease, stratify risk or prioritise treatment, with a particular focus on acute kidney injury (AKI).

Developing predictive models from clinical data promises to identify chronic disease early within the stable routine of healthcare provision. This allows prevention or early intervention to improve long term outcomes, with the potential to reduce the burden of chronic disease. For these predictions to be accurate the clinical context must be as similar as possible to that in which the model was developed.

Routinely collected observational health data are limited in quality compared to prospectively collected research data. Data standards for collection are either missing or variably adhered to over time, or over clinical setting, and are influenced by the nature of the user interfaces surrounding documentation (Mann and Williams, 2003). Operational clinical data recording is often an interpretation and may not always document the factual observation of the clinical picture, and may occur after a clinical encounter (Carpenter et al., 2007).

There are clinical safety risks in using observational clinical data to develop predictive models, which arise from variable biases in clinical data. We aimed to support the relationship between SFT and DeepMind, by developing a conceptual framework for evaluating the clinical safety and risk of bias of such algorithms based on an analysis of research describing existing data driven predictive models, and summarising previously identified risks of AI systems. In our papers “Artificial intelligence, bias and clinical safety” (Challen et al., 2019) and “Emerging safety issues in Artificial Intelligence” (Challen, 2019) we address potential risks during both technical development and in deployment into clinical practice. Risks are determined by the nature of the data available, the complexity of the task that the algorithm is addressing, the frequency of update of the algorithm, and the degree of human

supervision involved.

Although the safety of clinical support systems is described, the safety and ethics of machine learning in medicine was a relatively new field of study, reflecting the fact that research on the application of data driven predictive models in health care are in early stages of development. “Artificial intelligence, bias and clinical safety” (Challen et al., 2019) was published in the BMJ Quality and Safety in 2019 and as of September 2021 had been cited more than 230 times. It is referred to by Nature digital medicine as part of their best practices for authors of healthcare-related artificial intelligence manuscripts (Kakarmath et al., 2020). It is cited by policy papers on AI in healthcare from both the UK Parliament (Smeaton and Christie, 2021), and the World Health Organisation (Habli, Lawton and Porter, 2020).

## **1.2 Published paper: Factors influencing digital review of pathology test results in an inpatient setting: a cross-sectional study**

As part of the GDE programme, SFT has implemented a range of new electronic patient record systems, and enhanced existing ones, such as electronic whiteboards, mobile device based pathology results viewers, and electronic prescribing. SFT’s approach to the GDE has been the adoption of open digital platforms that enable the storage and management of clinical and system audit data, on site within SFT, providing a rich and detailed data set covering both the clinical state of patients, and also the digital clinical workflow.

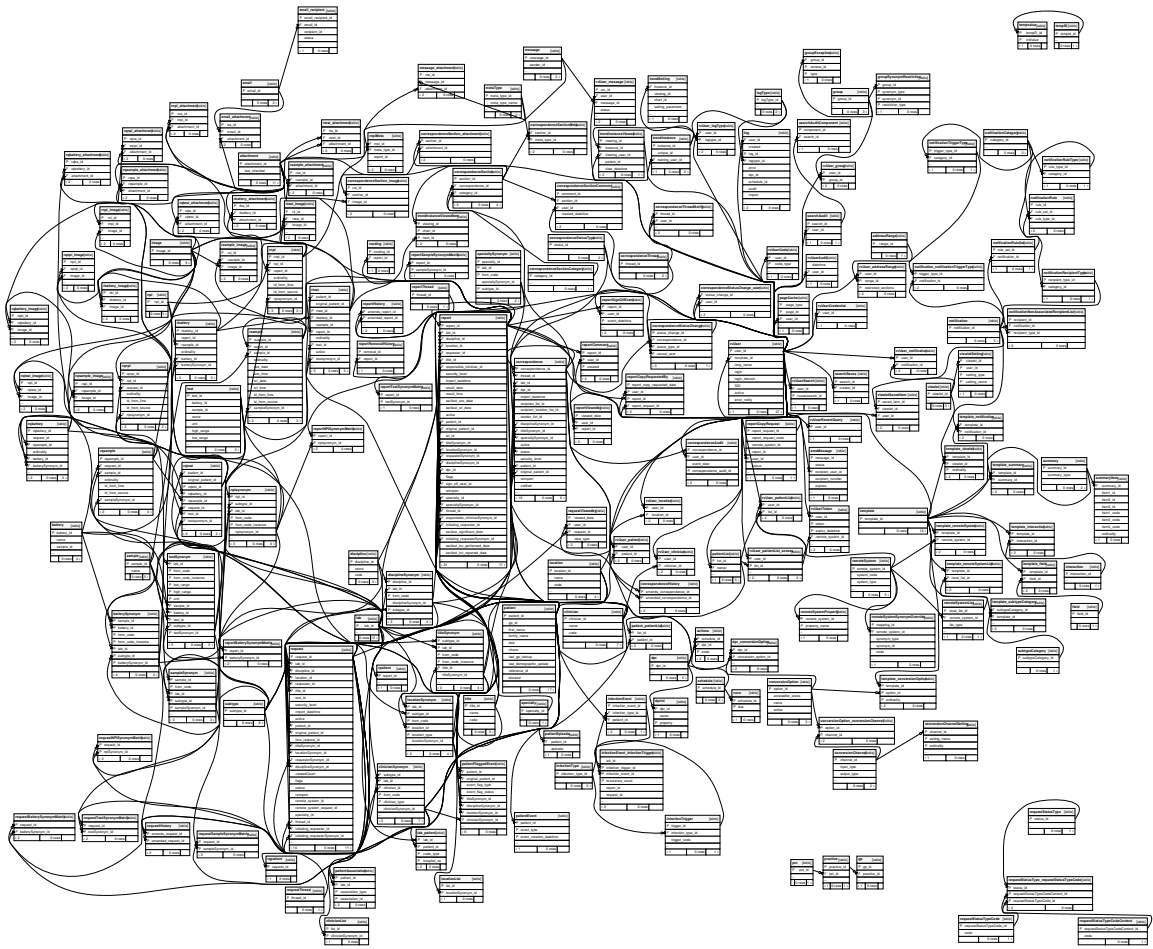
With the partnership with Google DeepMind there were plans to implement a mobile device based laboratory test results viewer. This was planned to be the platform on which to run clinical decision support of the kind described above, but also aimed



to improve clinical workflow around laboratory test results review. We investigated the workflow surrounding the clinical review of laboratory test results, with the aim of identifying a general baseline for measuring change, and discovering specific issues with workflow that the implementation should directly address. We extracted and de-identified data from the backend database supporting the existing desktop PC based clinical test viewer (see Figure 1.1). These data related to the timing of test publication and the timing of the first clinician review for over 1.7 million tests. We then performed time series analyses, identified correlations and performed accelerated failure time analysis on the test review time to determine factors that significantly influenced the timeliness of test review. The approach of using large scale data analysis to investigate workflow issues in this field was novel as previous investigation had largely relied on user surveys or small scale direct observation of workflow.

The details of the laboratory test review were published in the open access version of the Journal of the American Medical Association (JAMIA Open) under the title “Factors influencing digital review of pathology test results in an inpatient setting: a cross-sectional study” (Challen et al., 2020). Through the detailed analysis of the workflow surrounding review we were able to make a set of recommendations about the design of the new platform, and identified other opportunities to change laboratory test ordering workflow to prevent ineffective testing. However, due to a change in the commercial relationship between SFT and Google DeepMind the findings of this work were not implemented as planned.

The analysis of the periodicity of time delays in this analysis, highlights the potential for workflow factors to influence the recording of data, and hence the interpretation of that data. The influence of weekly social rhythms on workflow and the potential for this to impact on patient outcomes is a contextual bias that could influence data driven predictive models. Workflow factors are sensitive to organisational cultural and policy changes and may change suddenly in the face of external challenges, resulting



*Figure 1.1: An entity relationship diagram generated from the database of the laboratory test viewing application demonstrates the complexity of the typical electronic patient record system, with hundreds of tables supporting the internal functions of the application. A very small number of these tables contain clinically relevant data and extracting that data is a significant challenge.*

in apparent step changes in the data collection process, and hence distributional shift.

### **1.3 Literature review: Critical early COVID research**

In late 2019 an outbreak of a novel infectious disease was detected in Wuhan, China. It manifested principally with severe acute respiratory distress, and pneumonia (Huang et al., 2020), although many cases followed a mild course (Wu et al., 2020),

and some infected individuals were completely asymptomatic but able to pass the disease on (Bai et al., 2020). The pathogen was rapidly identified as a new species of coronavirus (severe acute respiratory syndrome coronavirus 2 - SARS-CoV-2), and the disease named COVID-19 (Sohrabi et al., 2020).

Global transmission of the virus followed and outbreaks were observed in Europe, beginning with a small cluster in France (Lescure et al., 2020) on the 24th January 2020. On the 31st Jan 2020 the first cases were identified in the UK (Moss et al., 2020). On the 5th February a large outbreak was confirmed on the Diamond Princess cruise ship, in Yokohama, Japan (Mizumoto et al., 2020). The largest early outbreak in Europe began with cases identified in the Lombardy region of Italy (Cereda et al., 2021) on the 20th February 2020, where hospitals quickly came under significant pressure (Remuzzi and Remuzzi, 2020). During the February half term, between the 17th and 21st February, travel to the ski resorts in the north of Italy resulted in significant importations to the UK. This established community transmission in the UK and following this a period of rapid growth of cases in the UK ensued. A step-wise implementation of social distancing measures were mandated by the government including voluntary self isolation of any symptoms & vulnerable people (UK Government, 2020b) from the 16th March 2020, then a ban on non essential travel worldwide and school closures. Finally on 23rd March 2020 the government mandated that everyone apart from essential workers should stay at home and away from others (UK Government, 2020c).

In the weeks leading up to 23rd March the priority was to predict the impact of COVID-19 on the UK. To do this a key set of parameters that describe the behaviour of the disease was needed, leading to an intense focus on the early research from the initial outbreak in Wuhan. These parameters were the basic reproduction number ( $R_0$ ) which describes the number of secondary cases that a primary case will infect in an unconstrained epidemic; the generation interval, which is the average time between a primary and secondary infection. A long generation time coupled with

a lower basic reproduction number might be expected to result in a an epidemic that grows slowly, whereas a higher reproduction number with lower generation time will leading to rapid growth. Other important measures that influence the demand on healthcare services are the severity of infections, in terms of the infection hospitalisation rate (IHR), the infection fatality rate (IFR), and the timing of demand in terms of the incubation period, the time delay from infection to hospitalisation, and from infection to death.

An assessment of the severity of this novel disease was also urgently required. In late January case reports of the first 41 patients hospitalised in Wuhan were published. They had a median age 49, and of these 32% cases were admitted to ICU, and 15% died. (Huang et al., 2020). By February a larger data set was published of 1099 patients of whom 6.1% had severe disease, defined as patients for whom intensive care was required during their admission or who died (Guan et al., 2020). By March enough studies had been published for a systematic review of three studies covering 3600 patients of whom 25.6% were counted as severe, and 3.6% died (Fu et al., 2020). These widely differing estimates were due in part to a lack of a consistent case definition of severe disease and also due to a widespread under ascertainment of disease, and in particular of mild cases. In the Diamond Princess cruise ship the opportunity for mass screening arose and in early February this showed that 18% of all cases were completely asymptomatic (Mizumoto et al., 2020), and a higher proportion had very mild disease. An analysis of the prevalence of disease in repatriated travellers from Wuhan, published in late March, coupled with 70117 case reports from China, showed that raw case fatality rates of 3.67%, were reduced to 0.66% when attempts were made to correct for under ascertainment, and that the rate of under ascertainment varied by age group (Verity et al., 2020).

The early estimates of the reproduction number were similarly hampered by the under ascertainment, but also by the uncertainty around the generation time. In late January early estimates of the basic reproduction number of 2.6 (1.5-3.5) based

on the original SARS generation time of 8.4 days were published in an Imperial college report (Imai et al., 2020) using data from Wuhan, by fitting a mathematical model to reported case numbers. Details of the model were not provided. Similar estimates of the reproduction number were made at this time, variously by using renewal equation methods (Li et al., 2020), a susceptible-exposed-infected-recovered (SEIR) meta-population model (Wu, Leung and Leung, 2020) and the methods of Wallinga and Lipsitch, 2007 involving direct measurement of the growth rate (Zhao et al., 2020). All of these were based on the same assumption that the generation time of SARS-CoV-2 was similar to that of SARS. In late January a re-analysis accommodating variable ascertainment rates and directly estimating the reproduction number and generation time (in the form of the infectious period and latent period) came to the conclusion that the reproduction number was somewhat higher (3.11 [2.39-4.13]) and generation time somewhat lower at 5.6 (4.35-7.23) days (Read et al., 2021). The potential for bias introduced by assumptions around the generation time were highlighted when in early March independent estimates based on data from Singapore suggested a generation interval of 5.2 days (Ganyani et al., 2020), using a Bayesian imputation of transmission events, or even as short as 4.0 days (Du et al., 2020), directly measured from contact tracing data.

The spontaneous cautious behaviour of people in Wuhan, who had recent experience of large scale infectious disease outbreaks, and rapidly complied with measures designed to control the outbreak, could give the appearance of an outbreak that is easy to control (Kucharski et al., 2020). When this effect was controlled for, a meta-analysis of estimates of the base reproduction number before social restrictions were put in place gave a summary estimate as high as 3.77 (Lin et al., 2020) however this meta-analysis estimate dropped to 1.88 in the subset of studies conducted following city lockdown. Similarly estimates of the reproduction number of the spread outside of Wuhan, in the contained environment of the Diamond Princess, using were very high, peaking at 12.1 for passengers and with a median value of 5.8 over the whole course of an epidemic within a closed population (Mizumoto and Chowell, 2020).

The lower estimates for the basic reproduction number potentially resulted from a degree of spontaneous reduced social mixing in Wuhan, and China more generally. As cases began to appear in Europe it became apparent that the longer doubling times seen in China (Volz et al., 2020), were not going to be replicated in Europe, with more rapid doubling up to 2-3 days (Pellis et al., 2021) and with estimates of the basic reproduction number slightly higher at 2.6-3.3 with an intermediate serial interval of 6.6 days (Cereda et al., 2021).

On this background of uncertainty various predictions had to be made on the trajectory of the outbreak in the UK. Using a spatiotemporal meta-population model in mid February Danon et al. estimated that an unconstrained outbreak could result in 1.2 million cases per day in the UK, 4-6 months into an outbreak (Danon et al., 2021) with a final epidemic size of 45 million infected people. Imperial produced a similar estimate in March (Ferguson et al., 2020), which additionally estimated the impact on hospitals and intensive care units, showing that they would quickly be overwhelmed in the absence of control measures. This was later supported by predictions from a stochastic age structured model which predicted an unmitigated final size of 23 million cases and examined the impact of shielding, school closures, self isolation and lockdown on the epidemic trajectory in the UK (Davies et al., 2020).

## **1.4 Submitted paper: Algorithmic hospital catchment area estimation using label propagation**

In the acute response to the pandemic, a key early priority was to increase the capacity of the NHS to provide intensive care and acute hospital beds. On 21st March 2020 the UK Government requisitioned private hospital beds (BBC News, 2020b), constructed field hospitals (BBC News, 2020a), and commissioned the

manufacture of emergency ventilators (UK Government, 2020a). Individual hospitals cancelled routine operations to free up operative ventilators, and reconfigured services to maximise capacity to deliver intensive care. Early modelling work suggested the potential for very large numbers of infections without significant control measures (Danon et al., 2021; Davies et al., 2020; Ferguson et al., 2020). The early geographic distribution of cases (see Figure 1.2) suggested that the first COVID wave would not hit all regions at the same time. Prior to the first lockdown areas such as the South West, North Wales and rural Scotland had initially lower case numbers than other parts of the country (Panel A), however there was widespread rapid growth observed across all areas (panel B). The trajectories of the smaller geographical areas (panel C) showed sustained growth, with increasing case numbers, but with some evidence of increasing regional variation in growth patterns seen as the spreading of the growth trajectories. The pressing need for local NHS trusts was a short term and localised prediction of the demand for intensive care and hospital beds, to develop mitigation strategies if needed (Lacasa et al., 2020).

We aimed to address this tactical need for a short term prediction of hospital demand. We combined a spatiotemporal prediction of infections (Danon, House and Keeling, 2009; Keeling et al., 2010) and the best available data on severity and delay distributions from the emerging pre-print grey literature, which is based on international experiences in China and Italy. To translate this to a prediction specific for individual NHS trusts required a model for the spatial relationship between the locations of predicted infections, and the intake of patients into individual hospitals.

Ideally such a model would be based on an analysis of existing data, however such data did not yet exist, as the sudden restructuring of care provision had caused a distributional shift, with alterations in bed capacity, rendering pre-existing data irrelevant.

We developed an algorithm for modelling the catchment areas of hospitals, given their

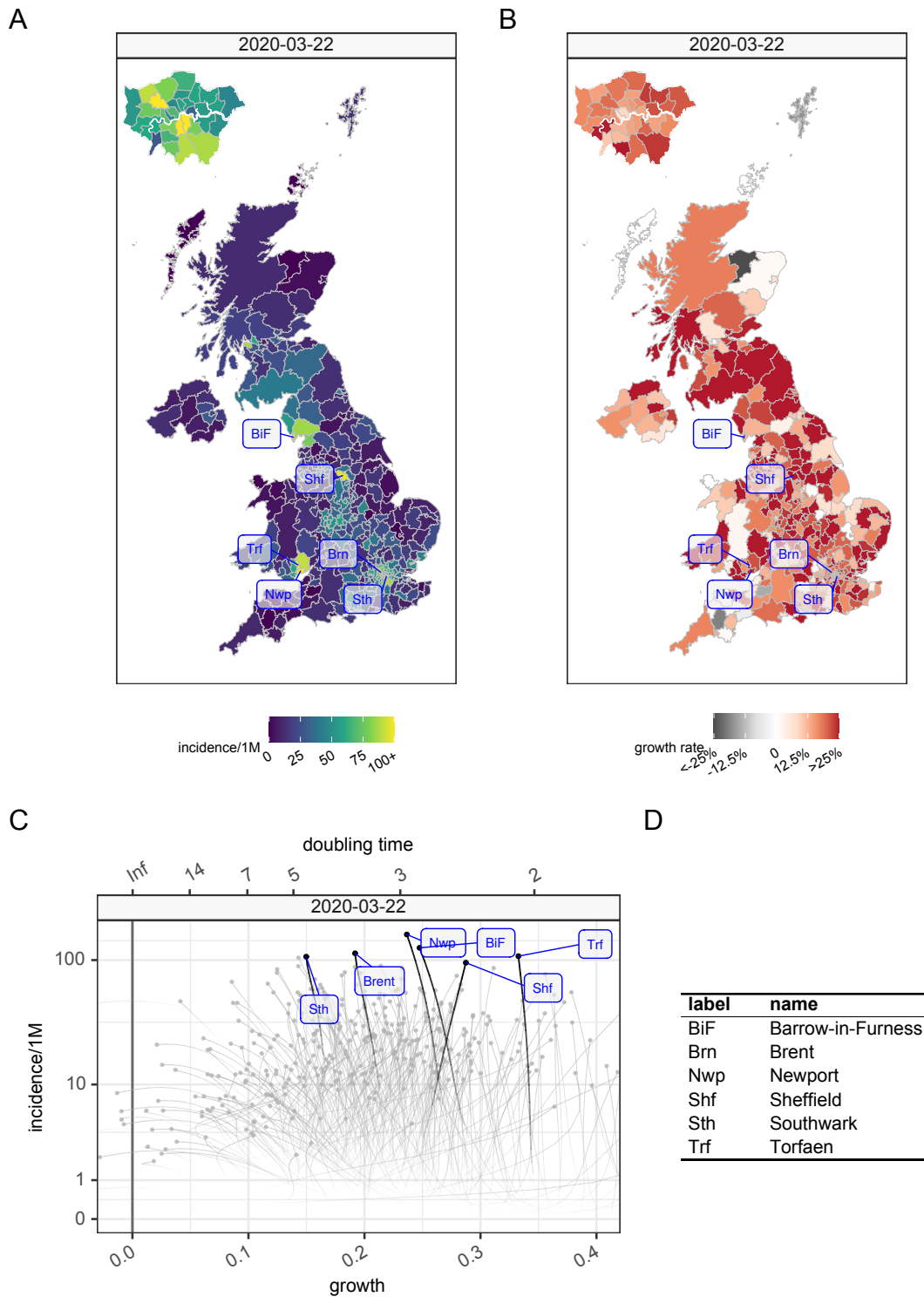


Figure 1.2: Lower tier local authority (LTLA) distribution of case incidence per 1M capita per day (A) and growth rates (B) on March 22, 2021. A nationwide lockdown was introduced the next day. Panel C shows a phase diagram with the growth rate on the X axis and the incidence per 1M capita on the Y axis. Each trace shows the trajectory of a LTLA over the 14 days prior to and including the 22nd March 2021, shown as the point on the head of the path. The 6 highlighted areas show the regions which were experiencing the most new cases per 1M capita each day on the 22nd. Case numbers were under-representative at this time as testing capacity was limited, and testing was being restricted to symptomatic cases presenting to hospital.



capacity and potential local demand (Challen et al., 2021a), representing geography as a network of small neighbouring regions, and using a technique from network clustering called label propagation. This enabled us to deliver pragmatic estimates of hospital demand to the local NHS trusts in the very early phase of the pandemic, prior to the imposition of control measures. These predictions were useful in the short term, but in the longer term the trajectory of the epidemic and particularly how it responded to the control measures imposed in March 2020 was the most important factor determining hospital load, and this was best described by the reproduction number.

## **1.5 Published paper: Meta-analysis of the SARS-CoV-2 serial interval and the impact of parameter uncertainty on the COVID-19 reproduction number**

The reproduction number gives us an indication of the degree of control of the epidemic, in terms of the number of secondary cases generated by one primary case, which is independent of how rapidly those infections happen (Vegvari et al., 2021). The overall growth rate of the epidemic is connected to the reproduction number by the time between primary and secondary infections, known as the generation time (Wallinga and Lipsitch, 2007). In reality the quantity that the reproduction number is measuring, i.e. infections, is never directly measured, and we infer their existence from proxies such as positive tests, hospitalisations, or deaths all of which are subject to a delay from infection. The generation time is also not observed directly, but instead we can measure the time between the detection of sequential cases, the serial interval. The reproduction number comes in a number of flavours (Fraser, 2007) which are formally defined in Appendix A:

1. the basic reproduction number ( $R_0$ ) which is the number of secondary infections for each primary infection in the particular circumstance of a freely mixing, and completely susceptible population.
2. The case reproduction number ( $R_t^c$ ) which is the number of secondary infections generated by a single primary infection which occurs at time  $t$ , and
3. The effective reproduction number ( $R_t$ ) which is the inverse ratio of the number of primary infections associated with secondary infections observed at time  $t$ . This is also known as the instantaneous reproduction number.

Both the case and effective reproduction number are time dependent measures and influenced by the degree of immunity and social mixing in the population at the time of the estimate. They are both similar and useful for tracking the progress of the epidemic, although the effective reproduction number averages out dynamics from the last generation and is slower to adjust to a new value following a step change in the case reproduction number. The case reproduction number however can only be determined in retrospect.

Responsibility for tracking the reproduction number for the UK fell to the Scientific Pandemic Influenza - Modelling subgroup (SPI-M), and a range of universities participated in weekly estimates of the reproduction number using a range of methods that were combined into an ensemble estimate (Maishman et al., 2021). Groups from the London School of Hygiene and Tropical Medicine, Imperial College, Warwick University, Manchester University and Cambridge University (in association with Public Health England) produced estimates from transmission models of varying sophistication and complexity (Vegvari, 2020). To provide a counterpoint to the complex models we aimed to produce estimates for the reproduction number which were as closely related to the observed data as possible, whilst making the minimal set of modelling assumptions. With an early interest in the spatial distribution of the epidemic growth, and limited computation resources available, we focused on

methods that could be easily scaled to include hundreds if not thousands of regions, on commodity hardware.

Initially we used the renewal equation method developed by Cori et al. (Cori et al., 2013; Thompson et al., 2019) for effective reproduction number estimation, as implemented in the R package EpiEstim (Cori et al., 2021). Later in the pandemic we extended this and developed other approaches as described in section 1.9.

The estimation of the reproduction number depends on the infectivity profile, which is a probability distribution defining the likelihood that a primary case was infected on a particular number of days before a secondary infection occurs. The infectivity profile is related to the distribution of the generation interval which is the time between infections in 2 sequential cases in a chain of transmission. The serial interval is the interval between sequential case detection in a chain of transmission and although somewhat different to the generation time can be directly measured. Quantification of the infectivity profile is the subject of our paper “Meta-analysis of the SARS-CoV-2 serial interval and the impact of parameter uncertainty on the COVID-19 reproduction number” (Challen et al., 2021b), which describes our early approach to estimating this key quantity. It documents the assumptions we made and analyses the impact these have on our estimates of the reproduction number. Over time better data on household transmission have become available, leading to more precise estimates of the generation time (Hart et al., 2021a) both for the original strain and for the novel SARS-CoV-2 variants with different characteristics (Hart et al., 2021b).

## **1.6 Published paper: Estimates of regional infectivity of COVID-19 in the United Kingdom following imposition of social distancing measures**

Over the course of the pandemic, real time monitoring of the transmission of COVID in the UK, and particularly in England, was needed to guide the Government's response. To meet this need, as part of SPI-M we performed a range of analyses on a weekly basis which provided insights into both the overall trajectory and the population dynamics of the pandemic including:

1. the weekly estimation of the reproduction number for SARS-CoV-2 as described above
2. analysis on the regional variation, of both the incidence of infection, and rate of growth, and
3. age stratification of both the incidence of infection, and rate of growth.

In the period leading up to and immediately after the March 2020 lockdown, reliable data that could be used to track the progress of the pandemic were lacking. The earliest data streams that were available were case counts of infections from PCR testing in Public Health England labs. The rapid increase in cases in early March rapidly saturated the capacity of laboratory testing and restrictions on the use of testing were imposed, limiting it to suspected in hospital cases only. Initially data were published on positive cases detected on a given day, however these related to specimens taken in the preceding few days, and the signal was delayed by a few days, varying not only by the lab processing speed and capacity, but also delays in the data aggregation and publishing. Similar to findings of our work on laboratory

test review (Challen et al., 2020), this was dependent on workflow factors, including weekend working, and varied from region to region, with, for example, no data at all being published on the weekend in Northern Ireland.

Positive case data may be biased if the testing effort varies over time resulting in a changing baseline. This may particularly happen if testing capacity changes (as was seen in the early phases) or local public health initiatives result in a surge in testing. Initially very limited data were available about the negative test results, and even in the later stages of the pandemic the precise relationship between positive and negative tests has been unclear as although the negative test rate later became available for analysis the relationship between repeat negative testing of individuals who subsequently test positive was never clear. To circumvent this we largely focused our analysis on the subset of positive cases where the test was sought due to symptoms of COVID. This subgroup was certainly less affected by biases caused by screening activity and were largely prioritised for testing when capacity was reduced.

Hospital admission and in-hospital case numbers were more potential data streams that could be monitored, and which promised less acquisition bias. The reality however of counting hospital admissions for COVID is complicated by the spectrum of severity, the natural history of the disease, and the variable context of the infection seen in hospital patients. By context we mean an inpatient with COVID may have acquired the infection in the community, be detected through community testing, but gradually deteriorate over the course of a few days, eventually requiring emergency hospital admission. Alternatively a care home resident may become acutely unwell and require emergency admission, receiving a COVID diagnosis subsequent to admission. In yet another scenario, a routine postoperative inpatient may acquire COVID in hospital as a complication of their care. In all these scenarios there are different relationships between the context of infection and admission. Case reporting is dependent on the definition of an in-hospital COVID case, which may be interpreted differently in different hospitals. As essentially a manually curated

data set, admissions reporting suffers the same, or worse, reporting delays as a result of workflow factors, and variable reporting delays between different hospitals made the aggregate UK wide signal difficult to interpret.

Alternative data streams include deaths, which although a concrete end point, are long delayed from infection, so difficult to use as a real time marker. Whether or not a death was caused by COVID, or contributed to by COVID, or simply occurred incidentally after COVID, is a judgement call that is subject to a great deal of interpretation. Reporting of deaths that occur out of hospital depends on death certification, processes in the Coroner's Office and the Office of National Statistics. The rules about how long after death a case can be attributed to COVID subject to change during the pandemic, initially including deaths to 60 days, but subsequently only to 28 days. Even though a death is a concrete end point, interpretation as to causation, and the potential for reporting delays make them equally unreliable.

Other data streams we investigated at various points were the 111 telephone triage data which tracked the number of calls made by people seeking advice on coronavirus symptoms, and the resulting triage category, given an indicator of severity. During the summer of 2020 this was the primary method by which coronavirus testing was accessed, and 111 call levels appeared to be an early indication of infection levels, as call logs were not subject to reporting delay. However during the autumn policy changed around the access to testing making direct booking possible, and 111 triage data became decoupled from the other data streams, again demonstrating the impact of workflow on the data.

Our paper "Estimates of regional infectivity of COVID-19 in the United Kingdom following imposition of social distancing measures" represents an early snapshot of this work, which highlights the regional variation of the growth of the pandemic at a coarse-grained geographical level. Figure 1.3 shows a different view of this problem by looking at the areas where case incidence remained high, relative to the

UK baseline over the course of the first lockdown. In early March the areas most impacted prior to the lockdown were London, South Wales, and the Morecambe bay area (including Barrow-in-Furness and Lancaster). London particularly responded quickly to lock down, and by May was under the UK average for cases per capita. Over the period in Figure 1.3 we see areas such as Thanet, Leicester, Pendle and Hyndburn, become established as regions with persistently high infection rates. This more detailed analysis informed local control efforts.

One aspect of the different data streams that we particularly focused on was the relationship between infection dynamics and age. Age is the key determinant of severity and hence hospital burden. Social contacts in different age groups are variable over time as social distancing policies changed, and school closures and holidays came and went. In the later stages of the pandemic the age stratification of vaccination roll out also affected the age distribution. As the different data streams of cases, hospitalisations and deaths are representative of different severity of outcome, they also are representative of different age groups. As social contact patterns vary during the year, it has been possible to observe the swing of the burden of infection, and hospitalisations (Figure 1.4) from old to young and back again.

The routine monitoring of the pandemic highlighted here was a response to the need of the national Government for information to base decision making on. Much of this was conducted with imperfect data on short timescales, and the evolution of both the biases in the data and our understanding of them, undoubtedly had an impact on the timing of decisions made particularly early in the pandemic. Much of the data relating to the beginning of the pandemic have subsequently been revised, and made more easily accessible through services such as the UK Government coronavirus website. It is important to remember that the data flows have evolved throughout the pandemic and that the best available information at the time was sometimes very different to the information available in retrospect.

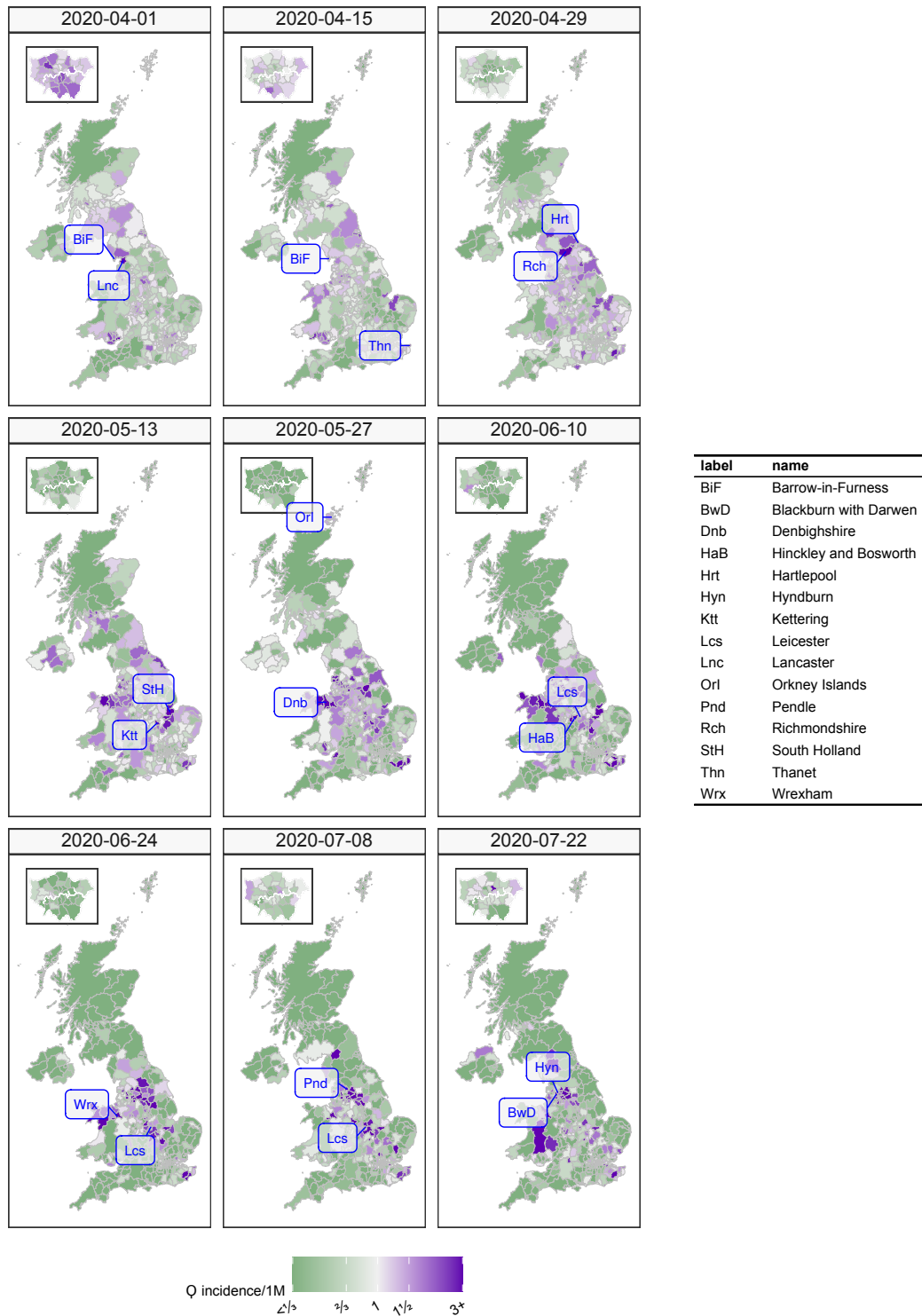


Figure 1.3: The incidence rate ratio by Lower Tier Local Authority throughout the period of April to late July 2020 shows the evolution of regions with high incidence compared to the UK baseline from areas where high levels of importation and social mixing lead to the initial outbreak to areas where infections persisted, possibly due to societal factors as high levels of key or factory workers, dense housing, close social networks, or high proportions of large and multi-generational households.



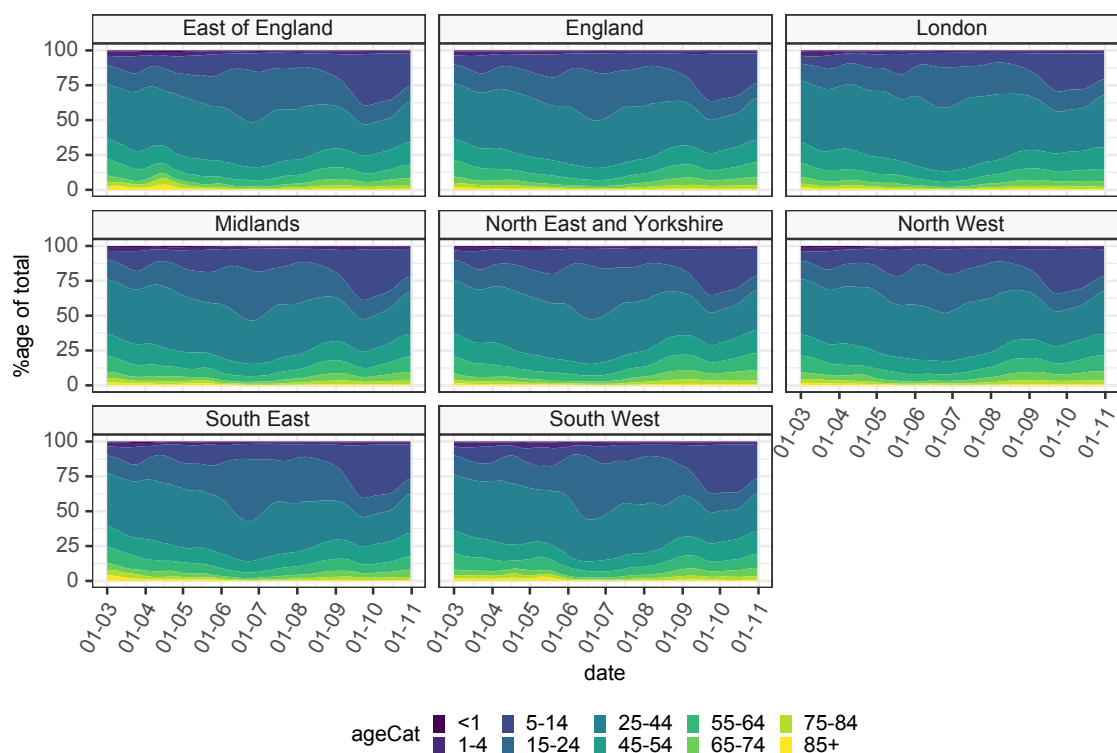


Figure 1.4: A 2 weekly rolling average proportion of hospital admissions by age category in the period from 1st December 2020, to early October 2021. As vaccination rates in older age groups increased over the course of the programme, which was rolled out on an age stratified basis, the proportion of hospitalisations in older age groups dropped. Gradually as younger age groups became eligible for vaccination the benefit inferred to the older age groups was lessened. This coupled with waning immunity in the older age groups results in a reversion to pre-vaccination age distributions.

## 1.7 Published paper: Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study

In early October 2020 a novel variant had been identified on genomic sequencing of cases in the South East of England. This variant possesses a deletion in the SARS-CoV-2 genome at site 69-70 in the spike protein, which happens to be the binding site for the Spike protein PCR target (S-gene) of the ThermoFisher TaqPath quantitative PCR assay, which was in widespread use at the time. The mutation caused the S-gene target to fail to amplify in contrast to other targets, producing a “S-gene negative” result. The S-gene negative test results provided a proxy marker for the new variant. Initially this was non specific, as small numbers of “wild-type” infections would also

exhibit S-gene target failure, particularly when in the recovery phase, however as Alpha variant case numbers increased it quickly became an excellent proxy. The new variant was subsequently given the Pango lineage (Rambaut et al., 2020) identifier B.1.1.7 (O’Toole et al., 2021), identified by Public Health England as VOC202012/1, and given the WHO designation “Alpha” (World Health Organisation, n.d.).

In October and November 2020, a progressive rise in case numbers became concerning enough for the Government to intervene. The associated increase in admissions were predicted to stretch the capacity of the NHS (Davies et al., 2021a), with the additional uncertainty of the impact of increased social mixing over Christmas. As a result the Government eventually re-imposed social distancing measures over a 4 week period in November, but without closing schools. At about the same time the proportion of test results with S-gene failure began to rise, spreading outwards from the South East (Figure 1.5) and caused a rapid increase in case numbers despite the November lock down. This rapid spread was the result of a transmission advantage for B.1.1.7 seen in Figure 1.6 where S-gene negative cases have an increased effective reproduction number, compared to S-gene positive cases. This transmission advantage resulted in a wave of cases, which in turn resulted in the Government implementing a full nationwide lockdown, including school closures, in January 2021.

The transmission advantage caused a rise in cases which was a driver for an increase in hospitalisations. We aimed to determine the relative severity of the B.1.1.7 admission to help the international community prepare for the spread of the Alpha variant. In parallel to groups in the London School of Hygiene and Tropical Medicine, we investigated the relative risk of mortality of the Alpha variant, compared to previously circulating variants. Death following community identified infection, remains a relatively rare outcome. The rapid propagation of the variant introduced potential confounding factors, including pressure on NHS hospitals. As a result it was challenging to get an unbiased estimate of any increase in severity. Using a matched cohort and constructing a data set that accounted for known biases, our

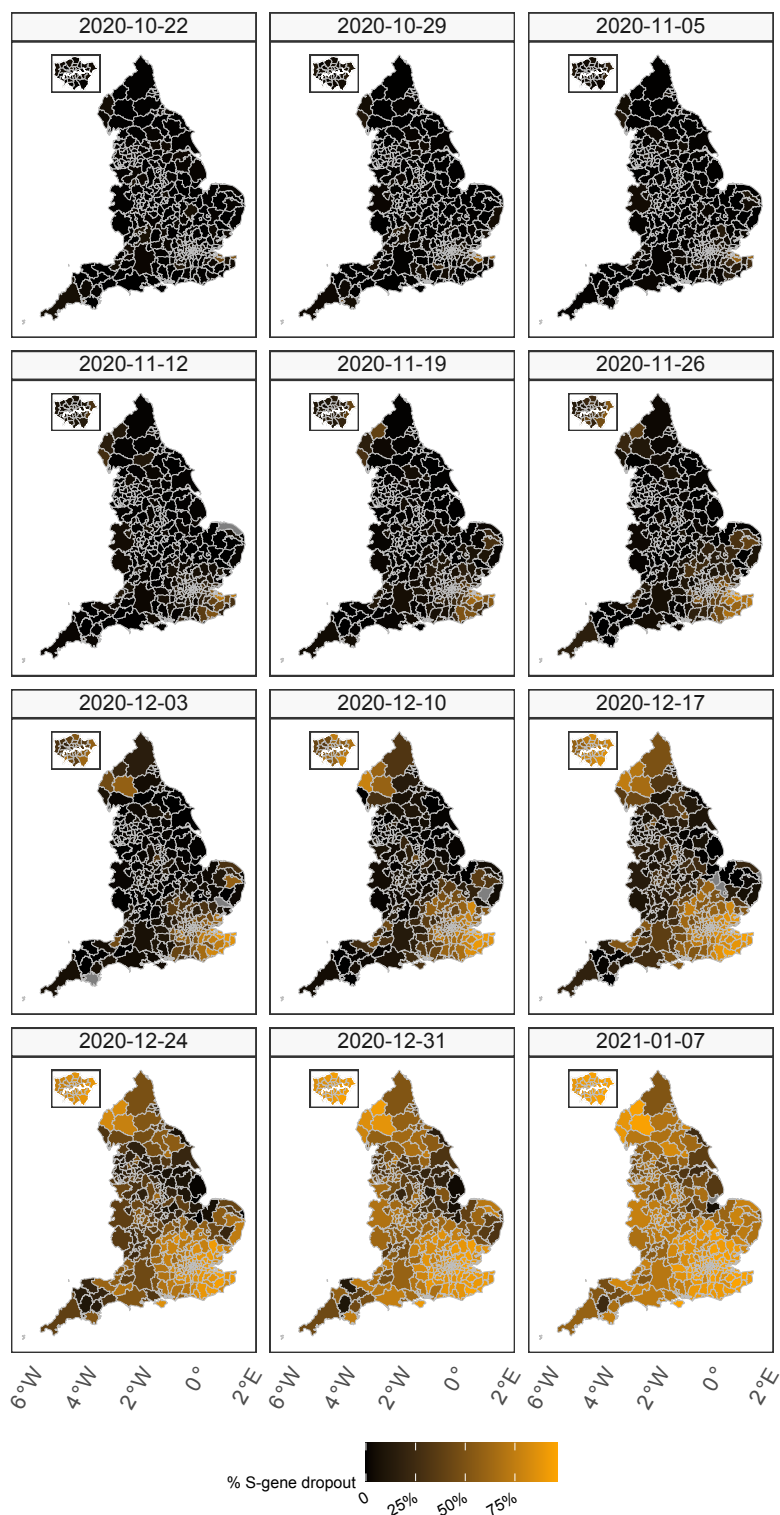


Figure 1.5: *S*-gene positivity by LTLA at 12 time points between late October 2020 and early January 2021. The data plotted are the 2 week rolling averages of the lower limit of the 95% binomial confidence interval, and represents the minimum proportion estimated at a given time point accounting for areas with small case numbers.

research complemented similar work from other groups that used different statistical approaches (Davies et al., 2021b), providing mutual reassurance of the validity of our findings.

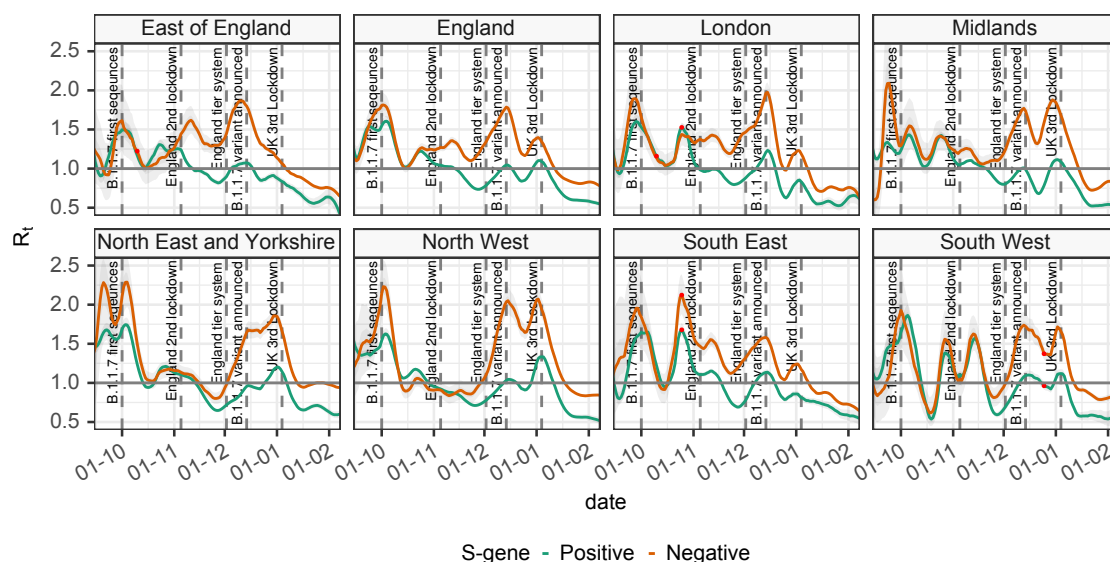


Figure 1.6: The reproduction number of *S*-gene positive and *S*-gene negative cases over the late autumn and winter 2020-2021. The earliest sign of an advantage for *S*-gene negative cases was seen in late October in the South East of England, where the growth of *S*-gene negative cases outpaced the others, leading to a separation of the *S*-gene negative and *S*-gene positive reproduction number curves. This pattern is repeated at different times in different regions as B.1.1.7 spread across the country, and once established, the advantage persists as the reproduction number varies.

The analysis was important globally as Alpha had already spread outside the UK. The finding that the Alpha variant does indeed lead to increased mortality, was picked up by the international media, whose interest helped warn the international community to plan ahead for the Alpha variant wave, and advised towards strengthening control measures whilst accelerating developing vaccination programmes.

One bias that was not completely possible to account for is if the time from infection to time to case acquisition is different for each variant. This has been shown to produce potentially biased estimates of severity (Seaman et al., 2021). We did however control for the viral load, by using the PCR cycle threshold value. This has a relationship, albeit an uncertain one, to the delay from infection to test result, and although controlling for this did reduce the effect we observed it did not remove it altogether.

## **1.8 Submitted paper: Early epidemiological signatures of novel SARS-CoV-2 variants: establishment of B.1.617.2 in England**

After the start of the UK national lockdown on the 5th January 2021 (BBC News, 2021c) case numbers began to fall rapidly, as schools and workplaces remained shut, and as the vaccination programme gathered pace (Public Health England, 2021a). In March the first step of a planned road-map of easing restrictions came into force, allowing schools to return (BBC News, 2021d). The existing Alpha variant was proven to be controlled by the lockdown in place, but over this period there were concerns of the possibility of the establishment of a range of possible variants including the South African (B.1.351) or Brazilian variants (P.1 & P.2) which had become dominant in other geographies.

The first detected case of a new mutation referred to locally as the “Triple mutant” was from the state of Maharashtra, in India in December 2020, which was subsequently designated as the lineage B.1.617. Over February, March and April, it and other lineages including B.1.1.7, began to spread rapidly, as restrictions were eased and various large events took place, such as the Kumbh Mela festival (BBC News, 2021a). On 23rd April travel from India to the UK was made subject to quarantine restrictions (UK Government, 2021). The impact of the new variants on the situation in India was difficult to assess due to the comparatively low number of sequenced cases, the rapid increase in cases of multiple different variants, concurrent changes in social behaviour, combined with relatively low vaccination rates and population immunity at the time in India.

We aimed to detect outbreaks of variants which posed a threat to the ongoing control of the epidemic in England and enable timely public health intervention. To this end we had already begun monitoring the routinely collected data for changes that

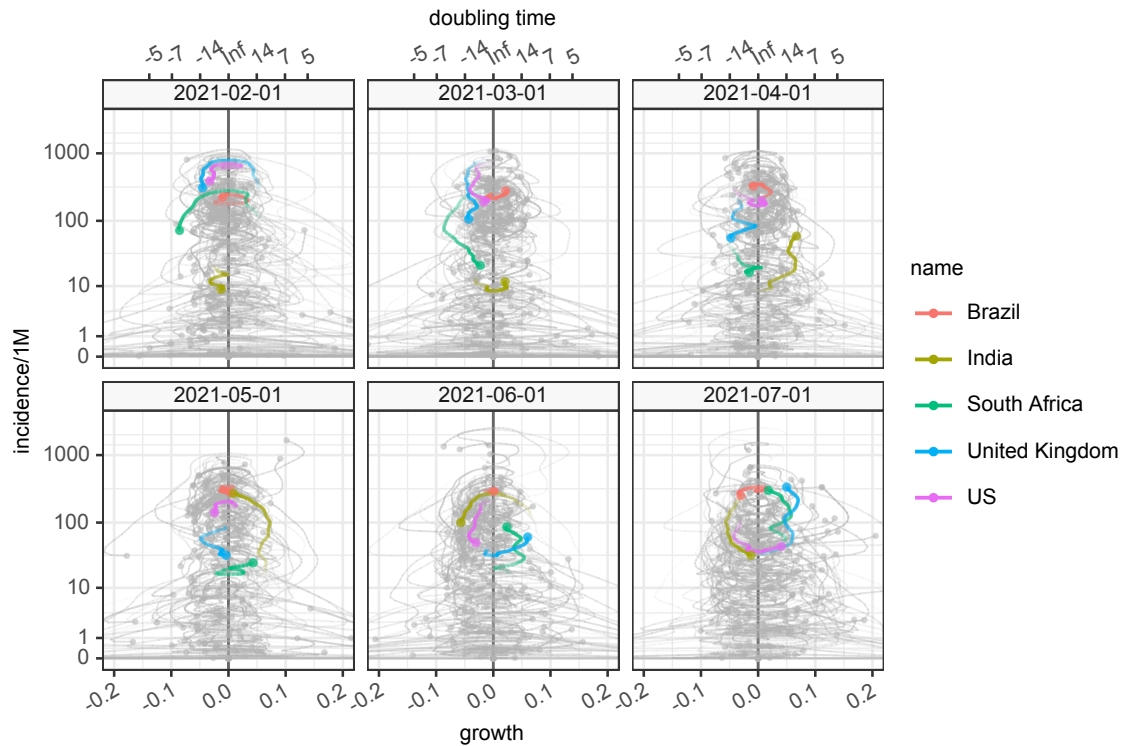


Figure 1.7: The growth rate of reported COVID cases versus incidence per 1 Million population over time in Brazil, India, South Africa, the UK, and the US between February to July 2021. Data are from John Hopkins University Coronavirus tracker. This phase diagram shows case incidence per capita on the y-axis, and growth on the x-axis, and the evolution of growth and incidence over the previous 28 days in each country is shown as a path, with the head on the date of each panel. In a new wave of infection increasing case numbers and growth rates are seen as the path moving up and to the right. As the wave peaks the path trajectory moves anti-clockwise, before crossing the y-axis where growth is zero at the peak. Subsequent resolution of a wave of infection is seen as the trajectory descending on the left hand side of the figure.

were suggestive of non-Alpha variants of concern over March 2021. This was later augmented by the addition of linked genomic data, when these streams became available. Sequencing results typically were available only 2-3 weeks from specimen collection so we used the S-gene status of the positive specimens as a proxy for variants of concern. In late March we saw a handful of cases that were incompatible with the Alpha variant, as they had a detectable S-gene PCR target on TaqPath assay (S-gene positive). These grew exponentially and as sequencing results became available, it gradually became clear that this was due to the community spread of the Delta variant, B.1.617.2 a sub-lineage of B.1.617. On 9th of April 2021 we presented a preliminary analysis to SPI-M which indicated that Bolton was an area of high risk of spread of variants of concern.

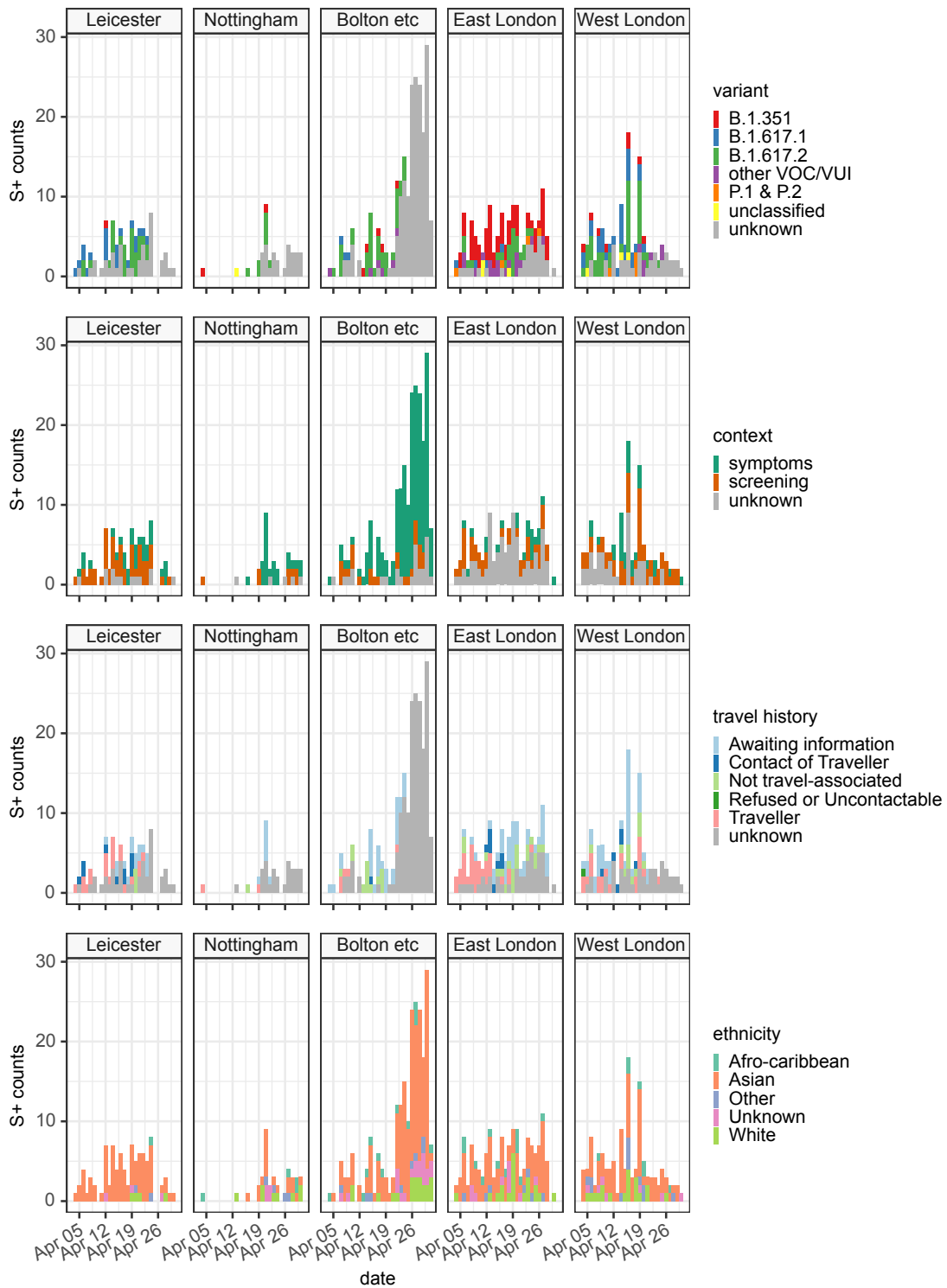


Figure 1.8: Epidemic curves of 5 limited geographic areas which were identified as having clusters of high numbers of S-gene positive cases and some Delta variant cases (B.1.617.2) identified through delayed genomic sequencing.

By 6th May as more sequencing data became available we were able to demonstrate that the S-gene positive cases in Bolton and the surrounding areas were demonstrating a growth advantage over the S-gene negative Alpha variant cases in the specific

geographic areas. This growth advantage in cases helped to make the case that B.1.617.2 should be upgraded to a variant of concern and the response of public health protection agencies strengthened. This analysis was presented to SAGE, the Government’s scientific advisory group on 11th May 2021 (Challen, 2021) by which time a robust set of public health measures was already being implemented in Bolton (Local Government Association, 2021). The SAGE analysis also led to the Government reconsidering the timing of the later steps of the proposed road-map of restriction easing (BBC News, 2021b).

After May 2021, the Delta variant spread to the rest of the UK resulting in a wave of cases that was still growing by the beginning of July, (Figure 1.7). The spread over a wider population allowed a better view of the epidemiological characteristics of the Delta variant and this enabled us to make estimates of the transmission advantage of the Delta variant over Alpha. The whole of this analysis is summarised in the paper “Early epidemiological signatures of novel SARS-CoV-2 variants: establishment of Delta (B.1.617.2) in England” (Challen et al., 2021c).

Following this analysis we have seen the further international spread of Delta in both sequencing results and in the growth patterns observed in case data and shown in Figure 1.8, which demonstrates how the spread of Delta to the US is following similar patterns to those established in India and the UK.

The analysis presented here seems relatively simple in retrospect. As the situation was unfolding however there were considerable challenges accessing and interpreting the data. S-gene statuses had become routinely available, following the emergence of Alpha, however sequencing data linked to S-gene data only became available in mid April. At this stage the Pango lineage B.1.617 (*Proposed New B.1 Sublineage Circulating in India* 2021) had only recently been assigned to the “Triple mutant” cases identified in India. The B.1.617.2 identifying the Delta variant was proposed on 21st April (*Potential Sequences That Should Be Included in B.1.617* 2021) and



although B.1.617.1 was a variant under investigation B.1.617.2 did not become formally designated as such until 6th May (Public Health England, 2021b). This meant that B.1.617.2 cases were not reported through the established data streams until around that date. Even after this point the 2-3 week delay in reporting genomic sequencing results made the explicit genomic signal too delayed for the purpose of outbreak identification, as seen in Figure 1.8. This necessitated the use of a proxy measure, in the S-gene status. As this alternative and more timely signal was not specific for B.1.617.2 we had to visualise the spatiotemporal patterns of S-gene and variant detection and later assemble a model to describe these formally. All of which had to take into account changing variant designations and data streams.

Ultimately, from this experience, we can learn that if the goal is to prevent outbreaks of a specific variant, which is believed to have a transmission advantage, it is important to react swiftly and disproportionately to even the weakest evidence, as the delay waiting for better data renders the interventions required far less effective. On the other hand we can also conclude that monitoring for a specific variant outbreak in the situation where there are high levels of background infection is challenging, as the signal from the novel variant is effectively hidden.

## **1.9 Methodology for Estimating the Reproduction Number and Growth Rate**

This body of work is underpinned by statistical methods for estimating the reproduction number and exponential growth rate. Initially we needed answers quickly and we adopted the renewal equation method developed by Cori et al. (Cori et al., 2013; Thompson et al., 2019) for effective reproduction number estimation, as implemented in the R package EpiEstim (Cori et al., 2021). As time went on we recognised specific limitations in this algorithm and sought to address them, with enhanced

implementations of the renewal equation method and other approaches.

To assess the validity of these new approaches we developed a framework for creating synthetic data designed to test the behaviour of the algorithms in particular circumstances such as large step changes in the reproduction number, and to quantify delay in the estimates. We also implemented estimate scoring rules and an associated validation methodology that quantifies the overall performance of any particular estimation method. These are described in Appendix A.

We considered some extensions to the implementation of the renewal equation method in EpiEstim. This method is described in detail in Appendix B, but in summary combines a prior assumption on the distribution of the reproduction number, the assumption that it is constant over a fixed time window, an estimate of the generation time, and observations of the incidence of cases within a Bayesian framework. From this it is possible to derive a closed form expression for the posterior distribution of the reproduction number given the evidence contained within the case time series.

The two additional areas of functionality we implemented concern the selection of the prior distribution for the effective reproduction number, the selection of posterior estimates including adjusting the time window over which the reproduction estimate is made, and the methods of combining multiple posterior estimates to a single result. These addressed issues introduced by statistical noise in situations when national case counts became low, or where we were concentrating on small geographical areas with low case numbers, and issues of over precision resulting from violation of some of the underlying assumptions of the method. More details of these issues and extensions are detailed in Appendix B. To make the methodology available to a wider set of technical platforms, we implemented the original algorithm and the extensions in Java, with an interface to R (Challen, 2022).

Estimates of the reproduction number are useful as they directly relate to the degree

of social interaction and acquired immunity. The exponential growth rate ( $r_t$ ) on the other hand is a reliable measure of increasing cases that is based solely on observed data, and not reliant on any other assumptions. For small area estimation we implemented a simple statistical model where daily count data ( $I_t$ ) is assumed to arise from a Poisson distribution ( $I_t \sim \text{Poisson}(\lambda_t)$ ).  $\lambda_t$  is estimated using a local regression using maximum likelihood of the nearest data points to a given time point (Loader, 1999). This uses a logarithmic link function and a Quasi-Poisson model to account for over-dispersion. This regression is evaluated for every time point and the gradient of the regression is a direct estimate of the exponential growth rate ( $d(\log I_t)/dt$ ). Depending on the situation this estimate of growth rate can be translated into a case reproduction number using the methodology described in Wallinga and Lipsitch, 2007, Further details of the growth rate methods and validation can be found in Appendix C.

Over time we developed additional frameworks for calculating both the growth rate and reproduction number and, in particular, a Bayesian approach to calculating the growth rate and reproduction number at the same time. This approach is described in Appendix D.

## 1.10 Conclusion

The COVID-19 pandemic delivered the largest acute challenge to the NHS in living memory. It fundamentally, albeit temporarily, diverted the priorities of the NHS away from the management of chronic disease to the provision of acute care. The pandemic also changed all aspects of the delivery of health care services, and existing pathways of care, as the NHS grappled with the dual impact of infection control measures and societal interventions designed to reduce transmission.

With COVID infection comes a large cohort of patients who have been exposed to a

new multi-system disease with uncertain medium and long term health effects, and unknown interactions with existing comorbidities. With the disruption to health care services comes an inevitable impact on the clinical outcomes of other chronic diseases, as a result of delayed care.

In our work on safety in machine learning based clinical decision support systems these paradigm shifts were identified as key risks for the accuracy of data driven systems. Similarly in our observational study of laboratory test review we drew conclusions based on the analysis of a workflow that has doubtless changed due to the far reaching impact of the pandemic.

The ongoing relevance of retrospective observational studies has been brought into question by the paradigm shift brought about by COVID-19. More concerning for clinical research, we can no longer assume the findings of research conducted prior to the pandemic are still relevant to the future. For ongoing research the effects of the disruption of the pandemic, of comorbid COVID-19 infection, and of the subject of the research itself, will be difficult to tease apart, particularly where we are investigating small effect sizes. We took a case cohort approach in our study of severity of the Alpha variant, matching a cohort on time and place of COVID-19 infection. Similar approaches may be needed in future research, to control for the dual impact of COVID-19 infection and disruption of clinical care.

Critical to re-establishing clinical research as we exit the pandemic will be quality data. The early stages of the pandemic highlighted how ill-prepared we were to collect even simple national level data-sets for public health purposes, despite the considerable infrastructure in the NHS to collect data for administrative purposes. Preexisting pandemic response plans relied on the comprehensive data within general practice patient record systems, however in the event of acute respiratory disease requiring hospital care, data simply do not flow into general practice in a timely fashion. With some of the analysis presented here, delays to data of even a few days

made accurate assessment of the trajectory of the pandemic complex. Re-evaluation of national data flows is needed in the future.

Data are only useful with analysis. The pragmatic and timely distribution of anonymised individual level data from Public Health England to academics, who volunteered their time to the COVID-19 response, enabled the Government to make an informed response to the pandemic. This was not without issue however. Strong cultural resistance to sharing information persisted despite the clear guidance of the National Data Guardian (Caldicott, 2020), and, for example, it was only much later that data about hospital stays began to flow. From personal experience, the process of obtaining access to protected research environments such as those of the Office of National Statistics through the University of Exeter proved impossibly complicated. It is issues such as these that need to be resolved if the Government is to benefit from the support of academic health data scientists in future emergencies.

There is clearly a risk of future waves of new variants of SARS-CoV-2 and we must develop effective strategies to monitor for these. However, as the acute phase of the pandemic resolves, we anticipate the NHS will have to focus on the mammoth task of clearing the disruption COVID has had on chronic disease management, whilst identifying and managing potential long term complications of COVID-19. Healthcare data science can provide support for these goals, but can only do so with analysis of up-to-date clinically rich data. The pandemic has provided an impetus to redesign the relationship between academia and clinical data owners, and there is an opportunity to develop this to be fit for both the immediate challenges facing the NHS and for the longer term future of clinical research.

This body of work demonstrates what can be achieved when pragmatic collaboration between Public Health, the NHS and academic healthcare data science research is facilitated. This collaboration has been instrumental in the Government's response to a national emergency, and maintaining it should be a priority for Public Health,

the NHS, academia, and the UK Government.

## 2. Artificial intelligence, bias and clinical safety

Status: published in BMJ Quality and Safety, vol 28, p231-237

Date: 12th Jan 2019

Candidate's contribution: All authors discussed the concept of the article and RC wrote the initial draft. KTA, JD, TE, MP and LG commented and made revisions, DC critically reviewed the draft. All authors agreed with the final manuscript. RC is the guarantor.



OPEN ACCESS

# Artificial intelligence, bias and clinical safety

Robert Challen,<sup>1,2</sup> Joshua Denny,<sup>3</sup> Martin Pitt,<sup>4</sup> Luke Gompels,<sup>2</sup> Tom Edwards,<sup>2</sup> Krasimira Tsaneva-Atanasova<sup>1</sup>

<sup>1</sup>EPSRC Centre for Predictive Modelling in Healthcare, University of Exeter College of Engineering Mathematics and Physical Sciences, Exeter, UK  
<sup>2</sup>Taunton and Somerset NHS Foundation Trust, Taunton, UK  
<sup>3</sup>Departments of Biomedical Informatics and Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA  
<sup>4</sup>NIHR CLAHRC for the South West Peninsula, St Luke's Campus, University of Exeter Medical School, Exeter, UK

## Correspondence to

Dr Robert Challen, EPSRC Centre for Predictive Modelling in Healthcare, University of Exeter College of Engineering Mathematics and Physical Sciences, Exeter EX4 4QF, UK; rc538@exeter.ac.uk

Received 23 May 2018  
Revised 23 November 2018  
Accepted 6 December 2018  
Published Online First  
12 January 2019



► <http://dx.doi.org/10.1136/bmjqs-2018-008551>



© Author(s) (or their employer(s)) 2019. Re-use permitted under CC BY. Published by BMJ.

**To cite:** Challen R, Denny J, Pitt M, et al. *BMJ Qual Saf* 2019;**28**:231–237.

## INTRODUCTION

In medicine, artificial intelligence (AI) research is becoming increasingly focused on applying machine learning (ML) techniques to complex problems, and so allowing computers to make predictions from large amounts of patient data, by learning their own associations.<sup>1</sup> Estimates of the impact of AI on the wider economy globally vary wildly, with a recent report suggesting a 14% effect on global gross domestic product by 2030, half of which coming from productivity improvements.<sup>2</sup> These predictions create political appetite for the rapid development of the AI industry,<sup>3</sup> and healthcare is a priority area where this technology has yet to be exploited.<sup>2,3</sup> The digital health revolution described by Duggal *et al*<sup>4</sup> is already in full swing with the potential to 'disrupt' healthcare. Health AI research has demonstrated some impressive results,<sup>5–10</sup> but its clinical value has not yet been realised, hindered partly by a lack of a clear understanding of how to quantify benefit or ensure patient safety, and increasing concerns about the ethical and medico-legal impact.<sup>11</sup>

This analysis is written with the dual aim of helping clinical safety professionals to critically appraise current medical AI research from a quality and safety perspective, and supporting research and development in AI by highlighting some of the clinical safety questions that must be considered if medical application of these exciting technologies is to be successful.

## TRENDS IN ML RESEARCH

Clinical decision support systems (DSS) are in widespread use in medicine and have had most impact providing guidance on the safe prescription of medicines,<sup>12</sup> guideline adherence, simple risk screening<sup>13</sup> or prognostic scoring.<sup>14</sup> These systems use predefined rules, which have

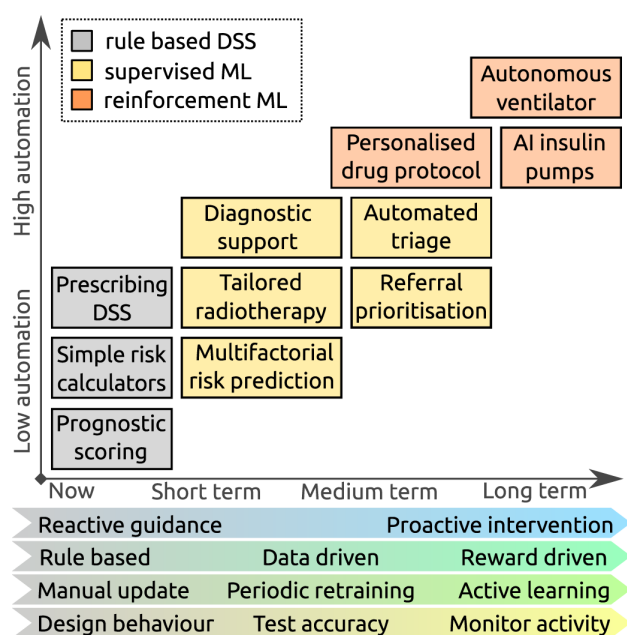
predictable behaviour and are usually shown to reduce clinical error,<sup>12</sup> although sometimes inadvertently introduce safety issues themselves.<sup>15,16</sup> Rules-based systems have also been developed to address diagnostic uncertainty<sup>17–19</sup> but have struggled to deal with the breadth and variety of information involved in the typical diagnostic process, a problem for which ML systems are potentially better suited.

As a result of this gap, the bulk of research into medical applications of ML has focused on diagnostic decision support, often in a specific clinical domain such as radiology, using algorithms that learn to classify from training examples (supervised learning). Some of this research is beginning to be applied to clinical practice, and from these experiences lessons can be learnt about both quality and safety. Notable examples of this include the diagnosis of malignancy from photographs of skin lesions,<sup>6</sup> prediction of sight-threatening eye disease from optical coherence tomography (OCT) scans<sup>7</sup> and prediction of impending sepsis from a set of clinical observations and test results.<sup>20,21</sup>

Outside of diagnostic support ML systems are being developed to provide other kinds of decision support, such as providing risk predictions (eg, for sepsis<sup>20</sup>) based on a multitude of complex factors, or tailoring specific types of therapy to individuals. Systems are now entering clinical practice that can analyse CT scans of a patient with cancer and by combining this data with learning from previous patients, provide a radiation treatment recommendation, tailored to that patient which aims to minimise damage to nearby organs.<sup>22</sup>

Other earlier stage research in this area uses algorithms that learn strategies to maximise a 'reward' (reinforcement learning). These have been used to test





**Figure 1** Expected trends in machine learning (ML) research: boxes show representative examples of decision support tasks that are currently offered by rule-based systems (grey), and hypothetical applications of ML systems in the future (yellow and orange), demonstrating increasing automation. The characteristics of the ML systems that support these tasks are anticipated to evolve, with systems becoming more proactive and reward driven, continuously learning to meet more complex applications, but potentially requiring more monitoring to ensure they are working as expected. AI, artificial intelligence; DSS, decision support systems.

approaches to other personalised treatment problems such as optimising a heparin loading regime to maximise time spent within the therapeutic range<sup>23</sup> or targeting blood glucose control in septic patients to minimise mortality.<sup>24</sup>

Looking further ahead AI systems may develop that go beyond recommendation of clinical action. Such systems may, for example, autonomously triage patients or prioritise individual's access to clinical services by screening referrals. Such systems could entail significant ethical issues by perpetuating inequality,<sup>25</sup> analogous to those seen in the automation of job applicant screening,<sup>26</sup> of which it is said that 'blind confidence in automated e-recruitment systems could have a high societal cost, jeopardizing the right of individuals to equal opportunities in the job market'. This is a complex discussion and beyond the remit of this article.

Outside of medicine, the cutting edge of AI research is focused on systems that behave autonomously and continuously evolve strategies to achieve their goal (active learning), for example, mastering the game of Go,<sup>27</sup> trading in financial markets,<sup>28</sup> controlling data centre cooling systems<sup>29</sup> or autonomous driving.<sup>30 31</sup> The safety issues of such actively learning autonomous systems have been discussed theoretically by Amodei *et al*<sup>32</sup> and from this work we can identify potential issues in medical applications. Autonomous systems

are long way off practical implementation in medicine, but one can imagine a future where 'closed loop' applications, such as subcutaneous insulin pumps driven by information from wearable sensors,<sup>33</sup> or automated ventilator control driven by physiological monitoring data in intensive care,<sup>34</sup> are directly controlled by AI algorithms.

These various applications of ML require different algorithms, of which there are a great many. Their performance is often very dependent on the precise composition of their training data and other parameters selected during training. Even controlling for these factors some algorithms will not produce identical decisions when trained in identical circumstances. This makes it difficult to reproduce research findings and will make it difficult to implement 'off the shelf' ML systems. It is notable in ML literature that there is not yet an agreed way to report findings or even compare the accuracy of ML systems.<sup>35 36</sup>

Figure 1 summarises expected trends in ML research in medicine, over the short, medium and longer terms, with the focus evolving from reactive systems, trained to classify patients from gold standard cases, with a measurable degree of accuracy, to proactive autonomous systems which continuously learn from experience, whose performance is judged on outcome. Translation of ML research into clinical practice requires a robust demonstration that the systems function safely, and with this evolution different quality and safety issues present themselves.

## QUALITY AND SAFETY IN ML SYSTEMS

In an early AI experiment, the US army used ML to try to distinguish between images of armoured vehicles hidden in trees versus empty forests.<sup>1</sup> After initial success on one set of images, the system performed no better than chance on a second set. It was subsequently found that the positive training images had all been taken on a sunny day, whereas it had been cloudy in the control photographs—the machine had learnt to discriminate between images of sunny and cloudy days, rather than to find the vehicles. This is an example of an unwittingly introduced bias in the training set. The subsequent application of the resulting system to unbiased cases is one cause of a phenomenon called 'distributional shift'.

### Short-term issues

#### Distributional shift

Distributional shift<sup>32</sup> is familiar to many clinicians, who find previous experience inadequate for new situations, and have to operate, cautiously, outside of a 'comfort zone'. ML systems can be poor at recognising a relevant change in context or data, and this results in the system confidently continuing to make erroneous predictions based on 'out-of-sample' inputs.<sup>32</sup>

A mismatch between training and operational data can be inadvertently introduced, most commonly, as above, by deficiencies in the training data, but also by inappropriate application of a trained ML system to an unanticipated patient context. Such situations can be described as ‘out-of-sample’ input, and the need to cater for many such edge cases is described as the ‘Frame problem’<sup>25</sup> of AI.

The limited availability of high quality data for training, correctly labelled with the outcome of interest, is a recurrent issue in ML studies. For example, when data are available it may have been collected as ‘interesting cases’ and not representative of the normal, leading to a sample selection bias.<sup>6</sup> In another example, the outcome may be poorly defined (eg, pneumonia) and variably assigned by experts, leading to a training set with poor reproducibility, and no ‘ground truth’ to learn associations.<sup>9</sup>

Inappropriate application of an ML system to a different context can be quite subtle. De Fauw *et al*<sup>7</sup> discovered their system worked well on scans from one OCT machine, but not another, necessitating a process to normalise the data coming from each machine, before a diagnostic prediction could be made. Similarly we anticipate that the system for diagnosing skin malignancy,<sup>6</sup> which was trained on pictures of lesions biopsied in a clinic, may not perform as well when applied to the task of screening the general population where the appearance of lesions, and patient’s risk profile, is different.

In some cases, distributional shift is introduced deliberately. ML systems perform best when index cases and controls are approximately equal in the training set,<sup>37</sup> and this is not common in medicine. Imbalanced data sets may be ‘rebalanced’ by under-sampling or over-sampling, and without correction the resulting system will tend to over-diagnose the rare case.<sup>38</sup> Alternative approaches may ‘boost’ the significance of true positive or false negative cases depending on the application, which can lead, for example, to a model good for screening but poor for diagnosis.<sup>39</sup>

Over time disease patterns change, leading to a mismatch between training and operational data. The effect of this on ML models of acute kidney injury was studied by Davis *et al*,<sup>40</sup> who found that over time decreasing AKI incidence was associated with increasing false positives from their ML system, an example of prediction drift.

There are many different ML algorithms, and they perform differently under the challenge of distributional shift, and this ‘may lead to arbitrary and sometimes deleterious effects that are costly to diagnose and address’.<sup>41</sup> It is notable however that the sepsis detection system mentioned above<sup>20</sup> has been successfully tested in the different context of a community hospital<sup>5</sup> despite being trained in intensive care, a potential distributional shift, and

thus shows some capability of adaptation through ‘transfer learning’.<sup>38 42</sup>

#### Insensitivity to impact

In the comparison between ML systems and expert dermatologists performed by Esteva *et al*,<sup>6</sup> both humans and machines find it difficult to discriminate between benign and malignant melanocytic lesions, but humans ‘err on the side of caution’ and over-diagnose malignancy. The same pattern was not observed for relatively benign conditions. While this decreases a clinician’s apparent accuracy, this behaviour alteration in the face of a potentially serious outcome is critical for safety, and something that the ML system has to replicate. ML systems applied to clinical care should be trained not just with the end result (eg, malignant or benign), but also with the cost of both potential missed diagnoses (false negatives) and over-diagnosis (false positives).<sup>43</sup> During learning ML systems assess and maximise their performance based on a measure of accuracy obtained on predictions made from training data. Often this accuracy measure does not take into account real-world impacts, and as a result the ML system can be optimised for the wrong task, and comparisons to clinician’s performance flawed.

#### Black box decision-making

One of the key differences between rule-based systems and the multitude of ML algorithms is the degree to which the resulting prediction can be explained in terms of its inputs. Some ML algorithms, particularly those based on artificial neural networks, make inscrutable predictions and for these algorithms it is harder to detect error or bias. This issue was demonstrated by the armoured vehicle detection system developed by the US army described above<sup>1</sup> and has been most studied in ML systems relying on image analysis.<sup>6 9</sup> To mitigate this, such systems can produce ‘saliency maps’ which identify the areas of, for example, the skin lesion<sup>6</sup> or the chest X-rays,<sup>9</sup> which most contributed to their prediction. However, outside of image analysis this inscrutability is harder to manage, and detection of bias in black box algorithms requires careful statistical analysis of the behaviour of the model in the face of changing inputs.<sup>44 45</sup>

#### Unsafe failure mode

The concept of confidence of prediction was mentioned in the context of distributional shift above. As with interpretability, not all ML algorithms produce estimates of confidence. If ML systems are opaque to interpretation, it becomes essential for the clinician to be aware whether the system believes its prediction is a sensible one. If the system’s confidence is low, best practice design would be to fail-safe<sup>46</sup> and refuse to make a prediction either way.

**Narrative review**

**Table 1** A general framework for considering clinical artificial intelligence (AI) quality and safety issues in medicine

Issue	Summary	Example
<b>Short term</b>		
Distributional shift	A mismatch between the data or environment the system is trained on and that used in operation, due to bias in the training set, change over time, or use of the system in a different population, may result in an erroneous 'out of sample' prediction.	The accuracy of a system which predicts impending acute kidney injury based on other health records data, became less accurate over time as disease patterns changed. <sup>40</sup>
Insensitivity to impact	A system makes predictions that fail to take into account the impact of false positive or false negative predictions within the clinical context of use.	An unsafe diagnostic system is trained to be maximally accurate by correctly diagnosing benign lesions at the expense of occasionally missing malignancy. <sup>6</sup>
Black box decision making	A system's predictions are not open to inspection or interpretation and can only be judged as correct based on the final outcome.	An X-Ray analysis AI system could be inaccurate in certain scenarios because of a problem with training data, but as a black box this is not possible to predict and will only become apparent after prolonged use. <sup>9</sup>
Unsafe failure mode	A system produces a prediction when it has no confidence in the prediction accuracy, or when it has insufficient information to make the prediction.	An unsafe AI decision support system may predict a low risk of a disease when some relevant data is missing. Without any information about the prediction confidence, a clinician may not realise how untrustworthy this prediction is. <sup>46</sup>
<b>Medium term</b>		
Automation complacency	A system's predictions are given more weight than they deserve as the system is seen as infallible or confirming initial assumptions.	The busy clinician ceases to consider alternatives when a usually predictable AI system agrees with their diagnosis. <sup>48</sup>
Reinforcement of outmoded practice	A system is trained on historical data which reinforces existing practice, and cannot adapt to new developments or sudden changes in policy	A drug is withdrawn due to safety concerns but the AI decision support system cannot adapt as it has no historical data on the alternative.
Self-fulfilling prediction	Implementation of a system indirectly reinforces the outcome it is designed to detect.	A system trained on outcome data, predicts that certain cancer patients have a poor prognosis. This results in them having palliative rather than curative treatment, reinforcing the learnt behaviour.
<b>Long term</b>		
Negative side effects	System learns to perform a narrow function that fails to take account of some wider context creating a dangerous unintended consequence.	An autonomous ventilator derives a ventilation strategy that successfully maintains short term oxygenation at the expense of long-term lung damage. <sup>34</sup>
Reward hacking	A proxy for the intended goal is used as a 'reward' and a continuously learning system finds an unexpected way to achieve the reward without fulfilling the intended goal.	An autonomous heparin infusion finds a way to control activated partial thromboplastin time (aPTT) at the time of measurement without achieving long-term control. <sup>23</sup>
Unsafe exploration	An actively learning system begins to learn new strategies by testing boundary conditions in an unsafe way.	A continuously learning autonomous heparin infusion starts using dangerously large bolus doses to achieve rapid aPTT control. <sup>23</sup>
Unscalable oversight	A system requires a degree of monitoring that becomes prohibitively time consuming to provide.	An autonomous subcutaneous insulin pump requires the patient to provide exhaustive detail of everything they have eaten before it can adjust the insulin regime. <sup>33</sup>

A similar fail-safe may be needed if the system has insufficient input information or detects an 'out-of-sample' situation as described above.<sup>46</sup>

**Medium-term issues**

**Automation complacency**  
 As humans, clinicians are susceptible to a range of cognitive biases which influence their ability to make accurate decisions.<sup>47</sup> Particularly relevant is 'confirmation bias' in which clinicians give excessive significance to evidence which supports their presumed diagnosis and ignore evidence which refutes it.<sup>25</sup> Automation bias<sup>48</sup> describes the phenomenon whereby clinicians accept the guidance of an automated system and cease searching for confirmatory evidence (eg, see Tsai *et al*<sup>49</sup>), perhaps

transferring responsibility for decision-making onto the machine—an effect reportedly strongest when a machine advises that a case is normal.<sup>48</sup> Automation complacency is a related concept<sup>48</sup> in which people using imperfect DSS are least likely to catch errors if they are using a system which has been generally reliable, they are loaded with multiple concurrent tasks and they are at the end of their shift.

Automation complacency can occur for any type of decision support, but may be potentiated when combined with other pitfalls of ML described above. For example, given the sensitivity to distributional shift described, the usually reliable ML system that encounters an out-of-sample input may not 'fail safely' but continue confidently to make an erroneous prediction of low malignancy risk and not be

questioned by the busy clinician who then ceases to consider alternatives.

#### Reinforcement of outmoded practice and self-fulfilling predictions

In the medium term, we expect to see systems emerging from research that use ML to recommend the most appropriate clinical actions, for example, by identifying patients who might benefit most from a specific treatment or for whom further referral and investigation is warranted.<sup>7</sup>

Such recommendation decision support already exists, but in systems whose behaviour is determined by explicitly designed rules. The shift to a data-driven approach introduces a new risk in the situation of a sudden change in clinical practice that requires the DSS to change, for example, a drug safety alert. While the rule-based system can be manually updated, as ML is predicated on the availability of appropriate data, it has the potential to reinforce outmoded practice, and a radical change that invalidates historical practice is difficult to absorb, as there are no prior data to retrain the system with. The need to periodically retrain and evaluate performance in response to technological evolution, new knowledge and protocol changes in medicine requires costly updating of gold standard data sets.

On the other hand, a related potential problem could arise in ML systems that are very frequently updated, and particularly those that continuously learn. Suppose a system predicts a prognosis, this may in turn influence therapy in a way that reinforces the prognosis and lead to a positive feedback loop. In this scenario, there is a self-fulfilling prediction, which then may be further reinforced as the ML system learns.

#### Longer-term issues

Table 1 incorporates Amodei *et al*'s framework for safety in AI,<sup>32</sup> which deals with issues more specific to continuously learning, autonomous systems. For obvious reasons, such systems will be challenging to deploy in the context of medicine and so their safety issues are less immediate. Rather than repeating Amodei *et al*'s detailed analysis,<sup>32</sup> we describe these issues using hypothetical scenarios based on the research into personalised heparin dosing mentioned above<sup>23</sup>:

- ▶ Negative side effects: The target of maximising the time in the therapeutic window requires careful management of heparin infusions that delay administration of other medications
- ▶ Reward hacking: An automated system may find ways in which to 'game' the goals defined by the reward function. The heparin dosing system, for example, might stumble on a strategy of giving pulses of heparin, immediately before activated partial thromboplastin time (aPTT) measurement, giving good short-term control,

### Box 1 - Quality control questions for short-term and medium-term issues in machine learning

#### Distributional shift

- ▶ Has the system been tested in diverse locations, underlying software architectures (such as electronic health records), and populations?
- ▶ How can we be sure the training data matches what we expect to see in real life and does not contain bias?
  - How can we be confident of the quality of the 'labels' the system is trained on?
  - Do the 'labels' represent a concrete outcome ('ground truth') or a clinical opinion?
  - How has imbalance in the training set been addressed?
  - Is the system applied to the same diagnostic context that it was trained in?
- ▶ How is the system going to be monitored and maintained over time to adjust for prediction drift?

#### Insensitivity to impact

- ▶ Does the system adjust its behaviour ('err on the side of caution') where there are high impact negative outcomes?
- ▶ Can the system identify 'out of sample' input and adjust its confidence accordingly?

#### Black box decision-making, unsafe failure and automation complacency

- ▶ Are the system's predictions interpretable?
- ▶ Does it produce an estimate of confidence?
- ▶ How is the certainty of prediction communicated to clinicians to avoid automation bias?

#### Reinforcement of outmoded practice and self-fulfilling predictions

- ▶ How can it accommodate breaking changes to clinical practice?
- ▶ What aspects of existing clinical practice does this system reinforce?

but without achieving the intended goal of stable long-term control. This is known as 'hacking the reward function' or 'wireheading'.<sup>32</sup>

- ▶ Unsafe exploration: As part of its continuous learning, the system may experiment with the dosing of heparin to try and improve its current behaviour. How do we set limits to prevent dangerous overdosing, and define what changes in strategy are safe for the system to 'explore'<sup>50</sup>?
- ▶ Unscalable oversight: As the system is learning new strategies for heparin management for novel patient groups, the management strategies it proposes require inconveniently frequent and expensive aPTT measurement.

At present these issues are merely theoretical in medicine, but they have been observed in ML test environments<sup>51</sup> and are increasingly becoming relevant in applications such as autonomous driving systems.<sup>31</sup>

**CONCLUSION**

Developing AI in health through the application of ML is a fertile area of research, but the rapid pace of change, diversity of different techniques and multiplicity of tuning parameters make it difficult to get a clear picture of how accurate these systems might be in clinical practice or how reproducible they are in different clinical contexts. This is compounded by a lack of consensus about how ML studies should report potential bias, for which the authors believe the Standards for Reporting of Diagnostic Accuracy initiative<sup>52</sup> could be a useful starting point. Researchers need also to consider how ML models, like scientific data sets, can be licensed and distributed to facilitate reproduction of research results in different settings.

As ML matures we suggest a set of short-term and medium-term clinical safety issues (see [table 1](#)) that need addressing to bring these systems from laboratory to bedside. This framework is supported by a set of quality control questions ([Box 1](#)) that are designed to help clinical safety professionals and those involved in developing ML systems to identify areas of concern. Detailed mitigation of these issues is a large topic that cannot be addressed here, but is discussed by *Amodei et al.*<sup>32</sup> and *Varshney et al.*<sup>46</sup>

Implementation of ML DSS in the short term is likely to focus on diagnostic decision support. ML diagnostic decision support should be assessed in the same manner and with the same rigour as the development of a new laboratory screening test. Wherever possible a direct comparison should be sought to existing decision support or risk scoring systems—ideally through a randomised controlled trial as exemplified by *Shimabukuro et al.*<sup>42 53</sup>

As with all clinical safety discussions we need to maintain a realistic perspective. Suboptimal decision-making will happen with or without ML support, and we must balance the potential for improvement against the risk of negative outcomes.

**Acknowledgements** The authors thank David Chalkley, Deputy CCIO & IT Clinical Safety Lead, TSFT, for comments that greatly enhanced this article.

**Contributors** All authors discussed the concept of the article and RC wrote the initial draft. KTA, JD, TE, MP and LG commented and made revisions, DC critically reviewed the draft. All authors agreed with the final manuscript. RC is the guarantor.

**Funding** This article was funded by Engineering and Physical Sciences Research Council and the grant number is EP/N014391/1.

**Competing interests** None declared.

**Patient consent for publication** Not required.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any

purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <http://creativecommons.org/licenses/by/4.0>

**REFERENCES**

- 1 Dreyfus HL, Dreyfus SE. What artificial experts can and cannot do. *AI Soc* 1992;6:18–26.
- 2 Rao A, Verweij G, Cameron E. Sizing the prize: what's the real value of AI for your business and how can you capitalise? PwC. 2017. Available: <https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf>
- 3 Hall W, Pesenti J. Growing the artificial intelligence industry in the UK - GOV.UK. Department for digital, culture, media & sport and department for business, energy & industrial strategy. 2017. Available: [https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/652097/Growing\\_the\\_artificial\\_intelligence\\_industry\\_in\\_the\\_UK.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf)
- 4 Duggal R, Brindle I, Bagenal J. Digital healthcare: regulating the revolution. *BMJ* 2018;360:k6.
- 5 McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Qual* 2017;6:e000158.
- 6 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
- 7 De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342–50.
- 8 Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci* 2017:1–13.
- 9 Rajpurkar P, Irvin J, Zhu K. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv [cs.CV]. 2017. Available: <http://arxiv.org/abs/1711.05225>
- 10 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402.
- 11 Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. *N Engl J Med* 2018;378:981–3.
- 12 Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. *Arch Intern Med* 2003;163:1409–16.
- 13 Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336:1475–82.
- 14 Bouch DC, Thompson JP. Severity scoring systems in the critically ill. *Cont Edu Anaesth Crit Care Pain* 2008;8:181–5.
- 15 Koppel Ret al. Role of computerized physician order entry systems in facilitating medication errors. *JAMA* 2005;293:1197–203.
- 16 Han YY et al. Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. *Pediatrics* 2005;116:1506–12.
- 17 Miller RA. Medical diagnostic decision support systems--past, present, and future: a threaded bibliography and brief commentary. *J Am Med Inform Assoc* 1994;1:8–27.
- 18 Nurek M, Kostopoulou O, Delaney BC, et al. Reducing diagnostic errors in primary care: a systematic meta-review of computerized diagnostic decision support systems by the

- LINNEAUS collaboration on patient safety in primary care. *Eur J Gen Pract* 2015;21(sup1):8–13.
- 19 Bond WF, Schwartz LM, Weaver KR, *et al.* Differential diagnosis generators: an evaluation of currently available computer programs. *J Gen Intern Med* 2012;27:213–9.
  - 20 Calvert JS, Price DA, Chettipally UK, *et al.* A computational approach to early sepsis detection. *Comput Biol Med* 2016;74:69–73.
  - 21 Desautels T, Calvert J, Hoffman J, *et al.* Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med Inform* 2016;4:e28.
  - 22 Thompson RF, Valdes G, Fuller CD, *et al.* Artificial intelligence in radiation oncology: a specialty-wide disruptive transformation? *Radiother Oncol* 2018;129:421–6.
  - 23 Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach 23. In: *2016 38th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016: 2978–81.
  - 24 Weng W-H, Gao M, He Z. Representation and reinforcement learning for personalized glycemic control in septic patients. arXiv [cs.LG]. 2017. Available: <http://arxiv.org/abs/1712.00654>
  - 25 K-H Y, Kohane IS. Framing the challenges of artificial intelligence in medicine. *BMJ Qual Saf* 2019;28:238–41.
  - 26 Faliagka E, Tsakalidis A, Tzimas G. An integrated e-recruitment system for automated personality mining and applicant ranking. *Internet Research* 2012;22:551–68.
  - 27 Silver D, Huang A, Maddison CJ, *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* 2016;529:484–9.
  - 28 Nuti G, Mirghaemi M, Treleven P, *et al.* Algorithmic trading. *Computer* 2011;44:61–9.
  - 29 Evans R (deepmind), Gao J (deepmind). DeepMind AI reduces google data centre cooling bill by 40%. Available: <https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/>
  - 30 Office of the Assistant Secretary for Research and Technology. Automated driving systems 2.0 A vision for safety. national highway traffic safety administration. 2017. Available: <https://www.nhtsa.gov/document/automated-driving-systems-20-voluntary-guidance>
  - 31 IIHS Status Report newsletter. 2018. Available: <https://www.iihs.org/externaldata/srdata/docs/sr5304.pdf>
  - 32 Amodei D, Olah C, Steinhardt J. *Concrete problems in AI safety*. arXiv [cs.AI]. 06565, 2016.
  - 33 Bothe MK, Dickens L, Reichel K, *et al.* The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Expert Rev Med Devices* 2013;10:661–73.
  - 34 Prasad N, Cheng L-F, Chivers C. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. arXiv [cs.AI]. 2017. Available: <http://arxiv.org/abs/1704.06300>
  - 35 Forman G, Scholz M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter* Published Online First. 2010. Available: <https://dl.acm.org/citation.cfm?id=1882479>
  - 36 Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr* 2008;17:145–51.
  - 37 Haixiang G, Yijing L, Shang J, *et al.* Learning from class-imbalanced data: review of methods and applications. *Expert Syst Appl* 2017;73:220–39.
  - 38 Storkey AJ. When Training and Test Sets are Different: Characterising Learning Transfer. In: Lawrence CSS, ed. *Dataset shift in machine learning*. MIT Press, 2013: 3–28.
  - 39 Bae S-H, Yoon K-J. Polyp detection via imbalanced learning and discriminative feature learning. *IEEE Trans Med Imaging* 2015;34:2379–93.
  - 40 Davis SE, Lasko TA, Chen G, *et al.* Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc* 2017;24:1052–61.
  - 41 Sculley D, Phillips T, Ebner D. Machine learning: the high-interest credit card of technical debt. 2018. Available: <https://research.google.com/pubs/pub43146.htmlhttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.675.9675> [Accessed 5 Mar 2018].
  - 42 Mao Q, Jay M, Hoffman JL, *et al.* Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open* 2018;8:e017833.
  - 43 Megler V, Gregoire S. Training models with unequal economic error costs using Amazon sagemaker. AWS machine learning blog. 2018. Available: <https://aws.amazon.com/blogs/machine-learning/training-models-with-unequal-economic-error-costs-using-amazon-sagemaker/> [Accessed 19 Oct 2018].
  - 44 Adler P, Falk C, Friedler SA, *et al.* *Auditing black-box models for indirect influence*. arXiv [stat.ML], 2016.
  - 45 Caruana R, Lou Y, Gehrke J. Intelligent models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. NY, USA: ACM, 2015: 1721–30.
  - 46 Varshney KR. Engineering safety in machine learning. In: *2016 Information Theory and Applications Workshop*. ITA, 2016: 1–5.
  - 47 Dawson NV, Arkes HR. Systematic errors in medical decision making: judgment limitations. *J Gen Intern Med* 1987;2:183–7.
  - 48 Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional Integration. *Hum Factors* 2010;52:381–410.
  - 49 Tsai TL, Fridsma DB, Gatti G. Computer decision support as a source of interpretation error: the case of electrocardiograms. *J Am Med Inform Assoc* 2003;10:478–83.
  - 50 Garcia J, Fernandez F, Fern F. Safe exploration of state and action spaces in reinforcement learning. *J Artif Intell Res* 2012;45:515–64.
  - 51 Leike J, Martic M, Krakovna V. AI safety gridworlds. arXiv [cs.LG]. 2017. Available: <http://arxiv.org/abs/1711.09883>
  - 52 Bossuyt PM, Reitsma JB, Bruns DE, *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clin Chem* 2003;49:1–6.
  - 53 Shimabukuro DW, Barton CW, Feldman MD, *et al.* Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir Res* 2017;4:e000234.

# **3. Factors influencing digital review of pathology test results in an inpatient setting: a cross-sectional study**

Status: published in JAMIA Open, Volume 3, Issue 2, p290-298

Date: 17th March 2020

Candidate's contribution: All authors discussed the concept of the article and RC conducted the analysis and wrote the initial draft. KT-A, TE, LG, MD, MP, and LD commented and made revisions. All authors read and approved the final manuscript. RC is the guarantor.

All the analysis in this paper was performed by me and findings validated by other team members.

## Research and Applications

# Factors influencing digital review of pathology test results in an inpatient setting: a cross-sectional study

Robert Challen <sup>1,2</sup> Krasimira Tsaneva-Atanasova,<sup>1,3</sup> Tom Edwards,<sup>2</sup> Luke Gompels,<sup>2</sup> Mark Dayer,<sup>2</sup> Martin Pitt,<sup>4</sup> and Leon Danon<sup>3,5</sup>

<sup>1</sup>EPSRC Centre for Predictive Modelling in Healthcare, University of Exeter, Exeter, Devon, UK, <sup>2</sup>Taunton and Somerset NHS Foundation Trust, Taunton, Somerset, UK, <sup>3</sup>The Alan Turing Institute, British Library, London, UK, <sup>4</sup>NIHR CLAHRC for the South West Peninsula, St Luke's Campus, University of Exeter Medical School, Exeter, UK, and <sup>5</sup>Data Science Institute, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

Corresponding Author: Robert Challen, MA, MBBS, Living Systems Institute, University of Exeter, Stocker Road, Exeter EX4 4QD, UK; rc538@exeter.ac.uk

Received 21 February 2019; Revised 27 November 2019; Editorial Decision 28 January 2020; Accepted 5 February 2020

### ABSTRACT

**Background:** Delay or failure to view test results in a hospital setting can lead to delayed diagnosis, risk of patient harm, and represents inefficiency. Factors influencing this were investigated to identify how timeliness and completeness of test review could be improved through an evidence-based redesign of the use of clinical test review software.

**Methods:** A cross-section of all abnormal hematology and biochemistry results which were published on a digital test review platform over a 3-year period were investigated. The time it took for clinicians to view these results, and the results that were not viewed within 30 days, were analyzed relative to time of the week, the detailed type of test, and an indicator of patient record data quality.

**Results:** The majority of results were viewed within 90 min, and 93.9% of these results viewed on the digital platform within 30 days. There was significant variation in results review throughout the week, shown to be due to an interplay between technical and clinical workflow factors. Routine results were less likely to be reviewed, as were those with patient record data quality issues.

**Conclusion:** The evidence suggests that test result review would be improved by stream-lining access to the result platform, differentiating between urgent and routine results, improving handover of responsibility for result review, and improving search for temporary patient records. Altering the timing of phlebotomy rounds and a review of the appropriateness of routine test requests at the weekend may also improve result review rates.

**Key words:** test result follow-up, quality improvement, clinical workflow, data quality, laboratory informatics

### INTRODUCTION

Large numbers of tests are undertaken in hospitals, and many health-care professionals are involved in the care of a patient from admission to discharge. There is evidence that some test results will never be reviewed, or if reviewed, never followed up on, in both primary and secondary care.<sup>1–3</sup> Estimates of the frequency of the failure to follow-up on test results are very variable, depending on setting,<sup>2,4,5</sup> and the

measurement methodology, but thought to occur somewhere between 1% and 22.9% of inpatient admissions,<sup>1</sup> and this is more frequent where there is a transfer of care between care settings. Other studies (summarized by Callen et al<sup>2</sup>) found even wider variation, with inappropriate or incomplete follow-up for 6.8% of abnormal test result alerts in 1 study, to 62% of abnormal glucose screening tests results in another study. This variability is in part due to differences in clinical



context, but also the exact nature of the definition of follow-up in these studies. Missed test results are found to be particularly problematic if there are test results which are pending after a patient is discharged.<sup>6–10</sup> The failure to review test results represents not only a certain degree of inefficiency, but also a potential clinical safety issue resulting in missed or delayed diagnoses,<sup>1</sup> for example, being a contributory cause for delayed diagnosis and treatment of lung cancer, and hepatitis C, among others.<sup>6,11</sup>

Investigations into the introduction of a mandatory test result acknowledgement system in a maternity unit in Australia<sup>12</sup> showed a wide variation in the length of time it took for clinicians to review and acknowledge test results. They found that the mean time it took for a newly available blood test result to be viewed was 19.1 h, but the median was just 3 h, whereas radiology results were reviewed on average 47.7 h after being published, with a median of 20.8 h,<sup>12</sup> suggesting that the distribution of time to view is heavily skewed. This variability can also contribute to a delay in diagnosis, and therefore action, as clinicians must first become aware of the abnormal result. The effect of directly alerting clinicians to abnormal pathology test results has been investigated in different studies<sup>9,13,14</sup> and, taken together, a reduction in the length of time taken for initiation of corrective therapy was seen, by reducing the delay before the clinician becomes aware of an abnormal result, or by making the clinician aware of abnormal results after the patient has gone home.<sup>7,9</sup>

This article describes a retrospective cross-sectional study of the clinical use of electronic pathology result review software in Taunton and Somerset NHS Foundation Trust (TSFT). This aims to describe the people, process and technology factors found to influence both the timeliness and the completeness of clinical review of test results. The study focused on a set of test results for which the overall completeness of review of test results is broadly in line with research above,<sup>2</sup> but with a focus on how this is affected by different system and workflow factors. The purpose of this is to identify how timeliness and completeness of test review could be improved through an evidence-based redesign of the use of clinical test review software within clinical workflow. This is relevant to people who design or implement software for pathology test review in a hospital setting. This study also serves as a baseline for assessing future changes to the software for pathology result review in TSFT.

The purpose of this article was not to determine the clinical significance of delays or failure to view test results, as there are numerous factors that affect clinical outcome that are not controlled for in this study, not least being that viewing a test result does not necessarily imply that an action has been taken as a result.

## METHODS

### Setting

The study was conducted on clinical tests from inpatients in Musgrove Park Hospital, which is part of TSFT. It is a district general hospital, providing care to a population of over 340 000. It also provides some specialist services for the whole of Somerset, making the catchment population around 544 000. The hospital has over 700 beds, 30 wards, 15 operating theaters, a fully equipped diagnostic imaging department, and a purpose-built cancer treatment center.<sup>15</sup> TSFT shares its pathology services with all the hospitals and primary care providers in Somerset.

### Information governance

The data required for this audit are routinely automatically collected as part of the pathology viewer software audit log. The data were

fully de-identified at source before being analyzed by TSFT staff on computers within the hospital's secure data center. As an internal audit of TSFT operations using nonidentifiable data it did not require patient consent. It was reviewed and approved by the hospital research and development office and Caldicott guardian (information governance lead). No patient identifiable information was retained by the data extraction.

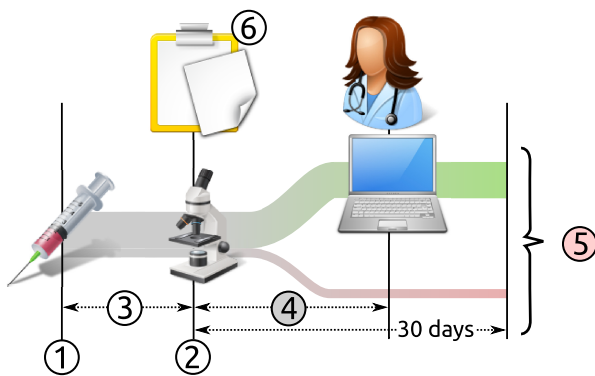
### Dataset

When a result is directly viewed on the system, the pathology viewer software records the time that the result was accessed in an audit log. There is also an indirect method for viewing test results in a timeline-based “grid view,” in which all of the most recent results are presented simultaneously. This indirect method of viewing results is also logged by the system, but not at the individual test result level. Based on feedback from clinical users of the system, the assumption was made that clinical review of results using the grid view is equivalent to the direct review of all test results released in the preceding 24 h.

The source database integrates results from many laboratories, and radiology departments, stemming from requests from many primary and secondary care providers. As such the data quality of patient records has issues, particularly in terms of the creation of temporary duplicate patient records which require subsequent merging. Such records may be created if a patient's contact details have changed or are miskeyed, or if a patient is admitted in an emergency where incomplete demographic details are available. During this subsequent merging process test results are reassigned to the canonical patient identifier when possible. Prior to the merge, test results filed under temporary patient record may be less easy to find than those correctly filed from the outset. Record merging may happen before, during, or after results become available on the system, but the results are only reassigned after they have been published. The reassignment of a test during this merge process was extracted as an experimental variable and is an indicator of the initial data quality of the patient record.

### Selection criteria and study size

The existing laboratory results viewer system contains results from the last 10 years for the whole of Somerset, involving inpatients and outpatients over multiple hospital sites and primary care providers. These were limited to tests that had been requested from inpatient wards, and emergency ambulatory care in TSFT between September 1, 2014, and September 30, 2017, during which no major changes occurred in either the test requesting and review processes, or in the laboratory test result review platform itself. The set was constrained to the 1 770 775 million biochemistry and hematology reports that were reported as abnormal on the system, and for these results, there were no missing data items. Abnormal results are visible as such on the test review system and such abnormal results represent investigations that should have clinician review. Finally, unlike radiology reports which are available on the picture archive and communications platform, biochemistry and hematology test results are only available digitally through the results review platform. They are less widely reviewed via paper reports in TSFT than microbiology, as the paper reports are only distributed some hours after the digital report is available. This delay makes the paper hematology and biochemistry reports less clinically useful than digital reports, but not so for microbiology given the longer turn-around times associated with cultures, and comparatively more telephone alerting of clinicians to abnormal results.



**Figure 1.** At time point (1) a specimen and associated request starts being processed at the laboratory. At (2) the result is made available on the electronic results system. The processing time (3) depends on various factors including the type of test. After a period of time (4) a result may be viewed by one or more clinicians—the “time to view.” Some results are not reviewed digitally or reviewed very late (5) (30-day cut-off was applied)—these results are deemed to be “unviewed.” Test results, however, may have been viewed through parallel routes such as paper reports (eg, microbiology) (6) or second dedicated systems (eg, radiology).

### Data analysis

Dependent and experimental variables were extracted from the source database using a custom structured query language routine. The main dependent variables investigated were the time for the first clinician to view a test result after it becomes available on the system (time to view) and the proportion of tests which were not digitally viewed on the pathology results viewing system within 30 days (unviewed tests) which included 99.9% of all test views. These outcomes are described in Figure 1. Qualitatively the timeliness of test review is related to the overall distribution of the time to view, but is essentially quantified here as the median of the time to view all the tests in a given sample (Timeliness = median of time to view a test result). Completeness of test result review for a given sample can be quantified as the complement of the unviewed tests (Completeness =  $1 - \text{Number of tests not viewed} / \text{Total number of test results}$ ). Test results were described at the level of the clinically ordered test battery, so this study regards the combined test of “Full Blood Count” as a single test, rather than the individual component “Hemoglobin level.” Only the most up to date revision of a test result was included.

The study stratified the sample by a number of different experimental variables and visualized the differences in time to view and percentage of test unviewed after 30 days by these experimental variables. Among many other factors considered as experimental variables, we report here on temporal variation, the detailed test type of the result, and the test reassignment status as defined above. Temporal variation was investigated using the time within the week when the test result was published as an experimental variable and it was hypothesized that tests may be reviewed less quickly when staffing levels are lower overnight, or during the weekend. The detailed type of test result was selected as clinical tests have different clinical impacts and it was expected that some test types will be monitored more closely by clinicians. Test reassignment was investigated as a marker of data quality of the patient record. It is expected that reassignment of tests to different patient identifiers is a proxy measure for the ease with which clinicians find test results, and hence would be related to how many results are missed.

All data were analyzed and visualized using R.<sup>16</sup> Correlation strengths are estimated with Cramér’s V coefficient<sup>16</sup> and Pairwise

Wilcoxon tests.<sup>17</sup> Distributions of time to view were compared with Kruskal–Wallis tests.<sup>18</sup> Instantaneous publish and view rates were calculated using piecewise polynomial fitting to cumulative rates using a Savitzky–Golay filter.<sup>18</sup> Continuous distributions of median time to view were calculated from a rolling 2 h window, and error estimates median time to view estimated using case resampling bootstrapping.<sup>17</sup> Estimates of error in unviewed tests were determined by the size of the group, assuming a binomial distribution.

## RESULTS

### Sample characteristics

There were 1 770 775 abnormal biochemistry and hematology test results released during the study period (approximately 1600 per day). Figure 2 shows the difference in daily test result rates between weekdays and weekends and demonstrates a steady increase over the study period.

Figure 3 shows the rates of test publication and test review during the day, split between weekday and weekend. A peak of activity in test publication is noted at midday. The weekend test publication pattern could represent a similar pattern with a smaller midday peak.

Of the test results that were published, 7.69% ( $N = 136\,175$ ) were subsequently reassigned to another patient record during a duplicate record merge process, as described in the “Methods.” This leaves 92.31% ( $N = 1\,634\,600$ ) which were correctly assigned to the canonical patient record at publication.

### Summary of dependent measures

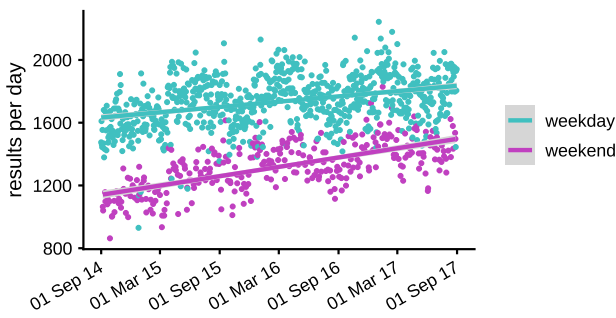
The time to view abnormal biochemistry and hematology test results is shown in Figure 4. The distribution is heavily skewed with a median of 89.8 min (interquartile range 33.9–213.8 min). This median has stayed constant during the study period (see Supplementary Table S1). Over the 30-day period after each of the 1 770 775 abnormal biochemistry and hematology test results were made available, 1 662 067 results were viewed on the digital platform (93.9%) with the remaining 108 708 (6.1%) of all test results being unviewed at 30 days (Figure 4). An unknown proportion of these 108 708 test results will have had clinical review of the paper copy. In the logarithmic time plot in Figure 4, 3 behavioral regimes are observed (1) a fast, exponential rise to a peak value at around 30 min, (2) an exponential decline between 30 min, and (3) a long tail with a small second peak of test viewing activity at 24 h.

### Detailed test result type

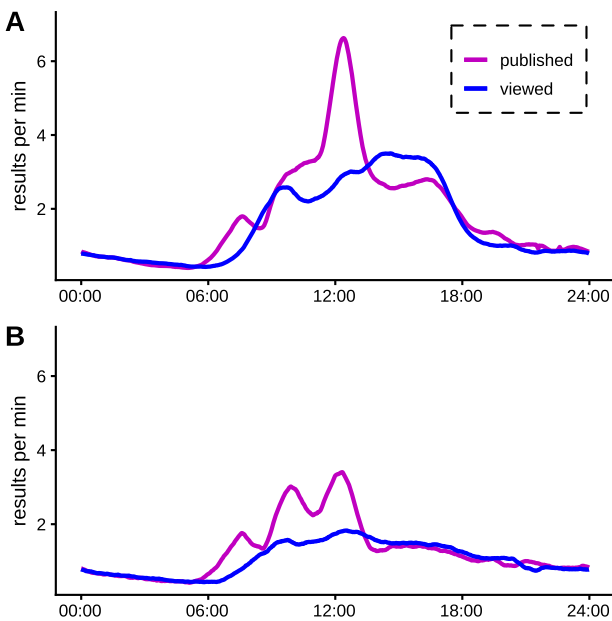
The 30 most frequently performed tests are shown in Figure 5 and a qualitative difference is apparent between tests such as D-dimer, or troponin I assays, which have a high clinical impact in the short term, and tests such as ferritin, or vitamin B12 levels, abnormalities of which have longer-term sequelae, and management. Both the speed and the likelihood that a clinician is to review a result varies depending on the test type. Correlation strengths are estimated with Cramér’s V coefficient<sup>19</sup> for the relationship between test type and whether a result has been viewed is 0.585. A Kruskal–Wallis<sup>20</sup> test demonstrated a significant difference between distributions of time to view ( $\chi^2 = 20\,476$ ,  $P < 2.2 \times 10^{-16}$ ). Pairwise Wilcoxon tests<sup>21</sup> demonstrated some correlated test types, where the results of a test are frequently reported together (eg, Cal/alb ratio and corrected calcium).

### Time of the week of test result publication

The rate at which tests are published and reviewed (Figure 4) demonstrated a daily rhythm of the test request and review process.



**Figure 2.** The rate of abnormal biochemistry and hematology test results per day during the study period.

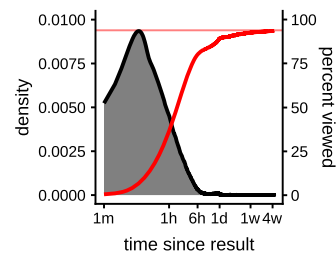


**Figure 3.** Publish and view rates over the course of the day for test results, for weekdays (A) and weekends (B).

The impact of this daily pattern on the time taken to view tests and the likelihood a test is not viewed was investigated and the results are shown in Figure 6. The density plot in the top panel has notable features. From Monday to Friday, the bulk of test results are viewed during the first six hours after becoming available, but the delay in reviewing a test increased throughout the morning. A test released at midday took twice as long to be reviewed as those made available at 6 AM (A). Through the afternoon and overnight, the delay reviewing a result dropped again, but the chance of the test being reviewed the following day increased (widening the interquartile range of time to view) (B). On Friday, the pattern is slightly different in that the results that became available in the afternoon and not seen the same day were more likely to be viewed on Monday morning (C).

The variation of the median time to view over the week is shown in the middle panel of Figure 7 for completeness as the linear plot better demonstrates the effect size and the relative errors of our estimates. It shows the increasing delay in test result viewing over the course of the morning, recovering by mid-afternoon, followed by a smaller second increase at the end of the afternoon as a proportion of results were left until the next day.

The proportion of tests that are not viewed is shown in the bottom plot of Figure 6. Tests conducted overnight when staffing levels were



**Figure 4.** The distribution of the time to view a result within the first 12 h of it becoming available (grey) shows a heavily skewed distribution with a long tail. The cumulative proportion of tests viewed is shown in red. Time is plotted on a logarithmic scale.

low were more likely to be looked at than those in the middle of the day (point D). As with the time to view, the chance of a test being missed increased if it was released later in the afternoon during the week. The most significant increase in unviewed tests occurred on Friday afternoon when a released result has just under 10% chance of not being seen (E). On the weekend the pattern was slightly different, with tests that were performed in the middle of the day being less likely to be viewed (F).

### Test result reassignment

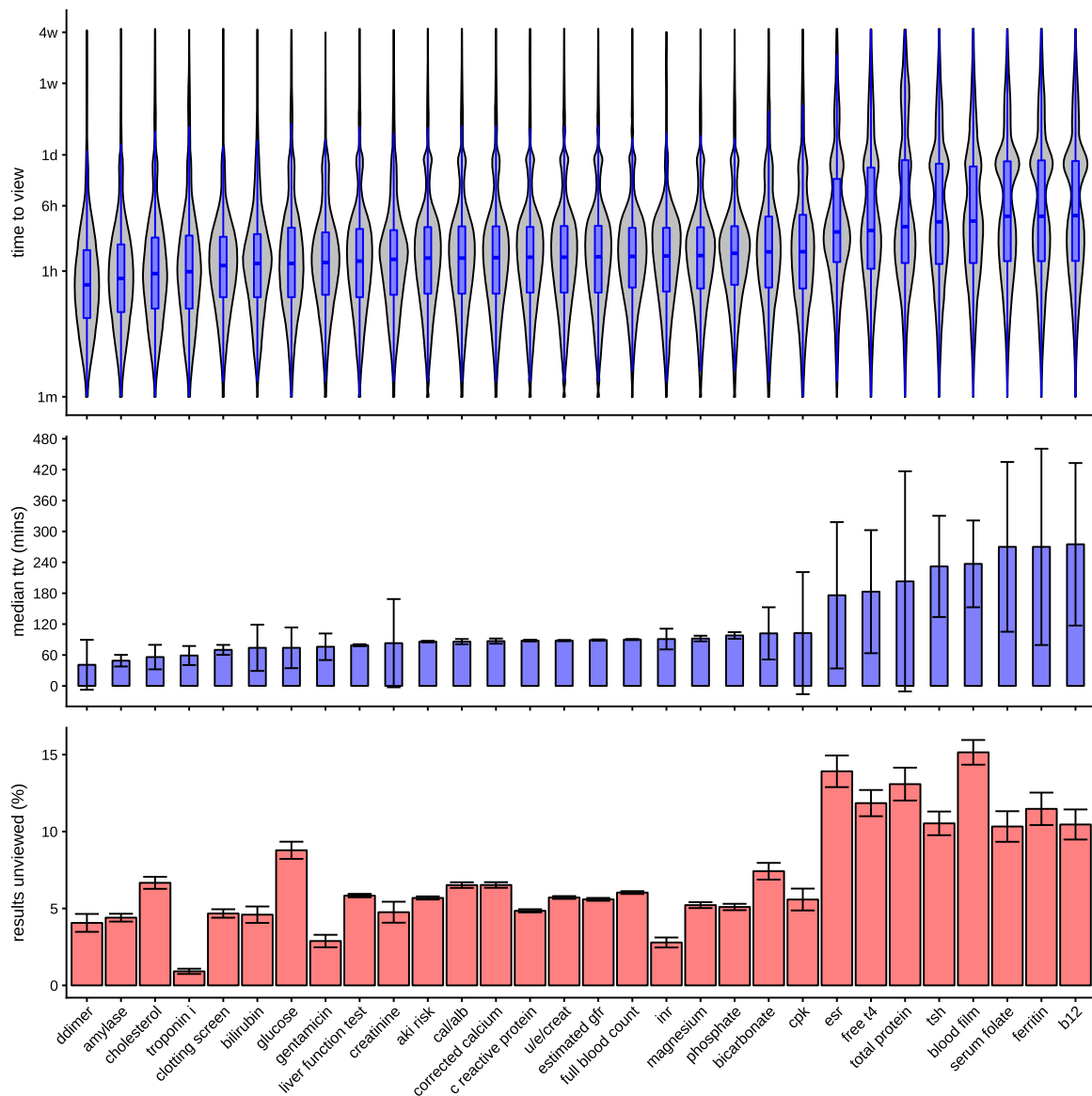
As described in the “Methods” above, test result reassignment may be an indication that a test result was difficult to find and less likely to be reviewed. The patient identifier was revised in 136 177 out of 1 770 775 of test results (7.69%) of which only 2386 (1.78%) represented revision to a completely different patient, for example, due to error ordering or processing the test. The remainder was revisions to a different identifier of the same patient and represents a data quality issue encountered during integration of new test results. Although this had only a small effect on the time to view the results, there was a significant effect on the proportion of results not viewed, with an additional 4% of the revised tests not being looked at within 30 days, compared to those which are not revised (see Figure 7).

## DISCUSSION

### Principal findings

For inpatients within TSFT, blood tests are typically performed during admission and throughout the patient’s stay as determined by the clinical team. Dedicated phlebotomists generally pick up requests, and draw blood for blood tests during the early part of the morning on both weekdays and a limited service at weekends. These will usually not be urgent test requests, and the request for the test may have been made the night before. If an urgent test is required blood will be drawn by a member of the clinical team. This system explains the morning peak in the volumes of tests pattern shown in Figure 3 when a high proportion of test results will be from routine requests.

Typically, particularly for general inpatients, the clinical team will conduct a ward round during the morning during which they will review patients, including the available blood test results if possible. This ward round is largely conducted using paper notes, and during this time the clinical team has variable access to desktop PC bound resources, including the laboratory test review software. Ward rounds generally finish during the course of the morning, after which the junior members of the clinical team have better access to digital systems.



**Figure 5.** Time to view distribution (top), median time to view (middle), and proportion unviewed (bottom) for the 30 most common tests. The top violin plot shows the distribution of time to view in the first 30 days from publication on a logarithmic timescale, for each test. In the middle plot, the same median is shown with 95% confidence intervals for the median calculated using a bootstrap estimation. In the lower plot, the absolute proportion of tests that are unviewed is shown with 95% confidence intervals determined by the size of the group, assuming a binomial distribution.

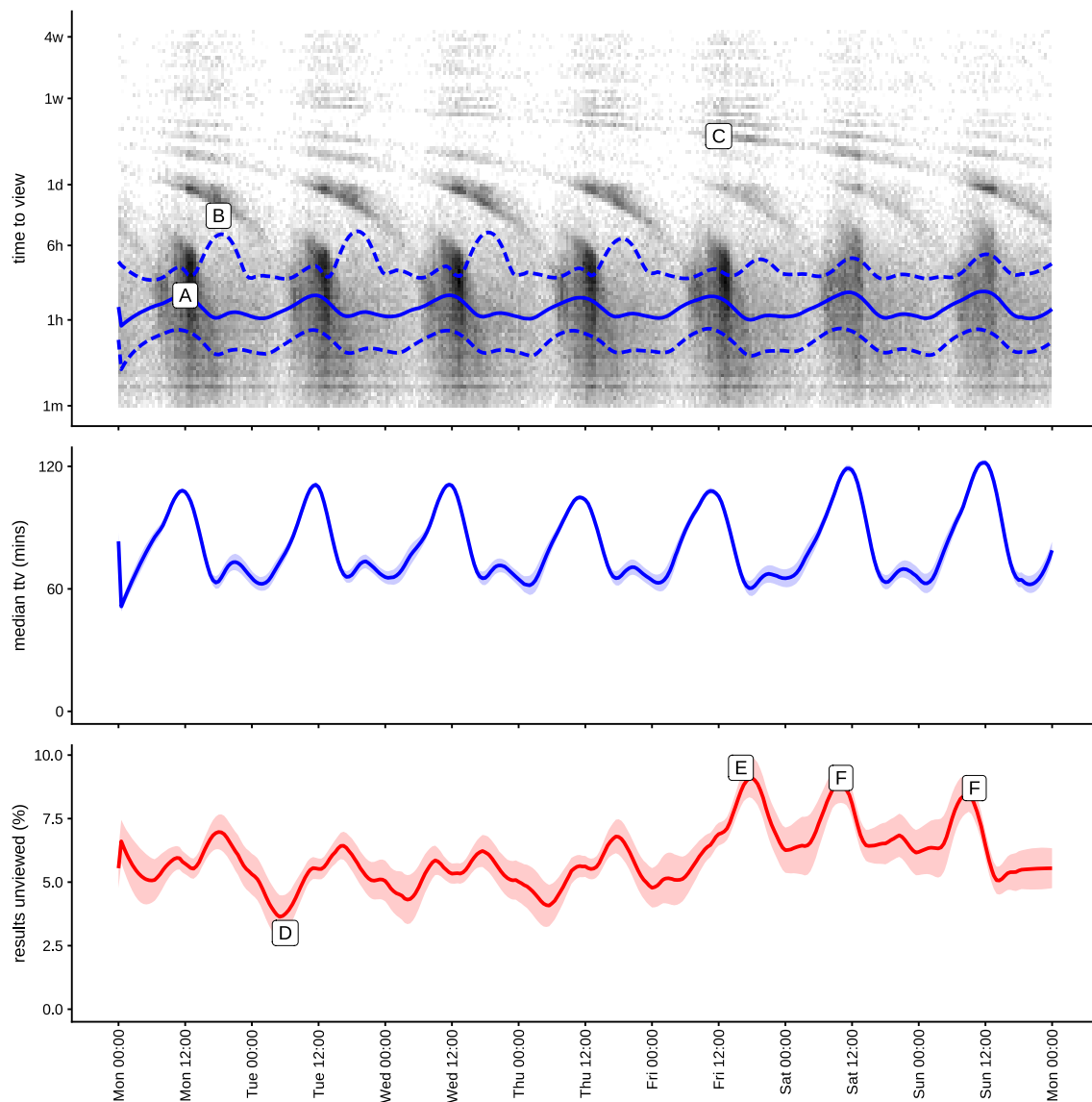
Test results may be reviewed by doctors, midwives, or any other clinical staff caring for the patient. However, the bulk of the results review, particularly of routinely requested tests from that day, will be performed by the junior clinical staff, typically during the afternoon. We believe that some of the increased delay in test result review in the middle of the day seen in Figures 3 and 6 is a result of the limited access that junior members of the clinical team have to the results platform during the ward round. Seriously abnormal results may be phoned through to ward staff, who will then alert a clinician to the result, who in turn will typically review the patient's results.

Outside of weekday working hours, the hospital runs a shift system for clinical staff.<sup>22</sup> This will typically involve a handover of clinical care around 5 PM to an evening shift, and again at approximately 10 PM to an overnight shift. In general, these teams will only be involved with acutely sick patients and will not consider routine test results. This handover of responsibility, and shift away

from routine, to urgent care, is likely to be the main driver for the increased delay for test results that are published later in the day seen in Figure 6.

The possibility of a strong effect of routine versus urgent testing on both completeness and timeliness of test result review is supported by the analysis of detailed test type shown in Figure 5. Tests for significant acute medical conditions such as D-dimer tests for thrombosis, or troponin tests for myocardial infarction, are more rapidly and completely reviewed than tests that are long term in nature, such as investigations for iron deficiency anemia. One possible conclusion is that investigation of longer-term health problems in inpatients is less efficient than the investigation of an acute health problem, but this also points to some prioritization of clinical review of test results that are urgent in nature.

Finally, this study found evidence in Figure 7 that data quality has a significant effect on the result review, in that test results which



**Figure 6.** Time to view distribution (top), median time to view (middle), and proportion unviewed (bottom) for the tests with relation to the time of the week. Darker colors of the top plot represent higher number of tests released at a given time of the week and how those are distributed over the time taken to view them. The blue lines showing the median (solid) and interquartile ranges (dashed) of time to view on a logarithmic timescale. The banding pattern represents results being reviewed on subsequent days. In the middle panel the same median time to view is shown in a linear time plot with very narrow 95% confidence intervals estimated using bootstrapping. The bottom plot shows the percent of tests that are not viewed, broken down by when they are published. We see high rates of test review at point (A), for test results published at midday on Monday. Results published later in the day may not be seen until the following day (B). This effect is more pronounced on Friday where results may not be seen until the following Monday (C). Tests are most likely to be viewed if they are published early on Tuesday (D) and least likely to be viewed if they are published on Friday evening (E) or during the day in the weekend (F).

have had to be reassigned to a duplicate patient identifier during a merge, were less likely to be reviewed. This is evidence that efforts should be directed at not only improving data quality, but also that further work is needed to understand what mitigations need to be in place for finding hard to identify and temporary patient records.

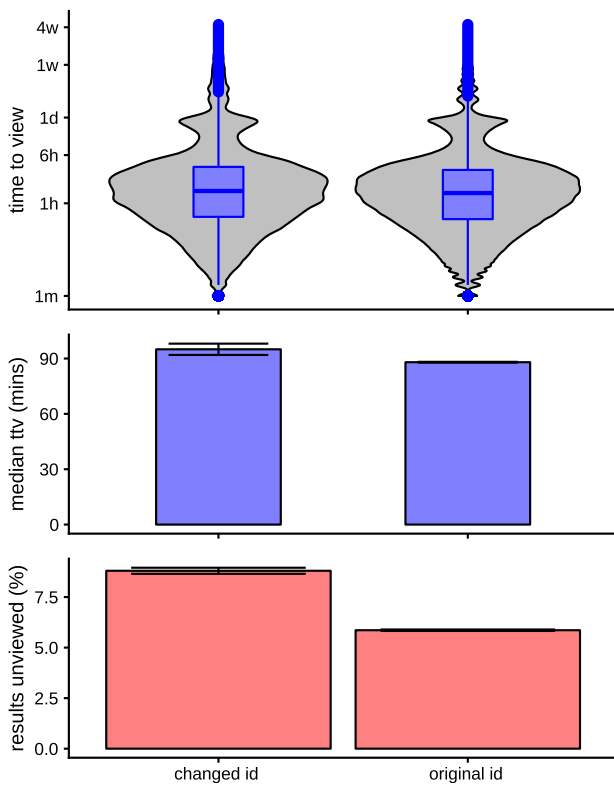
**Limitations**

This study has several drawbacks due to the nature of information available and quality of some fields in the source database. Particularly information about the identity of test requesters reviewers was

not well standardized, and neither was information about the precise location of ordering and reviewing results.

As a retrospective observational data study we were unable to collect hard data about the workflow of test requesting, sample collection, test reviewing, and clinical ward rounds. Instead all the assumptions about the nature of the workflow have been reviewed for accuracy by 6 independent TSFT staff, consisting of both frontline clinicians and members of the digital implementation team in TSFT involved in the deployment and maintenance of the digital test result platform.

Evidence suggests that there is a strong effect of the patient’s admission and discharge on the timeliness and completeness of test re-



**Figure 7.** Time to view distribution (top), median time to view (middle), and proportion unviewed (bottom) for the tests that have had their patient identifier changed versus those that retain their original patient identifier. Tests that have been updated and rematched to a “better” patient record are less likely to be viewed. 95% confidence intervals are shown on the graph. *P*-values for both comparisons are smaller than numerical precision of the calculating software due to the large numbers of tests involved.

sult review.<sup>6–10</sup> A linked patient admission, discharge, and transfer dataset was not available and so this study is unable to confirm previous findings about the significance of tests that are pending at discharge.<sup>5,7–9</sup> As patient discharge typically occurs in the later part of the day, a strong relationship to patient discharge could indirectly explain some of the observed patterns. This would warrant investigation as one possible confounder for the temporal patterns observed. Ineffective follow-up of test results pending at discharge has previously been shown to be a cause of diagnostic delay,<sup>2,6</sup> but as this was not available for analysis it should be a focus of further work.

As indicated in Figure 1, the digital review of test results is not the only mechanism through which results are viewed. Biochemistry and hematology results are printed on paper, delivered to requesting location, where the paper results may be reviewed, but these are subsequently shredded and not filed in the patient record. Paper-based review is quite variable across the hospital and is not recorded, but it should be expected that the estimate of unviewed tests presented here was at the upper end of the true value. Similarly, a test result can be clicked but not actually reviewed due to some distraction in practice, or a test result may need to be viewed multiple times to complete the review process.

Because of issues associated with parallel review processes we did not consider microbiology and radiology test results in our analysis. The complexity of result review in microbiology is described by Bruins et al<sup>23</sup> and this analysis focused on the simpler problem of

biochemistry and hematology test results. However, many of the examples in the literature that describe the clinical impact of missed test results focus on microbiology and radiology<sup>5,9</sup> and hence this reduces the comparability of our findings to previous studies.

Viewing a test result does not imply that appropriate clinical action has been taken, or that the viewing clinician has taken responsibility for follow-up. There were no clinical outcomes outside of lab results available in our dataset; investigation was conducted to see whether repeat testing could be used as a marker for clinical action as in Lin and Moore,<sup>24</sup> but this was uninformative. It has not been possible to follow-up on the clinical significance of abnormal test results, and there is no simple way to determine whether an unviewed test could have affected the outcome, and whether an unviewed test result itself would have provided useful information to a clinical decision-making process.

### Suggestions for improvement

It is to be expected that many of these findings, particularly around workflow, are specific either to TSFT itself or to hospitals in the National Health Service (UK). However, the findings suggest some improvement can be made in the efficiency of test review:

1. The competing demands for a clinician’s attention during the morning ward round and potential difficulty accessing a static desktop PC bound resource while at the bedside could be improved by providing access to test results as soon as they become available through mobile devices.
2. Test ordering and review systems should consider differentiating between urgent and routine requests in both alerting the clinician of the availability of a result and informing them of unexpected abnormality.
3. Further technical measures could be adopted to better alert clinicians to abnormal results at times of the day when test are likely to be missed, for example, by more proactive alerting clinicians to abnormal results on Friday afternoons, or by requiring mandatory sign off of abnormal results.<sup>12</sup>
4. Further improvement might be expected if the clinical review software was enhanced to allow better handover of clinical responsibility for review of abnormal test results between clinicians working different shifts. Suboptimal transfer of care between shift working teams within the hospital has previously been identified as an area of high risk.<sup>25–27</sup>
5. Improvements in patient record search prior to test ordering may improve data quality. This may reduce the number of tests reported under suboptimal duplicate patient records and which are subsequently lost in the system. However, better search for temporary records in the result viewing platform is also warranted.
6. Test result review would be more efficient if fewer results were pending during the clinical ward rounds. This might be achieved, for example, by changing the timing of the phlebotomy ward rounds to ensure results are available before clinical ward rounds.
7. Spikes in the rates of unviewed tests observed at the weekend could represent abnormalities in routine monitoring tests requested by the regular team, which are not being picked up by the weekend team, whose focus is on urgent care. These tests are arguably of low clinical value. A focus on appropriate test requesting at the weekends may be indicated.

## CONCLUSIONS

This study presents a detailed analysis of a large sample of electronic test results. The volume enables a closer look at workflow patterns during the day than has been presented previously, and this shows several patterns which are thought to be related to whole system workflow, information availability and handover of clinical responsibility, which have not, to our knowledge, been demonstrated before in the literature. To some degree the effect of shift handover may be explained by a transition from urgent to routine care, as previously described.<sup>28</sup>

The study identifies that the data quality of the patient record can influence the ease of locating a test result and increase the likelihood that a result will go unreviewed. This is a novel finding that is, so far, specific to TSFT but which would be useful to validate in another setting. Taken together these findings allow the recommendation of various changes to the systems and workflow of test review in TSFT as outlined above and the authors believe these are of relevance to other hospitals.

TSFT is in the process of implementing a mobile device-based pathology results viewer. The effect of this intervention on the workflow factors presented here may be significant as test results will be more easily available, and this will be monitored using this study as a baseline. We anticipate the recommendations presented in this study will inform future technical development and implementation of test result review platforms.

This is a retrospective observational study of a single site and our conclusions are limited by this. It would be useful to repeat this study in different hospitals to compare the results. It would be particularly valuable to look at a dataset that included the patient's admission, discharge, and transfer details to investigate the effect of handover and tests pending at discharge better. This study is designed to be a baseline to provide evidence for quality improvement initiatives and technical enhancements to the laboratory test reporting platform. Further studies should focus on periods before and after such interventions to more rigorously test the improvements suggested above.

## FUNDING

RC and KT-A gratefully acknowledges the financial support of the EPSRC via grant EP/N014391/1 and NHS England, Global Digital Exemplar Programme. LD gratefully acknowledges the financial support of The Alan Turing Institute under the EPSRC grant EP/N510129/1.

## AUTHOR CONTRIBUTIONS

All authors discussed the concept of the article and RC wrote the initial draft. KT-A, TE, LG, MD, MP, and LD commented and made revisions. All authors read and approved the final manuscript. RC is the guarantor. The views presented here are those of the authors and should not be attributed to TSFT or the GDE.

## ETHICS APPROVAL AND CONSENT TO PARTICIPATE

As an internal audit of TSFT operations using nonidentifiable data, this study did not require patient consent and has been exempted from formal NHS Health Research Authority approval and review by a research ethics committee. It was reviewed and approved by

TSFT research and development office and Caldicott guardian (information governance lead).

## AVAILABILITY OF DATA AND MATERIAL

The datasets generated and/or analyzed during the current study are not publicly available due to the remote possibility of targeted patient re-identification through a linkage attack. Limited data may be available from the corresponding author on reasonable request, and review by the information governance and ethics teams.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

## ACKNOWLEDGEMENTS

We thank Mark Hunt, IT Development manager, TSFT, for facilitating access to data and analysis environments.

## CONFLICT OF INTEREST STATEMENT

All authors declare: support for RC and KT-A's research is provided by the EPSRC via grant EP/N014391/1, RC is also funded by TSFT as part of the Global Digital Exemplar program (GDE); no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years, no other relationships or activities that could appear to have influenced the submitted work.

## REFERENCES

- Callen J, Georgiou A, Li J, *et al.* The safety implications of missed test results for hospitalised patients: a systematic review. *BMJ Qual Saf* 2011; 20 (2): 194–9.
- Callen JL, Westbrook JJ, Georgiou A, *et al.* Failure to follow-up test results for ambulatory patients: a systematic review. *J Gen Intern Med* 2012; 27 (10): 1334–48.
- Gandhi TK. Fumbled handoffs: one dropped ball after another. *Ann Intern Med* 2005; 142 (5): 352–8.
- Singh H, Thomas EJ, Mani S, *et al.* Timely follow-up of abnormal diagnostic imaging test results in an outpatient setting: are electronic medical records achieving their potential? *Arch Intern Med* 2009; 169 (17): 1578–86.
- Singh H, Thomas EJ, Sittig DF, *et al.* Notification of abnormal lab test results in an electronic medical record: do any safety concerns remain? *Am J Med* 2010; 123 (3): 238–44.
- Roy CL, Poon EG, Karson AS, *et al.* Patient safety concerns arising from test results that return after hospital discharge. *Ann Intern Med* 2005; 143 (2): 121.
- Dalal AK, Roy CL, Poon EG, *et al.* Impact of an automated email notification system for results of tests pending at discharge: a cluster-randomized controlled trial. *J Am Med Inform Assoc* 2014; 21 (3): 473–80.
- Kantor MA, Evans KH, Shieh L. Pending studies at hospital discharge: a pre-post analysis of an electronic medical record tool to improve communication at hospital discharge. *J Gen Intern Med* 2015; 30 (3): 312–8.
- Greenes DS, Fleisher GR, Kohane I. Potential impact of a computerized system to report late-arriving laboratory results in the emergency department. *Pediatr Emerg Care* 2000; 16 (5): 313–5.
- Dalal AK, Schaffer A, Gershanik EF, *et al.* The impact of automated notification on follow-up of actionable tests pending at discharge: a cluster-randomized controlled trial. *J Gen Intern Med* 2018; 33 (7): 1043–51.

11. Singh H, Hirani K, Kadiyala H, *et al.* Characteristics and predictors of missed opportunities in lung cancer diagnosis: an electronic health record-based study. *J Clin Oncol* 2010; 28 (20): 3307–15.
12. Georgiou A, Lymer S, Forster M, *et al.* Lessons learned from the introduction of an electronic safety net to enhance test result management in an Australian mothers' hospital. *J Am Med Inform Assoc* 2014; 21 (6): 1104–8.
13. Etchells E, Adhikari NKJ, Cheung C, *et al.* Real-time clinical alerting: effect of an automated paging system on response time to critical laboratory values—a randomised controlled trial. *Qual Saf Health Care* 2010; 19 (2): 99–102.
14. Kuperman GJ, Teich JM, Tanasijevic MJ, *et al.* Improving response to critical laboratory results with automation: results of a randomized controlled trial. *J Am Med Inform Assoc* 1999; 6 (6): 512–22.
15. Taunton, Somerset NHS Foundation Trust. About Your Hospital. <http://www.tsft.nhs.uk/about-your-hospital/>. Accessed February 19, 2020.
16. R Core Team. R: A Language and Environment for Statistical Computing. 2017. <https://www.R-project.org/>. Accessed February 19, 2020.
17. El-Shaarawi AH, Piegorch WW, eds. Bootstrap resampling. In: *Encyclopedia of Environmetrics*. Chichester, UK: John Wiley & Sons, Ltd; 2006: 43.
18. Savitzky A, Golay M. Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem* 1964; 36 (8): 1627–39.
19. Cramér H. *Mathematical Methods of Statistics*. 1946; 591.
20. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc* 1952; 47 (260): 583–621.
21. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull* 1945; 1 (6): 80–3.
22. Bonner J. Hospitals at night. *BMJ* 2007; 335 (7629): s189–91.
23. Bruins MJ, Ruijs G, Wolfhagen M, *et al.* Does electronic clinical microbiology results reporting influence medical decision making: a pre- and post-interview study of medical specialists. *BMC Med Inform Decis Mak* 2011; 11 (1): 19.
24. Lin JJ, Moore C. Impact of an electronic health record on follow-up time for markedly elevated serum potassium results. *Am J Med Qual* 2011; 26 (4): 308–14.
25. Menon S, Smith MW, Sittig DF, *et al.* How context affects electronic health record-based test result follow-up: a mixed-methods evaluation. *BMJ Open* 2014; 4 (11): e005985.
26. Petersen LA, Brennan TA, O'Neil AC, *et al.* Does housestaff discontinuity of care increase the risk for preventable adverse events? *Ann Intern Med* 1994; 121 (11): 866–72.
27. Dutra M, Monteiro MV, Ribeiro KB, *et al.* Handovers among staff intensivists: a study of information loss and clinical accuracy to anticipate events. *Crit Care Med* 2018; 46 (11): 1717–21.
28. Hamilton-Fairley D, Coakley J, Moss F. Hospital at night: an organizational design that provides safer care at night. *BMC Med Educ* 2014; 14 (Suppl 1): S17.



# SUPPLEMENTARY MATERIAL FOR FACTORS INFLUENCING DIGITAL REVIEW OF PATHOLOGY TEST RESULTS IN AN INPATIENT SETTING: A CROSS-SECTIONAL STUDY

*Table S1 - the baseline coefficients for a linear models of the form  $y = At + B$  for the daily estimates of time to view and proportion unviewed over the study period.  $\sigma$  is the residual standard error. The models are centred around the midpoint of the study period (and as such the intercept,  $B$ , is the mean for the study period). All measures are very slightly decreasing over time.*

<b>variable</b>	<b>A</b>	<b>B</b>	<b>Adj R<sup>2</sup></b>	<b><math>\sigma</math></b>
<i>Median time to view</i>	-0.017 mins day <sup>-1</sup>	89.8 mins	0.1216	14.21 mins
<i>Lower quartile</i>	-0.005 mins day <sup>-1</sup>	33.9 mins	0.0658	6.157 mins
<i>Upper quartile</i>	-0.053 mins day <sup>-1</sup>	213.8 mins	0.1676	37.59 mins
<i>% tests unviewed</i>	-5.1x10 <sup>-5</sup> % day <sup>-1</sup>	6.129 %	< 0.001	1.887 %

# 4. Algorithmic hospital catchment area estimation using label propagation

Status: This version submitted to BMC Health Services Research. A revision of this paper has since been published.

Date: 9th Aug 2021.

Candidate's contribution: All authors discussed the concept of the article and RC developed the algorithm, performed the analysis and wrote the initial draft. KTA, LL, GG commented and made revisions. All authors read and approved the final manuscript. RC is the guarantor.

The algorithm development and all the analysis in this paper was performed by me and findings validated by other team members.

# Algorithmic hospital catchment area estimation using label propagation

Robert Challen<sup>1,2</sup>; Gareth Griffith<sup>3,4</sup>; Lucas Lacasa<sup>5,6</sup>; Krasimira Tsaneva-Atanasova<sup>1,7,8</sup>;

1) EPSRC Hub for Quantitative Modelling in Healthcare, University of Exeter, Exeter, Devon, UK.

2) Somerset NHS Foundation Trust, Taunton, Somerset, UK.

3) Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, BS8 2BN, UK

4) Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, BS8 2BN, UK

5) School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, UK

6) Instituto de Física Interdisciplinar y Sistemas Complejos (IFISC) (CSIC-UIB), Campus UIB, 07122, Palma de Mallorca, Spain

7) The Alan Turing Institute, British Library, 96 Euston Rd, London NW1 2DB, UK.

8) Data Science Institute, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK.

## Abstract

**Background:** Hospital catchment areas define the primary population of a hospital and are central to assessing the potential demand on that hospital, for example, due to infectious disease outbreaks.

**Methods:** We present a novel algorithm, based on label propagation, for estimating hospital catchment areas, from the capacity of the hospital and demographics of the nearby population, and without requiring any data on hospital activity.

**Results:** The algorithm is demonstrated to produce a mapping from fine grained geographic regions to larger scale catchment areas, providing contiguous and realistic subdivisions of geographies relating to a single hospital or to a group of hospitals. In validation against an alternative approach predicated on activity data gathered during the COVID-19 outbreak in the UK, the label propagation algorithm is found to have a high

25 *level of agreement and perform at a similar level of accuracy.*

26 **Conclusions:** *The algorithm can be used to make estimates of hospital catchment areas in new situations*  
27 *where activity data is not yet available, such as in the early stages of a infections disease outbreak.*

## 28 Introduction

29 During the COVID-19 pandemic, the rapid assessment of the available capacity of a hospital and the potential  
30 demand on its services has been important in identifying geographical areas where hospital services are at risk  
31 of becoming overwhelmed. Along with epidemic dynamics, residual hospital capacity guides the imposition of  
32 public health measures such as social distancing. When assessing the load on a hospital due to COVID-19  
33 the demand may be unevenly distributed in space and rapidly changing in time. Available capacity may  
34 be influenced by multiple factors, including staff availability. At the same time there may be fundamental  
35 changes to health provision in the acute response of the pandemic, with for example the cancellation of  
36 routine operations. In the early epidemic in the UK, for example, there was block booking of private health  
37 care providers to assist the NHS (1), and the rapid creation of large scale field hospitals (2). In previous work  
38 we examined the potential for redirecting patients from one region to another to balance the load of health  
39 care provision (3) and we have observed this phenomenon as intensive care units reach capacity (4). When  
40 we consider both the change in provision of services and the redistribution of patients, there is a potential  
41 need to redefine the demographic and geographic profiles of health care service providers (“catchment areas”  
42 and “catchment populations”) (5) to allow for effective planning.

43 The catchment area or population of a hospital is a broad concept which serves a number of purposes, such  
44 as:

- 45 • Definition of the primary population of a hospital (and their demographics) for strategic planning  
46 purposes (6).
- 47 • Definition of higher level organizational structures and collaborative networks (7).
- 48 • Identification of areas with under, or over provision of services
- 49 • Calculation (and visualisation) of incidence and prevalence of disease from hospital reported statistics  
50 (identifying the denominator) (8) and hence admission rates per head of population.
- 51 • Preferred routing of patients to hospitals for optimizing specific services.

52 There are two general approaches to modelling catchment areas which we will discuss in detail - activity  
53 based or algorithmic approaches. Algorithmic approaches are based solely on regional population counts and

54 hospital capacity. Activity based approaches minimally require data on hospital activity across all the region  
55 at an individual level, such as individual patient admission records.

56 Either of these individual modelling approaches result in a hospital catchment area that is either overlapping  
57 or non-overlapping. An overlapping output may reflect the fact that patients may have a choice in the use of  
58 the services, and that a range of individually varying predictors influence individuals' capacity and willingness  
59 to adhere to arbitrarily imposed boundaries. It may also reflect a fundamental organization of the service,  
60 for example the networks of critical care (4), in which some activity of a hospital caters directly for the  
61 local population, but other activity is conducted supporting other regional hospitals. As such overlapping  
62 approaches may better reflect reality, but non-overlapping outputs are often a necessary simplification for  
63 secondary analyses, where cross-classification is not specifiable (9). It is often desirable for secondary analysis  
64 that boundaries align with geographical and organizational boundaries, but non-overlapping outputs may  
65 result in real world cases being incorrectly assigned to a hospital based on the catchment area, and this will  
66 tend to be spatially uneven, clustering at the fringes of the imposed boundaries (10).

67 The simplest algorithmic approaches involve a measure of the size of a hospital inversely weighted by straight  
68 line distance (11). This can be extended by models which use an analogy to gravity to calculate the potential  
69 field of every hospital, based on both capacity (e.g. beds) and demand (e.g. patients) (11–13). The resulting  
70 potentials may be cut off at a specified value, or where they are exceeded by another hospital's potential,  
71 to produce either overlapping or non-overlapping fields. Such algorithmic approaches may not respect  
72 geographical or existing organizational boundaries, but they can be used to model hypothetical scenarios,  
73 such as the impact of creating a new hospital. Further details of the range of different models that have been  
74 proposed have been previously published (5,8).

75 Activity based models began with the proportional flow, or Norris-Bailey, model (14,15) which examines the  
76 proportion of patients from an area visiting a particular hospital versus the proportion of patients in an area  
77 who visit any health care provider. An extension of this was recently used to define catchment areas for  
78 major injury following acute trauma (16). More recently modern statistical approaches have been applied  
79 to the same basic activity data including k-Means classification (8), Bayesian regression modelling. (6) or  
80 Markov Multiscale Community Detection (7,17). Whilst arguably providing a more accurate reflection of  
81 reality, activity based models are predicated on the availability and recency of activity data, which may  
82 exhibit historical or cultural biases. Depending on the purpose of the catchment area such historical bias  
83 may or may not be desirable (8).

84 Estimation of hospital catchment areas is a simplification of a complex logistical and organizational problem.

85 In England, for example, hospital sites are typically grouped into single organizational units (NHS trusts)  
86 which report combined activity. Thus a single unit of health-care provision (NHS trust) may have a range of  
87 physical locations, not all of which offer the full range of services. ICU provision is often focused in a single  
88 hospital in an NHS Trust, whereas acute or step-down beds may be distributed across multiple sites. Some  
89 specialist services, such as intensive care, also may be unevenly distributed, and larger units used as “tertiary  
90 referral centres” which take in more complex patients from a wider geographical area.

91 In the early phase of the COVID-19 pandemic, a rapid estimate was needed of the potential demand on  
92 intensive care services as a result of observed and forecast infections, in the context of a changing landscape  
93 of health service provision. At this point, there was no comparable data with which to drive activity based  
94 models, and volatile estimates of hospital capacity. In order to plan provision of additional ventilators and  
95 high dependency beds, we needed a model of geographical catchment areas that could be used to translate  
96 regional epidemiological models of infections into a prediction of future admissions to individual hospitals,  
97 taking into account the regional demographics, and an estimate of the expected level of care the patients  
98 would need. Such a catchment area model must interface with existing spatial boundaries implemented in  
99 epidemiological models and publicly available demographic estimates, and fulfil the following criteria:

- 100 • Allow a clean one way mapping from fine grained geographic regions (e.g. from regional demographic  
101 estimates or epidemiological models) to the coarse grained administrative hospital region.
- 102 • Provide contiguous and realistic subdivisions of geographies relating to a single hospital or to a hospital  
103 group.
- 104 • Provide areas that are determined by the capacity of hospital at different levels of care provision, and  
105 the size of the local population, or anticipated size of outbreak in the local population.
- 106 • Create regions of approximately equal local supply (e.g. beds) and demand (e.g. patients) at boundaries.
- 107 • Respect crude physical geographical boundaries, such as large rivers.
- 108 • Flexible in that it can be recomputed rapidly if the background parameters change, for example, a  
109 regional outbreak or provision of additional hospitals, in a way that is not dependant on individual  
110 level activity data.

111 In this work we present a solution we developed for this problem, and introduce a novel algorithmic catchment  
112 area model which is specifically designed to meet the needs of the COVID-19 pandemic as described above,  
113 but is globally applicable to the situation where we can quantify demand for a resource and a set of point  
114 locations that supply that resource, and could be used, for example, in retail. This model is inspired by  
115 label propagation techniques used for community detection in networks (18–20). The paper is presented as  
116 follows; firstly we introduce the algorithm, secondly we describe some illustrative examples, and thirdly we

117 qualitatively compare the output of the algorithm to both manually created organizational boundaries, and  
118 to observed patient ICU admissions during the first wave of the COVID 19 pandemic.

## 119 **Materials and Methods**

120 This section consists of 3 parts: a detailed description of the algorithmic catchment area model, a description  
121 of the data used to create initial outputs from the model, and a description of initial assessment of the model  
122 against available data.

### 123 **Algorithm**

124 The algorithm is inspired by label propagation network clustering, where labels correspond to the supply of a  
125 service, and the nodes in the network correspond to the demand for the service. For illustrative purposes  
126 in this paper we will focus on the example of hospitals, where the “supply” is provision of hospital beds,  
127 the “demand” is the population size, and the “network” is the neighbourhood of geographical areas under  
128 consideration.

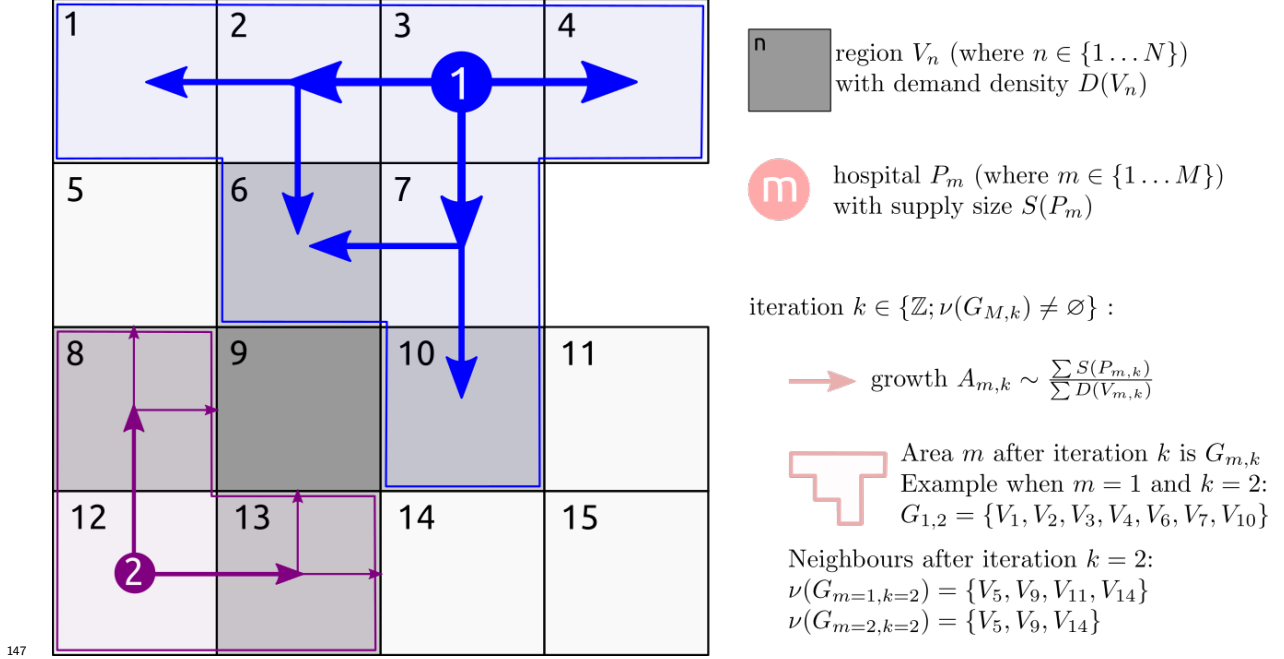
129 To connect supply and demand, or hospital beds to population size, the algorithm propagates a number of  
130 labels, each representing the source of supply (e.g. the hospital), through the geographical network, at a rate  
131 defined by both the size of the supply (e.g. beds in each hospital), and the demand for the service (e.g. the  
132 population) within the areas the label has already propagated to. Thus as demand outstrips supply from a  
133 particular source the rate of label propagation associated with that source decreases.

134 We assume the whole geographical region under consideration can be represented as a mathematical graph,  
135  $G$  and is divided into  $N$  smaller regions (parameterisation discussed below), represented by the vertices  $V$   
136 (where  $V = V_n, n = 1, 2, \dots, N$ ) each with known population of size  $D(V_n)$ .

137 We define  $M$  hospitals located at the geographical points  $P$  (where  $P = P_m, m = 1, 2, 3 \dots M$ ), and with  
138 capacity to supply  $S(P_m)$  beds. Typically there are fewer hospitals than regions ( $M \ll N$ ). We constrain  
139  $P_m$  such that no more than one  $P_m$  is found within any given  $V$ , i.e. each small region hosts no more than  
140 one hospital. In practice the assumption that a maximum of one hospital is found in each small region is  
141 occasionally not true. When this does happen, we preprocess the data to combine hospitals that are located  
142 together into a single entity.

143 The connections of neighbouring regions of any area  $V_x$  are defined by  $E_x = \nu(V_x)$ , and likewise the set of

144 neighbouring vertices of any subgraph  $G_y$  are defined by  $E_y = \nu(G_y)$ . These quantities are readily calculated  
 145 using the geographical intersection of different areas and various algorithms exist to calculate these from  
 146 geo-spatial data (21,22).



148 Figure 1: Schematic illustration of the proposed label propagation algorithm. The association of a hospital  
 149 with a region propagates from the hospital location (P) into the different regions (V) at a rate depending  
 150 on the hospital capacity  $S(P)$  and the population of the region,  $D(V)$ , at each round of the iteration (k)  
 151 until there are no more neighbours to propagate a label to. The direction of spread is determined by the  
 152 geographical neighbourhood of each region V

153 Our goal is to divide the graph  $G$  into  $M$  labelled sub-graphs  $G_m$  such that the sub-graphs are connected,  
 154 and that neighbouring sub-graphs have similar bed availability per unit population ( $\frac{\sum S_m}{\sum D_m}$ ). We do this by  
 155 assigning a score for each combination of region and hospital, which is initially zero. For every iteration  
 156 of the algorithm this score is incremented in any unlabelled region that neighbours a region that has been  
 157 labelled (i.e. assigned to a specific hospital). The score is increased by a small amount determined by the  
 158 ratio of supply (hospital beds) available, and demand (population to be served) in the regions assigned to  
 159 that hospital. Thus labels propagate more quickly from points with a high capacity, through regions with a  
 160 low population than vice-versa. The first label to propagate to a given area, and for which the score is above  
 161 a threshold is defined as the “supplier” for that area, which is labelled as such. This ensures that each region



162 is served by only one hospital.

---

**Algorithm 1:** A weighted label propagation algorithm for matching geographical supply to demand

---

**Input** :  $V_N$  - the  $N$  regions of demand as a set of geographical polygons  
**Input** :  $D(V_n)$  - the density of demand in any given region as a function of the region  $V_n$   
**Input** :  $P_M$  - a set of  $M$  labelled suppliers as a set of geographical points  
**Input** :  $S(P_m)$  - the capacity of supply at any given supply point as a function of the supplier  $P_m$   
**Input** :  $C_{growth}$  - a rate constant defining rate of label propagation  
**Output:**  $G_M$  -  $M$  labelled subgraphs of graph  $G$ , relating to the catchment areas of suppliers  $P_M$

- define  $G$  as the graph consisting of geographical regions  $V_N$ , connected by edges,  $E_N$ , given by their geographical neighbours  $\nu(V_N)$ :  
 $E_N \leftarrow \nu(V_N)$ ;  
 $G \leftarrow (V_N, E_N)$ ;
- define  $V_M$  and  $V_{M,0}^{new}$  as the geographic regions of  $G$  serviced by points  $P_M$ , and  $G_{M,0}$  as a set of labelled sub-graphs (also initially consisting solely of the vertices  $V_M$ ):  
 $V_M \leftarrow G \cap P_M$ ;  
 $V_{M,0}^{new} \leftarrow V_M$ ;  $G_{M,0} \leftarrow V_M$ ;
- define the initial unlabelled set of vertices:  
 $U_0 \leftarrow \neg V_M$ ;
- define the initial un-labelled neighbours of labelled sub-graphs,  $G_M$ :  
 $U_{M,0} \leftarrow \nu(V_M)$ ;
- define an accumulated growth score for each un-labelled neighbour  $U_{M,0}$  of each  $G_{M,0}$ :  
 $A_{U_{M,0}} \leftarrow 0$ ;
- $k \leftarrow 0$ ;
- execute the loop while there are still unlabelled vertices and there exist some unlabelled neighbours of labelled vertices

163 **while**  $|U_k| > 0$  and  $|U_{M,k}| > 0$  **do**

- $k \leftarrow k + 1$ ;
- define the un-labelled vertices as the set of  $V$  not contained in any of  $G_{M,k-1}$ :  
 $U_k \leftarrow \neg G_{M,k-1}$ ;
- define the un-labelled neighbours of  $G_{M,k-1}$  as  $U_{M,k}$  as the previously unlabelled neighbours and the neighbours of the most recently labelled neighbours  $V_{M,k-1}^{new}$ :  
 $U_{M,k} \leftarrow U_{M,k-1} \cup (U_k \cap \nu(V_{M,k-1}^{new}))$ ;
- define the reserve capacity,  $R_M$ , to supply existing labelled,  $G_{M,k-1}$ , and un-labelled neighbours  $U_{M,k}$ , as:  
 $R_M \leftarrow \frac{S(P_M)}{D(U_{M,k} \cup G_{M,k-1})}$ ;
- for unlabelled areas only, update the accumulated growth score,  $A_{U_{M,k}}$ , with the normalised rank of the reserve capacity and multiplied by a constant  $C_{growth} > 1$  representing the speed at which the accumulated growth score increases in all areas:  
 $R_{M,k} \leftarrow R_M \{m \in U_{M,k}\}$ ;  
 $A_{U_{M,k}} \leftarrow A_{U_{M,k-1}} + C_{growth} \times \text{rank}(R_{M,k}) / |R_{M,k}|$ ;
- for all the un-labelled vertices, select the label  $M$ , with the highest score, and if the accumulated score has reached the threshold of 1, incorporate it into the labelled sub-graph,  $G_{M,k-1}$ :  
 $A_{U_k}^{max} = \max(A_{U_{m,k}}, m \in M)$ ;  
 $V_{M,k}^{new} \leftarrow U_{M,k} \in \{A_{U_k}^{max} > 1\}$ ;
- $G_{M,k} \leftarrow G_{M,k-1} \cup V_{M,k}^{new}$ ;
- $U_{M,k+1} \leftarrow U_{M,k} \cap \neg V_{M,k}^{new}$ ;

**end**  
**return**  $G_{M,k}$

---

## 164 **Qualitative testing data**

165 The algorithm requires firstly an estimate of demand, for this we used population counts, secondly a  
166 geographical network and thirdly an estimate of supply, in this case hospital capacity data.

167 For Great Britain there are detailed estimates of the population at granular geographic detail (lower super  
168 output area - LSOA) available from the Office of National Statistics (ONS) for England and Wales, and  
169 population estimates by Data Zone (DZ) are provided by the National Records Service (NRS) in Scotland  
170 (23,24). These population estimates are available by single year of age for each area. These are combined to  
171 create a single figure for the adult population of each small geographic area.

172 Each geographical area is associated with a boundary files for lower super output areas and data zone from  
173 the 2011 census, which are provided by the ONS and NRS (25,26).

174 To estimate the capacity of hospitals we used a range of primary sources (described in the supplementary  
175 materials) to manually compile a list of NHS and independent hospital sites. When not provided in the  
176 primary sources, we identified their geographical locations from their postcode, and we estimated bed numbers  
177 from both a combination of published NHS statistics and from daily COVID-19 situation reports from early  
178 April 2020, provided by the NHS. The situation reports detailed both available beds at this point in time but  
179 also gave an indication of maximum surge capacity for high dependency beds. These data were manually  
180 curated and are indicative of the state of the NHS at maximal readiness. Bed state estimates for independent  
181 hospital providers were also available through the situation reports.

182 In Northern Ireland, population estimates were not available at a similar geographical resolution as the ONS  
183 and NRS sources, and we are unaware of any publicly available hospital capacity estimates. They were  
184 therefore not included in this analysis.

185 The detail of the original data sources we used is presented in the supplementary material, not all of  
186 which are publicly available. The algorithm is implemented as an R package **arear** (available from <https://terminological.github.io/arear/>), which also contains both the manually curated hospital capacity and data  
187 pertaining derived demographics data described here.  
188

## 189 **Validation**

190 There is no ground truth for the catchment areas for hospitals in the NHS during the COVID-19 pandemic.  
191 The rationale for original development of this algorithm was to make an estimate in absence of any activity

192 data, in the early stages of the pandemic. Since then activity data has become available and this allows us to  
193 validate the label propagation approach to the activity based approach.

194 The activity based mapping takes the form of a many-to-many probabilistic mapping between lower tier local  
195 authority districts (LTLA) and NHS Acute Trusts in England derived from Secondary Uses Service (SUS)  
196 health-care data for England (27). We create equivalent probabilistic associations between the coarse grained  
197 LTLA and NHS trusts by generating a fine grained lower super output area (LSOA) catchment area for NHS  
198 trusts using the label propagation algorithm, and the demographic and bed capacity estimates described  
199 above. This is aggregated to coarse grained local authority districts using mapping files provided by the  
200 ONS (28), weighted by LSOA population size (23) (Source: Office for National Statistics licensed under  
201 the Open Government Licence v.3.0). This equivalent mapping based on the label propagation algorithm  
202 is compared to the activity based mapping graphically. To determine the degree of agreement between  
203 approaches the expected number of admissions to each NHS trust from each LTLA was estimated using  
204 each method. These were compared to each other using the intra-class correlation coefficient (29,30) using a  
205 mean-of-raters, absolute-agreement, two-way random-effects model (31), as implemented in the R package  
206 `irr`(32).

207 Secondly we obtain the coarse location (partial UK postcode, also known as outcode) from a list of intensive  
208 care patients admitted between 20th October 2000 and 16th March 2021 from the CHESS data set (33), which  
209 is an anonymised patient level hospital admission data set. We use outcode boundary shapes (34), LSOA  
210 demographic estimates, and an areal interpolation (35) to generate an estimate of demographics for each  
211 outcode. Using this outcode based regional population estimate, outcode boundary shapes, and the manually  
212 curated high dependency unit capacity estimates we calculate an outcode based catchment area estimate  
213 from which we are able to predict the NHS trust each patient was admitted to based on their outcode, which  
214 we compare to the observed NHS trust from the CHESS data. For this comparison we calculate both the  
215 multinomial accuracy, and for each NHS trust, the one-versus-all binomial accuracy as follows:

$$\text{accuracy} = \frac{1}{|X|} \sum_{k \in G} \sum_{g_{obs}(x)=k} I(g_{pred}(x) = g_{obs}(x))$$

216 where  $X$  is the set of observations,  $G$  is the set of NHS trusts,  $g_{pred}$  and  $g_{obs}$  are the predicted and observed  
217 classes respectively and  $I$  is the indicator function which returns 1 if the predicted match observed and 0  
218 otherwise.

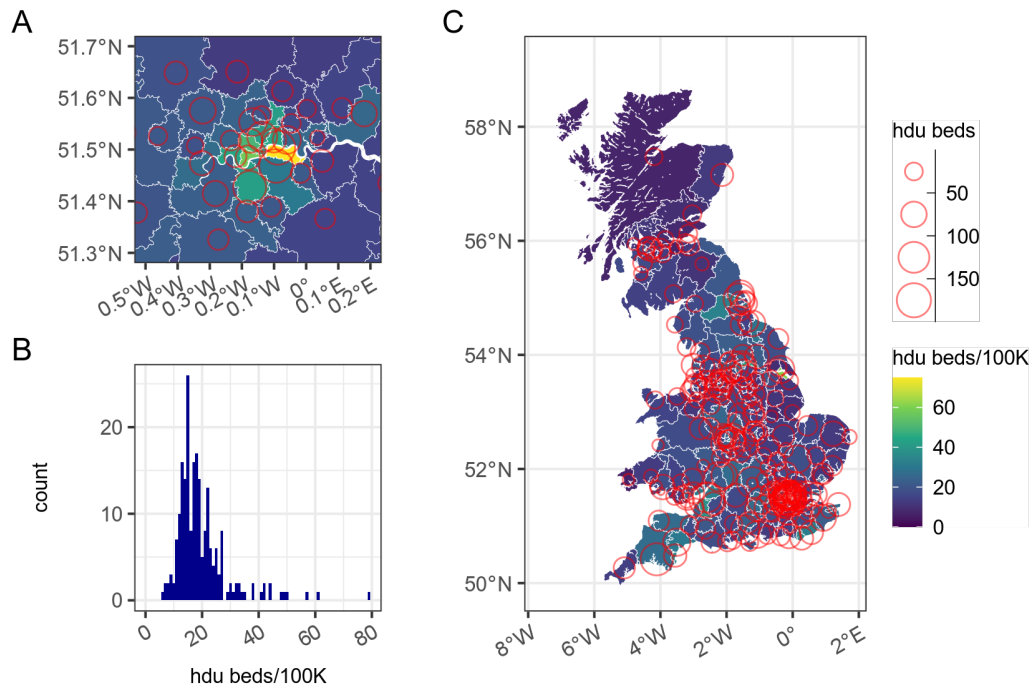
219 For the activity based approach we assign each patient to a LTLA by virtue of the geographical location of  
220 the centroid of their outcode shape and then determine the most probable NHS trust associated with that

221 LTLA. This forms a prediction of the NHS trust based on the patient's outcode, which we can compare to  
222 the observed NHS trust in the same manner as above.

## 223 **Results**

### 224 **Qualitative testing results**

225 The results presented in this section qualitatively test the algorithm to determine whether it is producing  
226 catchment area regions that are geographically contiguous, aligned with existing demographic boundaries,  
227 and respect coarse geographical boundaries such as large rivers. The catchment areas should also produce  
228 estimates that minimise differences in the level of service provision from area to area, and we expect the  
229 overall regional variation of supply versus demand to be locally smooth. Figure 2 shows a catchment area  
230 based on individual hospitals that offered high dependency beds during April 2020, and a regional demand  
231 based on population estimates of adults in lower super output areas. The resulting set of catchment areas  
232 presented in panel A and C behave as desired in terms of the geographical properties. They also produce  
233 a fairly uniform density of high dependency bed provision per capita population, from region to region, as  
234 seen in panel B. In areas where there are high densities of hospitals such as London where the algorithm, by  
235 design, cannot propagate from centrally located hospitals past more peripheral hospitals, leading to small  
236 numbers of areas with high provision per head of population. This is discussed further below.

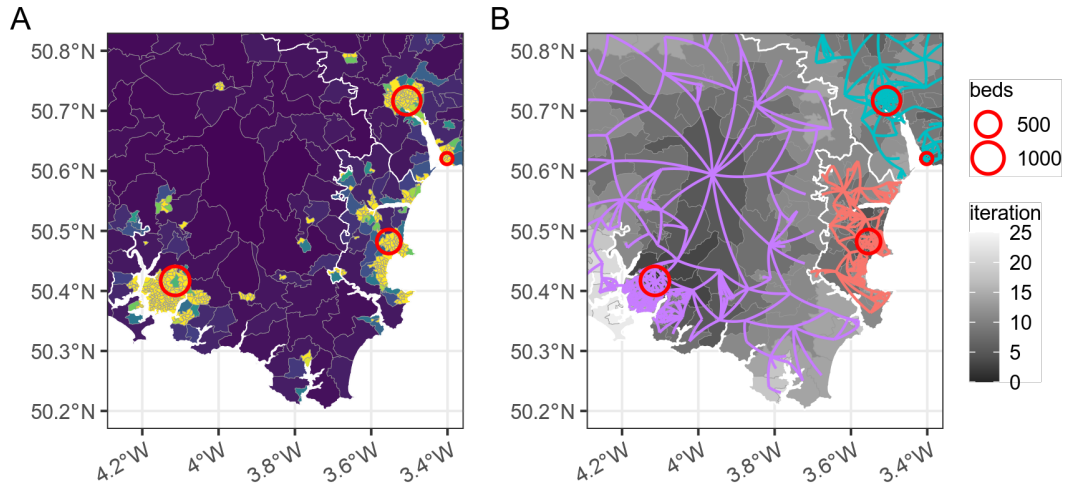


237

238 Figure 2: Panels A and C show a LSOA based catchment area map estimated from the high dependency  
 239 bed state in Great Britain in early April 2020, with catchment area boundaries shown in white. Red circles  
 240 are NHS hospital sites with size scaled to high dependency bed capacity. Map source: Office for National  
 241 Statistics licensed under the Open Government Licence v.3.0, Contains OS data © Crown copyright and  
 242 database right 2020. Panel B shows the distribution of high dependency beds per 100K population for each  
 243 of the catchment areas defined by the algorithm.

244 Further qualitative investigation of the properties of the algorithm are shown in Figure 3 where we see more  
 245 regional detail of the same algorithm applied this time to general hospital beds rather than high dependency  
 246 beds. Panel A shows the boundaries of the estimated catchment areas in white against the population density  
 247 of a small area of the South West of England containing three hospitals (Plymouth, Torbay and the Royal  
 248 Devon and Exeter hospitals). We can see in this example the extent of the catchment area to the South of  
 249 Torbay is defined by the Dart river estuary, thus respecting such geographical boundaries.

250 Figure 3 panel B shows details about the progression of the algorithm from one iteration to the next, as  
 251 labels propagate from each of the hospitals into the surrounding areas until encountering another catchment  
 252 area. As we expect from the design the algorithm is seen to spread from hospital sites quickly through areas  
 253 of low population (panel A), such as the countryside surrounding Plymouth in the bottom left, and more  
 254 slowly through areas of higher population such as the areas surrounding Torbay in the middle right.



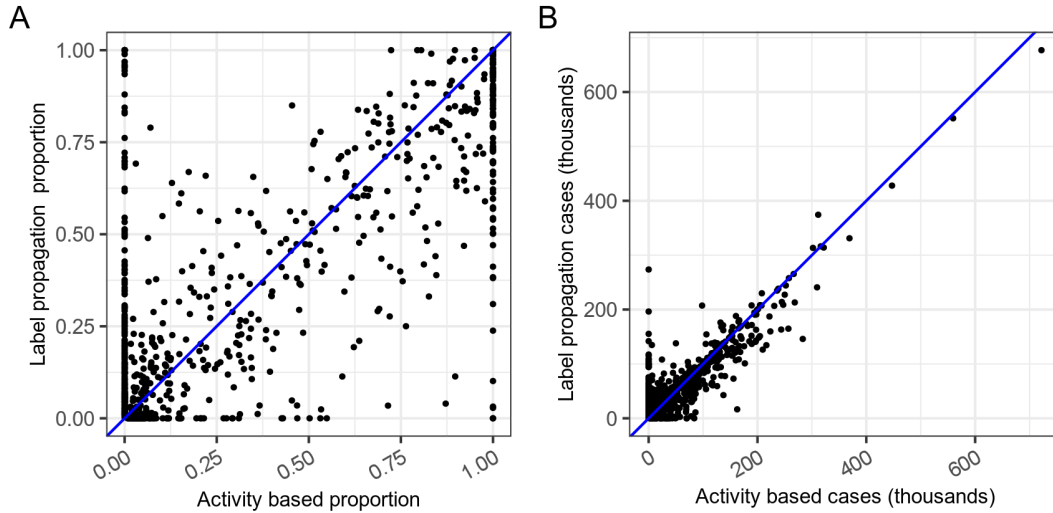
255

256 Figure 3: Detail LSOA based catchment area map for NHS trusts estimated from the general hospital bed  
 257 states in Great Britain in early April 2020. Red circles are NHS hospital sites. In panel A the fill represents a  
 258 relative measure of regional population density, with yellow areas being high density in and around cities. In  
 259 Panel B the same areas are shown but this time the fill shows the iteration number at which the algorithm  
 260 labelled a specific area, and the propagation of the algorithm by arrows. Map source: Office for National  
 261 Statistics licensed under the Open Government Licence v.3.0, Contains OS data © Crown copyright and  
 262 database right 2020

## 263 Validation

264 In comparing the label propagation mapping to the activity based mapping we see that the proportions of  
 265 any given LTLA that are assigned to any given trust are similar between the two methods (Figure 4, panel  
 266 A) with a clear trend to agreement. The major differences are seen in the extremes where, for example, in  
 267 the top left of panel A, the activity based approach may predict that no patients are observed in a given  
 268 hospital from a given LTLA, whereas the label propagation approach predicts the opposite. Panel B shows  
 269 the same relationship but this time scaled by the population size in each area, and this shows that the  
 270 impact of differences between predictions seen in panel A is in areas with smaller populations and is therefore  
 271 attenuated. Calculation of the intra-class correlation coefficient between the predicted number of cases from

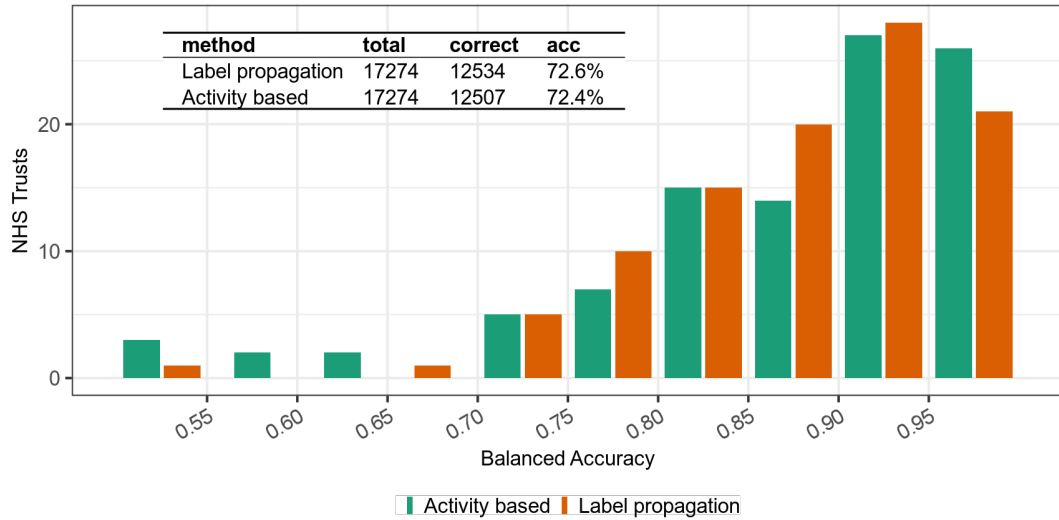
272 each method gives excellent agreement between the two methods, with a value of 0.94 (95% CI: 0.93 – 0.95)  
273 using a mean-of-raters, absolute-agreement, two-way random-effects model(31).



274

275 Figure 4: Classification agreement between activity based approach and label propagation algorithm. Each  
276 point is a unique combination of lower tier local authority and NHS trust and in panel A the proportion  
277 of the LTLA assigned to that trust is plotted for the activity based algorithm on the x-axis and the label  
278 propagation algorithm on the y-axis. In panel B the total number of cases assigned to each trust is plotted  
279 when the population size for the area is considered. The blue line represents perfect agreement.

280 In Figure 5 we compare observed admissions to ITU to predictions made by the label propagation algorithm  
281 and the activity based approach. As there are 178 trusts under consideration which form a large number  
282 of distractors for each prediction, a low value for the multinomial accuracy could be expected. The overall  
283 accuracy of both methods is comparable at 72.6%—72.4%. The distribution of the binomial one-versus-all  
284 accuracy in the histogram shows that the prediction performance is better for some trusts than others, and  
285 that the accuracy of the activity based approach has greater variability than that of the label propagation  
286 approach. Across the whole country exact agreement between the observed location of hospital admission  
287 and the predicted location of hospital admission based on the label propagation catchment area was seen in  
288 12534 out of 17274 cases, and the Matthew’s correlation coefficient was 0.72.



289

290 Figure 5: Accuracy measures for the predictions of activity based and label propagation approaches based on  
 291 UK postcode outcodes, and a subset of observed NHS trust of intensive care admissions in England between  
 292 20th October 2000 and 16th March 2021. The histogram shows the distribution of the balanced accuracy for  
 293 each NHS trust in a one-vs-all binomial evaluation, and the inset table shows the overall accuracy from the  
 294 multinomial evaluation, along with the raw counts of overall evaluations and correct predictions for each  
 295 method

296 In Table 1 we qualitatively examine the ten NHS Trusts that have the highest number of ITU patients that  
 297 the label propagation algorithm predicted to be admitted elsewhere, and mis-classified them. These represent  
 298 1833 (38.7%) of the total mis-classifications. The majority of these 10 hospitals are major tertiary referral  
 299 intensive care units, or specialist centres, as demonstrated by them being in the top quintile of NHS trusts by  
 300 ITU bed capacity. This result is consistent with both the possibilities that severely ill patients may end up in  
 301 specialist centres rather than their closest hospital for treatment, or that in the event of a large surge in cases,  
 302 patients may overflow from smaller to larger intensive care units. Both of these could lead to mis-classification  
 303 of these patients by the label propagation algorithm, as we see here.

304 Table 1: The NHS trusts with the ten most misclassified covid ITU cases as assigned by the label propagation  
 305 algorithm



Trust	April 2020 ITU beds (Centile)	Classification errors
St George's University Hospitals Nhs Foundation Trust	147 (91%)	282
Manchester University Nhs Foundation Trust	151 (93%)	272
Cambridge University Hospitals Nhs Foundation Trust	90 (73%)	233
Guy's And St Thomas' Nhs Foundation Trust	159 (94%)	206
London North West University Healthcare Nhs Trust	100 (77%)	188
The Newcastle Upon Tyne Hospitals Nhs Foundation Trust	128 (85%)	178
Walsall Healthcare Nhs Trust	47 (40%)	138
Imperial College Healthcare Nhs Trust	148 (92%)	131
Barnsley Hospital Nhs Foundation Trust	25 (14%)	103
King's College Hospital Nhs Foundation Trust	190 (97%)	102

306

## 307 Discussion

308 We have presented an algorithm for rapidly estimating hospital catchment areas for use when activity data is  
309 not available. We demonstrate how the output responds to the different capacities of the different levels of  
310 care provided (e.g. high dependency versus general hospital beds). We present catchment areas calculated  
311 using population size as demand, and total hospital beds as supply. This algorithm may be useful for longer  
312 term strategic planning, but was conceptualized as part of an acute response to COVID-19 outbreak. In this  
313 case we can use the different parameters for demand, for example local COVID-19 infection prevalence, and  
314 different parameters for supply, for example availability to staffed hospital beds. Our approach is novel in  
315 that it allows adaptation of local service provision to predictions of disease prevalence from epidemiological  
316 models of COVID-19 and real time bed states provided by NHS trusts. This allows us to model the degree  
317 of elasticity in the system to absorb localised shocks, caused by regional outbreaks, it helps us to develop  
318 a better concept of when services are being at risk of becoming overwhelmed, and allow routing of new  
319 admissions away from overloaded hospitals.

320 Benchmarking our algorithm against activity based approaches produced good to excellent agreement and  
321 application to both methods to real world patient admission data produces a very similar result. The finding  
322 that naive application of our algorithm to real world patient admission classifies only 72.6% correctly is  
323 explained by two things, firstly accuracy at boundaries decreases as the number of boundaries increases (10)  
324 and secondly the fact that many of the top 10 mis-classified trusts are major tertiary referral centres, which  
325 may take patients from distant regions for specialist care. This suggests that our constraint that catchment  
326 areas should be non overlapping is not borne out in reality for these cases.

327 Overlapping catchment areas could be modeled by multiple layers of non-overlapping catchment areas. When

328 we consider the provision of intensive care services in the UK during COVID-19, we propose there are at  
329 least 3 layers of hospital service provision: there is a local service, which provides care for patients from  
330 nearby. A subset of hospitals additionally provide a regional, or tertiary referral, service layer which takes  
331 sicker patients from neighbouring hospitals in larger areas. The final layer is a crisis overflow layer provided  
332 by the NHS Nightingale field hospitals (2). Each of these layers may be considered to have somewhat  
333 independent catchment areas. We propose that dividing the larger hospitals into local and regional services  
334 and considering the tertiary referral network as a second layer, with its own larger catchment area would  
335 improve the performance of the algorithm against real activity data. In such a layered model of catchment  
336 areas there is interplay between local layer demand for hospital beds and capacity for regional tertiary care  
337 provision, which will dynamically affect the “catchment area” for regional tertiary care provision, potentially  
338 on a day to day basis. In previous work we looked at the opportunities for balancing the load between different  
339 hospitals (3) when transferring COVID-19 patients away from overloaded areas, however moving unwell  
340 patients between hospitals is ideally minimized. With this algorithm we enable the dynamic re-specification  
341 of local service catchment areas and hospital tertiary referral networks, based on evolving demand. Coupled  
342 with flexible load sharing has interesting potential to model or influence patient admissions around the whole  
343 hospital network.

344 Hospital capacity is difficult to accurately estimate. During this work we encountered many of the uncertainties  
345 that influence capacity. The ability of a hospital to provide a bed to a patient depends on a multitude of  
346 factors, including staff availability, which may vary during the different stages of the pandemic. The ability  
347 of hospitals to absorb large numbers of emergency patients by re-configuring their service provision (e.g  
348 canceling routine operations) and providing overflow or “surge” high dependency capacity for short periods  
349 of time makes putting a single number on hospital capacity difficult. The ability to recalculate catchment  
350 areas based on changing assumptions around capacity is a strength of our approach, and in the future could  
351 be used to analyse the impact of introducing new capacity into the hospital system. One further limitation  
352 to note is that the algorithm does not consider travel time between regions which may increase both as the  
353 geographical size increases but also as the population density increases due to traffic and form a barrier to  
354 patients accessing services. Adding a travel time penalty to the rate of label spread into the model is possible  
355 given some estimate of the ease of transport within and between regions, and this is an area of future work.

356 There are opportunities to extend our algorithm. The general approach of label propagation in networks has  
357 been more widely studied and newer approaches described (19,36,37) which allow overlapping communities.  
358 This may address some of the issues described above. These are appealing and a possible avenue for future  
359 extension of the algorithm. There persists however an open question about whether the overlapping nature of

360 hospital service provision observed in activity data is not really a reflection of patient choice, but actually  
361 the result of subtly different services, or different levels of service, being provided by different hospitals to  
362 different catchment areas. Thus a specialist cancer hospital close to a specialist paediatric hospital will have  
363 geographically overlapping catchment areas, but in reality these hospitals are not providing the same service  
364 to the same population. This line of argument suggests that the concept of a single overlapping hospital  
365 catchment area is also an over-simplification, and when we take into account the heterogeneity of different  
366 services offered by a hospital, we propose that a hospital's overall catchment area may be well modeled by a  
367 collection of non-overlapping catchment layers.

## 368 **Conclusions**

369 This label propagation algorithm for estimating hospital catchment areas is a pragmatic solution to determining  
370 geographical and demographic subsets of the population when there is no previous activity data available. It  
371 suits situations where the level of service provision and demand on the hospital system is dynamic, as has  
372 been the case in the COVID-19 pandemic. The algorithm is simple and satisfies the major criteria we set out  
373 in the introduction, in that it provides a mapping from low level geographic regions which provide contiguous  
374 and realistic subdivisions of geographies relating to a single hospital or to a group of hospitals. The areas are  
375 determined by the capacity of the hospital and the size of local population, and are approximately equal in  
376 terms of local supply (e.g. beds) and demand (e.g. patients) at boundaries.

377 The algorithm depends solely on data reflecting supply and geographical demand for a service, and as such  
378 is quite generic and potentially more widely applicable outside of healthcare. Although we have discussed  
379 catchment areas in terms of the capacity of hospital beds, and demand of local populations, there is nothing to  
380 prevent us defining capacity in any other way - a heuristic on staffing levels may be appropriate, or in different  
381 contexts, availability of medical imaging devices. Likewise, demand may be refined to reflect sub-populations  
382 at risk of disease, or may even be the output of a predictive model. As such our approach is applicable to a  
383 wide variety of problems.

## 384 **References**

- 385 1. Coronavirus: Thousands of extra hospital beds and staff. BBC News: UK [Internet]. 2020 Mar 21 [cited  
386 2020 Dec 16]; Available from: <https://www.bbc.com/news/uk-51989183>

- 387 2. Coronavirus: Nightingale Hospital opens at London's ExCel centre. BBC News: UK [Internet]. 2020 Apr  
388 3 [cited 2020 Dec 17]; Available from: <https://www.bbc.com/news/uk-52150598>
- 389 3. Lacasa L, Challen R, Brooks-Pollock E, Danon L. A flexible method for optimising sharing of healthcare  
390 resources and demand in the context of the COVID-19 pandemic. PLOS ONE. 2020 Oct 21;15(10):e0241027.
- 391 4. Pett E, Leung HL, Taylor E, Chong MSF, Hla TTW, Sartori G, et al. Critical care transfers and COVID-19:  
392 Managing capacity challenges through critical care networks. Journal of the Intensive Care Society. 2020 Dec  
393 16;1751143720980270.
- 394 5. Jones S, Wardlaw J, Crouch S, Carolan M. Modelling catchment areas for secondary care providers: A  
395 case study. Health Care Manag Sci. 2011 Sep;14(3):253–61.
- 396 6. Wang A, Wheeler DC. Catchment Area Analysis Using Bayesian Regression Modeling. Cancer Inform.  
397 2015 Jan 1;14s2:CIN.S17297.
- 398 7. Clarke JM, Barahona M, Darzi AW. Defining Hospital Catchment Areas Using Multiscale Community  
399 Detection: A Case Study for Planned Orthopaedic Care in England. bioRxiv. 2019 May 3;619692.
- 400 8. Gilmour SJ. Identification of Hospital Catchment Areas Using Clustering: An Example from the NHS.  
401 Health Services Research. 2010;45(2):497–513.
- 402 9. Jones K, Johnston R, Manley D, Owen D, Charlton C. Ethnic Residential Segregation: A Multilevel,  
403 Multigroup, Multiscale Approach Exemplified by London in 2011. Demography. 2015 Dec 1;52(6):1995–2019.
- 404 10. Arcaya M, Brewster M, Zigler CM, Subramanian SV. Area variations in health: A spatial multilevel  
405 modeling approach. Health Place. 2012 Jul;18(4):824–31.
- 406 11. Reilly WJ. The law of retail gravitation, New York: W.J. Reilly; 1931.
- 407 12. Huff DL. Defining and Estimating a Trading Area. Journal of Marketing. 1964 Jul 1;28(3):34–8.
- 408 13. Stewart JQ. An Inverse Distance Variation for Certain Social Influences. Science. 1941 Jan 24;93(2404):89–  
409 90.
- 410 14. Bailey NTJ. Statistics in Hospital Planning and Design. Journal of the Royal Statistical Society: Series  
411 C (Applied Statistics). 1956;5(3):146–57.
- 412 15. Norris V. Role of Statistics in Regional Hospital Planning. Br Med J [Internet]. 1952 Jan 19 [cited 2020  
413 Dec 17];1(4750):129–33. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2022292/>
- 414 16. Alexandrescu R, O'Brien SJ, Lyons RA, Lecky FE, Trauma Audit, eSearch Network. A proposed approach

- 415 in defining population-based rates of major injury from a trauma registry dataset: Delineation of hospital  
416 catchment areas (I). *BMC Health Serv Res.* 2008 Apr 10;8(1):80.
- 417 17. Clarke J, Beaney T, Majeed A, Darzi A, Barahona M. Identifying naturally occurring communities of  
418 primary care providers in the English National Health Service in London. *BMJ Open.* 2020 Jul 1;10(7):e036504.
- 419 18. Fortunato S, Hric D. Community detection in networks: A user guide. *Physics Reports.* 2016 Nov  
420 11;659:1–44.
- 421 19. Xie J, Szymanski BK. Community detection using a neighborhood strength driven Label Propagation  
422 Algorithm. In: 2011 IEEE Network Science Workshop. 2011. pp. 188–95.
- 423 20. Xie J, Szymanski BK. LabelRank: A stabilized label propagation algorithm for community detection in  
424 networks. In: 2013 IEEE 2nd Network Science Workshop (NSW). 2013. pp. 138–43.
- 425 21. Bivand R, Rundel C, Pebesma E, Stuetz R, Hufthammer KO, Giraudoux P, et al. Rgeos: Interface  
426 to Geometry Engine - Open Source ('GEOS') [Internet]. 2020 [cited 2020 Dec 18]. Available from: <https://CRAN.R-project.org/package=rgeos>  
427
- 428 22. Pebesma E. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* [Internet].  
429 2018 [cited 2020 Dec 18];10(1):439–46. Available from: [https://journal.r-project.org/archive/2018/RJ-2018-  
430 009/index.html](https://journal.r-project.org/archive/2018/RJ-2018-009/index.html)
- 431 23. Population estimates - Office for National Statistics [Internet]. [cited 2020 Nov 17]. Available from:  
432 <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates>
- 433 24. Scotland Web Team NR of. National Records of Scotland [Internet]. National Records of Scotland; Na-  
434 tional Records of Scotland; 2013 [cited 2020 Dec 7]. Available from: [/statistics-and-data/statistics/statistics-  
435 by-theme/population/population-estimates/2011-based-special-area-population-estimates/small-area-  
436 population-estimates/time-series](https://statistics-and-data/statistics/statistics-by-theme/population/population-estimates/2011-based-special-area-population-estimates/small-area-population-estimates/time-series)
- 437 25. Open Geography Portal [Internet]. [cited 2020 Nov 17]. Available from: [https://geoportal.statistics.gov.  
438 uk/](https://geoportal.statistics.gov.uk/)
- 439 26. SpatialData.gov.scot SG. Data Zone Boundaries 2011 [Internet]. 2020 [cited 2020 Dec 7]. Available from:  
440 <https://data.gov.uk/dataset/ab9f1f20-3b7f-4efa-9bd2-239acf63b540/data-zone-boundaries-2011>
- 441 27. Meakin S, Abbott S, Funk S. NHS trust level Covid-19 data aggregated to a range of spatial scales. 2021.
- 442 28. Lower Layer Super Output Area (2011) to Ward (2019) Lookup in England and Wales - Office for National  
443 Statistics [Internet]. [cited 2021 Aug 6]. Available from: <https://geoportal.statistics.gov.uk/datasets/ons::>

- 444 lower-layer-super-output-area-2011-to-ward-2019-lookup-in-england-and-wales/about
- 445 29. Bartko JJ. The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychol Rep.* 1966 Aug  
446 1;19(1):3–11.
- 447 30. Fisher RA. Statistical Methods for Research Workers. In: Kotz S, Johnson NL, editors. Breakthroughs  
448 in Statistics: Methodology and Distribution. New York, NY: Springer; 1992. pp. 66–70. (Springer Series in  
449 Statistics).
- 450 31. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability  
451 Research. *J Chiropr Med.* 2016 Jun;15(2):155–63.
- 452 32. Gamer M, Lemon J, Gamer MM, Robinson A, Kendall’s W. Package “irr”. Various coefficients of  
453 interrater reliability and agreement. 2012;22.
- 454 33. SGSS and CHESS data [Internet]. NHS Digital; [cited 2020 Dec 18]. Available from:  
455 [https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/directions-and-data-](https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/directions-and-data-provision-notice/data-provision-notice-dpns/sgss-and-chess-data)  
456 [provision-notice/data-provision-notice-dpns/sgss-and-chess-data](https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/directions-and-data-provision-notice/data-provision-notice-dpns/sgss-and-chess-data)
- 457 34. Open Door Logistics - Intelligent software for vehicle routing & territory management [Internet]. [cited  
458 2020 Dec 18]. Available from: <https://www.opendoorlogistics.com/data/>
- 459 35. Prener C, Revord C, Fox B. Areal: Areal Weighted Interpolation [Internet]. 2020 [cited 2020 Dec 18].  
460 Available from: <https://CRAN.R-project.org/package=areal>
- 461 36. Gregory S. Finding overlapping communities in networks by label propagation. *New J Phys.* 2010  
462 Oct;12(10):103018.
- 463 37. Sun H-L, Huang J-B, Tian Y-Q, Song Q-B, Liu H-L. Detecting overlapping communities in networks via  
464 dominant label propagation. *Chinese Phys B.* 2015 Jan;24(1):018703.

## 465 **Declarations**

466 *Funding:* RC and KTA gratefully acknowledge the financial support of the EPSRC via grant EP/N014391/1.  
467 KTA gratefully acknowledge the financial support of The Alan Turing Institute under the EPSRC grant  
468 EP/N510129/1 and EP/T017856/1. RC is supported by NHS England, Global Digital Exemplar programme  
469 and the MRC (MC/PC/19067). LL acknowledges the financial support of the EPSRC via Early Career  
470 Fellowship EP/P01660X/1. GJG is supported by an ESRC postdoctoral fellowship ES/T009101/1.

471 *Authors contributions:* All authors discussed the concept of the article and RC wrote the initial draft. KTA,  
472 LL, GG commented and made revisions. All authors read and approved the final manuscript. RC is the  
473 guarantor.

474 *Competing interests:* The authors declare no financial relationships with any organisations that might have  
475 an interest in the submitted work in the previous three years, no other relationships or activities that could  
476 appear to have influenced the submitted work.

477 *Ethics approval and consent to participate:* Data on cases were obtained from the COVID-19 Hospitalisation  
478 in England Surveillance System (CHESS) data set that collects detailed data on patients infected with  
479 COVID-19. These data contain confidential information, with public data deposition non-permissible for  
480 socioeconomic reasons. The CHESS data resides with the National Health Service ([www.nhs.gov.uk](http://www.nhs.gov.uk)). Data  
481 from the CHESS database were supplied after anonymisation under strict data protection protocols agreed  
482 between the University of Exeter and Public Health England. The ethics of the use of these data for these  
483 purposes was agreed by Public Health England with the Government's SPI-M(O) / SAGE committees.

484 *Availability of data and materials:* The CHESS data that support the findings of this study are available  
485 from Public Health England but restrictions apply to the availability of these data, which were used under  
486 license for the current study, and so are not publicly available. Data are however available from the authors  
487 upon reasonable request and with permission of Public Health England. All other data and materials are  
488 available as an R package with source code available on GitHub.

489 *Consent for publication:* Not applicable

490 *Acknowledgments:* Many thanks to TJ McKinley at the University of Exeter, for feedback on an early draft  
491 of the paper and testing of the implementation of the algorithm.

## 492 **Supplementary material - Estimating surge hospital capacity in** 493 **Britain during the COVID-19 pandemic**

494 Identifying a set of capacity data for the NHS proved complex. After several attempts to integrate data  
495 from various sources, we ultimately performed a manual curation of the sources listed below, with gaps or  
496 inconsistencies filled in by consultation with the relevant hospital's website. The resulting list is a snapshot in  
497 time of capacity and not representative of up to date practice. During the course of the COVID-19 pandemic  
498 a small number of NHS trusts merged which had to be manually adjusted for. There are also significant  
499 limitations due to the different ways the devolved administrations of the UK (England, Wales, Scotland and  
500 Northern Ireland) reported situation report of bed capacity during the pandemic, which meant only England  
501 and Wales hospitals has assessments of surge capacity, and we had no reliable information about Northern  
502 Ireland at all, and hence it was excluded. This does not significantly alter our conclusions here about the  
503 nature of the algorithm, but should be borne in mind, if the data set is to be used for other purposes.

### 504 ***NHS and Trust GIS locations (England):***

- 505 • <https://www.nhs.uk/about-us/nhs-website-datasets/>
- 506 • Lists of independent and NHS hospitals and trusts with location data
- 507 • public

### 508 ***NHS Trusts (England)***

- 509 • <https://www.nhs.uk/ServiceDirectories/Pages/NHSTrustListing.aspx>
- 510 • Lists of NHS trusts and locations (as postcode) with information about services offered and hospital  
511 sites
- 512 • public

### 513 ***Beds open - NHS England:***

- 514 • [https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/bed-](https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/bed-data-overnight/)  
515 [data-overnight/](https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/bed-data-overnight/)
- 516 • [https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/bed-](https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/bed-data-day-only/)  
517 [data-day-only/](https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/bed-data-day-only/)
- 518 • Information at an NHS trusts level on hospital beds and icu beds available
- 519 • public

### 520 ***Critical care capacity in England (pre-pandemic):***



521 • <https://www.england.nhs.uk/statistics/statistical-work-areas/critical-care-capacity/critical-care-bed->  
522 [capacity-and-urgent-operations-cancelled-2019-20-data/](https://www.england.nhs.uk/statistics/statistical-work-areas/critical-care-capacity/critical-care-bed-capacity-and-urgent-operations-cancelled-2019-20-data/)

523 • Prepandemic NHS trust bed and ICU capacity

524 • public

525 ***Wales:***

526 Average daily beds by site:

527 • <https://statswales.gov.wales/v/Hg4K>

528 • Prepandemic ICU and general bed availability

529 • public

530 ***Scotland:***

531 Annual trends in available beds:

532 • [https://www.isdscotland.org/Health-Topics/Hospital-Care/Publications/data-tables2017.asp?id=](https://www.isdscotland.org/Health-Topics/Hospital-Care/Publications/data-tables2017.asp?id=2494#2494)  
533 [2494#2494](https://www.isdscotland.org/Health-Topics/Hospital-Care/Publications/data-tables2017.asp?id=2494#2494)

534 • Prepandemic Hospital and ICU bed capacity

535 • public

536 ***Sitrep (Situation reports) data:***

537 **England:**

538 • filename: Covid sitrep report incl CIC 20200408 FINAL.xlsx

539 • Acute and ICU beds available in England at site level

540 • ICU (SIT032) and HDU (SIT033) beds available - many data quality issues and missing trusts

541 • restricted

542 **Wales:**

543 • filename: NHSWalesCovid19Sitrep-20200408.csv

544 • Acute and ICU beds available in Wales

545 • restricted

546 N.B. No sitrep data for Scotland or for Northern Ireland

# 5. Meta-analysis of the SARS-CoV-2 serial interval and the impact of parameter uncertainty on the COVID-19 reproduction number

Status: published in Statistical Methods in Medical Research (Pandemic special issue)

Date: 21st Dec 2021.

Candidate's contribution: All authors discussed the concept of the article and RC performed the analysis and wrote the initial draft. KTA, EB-P, and LD commented and made revisions. All authors read and approved the final manuscript. RC is the guarantor.

All the analysis in this paper was performed by me and findings validated by other team members.

# Meta-analysis of the severe acute respiratory syndrome coronavirus 2 serial intervals and the impact of parameter uncertainty on the coronavirus disease 2019 reproduction number

Statistical Methods in Medical Research  
1–18

© The Author(s) 2021



Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/09622802211065159

[journals.sagepub.com/home/smm](https://journals.sagepub.com/home/smm)

Robert Challen<sup>1,2,3</sup> , Ellen Brooks-Pollock<sup>3,4</sup>,  
Krasimira Tsaneva-Atanasova<sup>1,5,6</sup>  and Leon Danon<sup>4,5,6,7</sup>

## Abstract

The serial interval of an infectious disease, commonly interpreted as the time between the onset of symptoms in sequentially infected individuals within a chain of transmission, is a key epidemiological quantity involved in estimating the reproduction number. The serial interval is closely related to other key quantities, including the incubation period, the generation interval (the time between sequential infections), and time delays between infection and the observations associated with monitoring an outbreak such as confirmed cases, hospital admissions, and deaths. Estimates of these quantities are often based on small data sets from early contact tracing and are subject to considerable uncertainty, which is especially true for early coronavirus disease 2019 data. In this paper, we estimate these key quantities in the context of coronavirus disease 2019 for the UK, including a meta-analysis of early estimates of the serial interval. We estimate distributions for the serial interval with a mean of 5.9 (95% CI 5.2; 6.7) and SD 4.1 (95% CI 3.8; 4.7) days (empirical distribution), the generation interval with a mean of 4.9 (95% CI 4.2; 5.5) and SD 2.0 (95% CI 0.5; 3.2) days (fitted gamma distribution), and the incubation period with a mean 5.2 (95% CI 4.9; 5.5) and SD 5.5 (95% CI 5.1; 5.9) days (fitted log-normal distribution). We quantify the impact of the uncertainty surrounding the serial interval, generation interval, incubation period, and time delays, on the subsequent estimation of the reproduction number, when pragmatic and more formal approaches are taken. These estimates place empirical bounds on the estimates of most relevant model parameters and are expected to contribute to modeling coronavirus disease 2019 transmission.

## Keywords

Severe acute respiratory syndrome coronavirus 2, coronavirus disease 2019, serial interval, incubation period, generation interval

## Introduction

The purpose of this paper is to determine the best estimates for the parameters we need to calculate the effective reproduction number ( $R_t$ ) for the UK, and in particular the key quantities of the serial interval and generation interval. The calculation of  $R_t$  can be made pragmatically using simplifying assumptions, or in a more formal manner, for which other key

<sup>1</sup>EPSRC Centre for Predictive Modelling in Healthcare, University of Exeter, UK

<sup>2</sup>Somerset NHS Foundation Trust, UK

<sup>3</sup>Joint Universities Pandemic and Epidemiological Research (JUNIPER) consortium, UK

<sup>4</sup>Bristol Medical School, Population Health Sciences, University of Bristol, UK

<sup>5</sup>The Alan Turing Institute, British Library, UK

<sup>6</sup>Data Science Institute, College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK

<sup>7</sup>Department of Engineering Mathematics, University of Bristol, UK

## Corresponding author:

Robert Challen, Department of Mathematics & Living Systems Institute, S01.01, University of Exeter, Stocker Road, Exeter, EX4 4QD, Devon, UK.  
Email: [rc538@exeter.ac.uk](mailto:rc538@exeter.ac.uk)

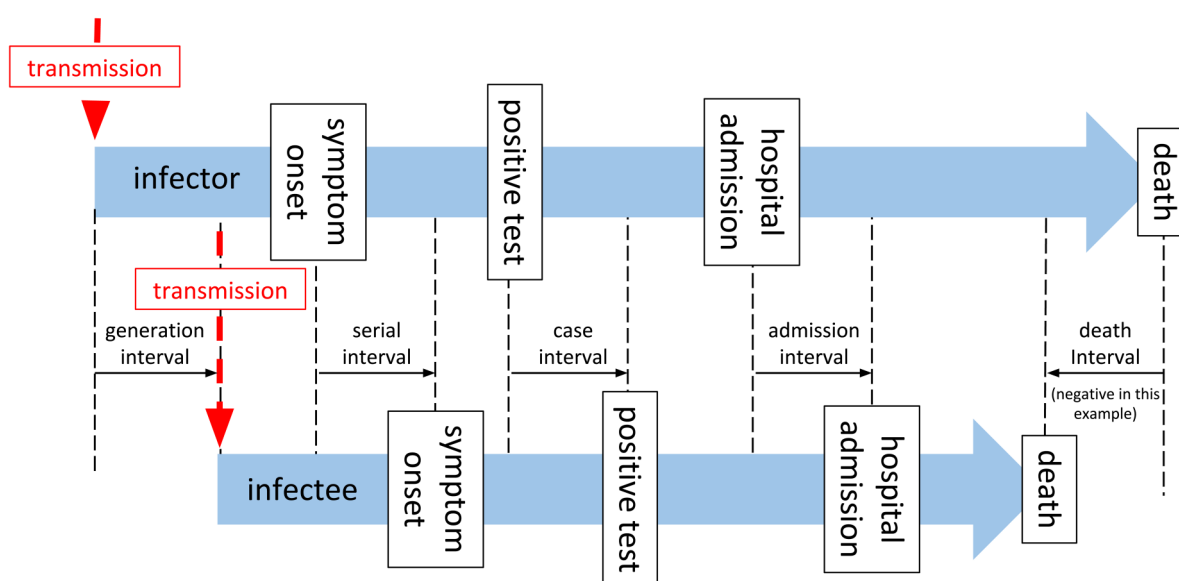
parameters are also required, in particular the delay between infection and diagnosis. Given that these parameters are all associated with uncertainty we investigate how this uncertainty may affect our estimates of the reproduction number, and we qualitatively compare the pragmatic approach compared to the more formal approach.

Since the end of 2019, the novel strain of coronavirus, Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused a global pandemic of disease. The speed with which the virus spreads is dependent on biological determinants that enable viral replication within individuals and onward transmission to others. The minimal set of parameters required for understanding the dynamics of any novel infectious disease pathogen include the potential for transmission, the duration of infectiousness (often captured in models as a recovery rate), and the generational interval: the time between two subsequent cases in a chain of infection.

Estimating the generation interval is particularly challenging due to the fundamentally hidden nature of transmission events. In Figure 1, we summarize the timeline of key events for two adjacent infected individuals (infector and infectee) in a chain of transmission.

As described by Svensson,<sup>1</sup> the generation interval is defined as the time between the infection of an infector and infectee, and in practice is not easy to observe, as infection goes through some latent period, during which it is undetectable,<sup>2</sup> and a pre-symptomatic phase during which an individual may be infectious, but possibly detectable through screening, before the disease manifests with clinical symptoms. The latent period and pre-symptomatic phase are together usually referred to as the incubation period ( $T_{incubation}$ ). From the onset of symptoms, the diagnosis will be confirmed by some canonical test after some time delay ( $T_{onset \rightarrow test}$ ), later the patient may require admission to hospital ( $T_{onset \rightarrow admission}$ ), or may die ( $T_{onset \rightarrow death}$ ). These subsequent events clearly may not happen in that order, and even diagnosis may occur after death. The time between these key events (onset of symptoms, test confirmation of case, hospital admission, and death) for two sequentially infected people in a chain of transmission are described as serial intervals,<sup>1</sup> although in general usage, and in the rest of this paper, the term “serial interval” is taken to mean the interval between onset of symptoms ( $SI_{onset}$ ). In Figure 1 and the rest of this paper, we use the terms “case interval” ( $SI_{case}$ ), “admission interval” ( $SI_{admission}$ ), and “death interval” ( $SI_{death}$ ) to differentiate the other intervals. The generation interval is by definition a non-negative quantity, but all the other measures may be negative if the variation of the delay from the event of infection from person to person exceeds the period between infections. This is more likely for death interval than for serial interval, and for diseases with long pre-symptomatic periods, for example, HIV.<sup>3</sup> The generation interval is defined for all transmission pairs, but the other intervals may or may not be, if for example one of the infector, or infectee is asymptomatic, or does not go to hospital, or does not die, in which case the related intervals are not defined for this pair.

The effective reproduction number ( $R_t$ ) is a key measure of the state of the epidemic. In its simplest form, for a given population, the effective reproduction number is the number of secondary infections that are expected to arise from one primary infection, at any given instant, and depends on the biological properties of the virus, the immune status, and the behavior of that population. Estimation of the effective reproduction number can either be done forward in time,



**Figure 1.** A timeline of events associated with a single infector–infectee pair in a transmission chain.

where we estimate the number of infections that arose from cases detected on a given day (case or cohort reproduction number<sup>4</sup>), or backward in time where we estimate the number of cases that caused the infections observed on a given day (instantaneous reproduction number<sup>5</sup>).

Estimation of the instantaneous reproduction number is implemented in the “renewal equation” method for estimating  $R_t$ , and depends on a time series of infections, and on the “infectivity profile”—a measure of the probability that a secondary infection occurred on a specific day after the primary case, given that a secondary infection occurred.<sup>6,7</sup> A Bayesian framework is then used to update a prior probabilistic estimate of  $R_t$  on any given day with both information gained from the time series of infections and the infectivity profile to produce a posterior estimate of  $R_t$ . From this point, we refer to  $R_t$  to specifically describe the instantaneous reproduction number estimated using the renewal equation method.

The infectivity profile in the renewal equation method is the probability a secondary infection occurred on a particular day after the primary infection, given a secondary infection occurred. This is the same as the probability density function of the generation interval distribution over all transmission pairs. However, in both the original<sup>5</sup> and revised<sup>7</sup> implementations of this method, the authors acknowledge the pragmatic use of the serial interval distribution, as a proxy measure for the infectivity profile, and the incidence of symptom onset or case identification as a proxy for the incidence of infection, with the caveat that these introduce a time lag into the estimates of  $R_t$ . It has been noted by various authors that the use of the serial interval distribution as a proxy for infectivity profiles is a pragmatic choice<sup>5,7</sup> but can introduce a bias into estimates of  $R_t$ .<sup>8,9</sup>

In the COVID-19 outbreak, a limited number of estimates of the serial interval distribution are available from studies of travelers from infected areas, and early contact tracing studies (detailed in Table 2). The infectivity profile is comprised of non-negative values by definition. The serial interval, on the other hand, can be measured as negative for several reasons. For example, if the incubation period of the infector is at the short end of the distribution and that of the infectee is at the tail, a negative serial interval would be observed. Negative values have been noted as a feature in at least one estimate of the serial interval of SARS-CoV-2 to date,<sup>10</sup> but cannot be used in renewal equation-based estimates of  $R_t$ .

Direct measurement of the serial interval distribution is further complicated by the fact that symptom onset is often not observed due to the scale of the outbreak and that infection may be asymptomatic.<sup>11,12</sup> The data available during an epidemic are a time series of counts of observations, such as confirmed cases (test results confirming diagnosis), hospital admissions, and deaths. As depicted in Figure 1 these events occur after infection, following a period of time. Therefore the observed time series of observed cases is a result of the unobserved time series of infections governed by the serial interval, convolved by the distribution of the time delay from infection to case identification.

Since renewal equation-based estimates of  $R_t$  are predicated on infections a “formal” approach to estimation is to infer the unobserved incidence of infection from the observations we have using backpropagation or de-convolution,<sup>3,9</sup> using the generation interval as the infectivity profile. However, this requires knowledge of the temporal relationship between unobserved infections and observed cases, admissions or deaths.

A pragmatic alternative<sup>9</sup> is to simply calculate  $R_t$  using a serial interval as the infectivity profile, and un-adjusted case numbers with a simple correction for the time delay between infection and cases by shifting our  $R_t$  estimates backward in time. As the renewal equation methods assume that the time between two infections cannot be negative, the serial interval distribution must be truncated at zero in this pragmatic approach.

Both formal and pragmatic approaches need estimates of the generation interval or serial interval, and the time delays between infection and symptom onset (incubation period), infection and case identification (test), infection and admission, and between infection and death. The formal method also needs an understanding of the shape of the distribution of these delays. In the rest of this paper, we estimate these key quantities and investigate the impact that uncertainty in these quantities has on an estimation of  $R_t$  (Table 1).

## Methods

### Serial interval estimation

Firstly, we conducted a literature review for studies that describe serial interval estimates using PubMed and the search terms “(SARS-CoV-2 or COVID-19) and ‘Serial interval’” and reviewed the abstracts of relevant original research papers. These were compared to papers reported on the MIDAS Online Portal for COVID-19 Modeling Research,<sup>13</sup> and with existing meta-analyses.<sup>14</sup> From these papers, serial interval mean and standard deviation estimates were extracted, along with information about assumed statistical distributions, and the sample size of the study. A random-effect meta-analysis was conducted<sup>15</sup> on the subset of papers that reported confidence intervals. The assumption of normal distribution underpinning the meta-analysis<sup>16–18</sup> is reasonable for the mean of the serial interval, given the central limit theorem, however, we cannot extend this to the standard deviation of the serial interval distribution and hence cannot

**Table 1.** Comparison of two approaches for estimating the reproduction number and parameters needed to support each approach.

Pragmatic approach	Formal approach
No adjustment prior to estimation of $R_t$ .	Deconvolution of observations (e.g. cases) to putative infection date (requires distribution of time from infection to observation delay requires incubation period estimate and delay from symptoms to observation).
Serial interval as proxy for infectivity profile, truncated at zero if necessary (requires estimate of serial interval distribution).	Generation interval as proxy for serial interval (requires generation interval estimate which depends on serial interval estimate; and incubation period estimate).
Simple shift of $R_t$ estimates to align to date of infection (requires point estimate of time delay from infection to observation, based on point estimate of mean delay of observation).	No adjustment after estimation of $R_t$ required.

assess the overall shape of the serial interval distribution. To address this and combine parameterized distribution estimates from multiple studies into a single distribution we undertook a re-sampling exercise, as described below, the goal of which was to reconstruct a data set of serial intervals that accurately reflect the distributions reported in each of the studies above, so that they could be combined and analyzed together.

Most studies reported a central estimate of a probability distribution for the serial interval, specified in terms of the mean and standard deviation of a given serial interval distribution. They also reported uncertainty on these mean and standard deviations as confidence or credible intervals. For these studies, we randomly selected one hundred probability distributions consistent with the central estimates and confidence intervals reported in each paper (further details are available in Supplemental table 1). From this family of probability distributions, we drew random samples based on the original sample size.<sup>19,20</sup> This gives us a set of samples that are consistent with the shape and uncertainty of the distributions reported by the source studies and which represent their sample size in a comparable way. Other studies reported empirical distributions and for these, we obtained original serial interval data were available and made 100 random bootstrap sub-samples with replacement, to a relative size determined by the original sample size of the study.

The 100 empirical and 100 probability distribution-based serial interval samples were combined into 100 groups, with each group containing serial intervals representative of each source article with numbers proportional to the size of the original study. This allows us to calculate 100 empirical probability distribution estimates from which we can derive summary statistics with confidence intervals (from here on referred to as the “re-sampled serial interval estimate”).

As much of the data is captured at the resolution of a single day, serial intervals appear as an integer number of days in the data. The discrete nature of the data was estimated using continuous distributions by replacing integer values,  $x$ , with interval-censored ranges spanning from the value  $x - 0.5$  to  $x + 0.5$  days. Normal, Weibull, and gamma probability distributions were then fitted to these 100 groups of interval-censored data using maximum likelihood estimation, implemented in  $R$ .<sup>21,22</sup> This additionally gives us 100 parametric probability distribution estimates for each distribution from which we can derive confidence intervals for the parameters and 100 empirical probability distributions estimates of the combination of all the source studies.

As a comparison, we also used data collected under the “First Few Hundred” (FF100) case protocols by Public Health England<sup>23,24</sup> which provides a limited number of linked cases of proven transmission, mostly within households, and interval-censored symptom onset dates. To this data, we fitted normal, gamma, and Weibull distributions using the same methodology as above.

In both cases, when fitting gamma and Weibull distributions we truncated the interval-censored data at zero, to prevent negative values, and for the gamma distributions we required that the shape parameter had a lower bound of 1, which enforces that the distribution density is zero at time zero. This however makes goodness of fit statistics not comparable between the normal distribution fit and the two other distributions.

## Incubation period estimation

The incubation period has been previously estimated by Lauer et al.<sup>25</sup> and Sanche et al.<sup>26</sup> for China in the early phase of the epidemic. The FF100 data contains interval-censored exposure data coupled to symptom onset; using this we derived a UK-specific estimate to assess if it was consistent. Furthermore, the Open COVID-19 Data Working Group<sup>27,28</sup> provides a large international data set that includes some travel history and symptom onset data.

**Table 2.** Sources of serial interval estimates from a literature search.

Reference	Statistic	Mean (95% CrI) days	Std (95% CrI) days	N	Distribution	Population
Bi, Q. et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. <i>The Lancet Infectious Diseases</i> 20, 911–919 (2020).	Serial interval	6.30 (5.20–7.60)	4.20 (3.10–5.30)	48	Gamma	Shenzhen
Cereda, D. et al. The early phase of the COVID-19 outbreak in Lombardy, Italy. arXiv:2003.09320 [q-bio] (2020).	Serial interval	6.60 (0.70–19.00)	4.88 (unk-unk)	90	Gamma	Italy
Du, Z. et al. Serial Interval of COVID-19 among Publicly Reported Confirmed Cases. <i>Emerg Infect Dis</i> 26, 1341–1343 (2020).	Serial interval	3.96 (3.53–4.39)	4.75 (4.46–5.07)	468	Norm	China
Ganyani, T. et al. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. <i>Eurosurveillance</i> 25, 2000257 (2020).	Serial interval	5.21 (– 3.35–13.94)	4.32 (4.06–5.58)	54	Empirical	Singapore
	Serial interval	3.95 (– 4.47–12.51)	4.24 (4.03–4.95)	45	Empirical	Taijin
Kwok, K. O., Wong, V. W. Y., Wei, W. I., Wong, S. Y. S. & Tang, J. W.-T. Epidemiological characteristics of the first 53 laboratory-confirmed cases of COVID-19 epidemic in Hong Kong, 13 February 2020. <i>Eurosurveillance</i> 25, 2000155 (2020).	Serial interval	4.58 (3.35–5.85)	3.28 (2.18–4.01)	26	Inorm	Hong Kong
Li, Q. et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. <i>New England Journal of Medicine</i> 382, 1199–1207 (2020).	Serial interval	7.50 (5.30–19.00)	3.40 (unk-unk)	5	Unknown	Wuhan
Nishiura, H., Linton, N. M. & Akhmetzhanov, A. R. Serial interval of novel coronavirus (COVID-19) infections. <i>Int. J. Infect. Dis.</i> 93, 284–286 (2020).	Serial interval	4.70 (3.70–6.00)	2.90 (1.90–4.90)	28	Inorm	SE Asia
Son, H. et al. Epidemiological characteristics of and containment measures for COVID-19 in Busan, Korea. <i>Epidemiol Health</i> 42, (2020).	Serial interval	5.54 (4.08–7.01)	3.90 (2.47–5.32)	28	Gamma	Korea
Tindale, L. C. et al. Evidence for transmission of COVID-19 prior to symptom onset. <i>eLife</i> 9, e57149 (2020).	Serial interval	4.17 (2.44–5.89)	1.06 (unk-unk)	93	Unknown	Singapore
	Serial interval	4.31 (2.91–5.72)	0.94 (unk-unk)	135	Unknown	Taijin
Xia, W. et al. Transmission of corona virus disease 2019 during the incubation period may lead to a quarantine loophole. medRxiv 2020.03.06.20031955 (2020) doi:10.1101/2020.03.06.20031955.	Serial interval	4.10 (unk-unk)	3.30 (unk-unk)	124	Empirical	China outside Hubei
Xu, X.-K. et al. Reconstruction of Transmission Pairs for novel Coronavirus Disease 2019 (COVID-19) in mainland China: Estimation of Super-spreading Events, Serial Interval, and Hazard of	Serial interval (household)	4.95 (unk-unk)	5.24 (unk-unk)	643	Empirical	China outside Hubei
	Serial interval (non-household)	5.19 (unk-unk)	5.28 (unk-unk)	643	Empirical	China outside Hubei

(continued)

Table 2. Continued

Reference	Statistic	Mean (95% CrI) days	Std (95% CrI) days	N	Distribution	Population
Infection. Clin Infect Dis doi:10.1093/cid/ciaa790.						
You, C. et al. Estimation of the time-varying reproduction number of COVID-19 outbreak in China. International Journal of Hygiene and Environmental Health 228, 113555 (2020).	Serial interval	4.27 (unk-unk)	3.95 (unk-unk)	71	Empirical	China outside Hubei
Zhang, J. et al. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. The Lancet Infectious Diseases 20, 793–802 (2020).	Serial interval	5.00 (0.80– 13.00)	3.22 (unk-unk)	28	Gamma	China outside Hubei
Zhao, S. et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. Int. J. Infect. Dis. 92, 214–217 (2020).	Serial interval	4.40 (2.90– 6.70)	3.00 (1.80– 5.80)	21	Gamma	Hong Kong

Compared to data on which earlier estimates of the incubation period were made<sup>25,26</sup> where travel principally originated from Wuhan, the travel cases in the Open COVID-19 data set include travelers between a far wider set of destinations, and in a later stage of the epidemic, which we believe reduces selection bias.<sup>29</sup> As we are performing this analysis retrospectively we also benefit from fewer issues due to the right censoring of the data.

For both data sets, we use random samples with replacement and fit 100 of each of gamma, Weibull and log-normal probability distributions to the interval between putative exposure and symptom onset, accounting for censoring were present, to estimate the incubation period distribution and uncertainty associated with its parameters. We remove from the data records where the time from exposure to symptom onset is negative or when it is longer than 21 days (on the basis of biological implausibility).

For subsequent phases of the analysis, we retain the 100 parameterized distributions of the distribution with the best overall fit to the data from the Open COVID-19 Data Working Group data set. From these, we can generate a parameterized bootstrap sample representative of the Open COVID-19 Data Working Group data set but without censoring.

## Generation interval estimation

The generation interval is the fundamental variable for modeling transmission. Under the assumption that the generation interval follows a gamma distribution we can infer its parameters using the bootstrap samples from the re-sampled serial interval data set, and the bootstrap samples from the incubation period data set from earlier stages, and the constraint that the generation interval is a non-negative quantity. We combined random samples from a parameterized gamma distribution for the generation interval with two of the bootstrap samples from the incubation period to satisfy the following relationship, thus simulating the serial interval:

$$SI_{onset,A \rightarrow B} = SI_{infection,A \rightarrow B} + T_{incubation,B} - T_{incubation,A}$$

$$E[SI_{onset}] = E[SI_{infection} + T_{incubation} - T_{incubation}]$$

The mean and standard deviation of the simulated serial interval distribution were then compared to the empirical re-sampled serial interval distributions we estimated in an earlier stage. The parameters for the generation interval distribution were then optimized by a recursive linear search on the standard deviation, with the constraints that the mean of the simulated and empirical distributions must be the same,<sup>8</sup> and the standard deviation must be smaller than the mean (ensuring the gamma function scale parameter is larger than one and hence the density is zero at time zero). The minimization function we employed was the absolute difference in inter-quartile ranges of simulated and observed distributions. This



process was repeated for 100 different simulated samples that were compared to the 100 different empirical re-sampled serial interval estimates from the previous stage of our analysis to get confidence intervals on our estimates of the generation interval distribution. This approach does not use the parameterized serial interval estimate and only the samples direct from the combination of data and literature. For the incubation period, we use parametric re-sampling of multiple fits to the original interval-censored incubation period data to generate bootstrap samples. This approach minimizes any assumptions we make about the distribution of these other quantities when estimating the generation interval.

## Impact of using the serial interval on an estimation of $R_t$

With various estimates of serial interval and generation interval, we wished to understand the qualitative impact this variation might have on our estimates of  $R_t$ . To investigate this we used the forward equation approach implemented in the R library EpiEstim.<sup>5-7</sup> We estimate values of  $R_t$  for 4 time points in the first wave of the COVID-19 pandemic in England representative of the ascending phase, the peak, the early descending phase, and the late descending phase. We used data retrieved from the Public Health England API<sup>30,31</sup> representing cases with positive test results, hospital admissions, and deaths within 28 days of a positive test, in the first wave of the outbreak in England. For this analysis we assume the infectivity profile can be represented using a parameterized gamma distribution, and estimate  $R_t$  for a wide range of combinations of mean and standard deviation, using a fixed calculation window of 7 days, at each of our 4 time points. The resulting relationship between  $R_t$ , mean infectivity profile, and standard deviation of infectivity profile were compared visually to qualitatively examine how the serial interval estimates influence the estimates of  $R_t$ .

## Time delays from infection to case identification, admission, and death

Estimation of the time interval between the onset of symptoms and the observations of the positive test result, hospital admission, and death was performed ( $T_{onset \rightarrow test}$ ,  $T_{onset \rightarrow admission}$ , and  $T_{onset \rightarrow death}$ ) using the CHES data set.<sup>32</sup> The CHS data set is hospital-based, and was initially limited to intensive care admissions, but a subset of hospitals have reported all admissions, and this is what we focused on (see Supplemental table 2). Within the CHES data set there are a set of patients who have symptom onset dates recorded, dates that a specimen was taken that subsequently was tested positive, hospital admission date, and date of death, if the patient died. We restricted cases to those in which a positive test was found no more than 14 days before and 28 days after symptom onset on the grounds of biological implausibility. Because we conducted this analysis at the end of the first wave of COVID-19 in the UK when patient numbers in hospitals had fallen to a low level there was minimal right censoring present in the data and this was not accounted for.

The time delay from infection to observation was obtained by combining our estimate of the incubation period distribution from the Open COVID-19 Data Working Group data set<sup>27,28</sup> with onset to observation delays from the CHES data set<sup>32</sup> using the following relationship:

$$\begin{aligned} T_{infection \rightarrow observation} &= T_{infection \rightarrow onset} + T_{onset \rightarrow observation} \\ &= T_{incubation} + T_{onset \rightarrow observation} \end{aligned} .$$

We combined the incubation period and onset to observation distributions using a random sampling approach, assuming the independence of the two variables. These random samples were then estimated as parameterized statistical distributions in the same manner as above, with the constraint that all the time delays from infection to observation are non-negative quantities, and their probability is zero at time zero.

The resulting time delays from infection to the different observations of the positive test result, hospital admission, or death were calculated and fitted to probability distributions using the same maximum likelihood estimation methods implemented in  $R$ <sup>22</sup> as described above.

## Impact of deconvolution

In the final piece of our analysis, we sought to qualitatively compare estimates of  $R_t$  based on data for England from the Public Health England API,<sup>30,31</sup> in each of the following two scenarios.

Firstly, following the ‘‘pragmatic’’ course outlined in the introduction, we based  $R_t$  estimates on observational data (cases, admissions, deaths) as a proxy for infection events, and used a truncated empirical serial interval distribution derived from our re-sampling procedure as a proxy for infectivity profile. In line with the pragmatic approach, simple to the dates of these  $R_t$  estimates was made to align the estimate to date of infection rather than the date of observation.

Secondly, the more formal approach was employed: using the time delay distributions from the previous stage, we used de-convolution to infer a set of time series of infections from the same observational data and used our estimate of the generation interval as a proxy for the infectivity profile. This second approach has been recommended by Gostic et al.<sup>9</sup> To do this, we applied a non-parametric back-projection algorithm from the surveillance R package,<sup>33</sup> based on work by Becker et al.<sup>3</sup> and Yip et al.,<sup>34</sup> to infer three putative infection time series from observed cases, admissions or deaths. The inferred time series were then used to estimate  $R_t$  through EpiEstim using the parametric gamma-distributed estimate of the generation interval from above. In applying the de-convolution we discovered it requires a full-time series beginning with zero cases for sensible results and this required we impute the early part of the hospital admission time series, which we did by assuming an early constant exponential growth phase.

In both cases, we used the renewal equation method with a 14-day sliding window to estimate a continuous time series of  $R_t$ . The resulting  $R_t$  time series were compared qualitatively.

## Results

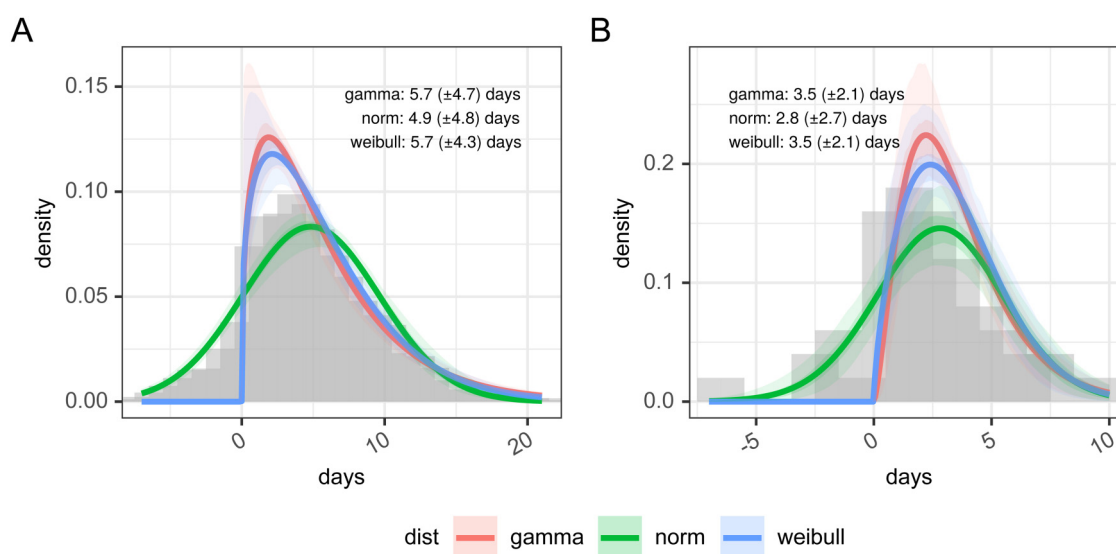
### Serial interval estimation

Our PubMed search retrieved 62 search hits of which 14 were original research articles containing estimates of serial intervals.<sup>10,35–48</sup> The mean and standard deviation of parameterized distributions were extracted and are presented in Table 2. The estimates of the mean range from 3.95 to 7.5 days. The majority of studies provided their results as gamma distributions defined by mean and standard deviation. Some studies, particularly Xu et al.<sup>10</sup> noted that the serial interval was not infrequently negative.

The random-effects meta-analysis on the subset of studies that reported modeled distributions, resulted in an overall estimate of the mean of the serial interval of 4.83 (95% CI 3.93–5.70) (for more details see Supplemental figure 1). This estimate is limited in its value by the fact that it only covers the subset of the studies in the table above, and does not represent the distributional nature of the serial interval.

In Figure 2 panel A, we present the results of the re-sampled serial interval estimate. The histogram shows the empirical distribution of the combination of all the studies, reinforcing the finding of a substantial proportion of the serial interval being negative. For the gamma and Weibull distribution fit the data is truncated at zero, and the full data is used for the normal distribution, resulting in mean values of 5.72 (gamma), 4.87 (norm), or 5.70 (Weibull) days. Full detail of the parameterization of this is available in Supplemental table 3.

In Figure 2 panel B, we present the distribution of the 50 linked cases in the FF100 data set for which onset dates are available for both infector and infectee. As with the re-sampled data the parameterized versions of this are based on truncated data for Weibull and gamma distributions, and hence show a poorer fit against the whole distribution (full detail of the parameterization of this is available in Supplemental table 4). Supporting the observations of Xu et al.,<sup>10</sup> the FF100 data



**Figure 2.** Panel A: days between infected infectee disease onset based on resampling of published estimates from the literature and panel B: estimates of the serial interval from FF100 data. The histogram in panel A shows the combined density of all sets of samples within the original research.

shows evidence of negative serial intervals. The mean of the serial interval from FF100 data was 3.50 (gamma) or 3.52 (Weibull) when data was truncated to exclude negative serial intervals and 2.81 (norm) with no truncation. This is on the lower end of the values reported in the literature.

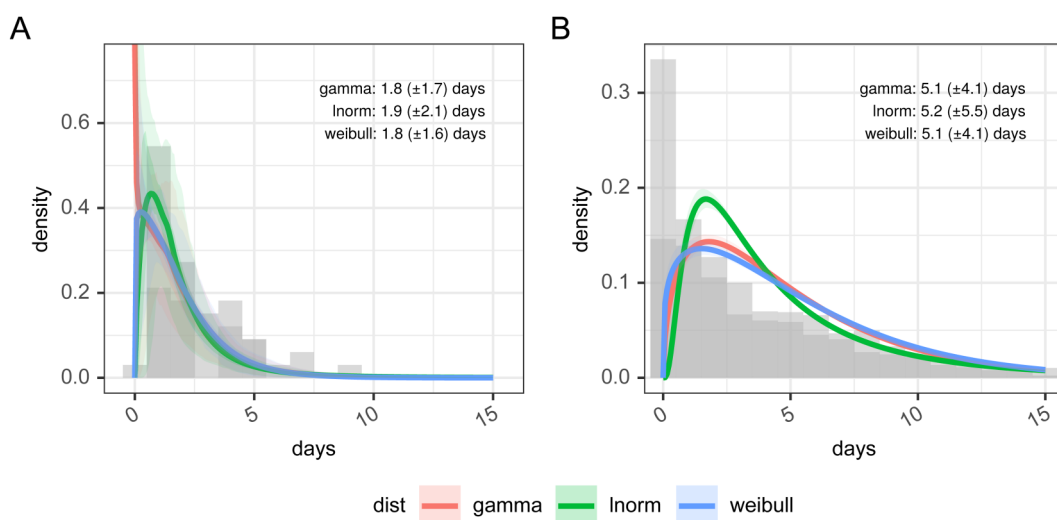
As noted in the methods, the re-sampling process allows us to estimate the serial interval as an empirical distribution. Within EpiEstim, our chosen framework for estimating  $R_t$ , however the use of negative serial intervals is not supported as a proxy for the infectivity profile. In the pragmatic approach to estimating  $R_t$ , we truncate the empirical distribution at zero to support this. Once truncated the resulting estimate, therefore, has a mode of 3.86 days, shorter than the normal distribution parameterization and is a truncated empirical distribution, with a mean plus 95% confidence interval of 5.88 days (5.22; 6.70), and a standard deviation of 4.12 days (3.79; 4.72), which is more in line with the gamma distribution parameterization.

### Incubation period estimation

Figure 3 and Table 3 show the results of estimating a parametric probability distribution to data from FF100 and data from the Open COVID-19 Data Working Group. Histograms of the data are not shown as it is interval-censored, which is not straightforward to represent graphically. There are only a small number of records from the FF100 data which suggest the mean of the incubation period is between 1.82 and 1.94 days. The data from the Open COVID-19 Data Working Group suggests the incubation period is longer with a mean of 5.19 days with best fitting distribution, and this agrees better with other estimates in the literature.<sup>14,25,26</sup> The best fit to the Open COVID-19 data is obtained with a log-normal distribution as shown in Table 3, with the lowest Akaike information criterion, Bayesian Information Criterion (both representing least information lost), and the largest value for the log-likelihood. (Full details of the fitting parameters and graphical assessment of the quality of fit are in Supplemental table 5 and Supplemental figure 2).

### Generation interval estimation

The generation interval is then inferred from the incubation period and empirical serial interval distribution prior to truncation. Our best estimate for this is a gamma distribution, with a mean plus 95% confidence interval of 4.87 days (4.24; 5.51), and a standard deviation of 1.98 days (0.53; 3.19), as shown in Figure 4. The mean of 4.87 days is identical to that of the empirical serial interval distribution prior to truncation in panel B, Figure 2 as a result of the constraints imposed during fitting. The standard deviation of our generation interval estimate is 1.98. This is within the confidence limits of estimates from the literature from both China (0.74–2.97) and Singapore (0.91–3.93).<sup>49</sup> The mean and standard deviation of these distributions are not independent, and this is explored further in Supplemental figure 3.



**Figure 3.** Incubation period distributions were reconstructed from the Open COVID-19 Data Working Group and from FF100 data. Histogram data is approximate due to interval censoring.

**Table 3.** Goodness of fit statistics for incubation period distributions reconstructed from open COVID-19 data working group and from FF100 data.

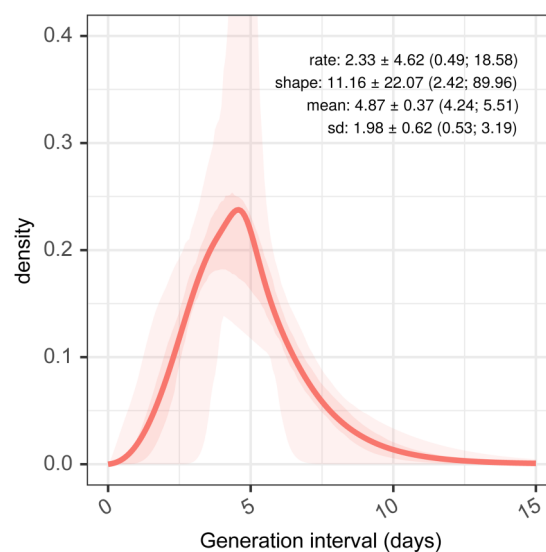
Source	N	AIC	BIC	Log-likelihood	Distribution
FF100	33	62.1	65.1	-29.1	Gamma
		62.1	65.1	-29.0	Weibull
		63.5	66.5	-29.7	Log-normal
Open COVID-19 Data Working Group	1062	5157.2	5167.1	-2576.6	Log-normal
		5191.7	5201.6	-2593.8	Gamma
		5216.6	5226.5	-2606.3	Weibull

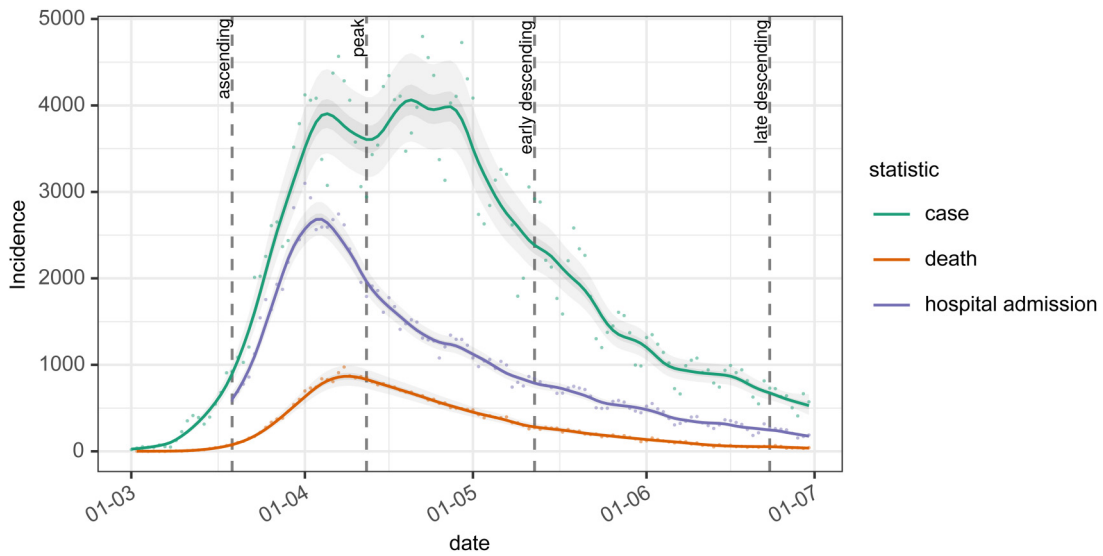
### Impact of using the serial interval on an estimation of $R_t$

With our three estimates of the serial interval and one generation interval and observed COVID-19 case counts, we investigate the impact on the estimates of  $R_t$ , of using these estimates as a proxy for the infectivity profile. This uses data at 4 time points on an epidemic curve from the first wave of the COVID-19 outbreak in England, as shown in Figure 5, which are 19 March, 12 April, 12 May, and 23 June, corresponding to the ascending, peak, early descending, and late descending phases, respectively.

At each of these 4 time points, Figure 6 shows the estimated  $R_t$  under the range of different assumptions about the mean and standard deviation of the infectivity profile, modeled as a gamma distribution. In the top left, bottom left and bottom right panels the effect of increasing the mean of the infectivity profile is to push the resulting estimate of  $R_t$  away from the critical value of 1 at which the epidemic is growing. In the top right panel, at the peak, the mean of the infectivity profile has a less clear-cut effect. The impact of changes to standard deviation is likewise varied. In the top left, bottom left and bottom right panels during ascending and descending phases there is relatively little impact of changing the standard deviation of the infectivity profile on the estimates of  $R_t$ , and any small changes that do occur depend on the shape of the preceding epidemic curve. At the peak, however, in the top right panel, the wider the standard deviation the more historical information influences the estimation of  $R_t$  and this acts to delay the estimated transition from positive to negative growth. The overall result of this is that estimates of the infectivity profile with a high standard deviation will predict  $R_t$  crossing 1 later than estimates based on an infectivity profile with a low standard deviation, but the point of crossing 1 is relatively insensitive to the value of the mean of the infectivity profile.

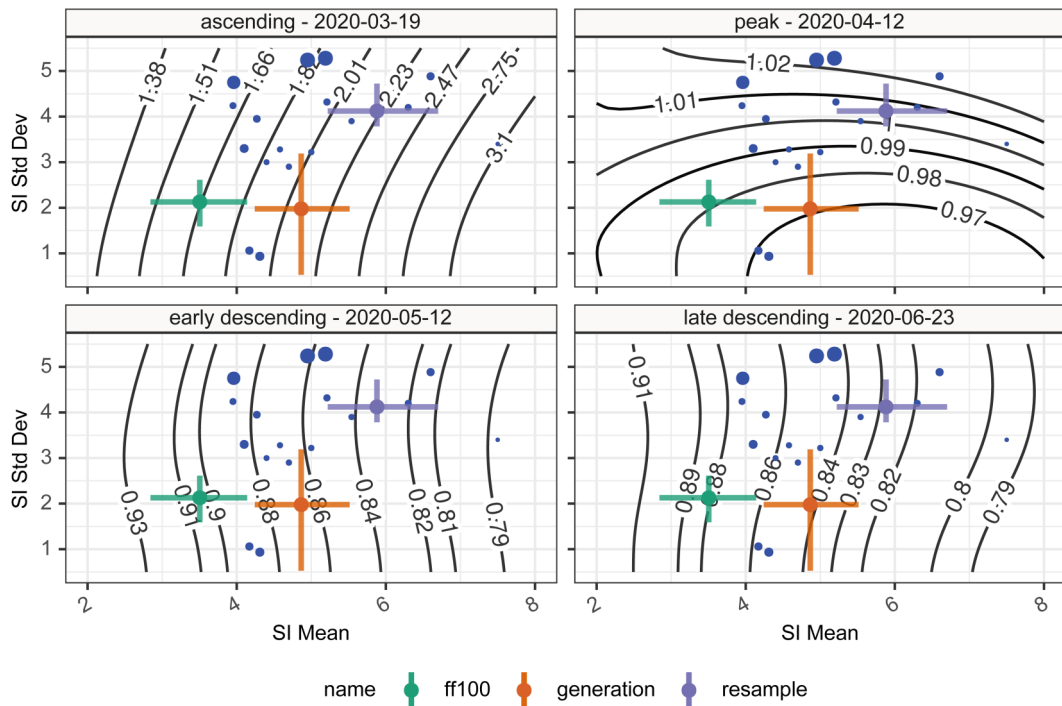
When we consider using the various estimates of the serial interval or generation interval, as a proxy for the infectivity profile, on the resulting estimates of  $R_t$  we can see from the colored crosses in Figure 6 representing the different estimates

**Figure 4.** Estimated generation interval distributions, from resampled serial intervals as a predictor, and estimated serial intervals from incubation period combined with samples from a generation interval assumed as a gamma-distributed quantity.



**Figure 5.** The epidemic curve for cases, deaths, and hospital admissions are used for analysis in this paper. Dashed vertical lines show dates at which we conduct our analysis, chosen to represent the ascending, peak, early, and late descending phases of cases during the first wave in the UK.

of serial or generation interval, that in the situations of dynamic change such as the ascending phase the variability may have quite a large impact on subsequent estimates of  $R_t$ , but at other times the impact is much smaller [ascending—28% variation ( $R_t$ : 1.66–2.20); peak—3% variation ( $R_t$ : 0.97–1.00); early descending—8% variation ( $R_t$ : 0.83–0.90); late descending—6% variation ( $R_t$ : 0.83–0.88)].



**Figure 6.** Time-varying reproduction numbers given various assumptions on the serial interval mean and standard deviation. The blue points show the central estimate of serial intervals from the literature, whereas the colored error bars show the mean and standard deviation of the two serial intervals (green, violet) and one generation interval (orange) estimates presented in this paper. Contours show the  $R_t$  estimate for that combination of mean and standard deviation serial interval. The four panels represent the four different time points investigated.

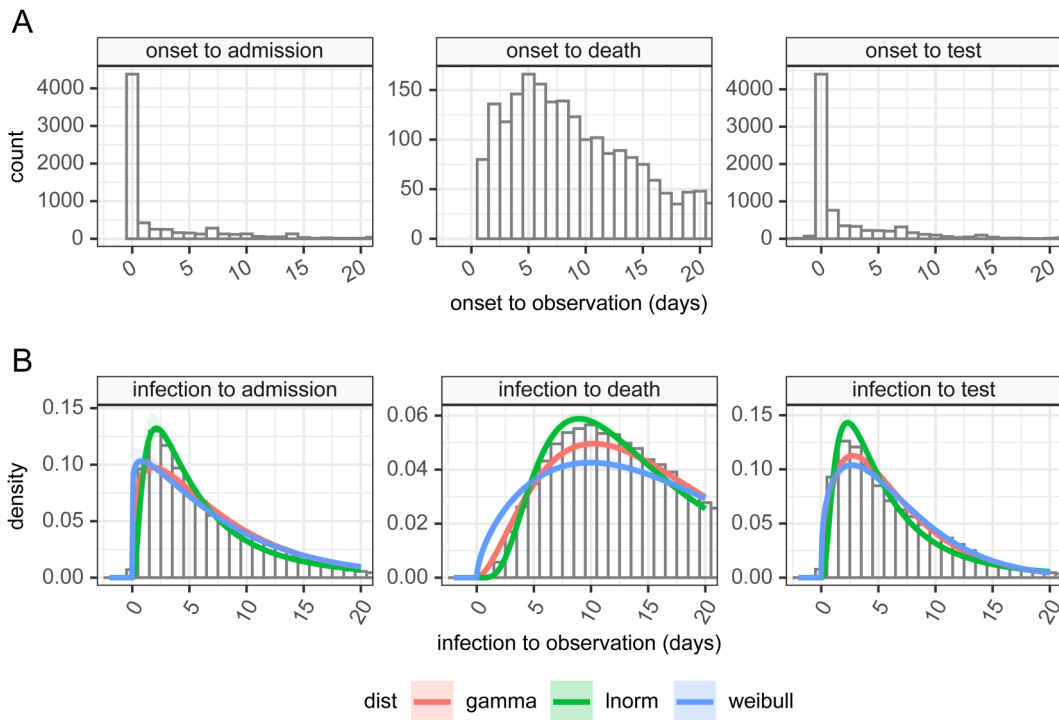
## Time delays from infection to case identification, admission, and death

There are 9902 patients in the subset of the CHES database we examined of which 82.3% had a symptom onset date. In Figure 7, we show probability distributions fitted to data from the CHES data set which define times from symptom onset (Panel A) to case identification ( $T_{onset \rightarrow case}$ ), admission ( $T_{onset \rightarrow admission}$ ), or death ( $T_{onset \rightarrow death}$ ). Symptom onset to test (case identification) can be a negative quantity if a swab is taken during disease screening and the patient is pre-symptomatic. In this data the time point that defines the time of test is the date when the specimen is taken, which will subsequently be tested positive for SARS-CoV-2, so does not include sample processing delays. However, in this hospital-based data source of admitted patients, the onset data were collected retrospectively. We also note peaks at 1 day, 1 week, 2 weeks, and so on which suggests approximation on data entry, and there may well be biases in the data collection.

It is more obvious from the clinical course of COVID-19 that admission should occur after disease onset. In the data, a large number of cases are reported to have symptom onset on the day of admission. This is potentially a reporting artifact as in the absence of certain knowledge about the onset, it is possible that the day of admission may have been captured instead, and we again see the peaks at 1 week, 2 weeks, and so on, suggesting approximations in data entry.

The time between the onset of symptoms and death can also be assumed as a positive quantity given this is based on an in-hospital cohort. This distribution shows a large tail, and some patients in that tail were noted to be admitted many months before the appearance of COVID-19. These patients likely represent hospital-acquired cases in chronically unwell patients. The extreme outlying values (with delay from admission to death greater than 100 days, or with admission before 1 January 2020) were removed as they prevented sensible estimation of the rest of the distribution.

By combining the incubation period distribution in panel B in Figure 3 with the time delay distributions in panel A of Figure 7, we can obtain probability distributions from infection to observation, and these are shown in panel B for the three observations of test (case identification), admission, and death. These distributions provide us with a means of estimating a time series of infection from observed case counts, admissions, and deaths. As above full details of their parameterizations are available in Supplemental table 6. The mean time from infection to the various time points described in the timeline in Figure 1 is presented in Table 4. The infection to onset is the incubation period, with a mean of 4.2 days. On average 6.4 days pass from infection to diagnosis, a subsequent 1.2 days until admission, and a further 8.3 days until death, however, it also shows considerable variation in these delays, exemplified by the 95% quantiles for the time from infection to death estimated as ranging from 3.6 to 42.9 days.



**Figure 7.** Panel A: time delay distributions from symptom onset to test (diagnosis or case identification), admission or death, estimated from CHES data set, plus in panel B estimated delays from infection to observation, and can be negative in certain cases, based on the incubation period and observation delay. These can be used for deconvolution.

## Estimation of $R_t$

With the estimates of a delay from infection to observation, we are able to use a non-parametric back-propagation as described in the methods to estimate a time series of infections as recommended by Gostic et al.<sup>9</sup> when using EpiEstim. The results of the back-propagation are shown in Figure 8, panel A which depicts the resulting point and smoothed infection curves associated with observation curves from England. The back-projection results in a sharper and narrower epidemic curve than the observation it is derived from and indicates additional structures (such as more pronounced fluctuations) which are not obvious from the underlying observations. Estimates of  $R_t$  are shown in panel B based on de-convolved time series plus generation interval, versus raw observation counts, re-sampled serial interval estimates, with time adjustment on the resulting  $R_t$  estimates to align  $R_t$  estimate date to putative date of infection. We have not calculated confidence intervals for the estimates of  $R_t$ . The estimates of  $R_t$  differ from their mean (using symmetric mean absolute percentage error, sMAPE) by between 3.33% and 6.51% [case: sMAPE 3.44% (IQR 1.86%; 5.17%); death: sMAPE 6.51% (IQR 2.62%; 12.24%); hospital admission: sMAPE 3.33% (IQR 1.46%; 5.60%)].

## Discussion

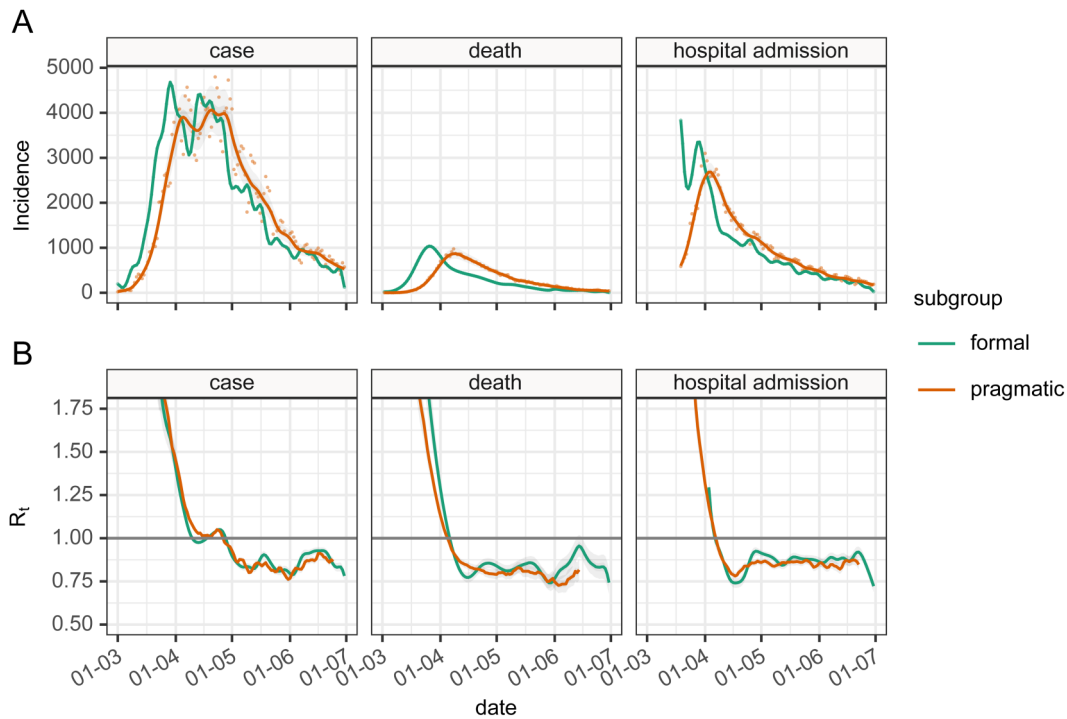
Our estimate of the serial interval from the FF100 UK data was found to be short, compared to international estimates. The FF100 data was collected in the early stage of the epidemic and is based mainly on household contacts of international travelers which may impart bias. As the participants in the study were put into self-isolation upon discovery observations the contact period tends to be shortened, leading to shorter serial interval estimates, and because of this, we do not believe the estimate from the FF100 study to be specific to the UK but rather that the data set is not broadly representative of the UK population.

In contrast, our literature search allowed us to estimate the serial interval from a much larger pooled sample drawn from multiple studies. Random effects and re-sampling meta-analyses produced comparable results (random effects: mean 4.83 (95% CI 3.93–5.70) days and re-sampling: mean 5.9 (95% CI 5.2; 6.7) and SD 4.1 (95% CI 3.8; 4.7) days), but our re-sampling study more clearly showed the potential for the serial interval to be negative, due to the relatively long and variable incubation period of SARS-CoV-2. This is seen in the unsatisfactory fits of parameterized probability distributions in Figure 2, and we choose not to use these in our pragmatic estimation of  $R_t$ , instead of which we rely on the set of empirical distributions resulting from our sampling analysis. Even so, the negative values of the serial interval within these are theoretically problematic for their use as a proxy for the infectivity profile of SARS-CoV-2 in transmission modeling, which requires the infectee is infected after the infector. As a way around this, we truncate the re-sampled serial interval at zero and use this as the infectivity profile, but as seen here this truncation increases the mean of the resulting serial interval distribution from 4.87 to 5.88 days.

An alternative to this is to use the generation interval as a proxy for the infectivity profile, but there are limited estimates of this quantity available in the literature.<sup>49</sup> Derivation of the generation interval from the serial interval is possible using knowledge of the incubation period. We again looked at the FF100 data for estimates of the incubation period and again found that the UK data suggests a value that is shorter than international estimates, for the same reasons as mentioned above.<sup>25,45,50–52</sup> We cross-referenced this with a second estimate derived from the Open COVID-19 Data Working Group data set,<sup>27,28</sup> which is based on a large international data set of people who tested positive for SAR-CoV-2 after traveling from areas with outbreaks. The resulting estimates of the mean incubation period of 5.5 days (log-normally distributed) are much closer to previous estimates, and we expect to be less influenced by right censoring. Again, we regard the short incubation period calculated from the FF100 data set to be a feature of the data set rather than a UK-specific finding. In fact, the estimate based on the Open COVID-19 Data may in itself be an under-estimate as the majority of travel histories only include the return date of the visit in question and not the start date.

**Table 4.** Estimated time delays between infection and various observations over the course of an infection, based on the combination of incubation period and symptom onset to observation delay.

Observation	Mean delay (days)	SD (days)	95% quantiles (days)
Onset	4.21	3.00	0.64; 11.99
Test	6.43	4.97	0.78; 19.18
Admission	7.64	8.73	0.78; 30.67
Death	15.98	10.90	3.59; 42.87



**Figure 8.** Panel A: The epidemic incidence curves in England for different observations (orange—formal) and inferred estimates of infection rates (green—pragmatic) based on deconvolution of the time delay distributions. Panel B: the resulting  $R_t$  values were calculated either using infection rate estimates and generation interval (formal subgroup) or unadjusted incidence of observation and serial interval (pragmatic). The  $R_t$  estimated directly from observed incidence curves (pragmatic) have their dates adjusted by the mean delay estimate.

With incubation period and serial interval estimates, we derived an estimate for generation interval, assuming a gamma distribution. This was comparable to previous estimates in that although it is based on a different serial interval, and hence has a different mean, the standard deviation of our estimate is in accordance with that estimated by Ganyani et al.<sup>49</sup> Although the confidence intervals for the mean and standard deviation of the generation interval are comparable, the variation in the shape and rate parameters of the underlying gamma distribution are quite large. Care should be taken when generating bootstrap samples from the generation interval when it is specified as an uncertain gamma distribution parameterized with mean and standard deviation, as it is possible that some combinations of parameters produce unrealistic distributions, particularly when the standard deviation is small and the mean is large, which could, in theory, result in posterior estimates for  $R_t$  being largely determined by the reciprocal of just a small number of observations.

Using estimates of the serial interval distribution and generation interval distribution as a proxy for infectivity profile, we investigated the resulting variation in the estimation of  $R_t$  using incident cases in England and the forward equation approach. We found the bias between the smallest and largest of our estimates to be as high as 20–25% of the central estimate of  $R_t$  when  $R_t$  was high, but somewhat smaller when  $R_t$  values were 1 or lower. Distributions with lower values of the mean tended to result in estimates of  $R_t$  that were closer to one. This suggests that biases introduced by the use of different serial interval distributions should not influence the answer to the key question, “is  $R_t$  greater or less than 1?” which defines whether the epidemic is expanding or contracting in size. It is also to be expected that the nature of change of  $R_t$  over time is not affected by this bias, so an increasing value of  $R_t$  will be increasing regardless of the infectivity profile that is used to estimate it.

Estimating  $R_t$  is based on knowledge of the incidence of infections. This is not a quantity that is readily observed in the SARS-CoV-2 epidemic, and pragmatically use of observations including symptom onset, case identification, admission, and death are expected to be used as proxy measures for infection.<sup>5,7</sup> However, as pointed out elsewhere<sup>9</sup> the variable time between infection and such observations causes the signal to be both delayed and blurred. By combining our estimates of the incubation period with data from hospital admissions in the CHES data set we were able to make estimates of the distribution of the delay from infection to observation for the UK. There are several caveats to this part of our analysis that must be kept in mind. The CHES data set relies on a retrospective report of onset of symptoms, and this field is not recorded for all patients. We select only those patients who have reported an onset date and it may be the case that



these patients represent a subgroup of patients whose symptom onset is significantly different from the average patient. The data collection around these dates as noted above to show patterns suggestive of rounding or approximation and this could also introduce some bias. The delay distributions are also unlikely to remain fixed during the outbreak, as we would hope to see the time from infection to case identification shorten during the epidemic, and the time from infection to death lengthen as treatment improves, and as the cohort of susceptible individuals changes. Our confidence in these distributions is therefore somewhat low, although we note acceptable agreement between our estimates produced using a mixture of international and UK data, and previously published estimates from different countries, most notably China<sup>52–55</sup> which cite an onset to an admission of 2.7–5.9 days<sup>26,44,52</sup> and onset to death delay of 16.1–17.8 days.<sup>26,52</sup> Given our estimate of the incubation period is log-normally distributed, it is unsurprising that the combination of incubation period and delay from onset to observation is also best described by log-normal distributions (Figure 7 panel B), and with these, we can apply non-parametric back-propagation to infer a time series of putative infections from the delayed observations we have available.

In Figure 8, we bring all the different parts of the analysis together and compare the two approaches of formal estimation of  $R_t$  using the generation interval and back-propagation of case, admission or death counts to putative infection, versus a pragmatic estimation using the truncated re-sampled serial interval, and direct use of observation numbers, combined with the simple-but-incorrect adjustment to the resulting  $R_t$  time series, shifting the date backward by the mean of the delay distribution.<sup>9</sup>

In comparing formal versus pragmatic methods, given the number of moving parts in this comparison it is encouraging to see the level of agreement between the  $R_t$  estimates from the two methods (Figure 8 panel B) in the early phase of the epidemic, and also to see that there is some similar structure of the  $R_t$  time series in the later parts. This is particularly the case for estimates based on cases and admissions, but less so for deaths where the de-convolution time series shows additional features that are not obvious from the data. We have no gold standard in comparing the  $R_t$  estimates for England, so we are limited in what we can conclude, but we do observe that  $R_t$  estimates based on our attempt at de-convolution have more variability than ones from un-adjusted observations, and de-convolved estimates have additional features not present in the un-adjusted estimates. The de-convolved estimates of infection are also noted to run to the end of the time series, which is somewhat surprising as estimates of infection rates at the end of the time series should depend on data that has not yet been observed. This is a feature of the back-propagation algorithm which needs to be used with caution as the resulting estimates of  $R_t$  based on de-convolution for the latter part of the time series appear inconsistent with each other.

## Limitations

We did not fully quantify uncertainty in our analysis and estimation of  $R_t$ . The informal approach to estimating  $R_t$  has uncertainty arising from the serial interval distribution, and stochastic noise in the value of the observation in question. The formal approach involves uncertainty in the initial estimation of the serial interval, uncertainty in the incubation period, resulting in uncertainty in the generation interval, the back-propagation itself is a source of uncertainty and involves uncertain time-delay distributions which are in turn based on uncertain incubation period, and finally the stochastic noise in the observation under consideration. Accurately tracking the uncertainty of all these components into a final estimate remains a challenge. One of the key reasons we choose to make simplifying assumptions in our use of the pragmatic approach in estimating  $R_t$  is that it makes comparatively transparent assumptions that can be backed up by experimental data, and for which the resulting biases are best understood.

Our estimates are based on the best available information at the present stage of the epidemic. However, the serial interval is not a fixed quantity and may be affected by behavioral changes such as case isolation, or social distancing. The assumption is constant is questionable although we have very little hard evidence about how it may vary over time. Similarly time distributions from infection to case identification, admission, and death are expected to be highly variable over the course of an outbreak. This has implications for their use in de-convolution, as changing time distributions will have a significant effect on the shape of inferred infection incidence curves, and the complexity of the de-convolution.

There are implicit selection biases in all the data sources we use. A large proportion of SARS-CoV-2 cases are asymptomatic.<sup>12,56,57</sup> It is highly likely that these people participate in transmission chains, and they may do so with a very different infectivity profile to those that are symptomatic, however, we have no information about these people in the data sets, and hence all the estimates presented here could be quite different when asymptomatic cases are taken into consideration.

The data used to assess delays to death rely on hospital-based data as this was the best UK-specific data we had. However, this means that delays to deaths are only assessed for the subset of patients that die in hospital, and we are forced to generalize this to the whole population.

Our approach in estimating key parameters has been to combine different data sets, which come from different international sources, and which have potentially different biases. In combining data sets we assume that time delays are independent of one another, and can be combined randomly, as we have no other evidence to the contrary. This assumption is questionable, as physiologically we can imagine that patients with a long incubation period, for example, may well have a longer period from symptom onset to admission. This could have unpredictable effects on our estimates of time delays but the most likely is that our estimated variance is too small as a result.

## Conclusions

We argue that our estimates for the statistical distributions of these key parameters or serial interval, incubation period, generation interval, for the UK, along with their uncertainty represent the best available estimates for the UK, at the time of writing (September 2020), given the current state of knowledge.

As there is a wide range of candidate values for these quantities, we have assessed the bias that the variation in choice of parameters introduces to  $R_t$  estimation, when using the forward equation method. Whilst these introduce significant variation in the worst-case scenario, we find that even large differences in infectivity profile can have only small impacts on estimates of  $R_t$  when  $R_t$  is close to 1. Larger values for the mean of the infectivity profile appear to result in  $R_t$  estimates that are further away from 1. This is a relatively reassuring finding in that the answer to the key question “is the epidemic under control?” is insensitive to the mean of the infectivity profile.

Using more formal methods for estimating  $R_t$  by back-propagation inference of infection rate and estimates of the generation interval, produces  $R_t$  estimates that are more variable when compared to the more pragmatic direct use of case counts as a proxy for infection. Both methods agree on when the epidemic crossed the  $R_t$  threshold of 1. However there is considerable uncertainty in the quantities needed to perform the back-propagation, and we are not able to ensure that all this uncertainty could be faithfully quantified in our resulting  $R_t$  estimates. We did not set out to assess whether one method is better than another, and this would be a natural extension to this work, however, we note that new methods for combining back-propagation with an estimation of  $R_t$  are under active development<sup>58,59</sup> and may very well address such further questions.

## Authors contributions

All authors discussed the concept of the article and RC wrote the initial draft. KTA, EB-P, and LD commented and made revisions. All authors read and approved the final manuscript. RC is the guarantor. The views presented here are those of the authors and should not be attributed to TSFT or the GDE.


## Declaration of conflicting interests


Support for RC and KTA’s research is provided by the EPSRC via grant EP/N014391/1, RC is also funded by TSFT as part of the NHS Global Digital Exemplar program (GDE); no financial relationships with any organizations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

## Funding

RC and KTA gratefully acknowledge the financial support of the EPSRC via grant EP/N014391/1 and NHS England, Global Digital Exemplar program. LD and KTA gratefully acknowledge the financial support of The Alan Turing Institute under the EPSRC grant EP/N510129/1. LD and EBP are supported by Medical Research Council (MRC) (MC/PC/19067). EBP was partly supported by the NIHR Health Protection Research Unit in Behavioural Science and Evaluation at the University of Bristol, in partnership with Public Health England (PHE).

## ORCID iDs

Robert Challen  <https://orcid.org/0000-0002-5504-7768>

Krasimira Tsaneva-Atanasova  <https://orcid.org/0000-0002-6294-7051>

## Supplemental Material

Supplemental material for this article is available online.

## References

1. Svensson Å. A note on generation times in epidemic models. *Math Biosci* 2007; **208**: 300–311.
2. Hao X, Cheng S, Wu D, et al. Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature* 2020; **584**: 420–424.

3. Becker NG, Watson LF and Carlin JB. A method of non-parametric back-projection and its application to AIDs data. *Stat Med* 1991; **10**: 1527–1542.
4. Wallinga J and Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am J Epidemiol* 2004; **160**: 509–516.
5. Cori A, Ferguson NM, Fraser C, et al. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am J Epidemiol* 2013; **178**: 1505–1512.
6. Cori A, Kamvar Z, Stockwin J, et al. EpiEstim v2.2-3: A tool to estimate time varying instantaneous reproduction number during epidemics. *GitHub repository*. <https://github.com/mrc-ide/EpiEstim>
7. Thompson RN, Stockwin JE, van Gaalen RD, et al. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics* 2019; **29**: 100356.
8. Britton T and Scalia Tomba G. Estimation in emerging epidemics: biases and remedies. *J R Soc Interface* 2019; **16**:20180670.
9. Gostic KM, McGough L, Baskerville E, et al. Practical considerations for measuring the effective reproductive number, Rt. *medRxiv* 2020; 2020.06.18.20134858.
10. Xu X, Liu X, Wang L, et al. Household transmissions of SARS-CoV-2 in the time of unprecedented travel lockdown in China.
11. Davies NG, Klepac P, Liu Y, et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat Med* 2020; **26**: 1205–1211.
12. Mizumoto K, Kagaya K, Zarebski A, et al. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess Cruise Ship, Yokohama, Japan, 2020. *Eurosurveillance* [Internet]. 2020 Mar 12 [cited 2021 Dec 2]; **25**(10): 2000180. <https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.10.2000180>
13. MIDAS Online Portal for COVID-19 Modeling Research, <https://midasnetwork.us/covid-19/> (accessed 20 August 2020).
14. Zhang P, Wang T and Xie SX. Meta-analysis of several epidemic characteristics of COVID-19. *J Data Sci.* 2020 Jul; **18**(3): 536–549.
15. Beath JK. Metaplust: an R package for the analysis of robust meta-analysis and meta-regression. *R J* 2016; **8**.
16. Jackson D and White IR. When should meta-analysis avoid making hidden normality assumptions? *Biom J* 2018; **60**: 1040–1058.
17. Tierney JF, Stewart LA, Ghersi D, et al. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007; **8**: 16.
18. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods* 2016; **7**: 55–79.
19. Adams DC, Gurevitch J and Rosenberg MS. Resampling tests for meta-analysis of ecological data. *Ecology* 1997; **78**: 1277–1283.
20. Chuang C-S and Lai TL. Hybrid resampling methods for confidence intervals. *Stat Sin* 2000; **10**: 1–33.
21. R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2017.
22. Delignette-Muller ML and Dutang C. Fitdistrplus: an R package for fitting distributions. *J Stat Softw* 2015; **64**: 1–34.
23. The first few hundred (ff100) enhanced case and contact protocol v12, [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/360190/2012\\_13\\_FF100\\_Protocol\\_H7N9\\_ver\\_12.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/360190/2012_13_FF100_Protocol_H7N9_ver_12.pdf) (accessed 8 July 2020).
24. Boddington NL, Charlett A, Elgohari S, et al. COVID-19 in Great Britain: Epidemiological and clinical characteristics of the first few hundred (FF100) cases: A descriptive case series and case control analysis. *medRxiv* 2020; 2020.05.18.20086157.
25. Lauer SA, Grantz KH, Bi Q, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann Intern Med* [Internet]. 2020 May 5 [cited 2021 Dec 2]; **172**(9): 577–582. <https://www.acpjournals.org/doi/10.7326/M20-0504>
26. Sanche S, Lin YT, Xu C, et al. High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerging Infectious Diseases* [Internet]. 2020 [cited 2021 Dec 2]; **26**(7): 1470–1477. [https://wwwnc.cdc.gov/eid/article/26/7/20-0282\\_article](https://wwwnc.cdc.gov/eid/article/26/7/20-0282_article)
27. Group OC-IDW. Detailed epidemiological data from the COVID-19 outbreak.
28. Xu B and Kraemer MUG. Open COVID-19 Data Curation Group. Open access epidemiological data from the COVID-19 outbreak. *Lancet Infect Dis* 2020; **20**: 534.
29. Zhao Q, Ju N, Bacallado S, et al. BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic, <http://arxiv.org/abs/2004.07743> (2020, accessed 29 June 2021).
30. Coronavirus (COVID-19) in the UK: Developers guide, <https://coronavirus.data.gov.uk/developers-guide> (accessed 2 September 2020).
31. Coronavirus (COVID-19) in the UK: About the data, <https://coronavirus.data.gov.uk/about-data> (accessed 2 September 2020).
32. Coronavirus (COVID-19): Using data to track the virus - public health matters, <https://publichealthmatters.blog.gov.uk/2020/04/23/coronavirus-covid-19-using-data-to-track-the-virus/> (accessed 8 July 2020).
33. Meyer S, Held L and Höhle M. Spatio-temporal analysis of epidemic phenomena using the R package surveillance. *J Stat Softw* 2017; **77**: 1–55.
34. Yip PSF, Lam KF, Xu Y, et al. Reconstruction of the infection curve for SARS epidemic in Beijing, China using a back-projection method. *Commun Stat - Simul Comput* 2008; **37**: 425–433.
35. Ali ST, Wang L, Lau EHY, et al. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. *Science* 2020; **369**: 1106–1109.
36. Bi Q, Wu Y, Mei S, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *Lancet Infect Dis* 2020; **20**: 911–919.

37. Cereda D, Tirani M, Rovida F, et al. The early phase of the COVID-19 outbreak in Lombardy, Italy, <http://arxiv.org/abs/2003.09320> (2020, accessed 20 August 2020).
38. Du Z, Xu X, Wu Y, et al. Serial interval of COVID-19 among publicly reported confirmed cases. *Emerg Infect Dis* 2020; **26**: 1341–1343.
39. Kwok KO, Wong VWY, Wei WI, et al. Epidemiological characteristics of the first 53 laboratory-confirmed cases of COVID-19 epidemic in Hong Kong, 13 February 2020. *Eurosurveillance* 2020; **25**: 2000155.
40. Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N Engl J Med* [Internet]. 2020 Mar 26 [cited 2021 Dec 2]; **382**(13): 1199–1207. <https://doi.org/10.1056/NEJMoa2001316>
41. Li M, Liu K, Song Y, et al. Serial interval and generation interval for imported and local infectors, respectively, estimated using reported contact-tracing data of COVID-19 in China. *Front Public Health* Epub ahead of print 2021; **8**. DOI: 10.3389/fpubh.2020.577431
42. Nishiura H, Linton NM and Akhmetzhanov AR. Serial interval of novel coronavirus (COVID-19) infections. *Int J Infect Dis* [Internet]. 2020 Apr 1 [cited 2021 Dec 2]; **93**: 284–286. [https://www.ijidonline.com/article/S1201-9712\(20\)30119-3/fulltext](https://www.ijidonline.com/article/S1201-9712(20)30119-3/fulltext)
43. Son H, Lee H, Lee M, et al. Epidemiological characteristics of and containment measures for COVID-19 in Busan, Korea. *Epidemiol Health* [Internet]. 2020 Jun 1 [cited 2021 Dec 2]; **42**: e2020035-0. <http://www.e-epih.org/journal/view.php?number=1102>
44. Tindale LC, Stockdale JE, Coombe M, et al. Evidence for transmission of COVID-19 prior to symptom onset. *eLife* 2020; **9**: e57149.
45. Xia W, Liao J, Li C, et al. Transmission of corona virus disease 2019 during the incubation period may lead to a quarantine loophole. *medRxiv* 2020; 2020.03.06.20031955.
46. You C, Deng Y, Hu W, et al. Estimation of the time-varying reproduction number of COVID-19 outbreak in China. *Int J Hyg Environ Health* 2020; **228**: 113555.
47. Zhang J, Litvinova M, Wang W, et al. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. *Lancet Infect Dis* 2020; **20**: 793–802.
48. Zhao S, Lin Q, Ran J, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: a data-driven analysis in the early phase of the outbreak. *Int J Infect Dis* 2020; **92**: 214–217.
49. Ganyani T, Kremer C, Chen D, et al. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance* 2020; **25**: 2000257.
50. Backer JA, Klinkenberg D and Wallinga J. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Eurosurveillance* 2020; **25**: 2000062.
51. Du ZC, Gu J, Li JH, et al. Estimating the distribution of COVID-19 incubation period by interval-censored data estimation method. *Zhonghua Liu Xing Bing Xue Za Zhi* 2020; **41**: 1000–1003.
52. Linton NM, Kobayashi T, Yang Y, et al. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data.
53. Jung S-m, Akhmetzhanov AR, Hayashi K, et al. Real time estimation of the risk of death from novel coronavirus (2019-nCoV) infection: Inference using exported cases. *medRxiv* 2020; 2020.01.29.20019547.
54. Sanche S, Lin YT, Xu C, et al. The novel coronavirus, 2019-nCoV, is highly contagious and more infectious than initially estimated. *medRxiv* 2020; 2020.02.07.20021154.
55. Verity R, Okell LC, Dorigatti I, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis* 2020; **20**: 669–677.
56. Bai Y, Yao L, Wei T, et al. Presumed asymptomatic carrier transmission of COVID-19. *JAMA* [Internet]. 2020 Apr 14 [cited 2021 Dec 2]; **323**(14): 1406–1407. <https://doi.org/10.1001/jama.2020.2565>
57. Hu Z, Song C, Xu C, et al. Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China. *Sci China Life Sci* [Internet]. 2020 May 1 [cited 2021 Dec 2]; **63**(5): 706–711. <https://doi.org/10.1007/s11427-020-1661-4>
58. Abbott S, Hellewell J, Sherratt K, et al. *EpiNow2: Estimate real-time case counts and time-varying epidemiological parameters*. 2020. Epub ahead of print 2020. DOI: 10.5281/zenodo.3957489.
59. Abbott S, Hellewell J, Thompson RN, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Res* 2020; **5**: 11211.

# Meta-analysis of the SARS-CoV-2 serial interval and the impact of parameter uncertainty on the COVID-19 reproduction number - Supplementary 1

Robert Challen<sup>1,2,3</sup>; Ellen Brooks-Pollock<sup>3,4</sup>; Krasimira Tsaneva-Atanasova<sup>1,5,6</sup>; Leon Danon<sup>4,5,6,7</sup>

- 1) EPSRC Centre for Predictive Modelling in Healthcare, University of Exeter, Exeter, Devon, UK.
- 2) Somerset NHS Foundation Trust, Taunton, Somerset, UK.
- 3) Joint Universities Pandemic and Epidemiological Research (JUNIPER) consortium.
- 4) Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK.
- 5) The Alan Turing Institute, British Library, 96 Euston Rd, London NW1 2DB, UK.
- 6) Data Science Institute, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK.
- 7) Department of Engineering Mathematics, University of Bristol, UK.

## Supporting tables and figures

### Supplemental table 1: An algorithm for resampling representative serial intervals based on parameterised distributions and raw data from published studies

The detail of the re-sampling algorithm that is used to combine estimates from a range of different studies is given below. The rationale for the definition of the sampling distributions is covered in more detail in Supplementary Materials 2. The algorithm combines raw data from original studies, with synthetic data from resampling the uncertain parameterised distributions provided in the literature. The combination of these is then used to compare different maximum likelihood parametric distributions fitting all the data, and later as a target for optimising the generation interval distribution when combined with a dataset of incubation period estimates.

*INPUTS:*

***parameterised serial interval distributions:***

list of parameterised serial interval distribution results from literature sources comprising:

- source***, the study the estimates come from
- distribution***, type of estimated serial interval distribution
- central estimate of mean***,
- lower CI of mean***, if available
- upper CI of mean***, if available
- central estimate of sd***,
- lower CI of sd***, if available
- upper CI of sd***, if available
- sample size***

***raw serial interval observations:***

raw data of empirical serial interval distributions from literature sources comprising:

*source*,  
set(*observed serial interval*)

OUTPUT:

**output serial interval observations:**

a set of 100 bootstrap replicates containing real and synthetic serial intervals observations comprising:

*iteration*,  
set(*simulated serial interval*)

ALGORITHM:

for each *source* in *raw serial interval observations*

define *sample size* as the count of *observed serial interval*

for *iteration* in 1 to 100

define set(*samples*) as random re-samples with replacement from *observed serial interval*

add (*iteration*, set(*samples*)) to *output serial interval observations*

for each *source* in *parameterised serial interval distributions*

— we are recreating a set of sampling distributions  $\mathcal{X}(\nu, \phi)$

— where  $\mathcal{X}$  as *distribution* from study *source* (e.g. gamma, log-normal, normal)

—  $\nu$  is a distribution of means compatible with study *source*

— and  $\phi$  is a distribution of std devns compatible with study *source*

—  $\nu$  is assumed to be a truncated normal distribution

— and is the sample distribution of the mean

define  $\mu_{mean}$  as *central estimate of mean*

define  $\sigma_{mean}$  as (*upper CI of mean-lower CI of mean*)/3.96 or zero if CI not given

define  $\nu \sim \mathcal{N}(\mu_{mean}, \sigma_{mean})$  truncated between *lower CI of mean* and *upper CI of mean*

—  $\phi$  is a Nakagami distributed quantity variously parameterised

— depending on whether we have known confidence limits

— this is the sample distribution of the standard deviation

define  $\mu_{sd}$  as *central estimate of sd*

define  $N$  as *sample size*

if *lower CI of sd* and *upper CI of sd* are defined

— if we know confidence limits and mean of SD we assume it is Nakagami distributed

— with  $\Omega$  parameter as  $\sigma^2$

fit  $\phi \sim \text{Nakagami}(m_{sd}, \sigma^2)$  to *lower CI of sd* and *upper CI of sd*

else

— if we know only know mean of SD (and sample size) we assume  $\phi$  has a different parameterisation

define  $\phi \sim \text{Nakagami}(\frac{N-1}{2}, \sigma^2)$

for *iteration* in 1 to 100

define  $\mu_{sample}$  as random sample from  $\nu$

define  $\sigma_{sample}$  as random sample from  $\phi$

define *sampling distribution*  $\sim \mathcal{X}(\mu_{sample}, \sigma_{sample})$ ; converting parameters as necessary

define set(*samples*) as  $N$  random samples from *sampling distribution*

add (*iteration*, set(*samples*)) to *output serial interval observations*

return *output serial interval observations*

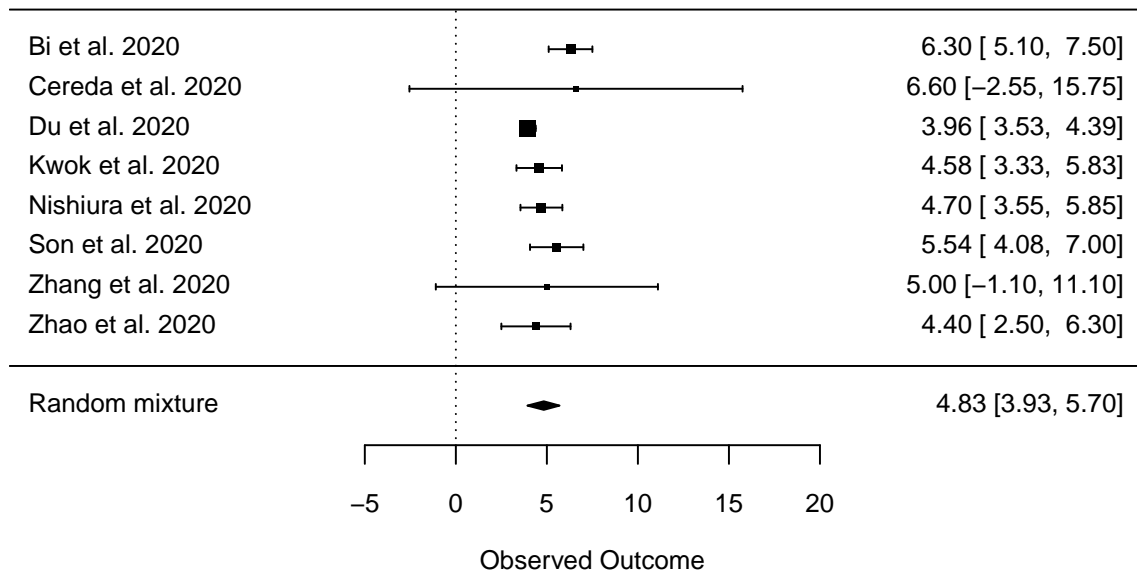
## Supplemental table 2: The NHS trusts in the CHESSE dataset who report all admissions

The CHESSE dataset includes a wide range of contributing hospitals with very varied data quality of submission. Only some hospitals continue to update their data, to include updates to patients as they are discharged or die, and only some submit data on all inpatients not just those who go to ITU. While investigating the delays to various events in the hospital stay we focussed on a subset of contributing hospitals that last updated records within 21 days of the date of analysis, and for which had a maximum of 20% of their cases had unknown or incomplete dates. This included the following contributing trusts and within these we excluded patients that had their COVID-19 diagnosis made more than 10 days after admission.

Trust	Patients
BARKING, HAVERING AND REDBRIDGE UNIVERSITY HOSPITALS NHS TRUST	1056
BLACKPOOL TEACHING HOSPITALS NHS FOUNDATION TRUST	558
CHESTERFIELD ROYAL HOSPITAL NHS FOUNDATION TRUST	440
GREAT WESTERN HOSPITALS NHS FOUNDATION TRUST	398
LIVERPOOL HEART AND CHEST HOSPITAL NHS FOUNDATION TRUST	523
NORTH CUMBRIA INTEGRATED CARE NHS FOUNDATION TRUST	360
NORTH WEST ANGLIA NHS FOUNDATION TRUST	758
ROYAL BERKSHIRE NHS FOUNDATION TRUST	629
SOUTH TEES HOSPITALS NHS FOUNDATION TRUST	85
TAMESIDE HOSPITAL NHS FOUNDATION TRUST	686
THE NEWCASTLE UPON TYNE HOSPITALS NHS FOUNDATION TRUST	514
THE QUEEN ELIZABETH HOSPITAL, KING'S LYNN, NHS FOUNDATION TRUST	401
UNIVERSITY HOSPITAL SOUTHAMPTON NHS FOUNDATION TRUST	571
WALSALL HEALTHCARE NHS TRUST	700
WEST HERTFORDSHIRE HOSPITALS NHS TRUST	688
WORCESTERSHIRE ACUTE HOSPITALS NHS TRUST	787
YEOVIL DISTRICT HOSPITAL NHS FOUNDATION TRUST	80
YORK TEACHING HOSPITAL NHS FOUNDATION TRUST	668

**Supplemental figure 1: Forest plot for serial interval studies for the mean of the serial interval, using the normal mixture random effect model, and from studies identified in the literature which give confidence intervals**

Assessing the serial interval distribution using a random effects meta-analysis of the studies that report a confidence limit for the mean has the benefit of using a standardised methodology but does not describe the distributional nature of the serial interval.





**Supplemental table 3: Parameterised serial interval distributions from resampling the literature. Gamma and weibull estimates are from data truncated at zero. AIC estimates are not comparable to those for Normal distribution which is fitting all data, including negative serial intervals, and hence has a lower mean.**

Maximum likelihood parameterised distributions fitted to the resampled serial interval data demonstrate variable quality of fits, as the serial interval data has a substantial component which is negative and therefore fitting continuous distributions with support in the positive real numbers requires truncating the data at zero as seen in Figure 2 in the main paper.

Distribution	AIC	N	Parameter / Moment	Mean $\pm$ SD (95% CI)
gamma	5081.1	92380	scale	$3.86 \pm 0.44$ (3.31; 4.83)
			shape	$1.50 \pm 0.20$ (1.09; 1.75)
			mean	$5.72 \pm 0.38$ (5.08; 6.51)
			sd	$4.69 \pm 0.27$ (4.28; 5.30)
normal	6106.4	101178	mean	$4.87 \pm 0.37$ (4.24; 5.52)
			sd	$4.79 \pm 0.26$ (4.45; 5.42)
			mean	$4.87 \pm 0.37$ (4.24; 5.52)
			sd	$4.79 \pm 0.26$ (4.45; 5.42)
weibull	5056.9	92380	scale	$6.19 \pm 0.48$ (5.32; 7.09)
			shape	$1.33 \pm 0.10$ (1.13; 1.46)
			mean	$5.70 \pm 0.38$ (5.06; 6.49)
			sd	$4.34 \pm 0.27$ (3.96; 4.91)

**Supplemental table 4: Parameterised serial interval distributions from FF100. Gamma and weibull estimates are from data truncated at zero. AIC estimates are not comparable to those for Normal distribution which is fitting all data, including negative serial intervals, and hence has a lower mean.**

The direct estimation of the serial interval from the FF100 data does not produce an estimate that is comparable with other international data sets.

Distribution	AIC	N	Parameter / Moment	Mean $\pm$ SD (95% CI)
gamma	182	43	shape	2.77 $\pm$ 0.49 (2.01; 3.78)
			scale	1.30 $\pm$ 0.24 (0.91; 1.77)
			mean	3.50 $\pm$ 0.33 (2.84; 4.14)
			sd	2.13 $\pm$ 0.25 (1.59; 2.61)
normal	251	50	mean	2.81 $\pm$ 0.41 (2.03; 3.54)
			sd	2.74 $\pm$ 0.34 (2.16; 3.41)
			mean	2.81 $\pm$ 0.41 (2.03; 3.54)
			sd	2.74 $\pm$ 0.34 (2.16; 3.41)
weibull	183	43	shape	1.75 $\pm$ 0.18 (1.43; 2.17)
			scale	3.94 $\pm$ 0.37 (3.21; 4.67)
			mean	3.52 $\pm$ 0.32 (2.85; 4.14)
			sd	2.08 $\pm$ 0.23 (1.62; 2.51)

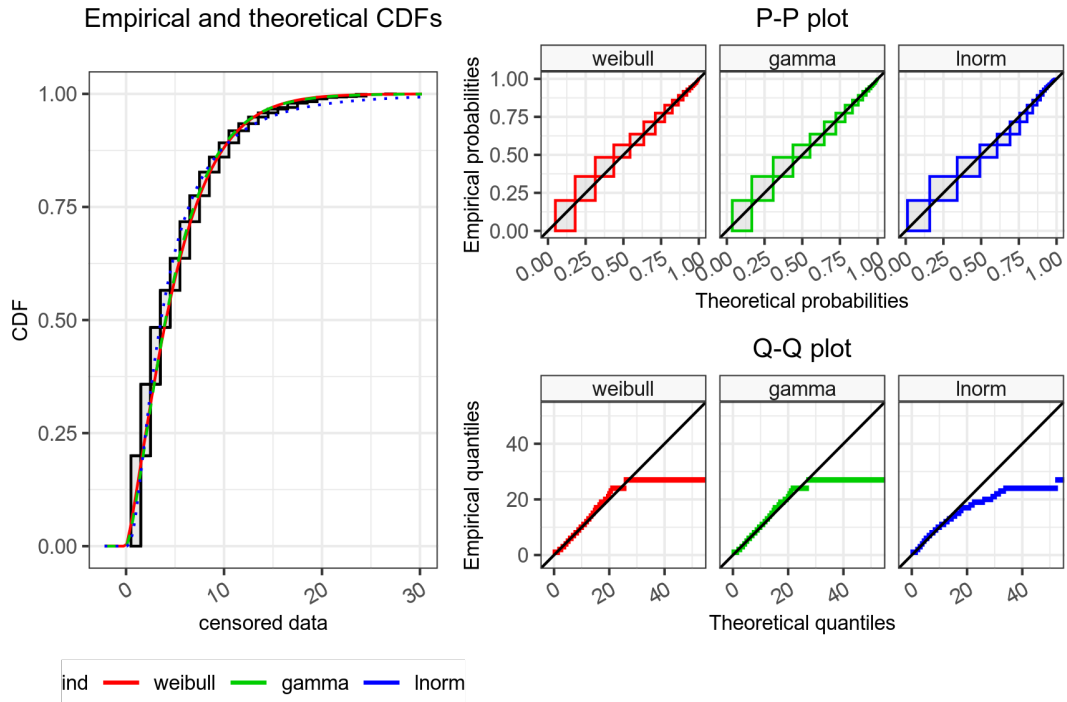
**Supplemental table 5: Distribution details for estimated incubation period distributions reconstructed from Open COVID-19 Data Working Group and from FF100 data**

Detail of the distribution parameterisation of the incubation period with the 3 investigated distributions on the 2 data sets, shows that in the larger Open COVID-19 Data working group data set the best fit is obtained with a log-normal distribution.

N	Source	AIC	Distribution	Parameter / Moment	Mean $\pm$ SD (95% CI)
33	FF100	62.1	gamma	shape	1.71 $\pm$ 4.14 (0.35; 4.81)
				scale	1.68 $\pm$ 0.85 (0.45; 3.71)
				mean	1.82 $\pm$ 0.35 (1.10; 2.50)
				sd	1.67 $\pm$ 0.42 (0.86; 2.42)
			weibull	shape	1.23 $\pm$ 0.39 (1.00; 1.99)
				scale	1.94 $\pm$ 0.37 (1.28; 2.62)
				mean	1.85 $\pm$ 0.32 (1.28; 2.51)
				sd	1.57 $\pm$ 0.33 (0.93; 2.20)
		63.5	log-normal	meanlog	0.25 $\pm$ 0.34 (-0.59; 0.68)
				sdlog	0.85 $\pm$ 0.24 (0.42; 1.36)
				mean	1.94 $\pm$ 0.34 (1.26; 2.60)
				sd	2.11 $\pm$ 0.93 (0.88; 4.43)
1062	Open COVID-19 Data Working Group	5157.2	log-normal	meanlog	1.27 $\pm$ 0.03 (1.21; 1.32)
				sdlog	0.87 $\pm$ 0.01 (0.84; 0.89)
				mean	5.19 $\pm$ 0.15 (4.90; 5.46)
				sd	5.49 $\pm$ 0.24 (5.05; 5.92)
		5191.7	gamma	shape	1.55 $\pm$ 0.05 (1.46; 1.65)
				scale	3.26 $\pm$ 0.15 (2.96; 3.52)
				mean	5.06 $\pm$ 0.14 (4.80; 5.29)
				sd	4.06 $\pm$ 0.14 (3.78; 4.30)
		5216.6	weibull	shape	1.25 $\pm$ 0.03 (1.20; 1.30)
				scale	5.45 $\pm$ 0.15 (5.17; 5.73)
				mean	5.08 $\pm$ 0.14 (4.82; 5.32)
				sd	4.09 $\pm$ 0.14 (3.80; 4.33)

## Supplemental figure 2: Graphical goodness of fit for parameterised incubation period distributions fitted to the Be Outbreak Prepared dataset

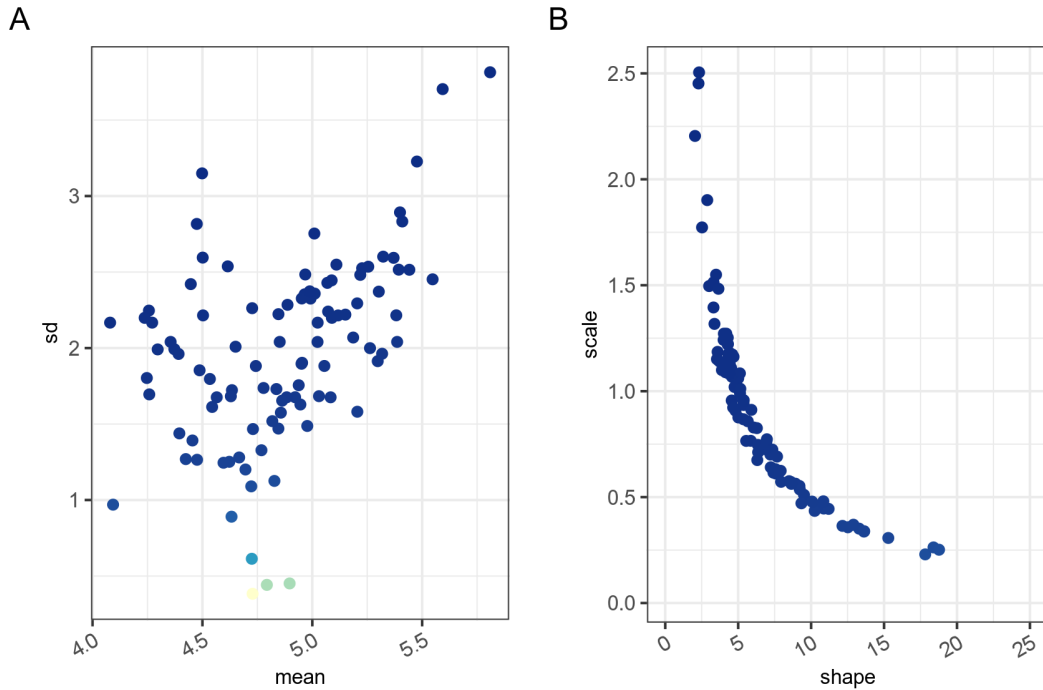
Further investigation of the fit of the log-normal incubation period demonstrates good alignment of the model fits in all models to the early part of the distribution but with some loss of fit towards the tails, due to a number of censored entries starting after day 28. These data are generally irrelevant given the incubation period distribution is largely only subsequently used in a discretised form up to a limited number of days. This poorly fitting tail could influence the upper confidence limit of the incubation period which may be slightly longer than anticipated.



### Supplemental figure 3: The distribution of the parameters of the fitted generation interval estimates

The range of generation distributions resulting from optimising the different combinations of bootstrapped serial interval and incubation periods shows a spread of possible generation interval distributions, but mean and standard deviations are not wholly independent as seen in panel A, with a tendency to increasing SD with increasing mean. The shape and scale parameterisation shows a close reciprocal relationship between them due in some part to the constraints placed on the mean of the distributions.

The finding that the generation time mean and standard deviation distributions are not independent raises a question about whether sampling from uncertainly specified distributions can ever produce a realistic sample of generation time distributions for subsequent use in analysis. For example in estimating  $R_t$  using the EpiEstim package uncertainly parameterised distributions are sampled using independent truncated normal distributions on the mean and standard deviation and their associated confidence intervals. This will tend to broaden the range of considered distributions beyond that seen here, and hence increase uncertainty. It is possible that re-sampling using independent truncated normal distributed means and standard deviation could also bias estimates of  $R_t$  but it is not obvious in which direction or by how much. This is an area for future investigation.



**Supplemental table 6: Time delay distributions estimated from CHES data set, for both transitions from disease onset to case, admission or death, and presumed infection and case, admission or death**

The combination of incubation period and delay from symptom onset to positive test, admission or death are graphically displayed in Figure 7. Here the details of these parameterisations are provided. In all cases the best fit is obtained with a log-normal distribution. The infection to test distribution is used in the de-convolution analysis presented in the main paper.

Transition	AIC	Distribution	Parameter	Mean $\pm$ SD (95% CI)
infection to admission	5262.3	log-normal	meanlog	1.61 $\pm$ 0.04 (1.51; 1.67)
			sdlog	0.93 $\pm$ 0.02 (0.88; 0.97)
			mean	7.70 $\pm$ 0.26 (7.22; 8.08)
			sd	9.05 $\pm$ 0.52 (7.93; 10.00)
	5361.7	gamma	scale	5.80 $\pm$ 0.40 (5.05; 6.57)
			shape	1.32 $\pm$ 0.07 (1.19; 1.44)
			mean	7.64 $\pm$ 0.26 (7.13; 8.07)
			sd	6.65 $\pm$ 0.32 (6.04; 7.36)
5388.0	weibull	scale	7.93 $\pm$ 0.25 (7.34; 8.34)	
		shape	1.09 $\pm$ 0.04 (1.02; 1.15)	
		mean	7.68 $\pm$ 0.26 (7.16; 8.10)	
		sd	7.05 $\pm$ 0.37 (6.30; 7.74)	
infection to death	6282.2	log-normal	meanlog	2.58 $\pm$ 0.02 (2.55; 2.61)
			sdlog	0.62 $\pm$ 0.01 (0.60; 0.65)
			mean	16.04 $\pm$ 0.28 (15.57; 16.56)
			sd	11.08 $\pm$ 0.38 (10.31; 11.85)
	6319.9	gamma	scale	5.77 $\pm$ 0.28 (5.25; 6.34)
			shape	2.77 $\pm$ 0.11 (2.58; 2.99)
			mean	15.98 $\pm$ 0.29 (15.49; 16.55)
			sd	9.60 $\pm$ 0.30 (9.05; 10.24)
6406.6	weibull	scale	17.97 $\pm$ 0.32 (17.43; 18.56)	
		shape	1.62 $\pm$ 0.04 (1.55; 1.71)	
		mean	16.09 $\pm$ 0.30 (15.59; 16.69)	
		sd	10.19 $\pm$ 0.37 (9.43; 10.97)	
infection to test	4936.5	gamma	scale	3.68 $\pm$ 0.18 (3.35; 4.07)
			shape	1.75 $\pm$ 0.08 (1.58; 1.90)
			mean	6.43 $\pm$ 0.15 (6.11; 6.69)
			sd	4.86 $\pm$ 0.15 (4.56; 5.18)
	4953.3	weibull	scale	7.06 $\pm$ 0.17 (6.68; 7.36)
			shape	1.36 $\pm$ 0.04 (1.28; 1.43)
			mean	6.46 $\pm$ 0.15 (6.14; 6.72)
			sd	4.80 $\pm$ 0.15 (4.51; 5.12)
4969.0	log-normal	meanlog	1.55 $\pm$ 0.03 (1.48; 1.60)	
		sdlog	0.85 $\pm$ 0.03 (0.80; 0.89)	
		mean	6.73 $\pm$ 0.17 (6.41; 7.03)	
		sd	6.89 $\pm$ 0.35 (6.30; 7.57)	

# Meta-analysis of the SARS-CoV-2 serial interval and the impact of parameter uncertainty on the COVID-19 reproduction number - Supplementary 2

Robert Challen<sup>1,2,3</sup>; Ellen Brooks-Pollock<sup>3,4</sup>; Krasimira Tsaneva-Atanasova<sup>1,5,6</sup>; Leon Danon<sup>4,5,6,7</sup>

- 1) EPSRC Centre for Predictive Modelling in Healthcare, University of Exeter, Exeter, Devon, UK.
- 2) Somerset NHS Foundation Trust, Taunton, Somerset, UK.
- 3) Joint Universities Pandemic and Epidemiological Research (JUNIPER) consortium.
- 4) Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK.
- 5) The Alan Turing Institute, British Library, 96 Euston Rd, London NW1 2DB, UK.
- 6) Data Science Institute, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK.
- 7) Department of Engineering Mathematics, University of Bristol, UK.

## Sampling from uncertain distributions

### Introduction

This is a brief note to explain the approach taken to producing samples from results given as parameterised distributions cited in the literature. The purpose for this is in combining multiple studies which report a quantity (e.g. the serial interval of an infection) as modelled by a parameterised statistical distribution (e.g. a Gamma distribution), with a central estimate of the mean, a central estimate of the standard deviation, and typically confidence limits on both of those quantities (or credible intervals if the distribution was estimated using a Bayesian framework). Combining such results into a single estimate through meta-analysis does not fit within the standard approaches, as these generally assume a normally distributed single dimensional effect, and whilst this is probably valid for the means of the parameterised distributions to be treated in this way, it is not necessarily a valid assumption for the parameter defining the spread (i.e. the standard deviation).

The challenge in combining such distributions is essentially that of estimating the mixture of all possible distributions that are compatible with the results published in all the studies. The resulting mixture distribution may then be further analysed in a range of ways.

For example an additional important capability for us is the ability to combine studies that present results as a parameterised distribution, with other studies where only empirical estimates are made on the quantities of interest, and the raw data is available. In this case generating representative samples from the original report is important, so that parameterised results can be combined with empirical results. This approach is akin to parametric bootstrapping, but where the bootstrapping is performed not on data but on the uncertain estimate of the parameterised distribution.

The key step of this re-sampling is the conversion of an uncertain parameterised distribution into a representative set of precisely specified distributions that can in turn be sampled. The set of studies described in Table 1 is a typical example of the kind of information analysed with this approach.

The specific problem of generating a set of representative set of precisely specified parameterised distributions from an uncertainly specified result is somewhat similar to that of sampling parameter values within a Bayesian framework where the mean and standard deviation of a parameter distribution are themselves

specified by prior distributions. In this scenario however the choice of distribution for the mean and spread parameter (usually variance) as hyper-parameters, can be assumed. Because of the Central Limit theorem, a sensible choice for the prior of the mean is a Gaussian distribution, but the spread parameter typically has support between zero and infinity, and often weakly informed priors chosen for this, from either uniform distributions or half- $t$  family (including the Cauchy distribution) (Gelman 2006)<sup>1</sup>.

In our situation, we are doing the reverse, and given a mean and standard deviation, and confidence intervals for each, but no knowledge of the distributions of these quantities, the challenge is to produce a set of sampling distributions that accurately reflect the study definition.

## The sampling distribution of the mean

To do this we need to make some assumptions about the nature of the sampling distribution of the mean. Fortunately this is rather simple, as a key finding of the central limit theorem, as regardless of the underlying distribution, as the number of samples of a distribution increases the sampling distribution of the mean ( $E_n$ ) is a Gaussian where  $\bar{\mu}$  is the central estimate of the mean<sup>2</sup>:

$$E_n \sim \mathcal{N}(\bar{\mu}, \frac{\bar{\sigma}}{\sqrt{n}})$$

Knowing that the sampling distribution of the mean is a Gaussian, we can use this assumption to estimate the  $\frac{\bar{\sigma}}{\sqrt{n}}$  quantity, which is the standard deviation of the sampling distribution of the mean, from the confidence intervals, giving us a fully specified sampling distribution of the mean.

$$E_n \sim \mathcal{N}(\bar{\mu}, \frac{CI_{\mu,upper} - CI_{\mu,lower}}{N_{upper} - N_{lower}})$$

## The sampling distribution of the variance and standard deviation

A normally distributed variable  $x$  with an expected value  $\mu$  and a standard deviation  $\sigma$  is sampled  $n$  times. the set of  $x_n$  observations has a mean of  $\bar{x}$  and observed variance of  $S_n^2$ , the sampling distribution of the variance can be shown to be a Chi-squared distribution<sup>2</sup> with  $n - 1$  degrees of freedom. Given that the Chi-squared distribution is a particular form of a Gamma distribution (here parameterised with shape,  $\alpha$  and rate,  $\beta$ ) and given the definition of the Nakagami-m distribution<sup>3</sup>, the following holds:

$$\begin{aligned} (n-1)S_n^2/\sigma^2 &= \frac{1}{\sigma^2} \sum_i (x_i - \bar{x})^2 \sim \tilde{\chi}_{n-1}^2 \\ \tilde{\chi}_{n-1}^2 &= \text{Gamma}\left(\frac{n-1}{2}, \frac{1}{2}\right) \\ X \sim \text{Gamma}(\alpha, \beta) &\implies yX \sim \text{Gamma}(\alpha, \beta/y) \\ S_n^2 = \frac{\sigma^2}{n-1} \left(\frac{(n-1)S_n^2}{\sigma^2}\right) &\sim \text{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2\sigma^2}\right) \\ m = \kappa = \alpha &= \frac{n-1}{2} \\ \Omega = \kappa\theta = \frac{\alpha}{\beta} &= \sigma^2 \\ X \sim \text{Gamma}(\alpha, \beta) &\implies \sqrt{X} \sim \text{Nakagami}(m, \Omega) \\ S_n &\sim \text{Nakagami}\left(\frac{n-1}{2}, \sigma^2\right) \end{aligned}$$



With no information about the nature of the underlying distribution this expression is a bounding limit on the sampling distribution of the standard deviation, and we use this in the situation where the central estimate of the standard deviation is given, alongside the sample size, but with no other information. However this estimate is unreliable in the situation where there are small numbers or there is kurtosis in the distribution<sup>4</sup>, and could lead to a broader range of samples than would be compatible with the reported results when these are Gamma, or Log-normally distributed.

In O’Neill (2014)<sup>5</sup> the asymptotic sampling distribution of the variance is explored with respect to the kurtosis of the underlying distribution, and this modifies the degrees of freedom applied to the Chi-squared distribution above, to the following expression, where  $\kappa$  is the kurtosis of the underlying distribution (this is their result 14).

$$DF_n S_n^2 / \sigma^2 \sim \tilde{\chi}_{DF_n}^2$$

$$DF_n = \frac{2n}{\kappa - (n-3)/(n-1)}$$

$$S_n^2 = \frac{\sigma^2}{DF_n} \left( \frac{(DF_n) S_n^2}{\sigma^2} \right) \sim \text{Gamma} \left( \frac{DF_n}{2}, \frac{DF_n}{2\sigma^2} \right)$$

$$S_n \sim \text{Nakagami} \left( \frac{DF_n}{2}, \sigma^2 \right)$$

Information about the kurtosis of the underlying distribution is available from the confidence limits on the standard deviations quoted in source studies and a closed form expression for these is given in O’Neill (2014)<sup>4</sup>. This involves the population size from which the sample is taken which is information we do not generally have. With both confidence intervals, it would be possible to eliminate the unknown population size (or we could reasonably assume it is very much larger than our sample size), but it is also possible to estimate the associated Nakagami distribution numerically from the confidence intervals and central estimate of the standard deviation ( $\sigma$ ) from the expression above. These again describe bounding distributions for the sampling distribution of the standard deviation.

## Generating samples from uncertain distributions

The main purpose of this approach is to generate a representative sample set from uncertainly specified parameterised distributions such that they can be combined. To test this we investigate a list of published studies that give estimates of the serial interval of SARS-CoV-2 as a parameterised distribution (see Table S2.1).

Table S2.1: A set of uncertainly specified parameterised distributions extracted from the SARS-CoV-2 literature.

source	dist	N	param	summary
Bi et al. 2020	gamma	48	mean sd	6.30, (95% CI 5.20–7.60) 4.20, (95% CI 3.10–5.30)
Cereda et al. 2020	gamma	90	mean sd	6.60, (95% CI 0.70–19.00) 4.88, (95% CI NA–NA)
Du et al. 2020	norm	468	mean sd	3.96, (95% CI 3.53–4.39) 4.75, (95% CI 4.46–5.07)
Kwok et al. 2020	lnorm	26	mean sd	4.58, (95% CI 3.35–5.85) 3.28, (95% CI 2.18–4.01)
Nishiura et al. 2020	lnorm	28	mean sd	4.70, (95% CI 3.70–6.00) 2.90, (95% CI 1.90–4.90)
Son et al. 2020	gamma	28	mean sd	5.54, (95% CI 4.08–7.01) 3.90, (95% CI 2.47–5.32)
Zhang et al. 2020	gamma	28	mean sd	5.00, (95% CI 0.80–13.00) 3.22, (95% CI NA–NA)
Zhao et al. 2020	gamma	21	mean sd	4.40, (95% CI 2.90–6.70) 3.00, (95% CI 1.80–5.80)

Using the methodology described above, for each of these studies a sampling distribution for the mean and standard deviation is estimated. This is used to generate 1000 representative parameterised distributions for

each study. From these 1000 distributions, 1000 random samples are taken representing 1,000,000 generated samples per study. With all studies taken together and with equally weighting, the combined sample has a mean and SD of  $5.32 \pm 4.34$ , (95% CI -0.03–15.99) however in reality we would take a number of samples proportional to the study size when combining.

More relevant though is comparing the distribution of means and standard deviations recovered from the re-sampling process. In this case keeping 1000 sample from each of the 1000 inferred distributions from each study separate, and summarizing the samples shows us how well the distribution sampling is performing. In the following figure, for the Bi et al. 2020 study, we see the sampled mean and standard deviation in each of the 1000 precisely specified distributions, compared to the quoted central estimates (solid red) and confidence intervals (dashed lines) in Panel A, and in panel B the associate shape and rate parameters.

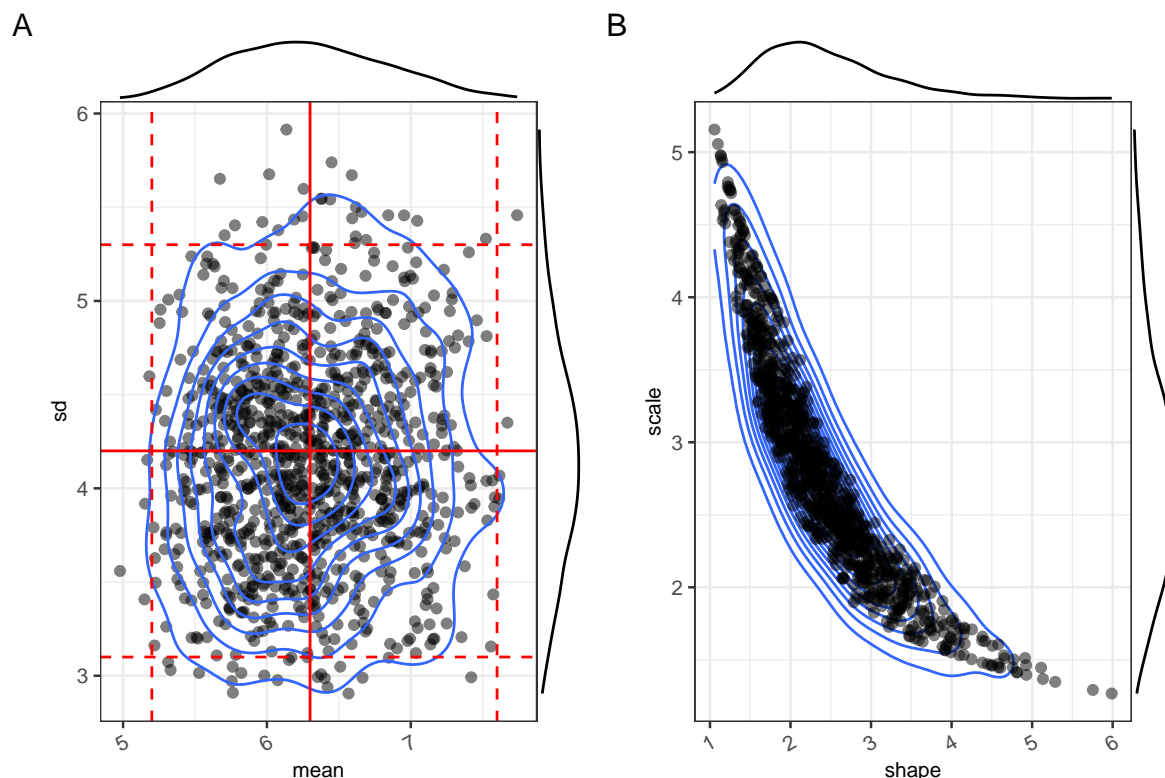


Figure S2.1: Panel A shows the first 2 moments of a set of precisely specified parameterised distributions compatible with the findings of Bi et al. 2020. Central estimates and confidence intervals from that study are marked in red lines in panel A. Panel B is the equivalent distributions expressed in shape and scale parameters of the Gamma distribution.

Combining the summaries from the Figure S2.1 above, allows us to reconstruct the uncertainty in the mean and standard deviation in our sampled data, and reconstruct central estimates and confidence intervals, for each source, which are shown in Table S2.2. These compare well with the originally reported values from the papers, where the numbers of cases is sufficiently large, or the reported confidence intervals are not excessively wide. It is less accurate where the very small numbers in some of the studies leads to wide confidence intervals, for example Zhang et al. 2020. In such cases the ability to replicate the exact shape is arguably less important for our intended purpose as such small studies will be relatively down-weighted during meta-analysis.

Table S2.2: The comparison between original uncertain distributions, and the result of aggregating samples generated using the procedure described in this paper. A good agreement is shown when source distributions are reasonably well constrained to begin with

source	dist	N	param	type	summary
Bi et al. 2020	gamma	48	mean	original	6.30, (95% CI 5.20–7.60)
				sampled	6.29 ± 0.53, (95% CI 5.35–7.33)
			sd	original	4.20, (95% CI 3.10–5.30)
				sampled	4.16 ± 0.56, (95% CI 3.13–5.31)
Cereda et al. 2020	gamma	90	mean	original	6.60, (95% CI 0.70–19.00)
				sampled	7.41 ± 3.85, (95% CI 1.17–15.11)
			sd	original	4.88, (95% CI NA–NA)
				sampled	4.86 ± 0.45, (95% CI 4.07–5.70)
Du et al. 2020	norm	468	mean	original	3.96, (95% CI 3.53–4.39)
				sampled	3.96 ± 0.23, (95% CI 3.52–4.42)
			sd	original	4.75, (95% CI 4.46–5.07)
				sampled	4.75 ± 0.18, (95% CI 4.41–5.12)
Kwok et al. 2020	lnorm	26	mean	original	4.58, (95% CI 3.35–5.85)
				sampled	4.62 ± 0.55, (95% CI 3.56–5.71)
			sd	original	3.28, (95% CI 2.18–4.01)
				sampled	3.22 ± 0.50, (95% CI 2.34–4.25)
Nishiura et al. 2020	lnorm	28	mean	original	4.70, (95% CI 3.70–6.00)
				sampled	4.73 ± 0.51, (95% CI 3.82–5.74)
			sd	original	2.90, (95% CI 1.90–4.90)
				sampled	2.85 ± 0.77, (95% CI 1.83–4.61)
Son et al. 2020	gamma	28	mean	original	5.54, (95% CI 4.08–7.01)
				sampled	5.52 ± 0.67, (95% CI 4.25–6.78)
			sd	original	3.90, (95% CI 2.47–5.32)
				sampled	3.84 ± 0.71, (95% CI 2.53–5.23)
Zhang et al. 2020	gamma	28	mean	original	5.00, (95% CI 0.80–13.00)
				sampled	5.51 ± 2.66, (95% CI 1.17–11.16)
			sd	original	3.22, (95% CI NA–NA)
				sampled	3.17 ± 0.45, (95% CI 2.36–4.07)
Zhao et al. 2020	gamma	21	mean	original	4.40, (95% CI 2.90–6.70)
				sampled	4.49 ± 0.84, (95% CI 3.09–6.21)
			sd	original	3.00, (95% CI 1.80–5.80)
				sampled	2.89 ± 0.97, (95% CI 1.75–5.29)

## Conclusion

We have presented a short summary on the method we use to generate samples from uncertainly specified parameterised distributions. We have demonstrated it is able to produce both a set of exactly specified parameterised distributions that is representative of the uncertainty in the original specification, but also that from this approach we can generate a set of samples that cover the range of possibilities described in the original source in a representative way. When aggregated these samples recover the original uncertainty with a reasonable degree of fidelity. This method is used in our approach to meta-analysis of quantities such as the serial interval that are reported in the literature as uncertainly defined parameter distributions, rather than as single dimensional effect sizes.

## References

1. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* [Internet]. 2006 Sep [cited 2021 Jul 5];1(3):515–34. Available from: <https://projecteuclid.org/journals/bayesian-analysis/volume-1/issue-3/Prior-distributions-for-variance-parameters-in-hierarchical-models-comment-on/10.1214/06-BA117A.full>
2. Sampling Distribution of Sample Variance | STAT 414 [Internet]. PennState: Statistics Online Courses. [cited 2021 Jul 3]. Available from: <https://online.stat.psu.edu/stat414/lesson/26/26.3>
3. Nakagami M. The m-Distribution—A General Formula of Intensity Distribution of Rapid Fading. In: Hoffman WC, editor. *Statistical Methods in Radio Wave Propagation* [Internet]. Pergamon; 1960 [cited 2021 Jul 6]. p. 3–36. Available from: <https://www.sciencedirect.com/science/article/pii/B9780080093062500054>
4. Douillet PL. Sampling distribution of the variance. In: *Proceedings of the 2009 Winter Simulation Conference (WSC)* [Internet]. Austin, TX, USA: IEEE; 2009 [cited 2021 Jun 29]. p. 403–14. Available from: <http://ieeexplore.ieee.org/document/5429351/>

5. O'Neill B. Some Useful Moment Results in Sampling Problems. *The American Statistician* [Internet]. 2014 Oct 2 [cited 2021 Jul 3];68(4):282–96. Available from: <https://doi.org/10.1080/00031305.2014.966589>

#### **Serial interval source citations**

- Bi et al. 2020** Bi, Q. et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *The Lancet Infectious Diseases* 20, 911–919 (2020).
- Cereda et al. 2020** Cereda, D. et al. The early phase of the COVID-19 outbreak in Lombardy, Italy. arXiv:2003.09320 [q-bio] (2020).
- Du et al. 2020** Du, Z. et al. Serial Interval of COVID-19 among Publicly Reported Confirmed Cases. *Emerg Infect Dis* 26, 1341–1343 (2020).
- Kwok et al. 2020** Kwok, K. O., Wong, V. W. Y., Wei, W. I., Wong, S. Y. S. & Tang, J. W.-T. Epidemiological characteristics of the first 53 laboratory-confirmed cases of COVID-19 epidemic in Hong Kong, 13 February 2020. *Eurosurveillance* 25, 2000155 (2020).
- Nishiura et al. 2020** Nishiura, H., Linton, N. M. & Akhmetzhanov, A. R. Serial interval of novel coronavirus (COVID-19) infections. *Int. J. Infect. Dis.* 93, 284–286 (2020).
- Son et al. 2020** Son, H. et al. Epidemiological characteristics of and containment measures for COVID-19 in Busan, Korea. *Epidemiol Health* 42, (2020).
- Zhang et al. 2020** Zhang, J. et al. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. *The Lancet Infectious Diseases* 20, 793–802 (2020).
- Zhao et al. 2020** Zhao, S. et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *Int. J. Infect. Dis.* 92, 214–217 (2020).

# 6. Estimates of regional infectivity of COVID-19 in the United Kingdom following imposition of social distancing measures

Status: published in Philosophical Transactions B

Date: 2nd Dec 2020.

Candidate's contribution: All authors discussed the concept of the article and R.C. performed the analysis and wrote the initial draft. K.T.A., M.P., T.E., L.G., L.L., E.B.P. and L.D. commented and made revisions. All authors read and approved the final manuscript. R.C. is the guarantor.

All the analysis in this paper was performed by me and findings validated by other team members.

Supplementary material:

Research



**Cite this article:** Challen R, Tsaneva-Atanasova K, Pitt M, Edwards T, Gompels L, Lacasa L, Brooks-Pollock E, Danon L. 2021 Estimates of regional infectivity of COVID-19 in the United Kingdom following imposition of social distancing measures. *Phil. Trans. R. Soc. B* **376**: 20200280.  
<https://doi.org/10.1098/rstb.2020.0280>

Accepted: 2 December 2020

One contribution of 21 to a theme issue 'Modelling that shaped the early COVID-19 pandemic response in the UK'.

**Subject Areas:**  
health and disease and epidemiology

**Keywords:**  
COVID-19, SARS-CoV-2, reproduction number, regional variation

**Author for correspondence:**  
Robert Challen  
e-mail: [rc538@exeter.ac.uk](mailto:rc538@exeter.ac.uk)

Electronic supplementary material is available online at <https://doi.org/10.6084/m9.figshare.c.5401677>.

# Estimates of regional infectivity of COVID-19 in the United Kingdom following imposition of social distancing measures

Robert Challen<sup>1,3</sup>, Krasimira Tsaneva-Atanasova<sup>1,4</sup>, Martin Pitt<sup>5</sup>, Tom Edwards<sup>3</sup>, Luke Gompels<sup>3</sup>, Lucas Lacasa<sup>6</sup>, Ellen Brooks-Pollock<sup>7</sup> and Leon Danon<sup>2,4,7</sup>

<sup>1</sup>EPSRC Centre for Predictive Modelling in Healthcare, and <sup>2</sup>Data Science Institute, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4SB, UK

<sup>3</sup>Taunton and Somerset NHS Foundation Trust, Musgrove Park Hospital, Taunton TA1 5DA, UK

<sup>4</sup>The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK

<sup>5</sup>NIHR CLAHRC for the South West Peninsula, University of Exeter Medical School, St Luke's Campus, Exeter, UK

<sup>6</sup>School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, UK

<sup>7</sup>Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK

RC, 0000-0002-5504-7768; EB-P, 0000-0002-5984-4932; LD, 0000-0002-7076-1871

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) reproduction number has become an essential parameter for monitoring disease transmission across settings and guiding interventions. The UK published weekly estimates of the reproduction number in the UK starting in May 2020 which are formed from multiple independent estimates. In this paper, we describe methods used to estimate the time-varying SARS-CoV-2 reproduction number for the UK. We used multiple data sources and estimated a serial interval distribution from published studies. We describe regional variability and how estimates evolved during the early phases of the outbreak, until the relaxing of social distancing measures began to be introduced in early July. Our analysis is able to guide localized control and provides a longitudinal example of applying these methods over long timescales.

This article is part of the theme issue 'Modelling that shaped the early COVID-19 pandemic response in the UK'.

## 1. Introduction

In late 2019, an outbreak of a novel infectious disease was detected. It manifested principally with severe acute respiratory distress and pneumonia, [1] although many cases followed a mild course [2]. The pathogen was rapidly identified as a new species of coronavirus (severe acute respiratory syndrome coronavirus 2, SARS-CoV-2) and the disease named COVID-19 [3]. Global transmission of the virus followed and major outbreaks have been observed in Europe, beginning with Italy [4]. On 31 January 2020, the first cases were identified in the UK [5]. This was initially managed using testing of suspected individuals in the community, contact tracing and isolation of affected cases. However, this was successful only in delaying the spread of the disease and on 13 March 2020, the UK government moved towards a mitigation strategy reserving testing for hospital inpatients only [6]. Following this, a stepwise implementation of social distancing measures was mandated by the government including voluntary self-isolation of any symptoms and vulnerable people [7], a ban on non-essential travel worldwide [8] and school closures

[9]. Finally on 23 March 2020, the government mandated that everyone apart from essential workers should stay at home and away from others [10], instituting a countrywide 'lock-down'.

Epidemiological studies conducted during the outbreak in China have provided us with a number of estimates of the parameters describing the virus's spread through the population including a reproduction number between 2.24 and 3.58 [11] and a median incubation period of 5.1 days (credible interval 4.5 to 5.8) [12]. It is estimated that fewer than 2.5% of people will show signs before 2.2 days and 97.5% of people who will develop symptoms will have done so by 11.2 days after exposure [12].

We investigated the reproduction number of SARS-CoV-2 in the UK to determine whether there are any spatial or temporal patterns beyond those resulting from the imposition of social distancing measures, and to track the progression of the outbreak. This article summarizes the methodology and interpretation of national and regional estimates of the time-varying reproduction number ( $R_t$ ) of the SARS-CoV-2 outbreak in the UK. These estimates were provided to the Scientific Pandemic Influenza Group on Modelling (SPI-M) [13] and formed part of the weekly UK  $R_t$  estimates [14].

## 2. Methods

This section describes the data sources, their processing and combination, and our methodology for estimating the serial interval of SARS-CoV-2 infections and calculation of  $R_t$  estimates.

### (a) Data

We integrate data from a variety of sources, both publicly available and provided to the SPI-M [13] by Public Health England (PHE) and the Defence Science and Technology Laboratory (DSTL) [15]. We use these data to estimate  $R_t$  and the exponential growth rate for the UK as a whole, four nations of the UK (CTRY) and seven NHS regions in England (NHSER).

At the UK level, data are available on cases and deaths through the PHE coronavirus tracker [16]. The tracker publishes an overall number of cases and deaths in the UK on a daily basis. It also provides a regional breakdown of the four nations, which exclude tests performed in private laboratories (Pillar 2 tests). At the time of this analysis, the historical time series of the UK level data was not made available through the PHE site; however, this information was collected prospectively and curated by Tom White's aggregated COVID 19 UK data github site [17]. Cumulative case counts from both the PHE headline UK figure and the combined sum of the four nations are compared as these numbers differ.

Hospital admission data are available from NHS trusts across the UK via the DSTL, which in turn aggregate the situation reports provided by NHS hospital trusts. These are aggregated to UK level. Although referred to here as 'hospital admission', it includes admissions of patients who are subsequently identified as COVID-19 cases in hospital, but identified in the hospital, and patients with known COVID-19 who are then admitted to hospital.

For the four devolved nations of the UK, we used different data sources for each nation. In England, we used data from the SPI-M provided line lists for cases and deaths in England. Cases are restricted to those processed in NHS labs (Pillar 1) and are available by date of specimen collection. For the other three nation states of the UK, as above, both historical cases and death data are aggregated from Public Health Wales [18] and Scotland's [19] sites, and from Northern Ireland's HSC site

[20] respectively, using a time series retrieved from Tom White's aggregated COVID 19 UK data github site [17] for Scotland, Wales and Northern Ireland. Death data from the four nations are provided via the CHESS (COVID-19 hospitalisation in England surveillance system) dataset [21]. Death data are subject to a weekly periodicity due to reporting delay over the weekend which is mitigated by using the date of death rather than the date of the report.

At the level of the NHS England regions, we take the case data from anonymized test result line listings. Mortality data for NHS England are provided by PHE in a canonical line list of deaths and is used alongside the CHESS dataset [21]. Admissions data are available from DSTL feeds and ICU admissions from CHESS dataset [21].

For England as a nation and for the NHS England regions, we also have data available from triage telephone calls from NHS 111 and 999 services. The data on the calls made to 111 and 999 as well as the outcome of that call are provided as aggregate numbers broken down by age.

For all data sources, cumulative case figures are converted into daily incidence figures and any data which are broken down by age, or by gender, are combined. In case and death data, the final 5 days of the time series are discarded to account for possible reporting delay. The resulting time series are analysed for outlying data points, which are more than five standard deviations away from the mean of the nearest 14 data points. Outlying or missing data are imputed from a linear interpolation of the logarithm of incidence figures, implemented in the R 'forecast' library [22], and the results are truncated to ensure no negative incidence figures. A smoothing function (a linear spline interpolation) is then fitted to the logarithm of incidence applied over a 7-day window [23]. This is needed as all the data sources have some degree of weekly periodicity regardless of source.

### (b) Combination of data sources

Our estimates are based on an aggregation of the various data sources described above. We generate four estimates of  $R_t$  based on single time series from each data stream:

- (a) deaths (EpiEstim/Deaths),
- (b) cases (EpiEstim/4NationsCases),
- (c) telephone triage (EpiEstim/Triage),
- (d) hospital admissions.

The aggregation procedure is kept simple and combines multiple data sources when applicable using the mean. Any possible biases this introduces are consistent throughout the time series so that it does not affect relative changes and hence either estimates of  $R_t$  or growth rates. The resulting time series are manually inspected for consistency and to check there are no abrupt changes in the data streams. The source of the combined datasets and more information about the processing steps used is shown in the electronic supplementary material, table S1.

### (c) The serial interval of SARS-CoV-2 infections, $R_t$ and time delay of estimates

The serial interval and the generation interval are closely related measures. The generation interval is a measure of the time taken for an infection to pass from one person (infector) to another (infectee) in a chain of transmission. The serial interval, on the other hand, is a measure of the time between the appearance of clinical symptoms in the infector and infectee. The generation interval cannot be observed directly as both infection events are only detectable once the virus has incubated and become symptomatic. The serial interval is often used as an observable proxy

for the generation interval. The use of the serial interval as a proxy for the generation interval is known to produce biased estimates for the reproduction number [24,25], but has the advantage of being directly observed through contact tracing.

Assumptions about the serial interval of SARS-CoV-2 have an impact on the absolute level of our estimates of  $R_t$ . We have used multiple approaches to estimate the serial interval. First, a UK specific serial interval was calculated from early case tracking data (the FF100 case data provided by DSTL). Second, a literature review was conducted and the serial intervals from a range of sources [4,26–35] pooled [36]. The serial interval can be well described by a truncated empirical distribution, with a mean plus 95% credible interval of 5.59 days (5.09; 6.20), and a standard deviation of 4.15 days (3.94; 4.46). This uncertainty in the serial interval distribution is propagated to our estimates of  $R_t$  and used to determine confidence intervals. In other work, we have analysed the effect of the pragmatic use of serial interval instead of generation interval and find the effect on estimates of  $R_t$  to be small (approx. 5% of the absolute value) when it is close to one. This choice does not influence the estimate of time when  $R_t$  transitions from growth to decay [36].

Using the inferred serial interval distribution, we analysed the time-series data using the forward equation method [37], implemented in the EpiEstim R library [37–39], to estimate the  $R_t$  during the outbreak.

The renewal equation method is predicated on a time series of infections and on the infectivity profile—a measure of the probability that a secondary infection occurred on a specific day after the primary case, given a secondary infection occurred [37]. A Bayesian framework is then used to update a prior probabilistic estimate of  $R_t$  on any given day with both information gained from the time series of infections in the epidemic to date and the infectivity profile to produce a posterior estimate. In both the original [38] and revised [39] implementations of this method, the authors acknowledge the pragmatic use of the serial interval distribution as a proxy measure for the infectivity profile, and the incidence of symptom onset or case identification as a proxy for the incidence of infection, with the caveat that these introduce a time lag into the estimates of  $R_t$ . We make the simplifying assumption that there is negligible mixing of populations between each geographical area and treat each location independently.

The method uses a sliding time window during which the instantaneous reproduction number is assumed to be constant. We used both a 7-day and a 28-day sliding window for calculations of the  $R_t$  which provides two estimates with alternative trade-offs between noise and loss of detail. Our  $R_t$  estimate is calculated using a loosely informed prior estimate of  $R_t$  as a gamma distribution with a mean of 1 and a standard deviation 2. This prior distribution is based on an assumption of the approximate value of  $R_t$  on the conditions following lock-down, rather than reflecting values of  $R_0$  commonly described in the literature [11] as those are based on the situation without social distancing measures in place.

Other events in the timeline of infection also serve as a proxy for observations of infections in the past including positive testing, hospitalization for severe disease, or death. Although best practice is to calculate  $R_t$  using a generation interval distribution and infection events [25], neither the generation interval distribution nor the infection event data can be directly observed, and inferring them can also introduce potential bias and uncertainty. As a pragmatic initial step, we use the serial interval distribution described above in lieu of the generation interval, and the various observations available to us, including triage contacts, cases, admissions and deaths as a proxy of prior transmission events, and compare those results.

There is a time delay between infection, symptom onset, case identification, hospital admission and ultimately death, and this affects the timing of our estimation of  $R_t$ . In a separate

**Table 1.** Estimates of the value of  $R_t$  in the UK on 4 July 2020.

observation	$R_t$ (95% CI)	count per 1M per day (95% CI)
cases	0.87 (0.81; 0.92)	2.4 (2.0; 2.9)
deaths	0.91 (0.84; 0.99)	1.4 (1.3; 1.5)
admissions	0.81 (0.76; 0.86)	2.2 (2.0; 2.4)

analysis, we estimated these time delays (shown in electronic supplementary material, table S2) and apply these estimates as a correction to the time of our estimates of  $R_t$  to align them to the date of presumed infection.

Code and processed data involved in this analysis are available on GitHub [40].

### 3. Results

#### (a) UK overview

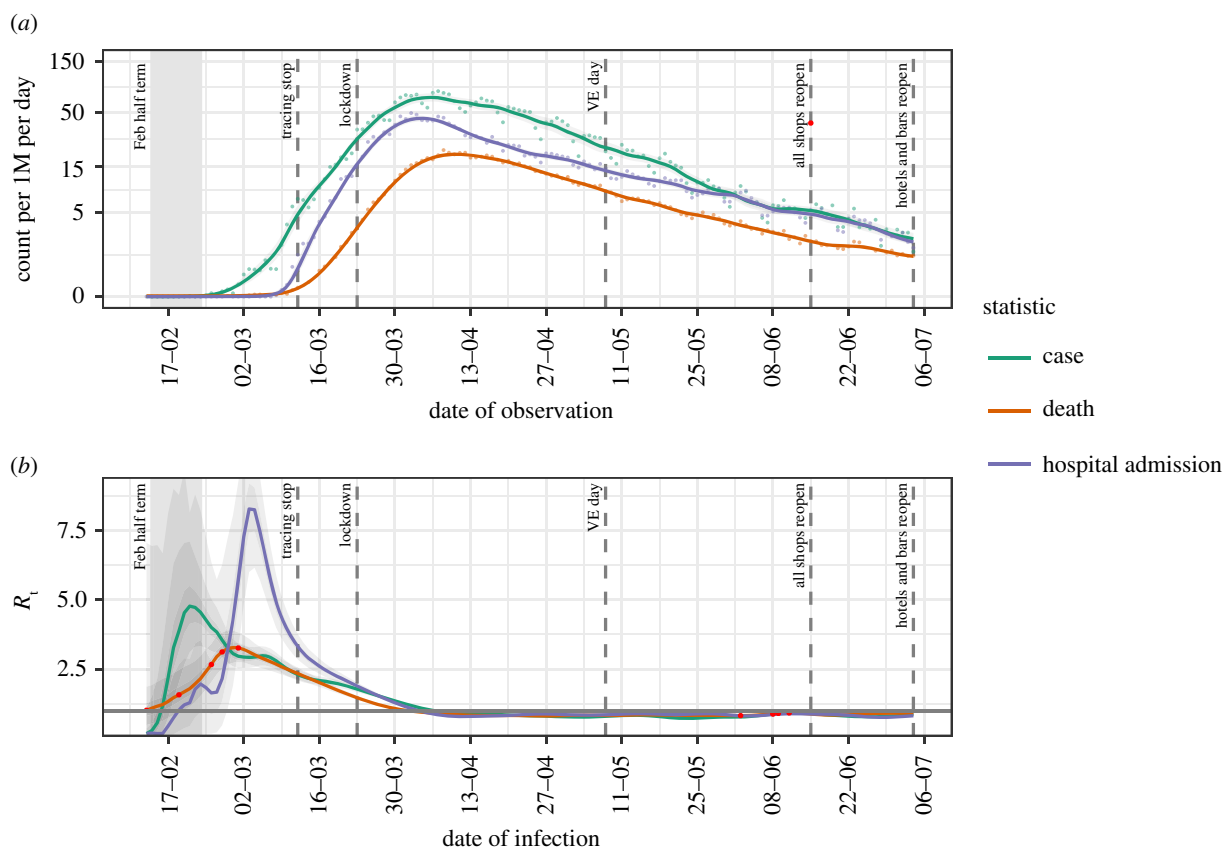
Our estimate of the median value of  $R_t$  for the UK based on cases, deaths and hospital admissions from 4 July 2020 is presented in table 1.

In figure 1*a*, we show the incidence per million people of cases, deaths and hospital admissions due to COVID-19 in the UK over the outbreak; figure 1*b* shows the associated values for  $R_t$ . The three data sources show similar patterns of exponential increase, with admissions and deaths lagging cases followed by a slower phase of exponential decline, beginning 2–3 weeks after the lock-down. In figure 1*b*, we see the estimates of  $R_t$  corrected to date of infection. There is a prominent single initial peak in  $R_t$  from admissions and deaths in late February, followed by a decline over the course of the next few months.  $R_t$  crosses 1 at the beginning of April, after which it remains below 1 during the lock-down period. Estimates based on cases show a biphasic pattern with an initial peak in mid-February and a second smaller peak in early March, following the same pattern as other estimates. The peak values of  $R_t$  vary by observation with peak  $R_t$  by admissions being 8.3, by cases 4.8 and by deaths 3.3. These are within the confidence limits of estimates described in other countries [11].

#### (b) Countries in the UK

Estimates of  $R_t$  in the different countries of the UK, based on cases, deaths, hospital admissions or triage calls for 4 July 2020 are presented in table 2 and figure 2. Triage figures based on 111 and 999 coronavirus pathway calls were available for England only; we do not have access to the full time series of all information for all countries. We show estimates based on a 28-day rolling window as cases and deaths in Northern Ireland and Scotland have fallen to a level which makes estimates over a shorter window unreliable. With insufficient information, the Bayesian method used reverts to the prior value of  $R_0$  supplied which we set to 1. These estimates show the median value of  $R_t$  is below 1 for all four nations, using all data sources. The pattern in all nations is similar with  $R_t$  rapidly decreasing following lock-down on 23 March and becoming less than 1 in early April. Northern Ireland and Scotland have maintained a lower  $R_t$  for a longer period of time than England and Wales, with  $R_t$  values in Northern Ireland and Scotland at or below 0.75 for much of





**Figure 1.** Timeline of cases and estimates of  $R_t$  based on cases reported by PHE and NHS laboratories (green), deaths reported in NHS trusts (red) and best available data for hospital admissions (blue). (a) Number of cases (Pillar 1), deaths and admissions per million; (b) estimates of  $R_t$ . Red points are either missing values or identified as anomalies and replacements imputed.

**Table 2.** Estimates of mean  $R_t$  and 95% confidence intervals for the individual countries in the UK, based on cases, deaths and hospital admissions on 4 July 2020.

country	observation	$R_t$ (95% CI)	count per 1M per day (95% CI)
England	cases	0.85 (0.83; 0.88)	2.4 (2.0; 2.8)
	deaths	0.90 (0.86; 0.93)	1.6 (1.5; 1.7)
	admissions	0.84 (0.81; 0.86)	2.5 (2.3; 2.7)
	triage	0.94 (0.93; 0.96)	14.6 (13.2; 16.1)
Northern Ireland	deaths	0.70 (0.31; 1.32)	0.2 (-0.0; 0.4)
Scotland	admissions	0.72 (0.44; 1.09)	0.4 (0.2; 0.7)
Wales	cases	0.83 (0.75; 0.93)	1.3 (0.9; 1.8)
	admissions	0.74 (0.53; 1.00)	0.1 (0.0; 0.2)
	deaths	0.85 (0.81; 0.90)	6.5 (4.7; 8.8)
Wales	deaths	0.79 (0.63; 0.99)	1.0 (0.7; 1.3)
	admissions	0.87 (0.76; 0.98)	1.7 (1.4; 2.2)

May, June and July compared to those in England and Wales, which have been between 0.75 and 1 for the same period.

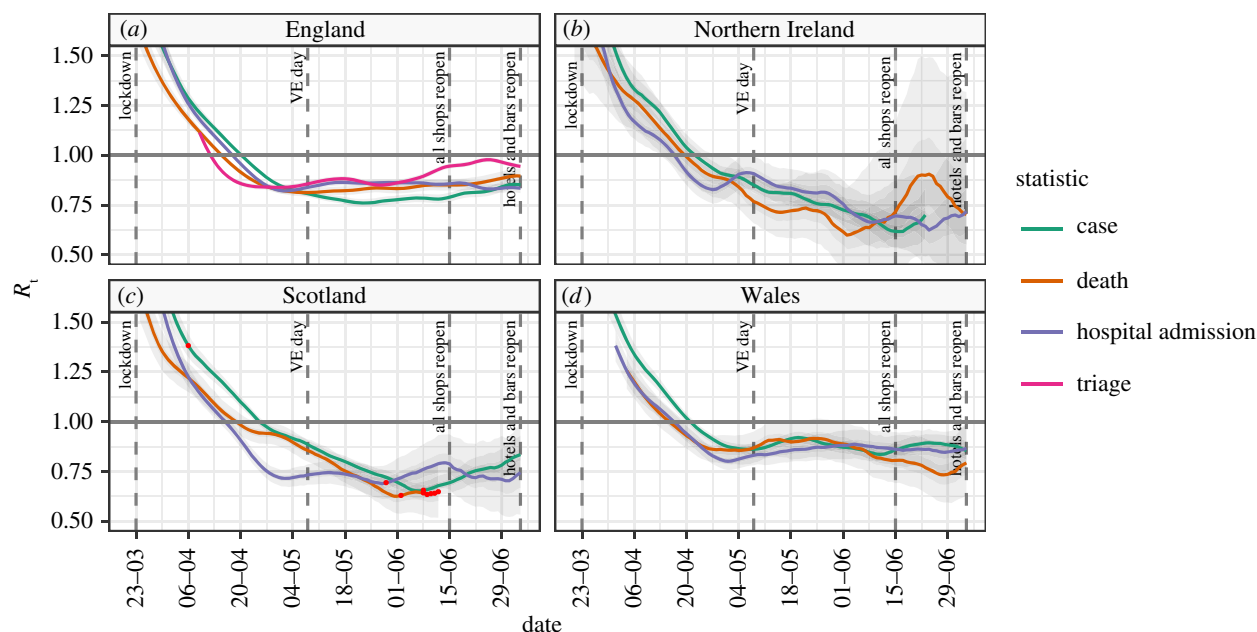
### (c) $R_t$ by England NHS region

In figure 3 (and, for completeness, electronic supplementary material, table S3), we present the same results as above but with a focus on the NHS regions in England. On 4 July 2020, the estimates of the mean of  $R_t$  were largely below 1

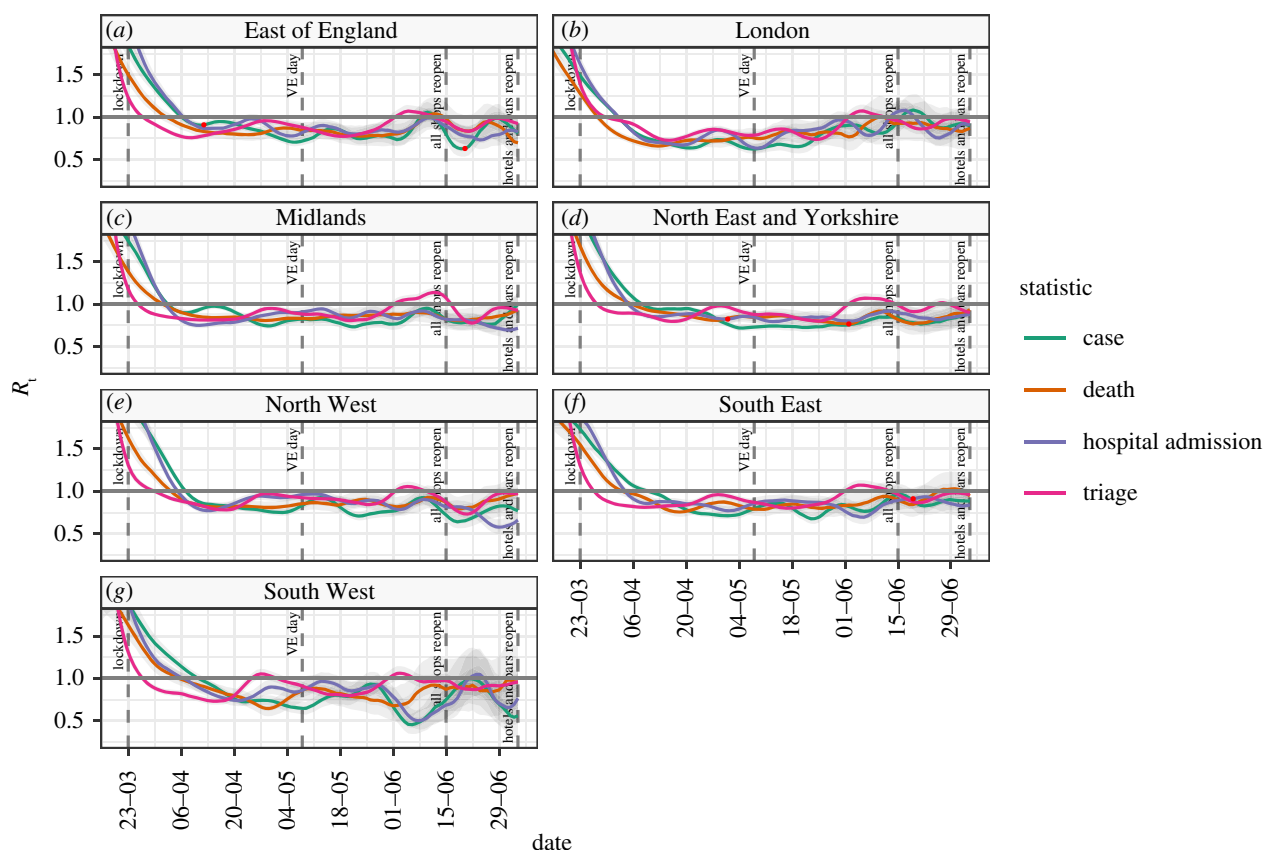
in the individual NHS regions. The observed count of different observations demonstrates a clear regional difference between London and the South West and the rest of the country with lower rates of all indicators than other regions. The time series of regional estimates of  $R_t$  reflect the overall patterns of England, albeit with more volatility due to the different windowing size between figures 2 and 3. It is, however, possible to see higher levels of uncertainty in the South West reflecting the smaller case numbers and to a lesser extent the same in London.

### (d) $R_t$ differences from England baseline

In figure 4, we plot the absolute difference of  $R_t$  in the seven NHS England administrative regions, from the  $R_t$  of England overall, as a baseline for each data source. This is based on 28-day window estimates due to the volatility observed in figure 3. This analysis highlights the regional differences in  $R_t$  over time. Over the period of the lock-down, the East of England, Midlands and South East regions approximately tracked the England baseline. The South West was seen to be initially following the national average but from June demonstrates a trend towards a lower value, although our confidence in these estimates is low. When lock-down started, London was initially well below the England average, by most indicators, and this continued until some point in late May, after which the trend reversed. On the other hand, the North West and less so North East and Yorkshire were consistently above the rest of the country until late May when they eventually reached the England average, and since mid-June, the  $R_t$  estimates in the North West have been lower than this



**Figure 2.** The median value of  $R_t$  and 95% confidence intervals for the individual countries in the UK, based on cases, deaths and hospital admissions, and a 28-day rolling window. For each data source, red points represent data points that were missing and have been imputed.



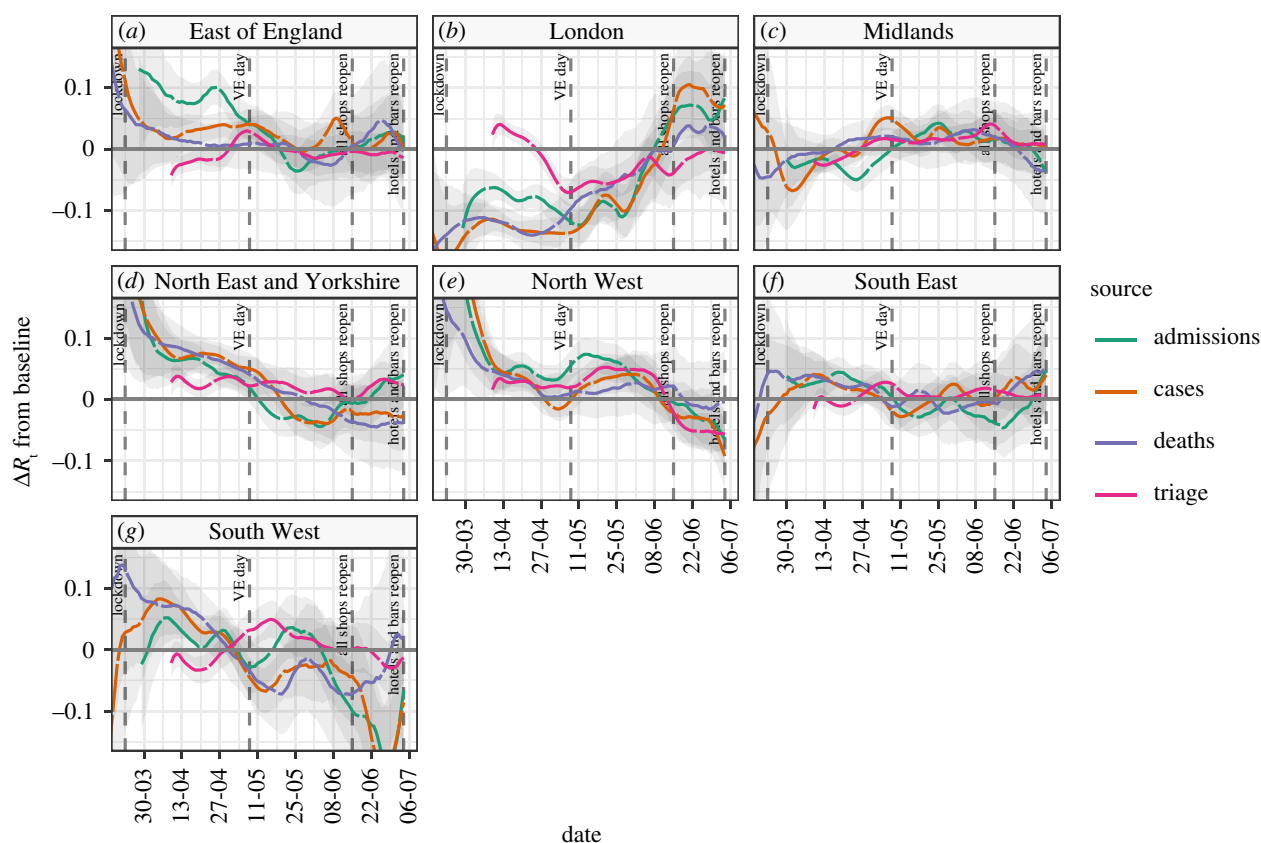
**Figure 3.** The median value of  $R_t$  and 95% confidence intervals for the sub-national regions of NHS England, based on cases, deaths and hospital admissions, and a 7-day rolling window.

average. The impact this variation has had on case loads is seen in table 3 which shows the pre- and post-lock-down rates of newly identified cases in each NHS region.

## 4. Discussion

In this paper, we estimated the reproduction number,  $R_t$  for COVID-19 in the UK using multiple data sources over a range

of geographies and time points, and with the same methodology. We find that, after the initial peak, different data sources produce similar results, without evidence of systematic bias for estimates based on more immediate measures (cases), compared to those based on longer term measures (deaths). In the UK,  $R_t$  peaked in mid-February 2020, coinciding with the end of the school holidays. We find that using different data sources affects the estimated size of peak burden although with considerable uncertainty.



**Figure 4.** The difference of  $R_t$  estimates for NHS regions and baseline  $R_t$  estimates for England, based on cases, deaths and hospital admissions, and a 28-day rolling window.

**Table 3.** Estimates of the burden of disease before and after lock-down in the different NHS regions.

NHS region	cases per 1M per day (95% CI) on 21 Mar	cases per 1M per day (95% CI) on 4 Jul
London	49.6 (45.0; 54.6)	1.7 (1.3; 2.1)
South East	17.4 (16.0; 19.1)	2.5 (2.0; 3.3)
South West	7.8 (6.6; 9.1)	0.4 (0.3; 0.6)
East of England	14.7 (12.4; 17.4)	2.7 (2.1; 3.6)
Midlands	20.8 (18.9; 23.0)	2.3 (1.8; 3.0)
North East and Yorkshire	13.5 (12.3; 14.8)	2.4 (2.1; 2.8)
North West	16.0 (14.3; 17.9)	2.6 (2.1; 3.2)

We find that  $R_t$  declined rapidly following lock-down on 23 March but did not reach the critical threshold ( $R_t$  less than one) separating growth and decline until early April. This delay is likely to be the combination of ongoing transmission within households, care homes and hospitals, or outbreaks in factories in key industries such as food production. Other factors that may influence the timing are delays in case identification and reporting, and limitations in the estimation methods. Subsequent to mid-April, we estimate  $R_t$  largely remained below one until the end of the lock-down period.

The multiple data sources we considered each have their advantages and drawbacks. Counts of test-positive cases and

telephone triage calls provide a rapid indication of infection risk and capture a broad representation of age groups, but may be influenced by changes in behaviour and testing policy. In the UK, initial attempts at community tracing were abandoned when case numbers started to outstrip test availability and afterwards testing was only performed on hospital admissions for suspected COVID-19 [6]. Later, test capacity was increased and the policy reversed to include more community cases, again altering the nature of the population being tested. Although a regional breakdown of testing capacity was not available at the time of this analysis, we do know that capacity was exceeded in the early phase of the epidemic, and this is one reason why case-based  $R_t$  estimates must be interpreted with caution until the middle of April. Hospital admissions and death data are less subject to changes in sampling strategy, although are subject to reporting delays and biases in ascertainment. As COVID-19 mortality is overwhelmingly in the elderly, statistics based on deaths mainly represent older groups. Due to reduced contact in the elderly, we propose the outbreak took longer to become established in those age groups. Counts of admissions may be unreliable when there is a delay in identifying a COVID-19 case, or when there is significant hospital transmission, as was the case in the early outbreak.

We took a pragmatic approach for sub-national analyses, based on data availability. In early April 2020, immediately following lock-down, estimates of  $R_t$  were lowest in London and highest in the North West and North East of England. At this time, the burden of cases was highest in London at nearly 50 cases per million people per day, but this reduced 25-fold over the lock-down period. In early July 2020, at the relaxation of lock-down, the case rates per capita were

lowest in the South West, due to the smaller initial outbreak size, followed by London due to the larger impact of social distancing. However, the benefit of lower case rates in London was offset by relative increases in  $R_t$ . When case numbers are low,  $R_t$  ceases to be a uniform statistic over a geographical area, because significant town to town variation will exist as clusters of infection become apparent. This was seen in the South West of England towards the end of the lock-down where increasing variability in regional estimates of  $R_t$  became less obviously significant. This reinforces the point that  $R_t$  is a relative measure and should not be interpreted without information about the incidence rate.

Our approach has a number of limitations. Our method treats the whole system and each region as independent and isolated. In reality, regions are connected via travel, although this was reduced during early social distancing measures in March and April. Furthermore, in the early phases, importation rates were high [41], and this would lead to our approach overestimating the true  $R_t$ .

## 5. Conclusion

We present a description of the methodology and data sources used in providing estimates of  $R_t$  in the UK for SPI-M [14]. Our approach is pragmatic and designed to produce timely, useful information to policy makers. Despite this, we find that using a number of data sources and careful interpretation helps elucidate the regional differences in  $R_t$  and shows they existed from the outset of lock-down and persisted during lock-down. Due to the compound nature of  $R_t$ , the result of this variation is higher case loads in Northern regions in the UK exiting lock-down.

As we move forward, early detection and prevention of the spread of emerging clusters of SARS-CoV-2 infections are critical to prevent large-scale outbreaks. This will be challenging as the long incubation period and high rate of

asymptomatic individuals makes undetected rapid spread easy. Prediction at a more localized level is needed to focus both community testing and more targeted social interventions on high-risk areas in the future.

Critical to this work continues to be rapid access to information about the spread of SARS-CoV-2 in the community both with high spatial resolution, but also with a short-time lag from infection to observation. As test and trace activities ramp up, we expect to see similar biases in case related data as testing volumes change locally in response to outbreaks. Our existing approaches to estimating  $R_t$  principally use hospital-based metrics and as such may not provide the perspective on the outbreak that is needed in the future. Telephone triage data are one potential source of information about local outbreaks and an area for future investigation [42].

**Data accessibility.** Data and code supporting this article can be accessed on GitHub at the following link: <https://github.com/terminological/uk-covid-datatools/tree/master/vignettes/current-rt>.

**Authors' contributions.** All authors discussed the concept of the article and R.C. wrote the initial draft. K.T.A., M.P., T.E., L.G., L.L., E.B.P. and L.D. commented and made revisions. All authors read and approved the final manuscript. R.C. is the guarantor. The views presented here are those of the authors and should not be attributed to TSFT or the GDE.

**Competing interests.** We declare we have no competing interests.

**Funding.** Support for R.C. and K.T.A.'s research is provided by the EPSRC via grant no. EP/N014391/1; R.C. is also funded by TSFT as part of the NHS Global Digital Exemplar programme (GDE); there were no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years, no other relationships or activities that could appear to have influenced the submitted work. L.D. and K.T.A. gratefully acknowledge the financial support of The Alan Turing Institute under the EPSRC grant no. EP/N510129/1. L.L. acknowledges the financial support of the EPSRC via Early Career Fellowship EP/P01660X/1. L.D. and E.B.P. are supported by Medical Research Council (MRC) (MC/PC/19067). E.B.P. was partly supported by the NIHR Health Protection Research Unit in Behavioural Science and Evaluation at University of Bristol, in partnership with PHE.

## References

- Huang C *et al.* 2020 Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506. (doi:10.1016/S0140-6736(20)30183-5)
- Wu Z, McGoogan JM. 2020 Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *JAMA* **323**, 1239–1242. (doi:10.1001/jama.2020.2648).
- World Health Organisation. In press. *Naming the coronavirus disease (COVID-19) and the virus that causes it*. [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it).
- Cereda D *et al.* 2020 *The early phase of the COVID-19 outbreak in Lombardy, Italy*. <http://arxiv.org/abs/2003.09320> (accessed 20 August 2020).
- Moss P, Barlow G, Easom N, Lillie P, Samson A. 2020 Lessons for managing high-consequence infections from first COVID-19 cases in the UK. *Lancet* **395**, e46. (doi:10.1016/S0140-6736(20)30463-3)
- Public Health England. 2020 *COVID-19: investigation and initial clinical management of possible cases*. GOV.UK. <https://www.gov.uk/government/publications/wuhan-novel-coronavirus-initial-investigation-of-possible-cases/investigation-and-initial-clinical-management-of-possible-cases-of-wuhan-novel-coronavirus-wn-cov-infection>.
- Public Health England. 2020 *COVID-19: guidance on social distancing and for vulnerable people*. GOV.UK. <https://www.gov.uk/government/publications/covid-19-guidance-on-social-distancing-and-for-vulnerable-people>.
- Foreign, Commonwealth Office. 2020 *Travel advice: coronavirus (COVID-19)*. GOV.UK. <https://www.gov.uk/guidance/travel-advice-novel-coronavirus>.
- Department for Education. 2020 *Closure of educational settings: information for parents and carers*. GOV.UK. <https://www.gov.uk/government/publications/closure-of-educational-settings-information-for-parents-and-carers>.
- Cabinet Office. 2020 *Full guidance on staying at home and away from others*. GOV.UK. <https://www.gov.uk/government/publications/full-guidance-on-staying-at-home-and-away-from-others>.
- Zhao S *et al.* 2020 Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: a data-driven analysis in the early phase of the outbreak. *Int. J. Infect. Dis.* **92**, 214–217. (doi:10.1016/j.ijid.2020.01.050)
- Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, Azman AS, Reich NG, Lessler J. 2020 The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann. Intern. Med.* **172**, 577–582. (doi:10.7326/M20-0504)
- UK Government. *Scientific pandemic influenza group on modelling (SPI-M)*. GOV.UK. <https://www.gov.uk/government/groups/scientific-pandemic-influenza-subgroup-on-modelling> (accessed 8 July 2020).
- UK Government. *The R number and growth rate in the UK*. GOV.UK. <https://www.gov.uk/>

- guidance/the-r-number-in-the-uk (accessed 19 August 2020).
15. UK Government. *Defence science and technology laboratory*. GOV.UK. <https://www.gov.uk/government/organisations/defence-science-and-technology-laboratory> (accessed 19 August 2020).
  16. Public Health England. *Total UK COVID-19 cases update*. <https://coronavirus.data.gov.uk/>
  17. White T. In press. *Covid-19-uk-data*. Github.
  18. Public Health Wales. *Public Health Wales statement on Novel Coronavirus (COVID-19) outbreak*. <https://covid19-phwstatement.nhs.wales/>.
  19. Scottish Government. *Coronavirus in Scotland*. <https://www.gov.scot/coronavirus-covid-19/>.
  20. HSC Public Health Agency. *COVID-19 surveillance reports*. <https://www.publichealth.hscni.net/publications/covid-19-surveillance-reports>.
  21. Public Health England. *Coronavirus (COVID-19): using data to track the virus - Public health matters*. <https://publichealthmatters.blog.gov.uk/2020/04/23/coronavirus-covid-19-using-data-to-track-the-virus/> (accessed on 8 July 2020).
  22. Hyndman R *et al.* *forecast: Forecasting functions for time series and linear models*. <https://pkg.robjhyndman.com/forecast/> (accessed on 8 July 2020).
  23. Loader C, Sun J, Lucent Technologies, Liaw A. 2020 *Locfit: local regression, likelihood and density estimation*. <https://CRAN.R-project.org/package=locfit>.
  24. Britton T, Scalia Tomba G. 2019 Estimation in emerging epidemics: biases and remedies. *J. R. Soc. Interface* **16**, 20180670. (doi:10.1098/rsif.2018.0670)
  25. Gostic KM *et al.* 2020 Practical considerations for measuring the effective reproductive number, Rt. *medRxiv*, 2020.06.18.2013485. (doi:10.1101/2020.06.18.20134858)
  26. Bi Q *et al.* 2020 Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *Lancet Infect. Dis.* **20**, 911–919. (doi:10.1016/S1473-3099(20)30287-5)
  27. Du Z, Xu X, Wu Y, Wang L, Cowling BJ, Meyers LA. 2020 Serial interval of COVID-19 among publicly reported confirmed cases. *Emerg. Infect. Dis.* **26**, 1341–1343. (doi:10.3201/eid2606.200357)
  28. Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, Hens N. 2020 Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance* **25**, 2000257. (doi:10.2807/1560-7917.ES.2020.25.17.2000257)
  29. Li Q *et al.* 2020 Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207. (doi:10.1056/NEJMoa2001316)
  30. Nishiura H, Linton NM, Akhmetzhanov AR. 2020 Serial interval of novel coronavirus (COVID-19) infections. *Int. J. Infect. Dis.* **93**, 284–286. (doi:10.1016/j.ijid.2020.02.060)
  31. Tindale LC *et al.* 2020 Evidence for transmission of COVID-19 prior to symptom onset. *eLife* **9**, e57149. (doi:10.7554/eLife.57149)
  32. Xu X, Liu X, Wang L, Ali ST, Du Z, Bosetti P, Cowling BJ, Wu Y. 2020 Household transmissions of SARS-CoV-2 in the time of unprecedented travel lockdown in China. *medRxiv* 2020.03.02.20029868. (doi:10.1101/2020.03.02.20029868)
  33. You C *et al.* 2020 Estimation of the time-varying reproduction number of COVID-19 outbreak in China. *Int. J. Hyg. Environ. Health* **228**, 113555. (doi:10.1016/j.ijheh.2020.113555)
  34. Zhang J *et al.* 2020 Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. *Lancet Infect. Dis.* **20**, 793–802. (doi:10.1016/S1473-3099(20)30230-9)
  35. Zhao S *et al.* 2020 Estimating the serial interval of the novel coronavirus disease (COVID-19): a statistical analysis using the public data in Hong Kong from January 16 to February 15, 2020. *Front. Phys.* 2020. (doi:10.3389/fphys.2020.00347)
  36. Challen R, Brooks-Pollock E, Tsaneva-Atanasova K, Danon L. 2020 Impact of uncertainty in serial interval, generation interval, incubation period and delayed observations in estimating the reproduction number for COVID 19. *Stat. Methods Med. Res.* Pending.
  37. Cori A. In press. Estimate time varying reproduction numbers from epidemic curves [R package EpiEstim version 2.2-1].
  38. Cori A, Ferguson NM, Fraser C, Cauchemez S. 2013 A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.* **178**, 1505–1512. (doi:10.1093/aje/kwt133)
  39. Thompson RN *et al.* 2019 Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics* **29**, 100356. (doi:10.1016/j.epidem.2019.100356)
  40. Challen R. 2020 *Terminological/uk-covid-datatools: Zenodo initial release*. Zenodo. (doi:10.5281/zenodo.4271338)
  41. Pybus O *et al.* 2020 *Preliminary analysis of SARS-CoV-2 importation & establishment of UK transmission lineages*. <https://virological.org/t/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/507/2> (accessed on 14 August 2020).
  42. Leclerc QJ *et al.* 2020 Analysis of temporal trends in potential COVID-19 cases reported through NHS pathways England. *Sci. Rep.* **11**, 7106. (doi:10.1038/s41598-021-86266-3)

# **7. Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study**

Status: published in BMJ

Date: 25th Feb 2021.

Candidate's contribution: RC and LDanon designed the study. All authors conceived the article. RC analysed the data and wrote the initial draft of the manuscript. KTA, EB-P, LDanon, JMR, and LDyson commented on, and revised, the manuscript. All authors read and approved the final manuscript. RC is the guarantor.

All the analysis in this paper was performed by me and findings validated by other team members.



FAST TRACK

# Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study

Robert Challen,<sup>1,2,3</sup> Ellen Brooks-Pollock,<sup>3,4,5</sup> Jonathan M Read,<sup>3,6</sup> Louise Dyson,<sup>3,7</sup> Krasimira Tsaneva-Atanasova,<sup>1,8</sup> Leon Danon<sup>3,5,8,9</sup>

For numbered affiliations see end of the article.

Correspondence to: R Challen rc538@exeter.ac.uk (or @rjchallen on Twitter: ORCID 0000-0002-5504-7768)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2021;372:n579 <http://dx.doi.org/10.1136/bmj.n579>

Accepted: 25 February 2021

## ABSTRACT

### OBJECTIVE

To establish whether there is any change in mortality from infection with a new variant of SARS-CoV-2, designated a variant of concern (VOC-202012/1) in December 2020, compared with circulating SARS-CoV-2 variants.

### DESIGN

Matched cohort study.

### SETTING

Community based (pillar 2) covid-19 testing centres in the UK using the TaqPath assay (a proxy measure of VOC-202012/1 infection).

### PARTICIPANTS

54 906 matched pairs of participants who tested positive for SARS-CoV-2 in pillar 2 between 1 October 2020 and 29 January 2021, followed-up until 12 February 2021. Participants were matched on age, sex, ethnicity, index of multiple deprivation, lower tier local authority region, and sample date of positive specimens, and differed only by detectability of the spike protein gene using the TaqPath assay.

### MAIN OUTCOME MEASURE

Death within 28 days of the first positive SARS-CoV-2 test result.

### RESULTS

The mortality hazard ratio associated with infection with VOC-202012/1 compared with infection with previously circulating variants was 1.64 (95% confidence interval 1.32 to 2.04) in patients who tested positive for covid-19 in the community. In this comparatively low risk group, this represents an increase in deaths from 2.5 to 4.1 per 1000 detected cases.

## CONCLUSIONS

The probability that the risk of mortality is increased by infection with VOC-202012/01 is high. If this finding is generalisable to other populations, infection with VOC-202012/1 has the potential to cause substantial additional mortality compared with previously circulating variants. Healthcare capacity planning and national and international control policies are all impacted by this finding, with increased mortality lending weight to the argument that further coordinated and stringent measures are justified to reduce deaths from SARS-CoV-2.

## Introduction

A new lineage of the SARS-CoV-2 virus (named B.1.1.7) was identified from genomic sequencing of samples from patients with covid-19 in the south east of England in early October 2020. In December 2020, Public Health England identified this virus as a variant of concern (VOC-202012/1).<sup>1</sup> During December this new variant spread from the south east to London and the rest of the UK, with three quarters of infections being attributable to the new variant by 31 December 2020.<sup>2</sup> The UK implemented a second national lockdown (5 November to 2 December 2020), which coincided with the relative growth of VOC-202012/1. After the lockdown, additional control measures were implemented as the increased rate of spread of the new variant became apparent and was made public.<sup>3</sup> International restrictions on travel from the UK quickly followed, in particular to France and to the rest of Europe late in December 2020 to curb spread of the new variant to other countries, despite evidence that it was already present outside the UK. Since then, the prevalence of VOC-202012/1 has been observed to be increasing in both Europe and the US.<sup>4-6</sup>

Multiplex target polymerase chain reaction (PCR) tests used in parts of the UK national testing system can distinguish VOC-202012/1 from other SARS-CoV-2 variants. Testing using the Thermo TaqPath system in the UK has shown a close correlation between VOC-202012/1 cases confirmed by genomic sequencing and TaqPath PCR results where the spike protein gene PCR target has not been detected but other PCR targets (N gene and ORF1ab gene) have been detected.<sup>2 7 9</sup> Such a result is referred to as S gene negative, or S gene target failure, and has a strong association to infection with the B.1.1.7 variant in the UK. S gene negative results have subsequently been used as a proxy to track the progression of this variant in the UK.<sup>2 7-9</sup> This association is not necessarily as strong in other countries as variants there can also produce S gene negative results.

## WHAT IS ALREADY KNOWN ON THIS TOPIC

The SARS-CoV-2 variant of concern 202012/1, first detected in the south east of England in autumn 2020, is more transmissible than previously circulating variants

The emergence of this variant coincided with high hospital occupancy, which is known to increase mortality

Before this study, unbiased estimates of the mortality of the variant of concern were not available

## WHAT THIS STUDY ADDS

Individuals infected with the variant of concern, identified at UK community test centres, were between 32% and 104% (central estimate 64%) more likely to die than equivalent individuals infected with previously circulating variants

The absolute risk of death in this largely unvaccinated population remains low, but clinicians and public health officials should be aware that a higher mortality rate is likely even if practice remains unchanged

Sequencing of VOC-202012/1 revealed 14 genetic mutations, eight of which occurred in parts of the genome that code for the spike protein responsible for cell binding,<sup>10</sup> and which impairs detection of the S gene. These mutations seem to have imparted a phenotypic change to the cell binding mechanism,<sup>27-911</sup> with the potential for increased infectivity.<sup>12 13</sup> The impact of the change on clinical presentation, patient outcome, and mortality remains poorly understood.

We used linked data from syndromic community testing and death records to assess whether the new SARS-CoV-2 variant is associated with a different risk of mortality compared with previously circulating variants.

### Methods

The study primarily set out to determine if mortality was different in patients testing positive for SARS-CoV-2 with PCR test results compatible with those for VOC-202012/1 compared with other variants. This objective was problematic because during the period under study rates of covid-19 cases in the UK increased steeply, putting hospital services under strain, which in turn affected mortality<sup>14</sup> and potentially biased observations of mortality.

We conducted a matched cohort study. To deal with bias from the varied geographical and temporal incidence of covid-19 and its burden on hospitals we matched patients closely on time and geographical location, and we also assessed the variability of our estimates when relaxing the matching criteria.

### Inclusion criteria

People were eligible for study inclusion if they were older than 30 years and had a single positive test result for covid-19 from 1 October 2020 to 29 January 2021. We restricted our sampling to test results that reported a PCR cycle threshold value. Antigen swab tests in the UK are carried out through two routes: pillar 1 represents National Health Service testing of healthcare workers and those with a clinical need, and pillar 2 represents community testing of people with symptoms. Community based covid-19 diagnoses are generally in a younger population with less severe disease than hospital based covid-19 diagnoses, as elderly people or those with severe disease tend to present directly to hospital (see supplementary file for details). We consider only the subset of pillar 2 tests that were processed in the high throughput Lighthouse laboratories that employ the Thermo TaqPath covid-19 multiplex PCR assay, which amplifies the open reading frame 1a/b junction (ORF1ab) and the N gene and S gene of SARS-CoV-2. We included people with a single positive PCR test using the TaqPath assay and with available PCR cycle threshold values for the S, N, and ORF1ab components of SARS-CoV-2.

### Data processing

We classified SARS-CoV-2 positive test results as S gene positive (compatible with previous variants) when cycle threshold values were: S gene <30, N gene <30,

and ORF1ab gene <30. We classified test results as S gene negative (compatible with VOC-202012/1) when cycle threshold values were: S gene not detected, N gene <30, and ORF1ab gene <30. Other combinations of known cycle threshold values were classified as equivocal and excluded from further analysis.

We used a unique study identifier to link the line list of positive test result details and line list of death details, when relevant. The line list of deaths records fatalities in both hospital and community settings within 28 days of a positive covid-19 test result, and follows the PHE definition of “a death in a person with a laboratory-confirmed positive covid-19 test and who died within (equal to or less than) 28 days of the first positive specimen date.”<sup>15</sup> This list is maintained by PHE and represents the most timely and complete record of deaths due to covid-19 in England.<sup>15</sup> The deaths line list also contains some details about the timing of hospital admission in those people who died. Patients who could not be linked and were therefore uninformative for S gene status were classified as “unknown” and were also excluded; these are generally samples not processed in Lighthouse laboratories, and include hospital cases.

During the study, hospitals experienced a period of intense demand in areas with large outbreaks of VOC-202012/01, which potentially could have adversely impacted patient outcomes. To control for any systematic bias this could have introduced, we matched people with S gene positive test results to individuals with S gene negative test results (highly likely to be VOC-202012/01) with exact matches on sex, ethnicity, index of multiple deprivation, location (as lower tier local authority region of about 190 000 people), and close matches on age (five years either way), and date of specimen collection (one day either way).

Some patients who were S gene negative matched multiple people who were S gene positive and vice versa, so we sampled participants randomly within our framework to generate 50 replicates, ensuring no S gene negative or S gene positive participant was present more than once in each replicate. All analyses were conducted on each replicate as a separate sample and the results pooled by combining the  $\beta$  coefficient estimates as a mixture of normal distributions and calculating combination mean and confidence intervals numerically from the mixture distribution (see supplementary file for details).

### Statistical analysis

Participants were followed-up for 28 days after infection or until 12 February 2020, after which point we censored those with no record of death. In these data more than 50% of covid-19 related deaths were reported within three days of the date of death, and more than 95% within 14 days<sup>16</sup> (see supplementary file for details). The delay in reporting deaths for participants who were S gene negative and S gene positive are the same. The deaths line list is constructed from multiple sources and is considered to



be the gold standard list of covid-19 related mortality in England. This list will ultimately include all deaths with covid-19 mentioned on the death certificate. We compared the rates of death in our community based dataset between participants who were S gene positive with those who were S gene negative. Using a Cox proportional hazards model we calculated the hazard ratio of death given an S gene negative test result versus death given an S gene positive test result<sup>17</sup> with age (years) as a linear covariate, taking into account censoring. All analyses were performed in R (version 3.6.3).<sup>18-20</sup>

### Sensitivity analyses

We examined different inclusion criteria for sources of systematic bias. We systematically adjusted values for cycle thresholds for the S, N and ORF1ab genes, and the tolerances of our algorithm to match both inexact age and inexact specimen dates.

### Patient and public involvement

Owing to the nature of this research, no patients or members of the public were involved in the design or reporting of this study.

### Results

Overall, 941 518 patients older than 30 had a single positive TaqPath test result between 1 October 2020 and 28 January 2021 (fig 1). From these, 214 082 people were identified who matched with at least one other individual on age, date of specimen collection, sex, ethnicity, geographical location, and index of multiple deprivation, and differing only by S gene status. Sampling these pairs to ensure they represented unique people resulted in 50 replicates with an average of 54 906 S gene positive people and 54 906 S gene negative people in each replicate. Every person was followed-up for a minimum of 14 days after their first positive test result, and more than 85% of the cases were followed for the whole 28 day period (see supplementary file for further details). Of these 109 812 participants, 367 died (averaged over the 50 replicates) within 28 days of a positive covid-19 test result (0.3%) (table 1). The matching and sampling process is observed to control well for all personal and geographical variables considered (with slight mismatches owing to differences in scale from matching and reporting). When a tolerance of five years was allowed for matching age the average difference between study arms was 0.0 years, and when a tolerance of one day was allowed for matching specimen date a mean difference of 0.2 days was observed (with S gene negative specimens taken later than S positive specimens).

The subset of participants who died were generally older (mean 66.9 v 46.3 years) and a higher proportion were men, as has been reported previously.<sup>21</sup> Both cases and deaths were underrepresented in the south west and east of England—these areas had only recently used TaqPath assays and thus did not report S gene status.

Of the 54 906 participants in the S gene negative arm, an average of 227 deaths occurred compared with 141 of 54 906 in the S gene positive arm (hazard ratio 1.64, 95% confidence interval 1.32 to 2.04;  $P < 0.001$ ) over the study period (table 2). The rate of death of S gene negative and S gene positive participants diverged after 14 days (fig 2). The proportional hazards assumption of the Cox model was therefore violated as the hazard ratio was not constant over time. This was investigated further (see supplementary file), and the violation might be corrected by considering the hazard ratio in days 0 to 14 compared with days 15 to 28 of follow-up. The hazard ratio in the first period was not significantly increased, but in days 15 to 28 the hazard ratio was 2.40 (1.66 to 3.47).

The matched cohort design controls for most potential biases, including variations in hospital capacity, as it pairs patients by personal characteristics, geography, and time of testing. Other further potential biases that might be present were investigated. One possibility for bias could be a difference in the timing of presentation of S gene negative and S gene positive people for testing, with, for example, S gene positive people presenting earlier, and thus seeming to progress slower. Hospital admission data were only available for patients who ultimately died, but there was no evidence for asymmetrical delays in time from test to hospital admission (fig 3). The Office for National Statistics also investigated this and found that S gene negative patients are more likely to present earlier for testing.<sup>22</sup>

The paired cases in this study were spread over time but concentrated around the end of December 2020 and beginning of January 2021 (fig 3). As the ratio of S gene negativity to S gene positivity changed over this period, in the early stages it was comparatively difficult to match S gene negative people with S gene positive equivalents, and in the later stages it was difficult to match S gene positive people with S gene negative equivalents, with the bulk of matching occurring during the time of transition from dominance of the S gene positive variant to dominance of the S gene negative variant (see supplementary file).

Cycle threshold values for the N gene were lower in participants who were S gene negative than in those who were S gene positive, and this effect was potentiated in those who died (table 1 and fig 3). Low values for the N gene cycle threshold implied that the viral load in participants at the time of sampling was higher. The higher mortality could be associated with the higher viral load in S gene negative participants because of the intrinsic properties of the VOC202012/1 mutation. Alternatively, it could be an indication of the timing of testing, with people who were S gene negative presenting at peak infectiousness, for some as yet unknown reason. Thus, cycle threshold values for the N gene could be regarded either as an indication of bias or as a feature of S gene negative infection. If this is interpreted as a source of bias, the Cox proportional hazards model can control for the N gene cycled threshold value (table 2, second model), which for S

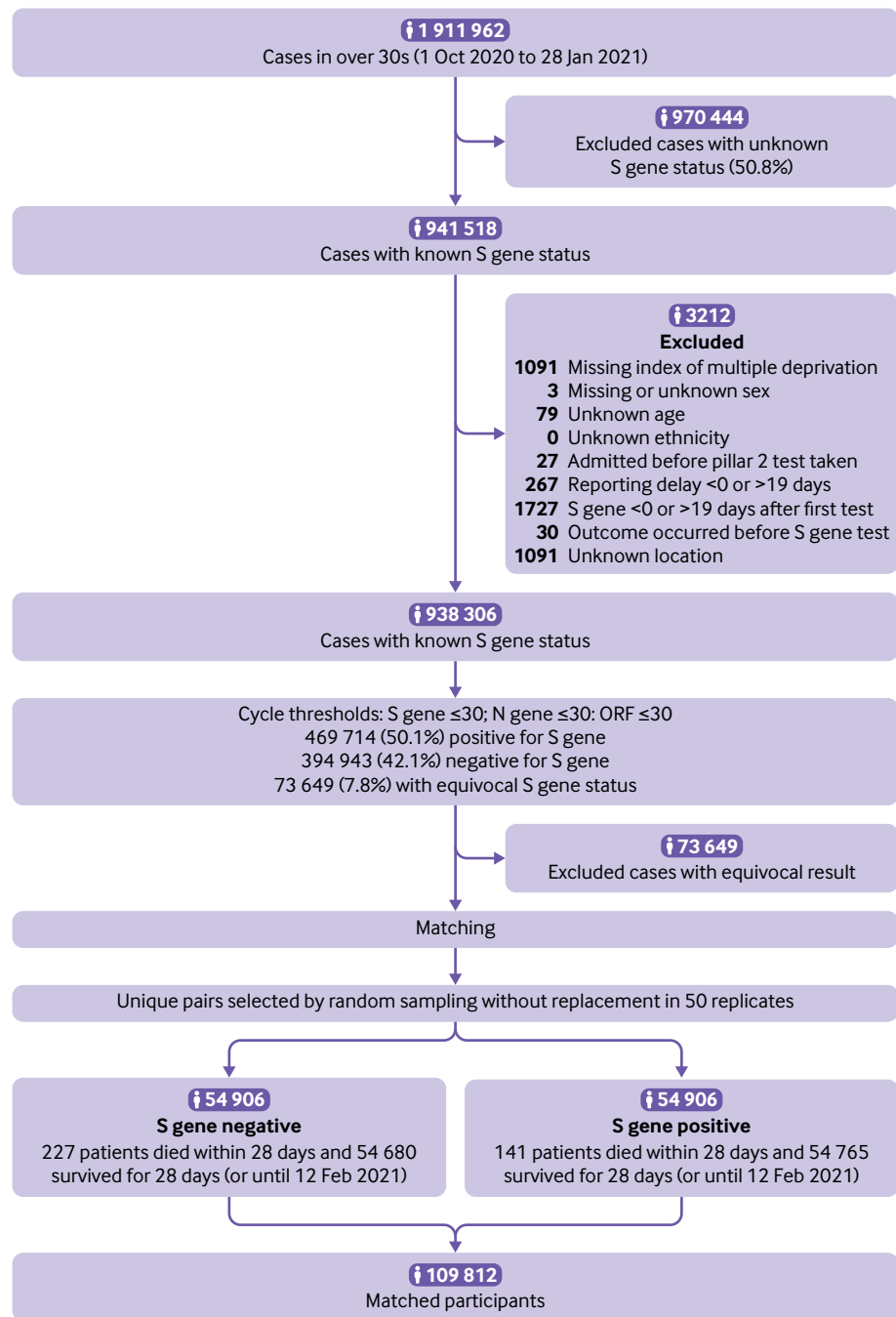


Fig 1 | Sample selection algorithm showing average figures for numbers of participants in each study arm. Matching involved random sampling to create 50 replicates. Some cases were excluded for more than one reason

gene negativity showed a hazard ratio of 1.37 (95% confidence interval 1.09 to 1.72). Even if increased viral load as a biological feature of S gene negative infection is not considered, the residual increase in hazard ratio implies a mortality effect not explained by viral load alone.

**Sensitivity analysis**

The cut-off value of cycle threshold used in definition of gene positivity mildly affects the central estimate of the hazard ratio, such that when the cycle threshold value for identifying a particular gene was reduced,

the central estimate of hazard ratios was observed to decrease (fig 4). Lower cycle threshold values were associated with a reduction in the number of certain S gene positive and S gene negative results and an increase in the number of equivocal results, which were subsequently excluded from analysis, resulting in an effective reduction in overall case numbers. Given that cycle threshold values were generally higher in patients with S gene positivity, as defined at the cycle threshold value of less than 30, further reductions in the cut-off value of the cycle threshold tend to be associated with a reclassification of S gene positive

**Table 1 | Matched S gene positive and S gene negative participants. Values are numbers (percentages) unless stated otherwise**

Characteristics	S gene positive (n=54 906)	S gene negative (n=54 906)	Death (n=367)
Mean (SD) age (years)	46.3 (11.0)	46.3 (11.0)	66.9 (14.1)
Age category:			
30-59	48 486 (88.3)	48 486 (88.3)	114 (31.0)
60-69	4973 (9.1)	4973 (9.1)	96 (26.1)
70-79	1175 (2.1)	1175 (2.1)	89 (24.2)
≥80	273 (0.5)	273 (0.5)	69 (18.8)
Ethnicity:			
White	45 698 (83.2)	45 698 (83.2)	325 (88.3)
Asian	6930 (12.6)	6930 (12.6)	38 (10.3)
Other	1167 (2.1)	1167 (2.1)	1 (0.3)
Unknown	127 (0.2)	127 (0.2)	
Afro-Caribbean	985 (1.8)	985 (1.8)	4 (1.1)
Sex:			
Women	29 378 (53.5)	29 378 (53.5)	141 (38.3)
Men	25 528 (46.5)	25 528 (46.5)	227 (61.7)
Index of multiple deprivation 10th:			
1st	5005 (9.1)	5005 (9.1)	26 (7.0)
2nd	9413 (17.1)	9413 (17.1)	93 (25.0)
3rd	7262 (13.2)	7262 (13.2)	65 (17.5)
4th	6241 (11.4)	6241 (11.4)	34 (9.1)
5th	5344 (9.7)	5344 (9.7)	36 (9.7)
6th	4402 (8.0)	4402 (8.0)	31 (8.3)
7th	4421 (8.1)	4421 (8.1)	24 (6.5)
8th	4336 (7.9)	4336 (7.9)	26 (7.0)
9th	4364 (7.9)	4364 (7.9)	20 (5.4)
10th	4123 (7.5)	4123 (7.5)	17 (4.6)
Mean (SD) N gene cycle threshold	21.3 (4.2)	19.0 (4.4)	18.3 (4.3)
Region:			
East of England	3634 (6.6)	3637 (6.6)	18 (4.9)
London	8874 (16.2)	8874 (16.2)	26 (7.0)
Midlands	10 550 (19.2)	10 563 (19.2)	88 (23.7)
North East and Yorkshire	10 733 (19.5)	10 740 (19.6)	83 (22.4)
North West	14 711 (26.8)	14 693 (26.8)	123 (33.2)
South East	5105 (9.3)	5106 (9.3)	22 (5.9)
South West	1301 (2.4)	1297 (2.4)	11 (3.0)
S gene:			
Positive	54 906 (100.0)		141 (38.3)
Negative		54 906 (100.0)	227 (61.7)
Status:			
Dead <28 days of positive covid-19 result	141 (0.3)	227 (0.4)	367 (100.0)
Survived 28 days or until 12 Feb 2020	54 765 (99.7)	54 680 (99.6)	

Participants were matched on age, ethnicity, sex, index of multiple deprivation, geography, and specimen date (not shown).

rather than S gene negative patients with more mild disease as equivocal. This could explain the small reduction in hazard ratio associated with reducing the cut-off value of the cycle threshold. A marginal, non-significant increase was observed at a cycle threshold cut-off value of 30; as most laboratories use this as a standard cycle threshold cut-off value, this was chosen as the central estimate.

When matching patients for the cohort, allowing larger mismatching led to small changes in the associated hazard ratio estimate. The effect of mismatching on age between S gene negative and S gene positive participants did not create a systematic bias, and the mean age difference between both study arms was less than 0.005 years (fig 4). Age was observed to be a strong predictor of mortality in covid-19, so some potential bias might be expected; this is controlled for by including age as a covariate in the calculation of hazard ratios (table 2).

A dilution of hazard ratios was observed when a greater degree of mismatching was allowed between sample dates in S gene positive patients matched to S gene negative patients (fig 4). Because of the change in prevalence over the study period from predominantly S gene positive to predominantly S gene negative, increasing the degree of mismatching of sample date was associated with a systematic pairwise bias in the dates of the original positive test result (fig 4), with S gene negative patients generally being identified after S gene positive patients. Given that over the study period the number of cases was observed to exponentially increase, this could have affected the overall result as hospital capacity generally worsened during the study period. To avoid this the study minimised the sample date tolerance, trading off the reduction in bias against the variance introduced by the reduced number of cases resulting from tight matching criteria.

**Table 2 | Risk of death in S gene negative compared with S gene positive (reference category) participants**

Model, predictor, value	Hazard ratio (95% CI)	P value
<b>S gene+age</b>		
S gene status:		
Positive (ref)	—	—
Negative	1.64 (1.32 to 2.04)	<0.001
Age (per decade)	3.55 (3.28 to 3.84)	<0.001
<b>S gene+N gene cycle threshold+age</b>		
S gene status:		
Positive (ref)	—	—
Negative	1.37 (1.09 to 1.72)	0.004
Age (per decade)	3.51 (3.24 to 3.80)	<0.001
N gene cycle threshold (per 10 units)	0.50 (0.39 to 0.65)	<0.001

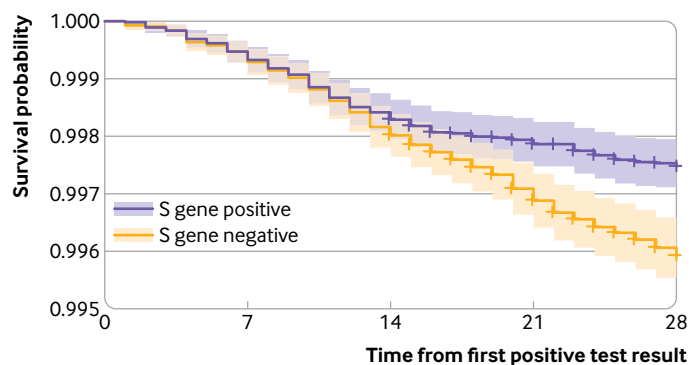
Hazard ratios >1 are indicative of an increased rate of death in people with infections compatible with VOC-202012/01. In the first model the S gene status is assessed as an indicator with age as a covariate, in the second model variability is included in the N gene cycle threshold value measured in the original specimen as a continuous predictor.

Despite the differences between the combinations investigated, all studies reported a statistically significant increase in the risk of mortality associated with VOC-202012/1, suggesting a real effect, and most central estimates were within the range of 1.5 to 1.7. The supplementary file discusses other potential covariates.

### Discussion

Infections with the new variant VOC-202012/1 (as measured by S gene negativity) were associated with an increased risk of death ( $P<0.001$ ) in people testing positive for covid-19 in the community. The increased hazard ratio between 1.32 and 2.04, higher than for other variants, translates to a 32% to 104% increased risk of death, with the most probable hazard ratio estimate of 1.64, or a 64% increased risk of death. The absolute risk of death in this group of community identified participants, however, remains relatively low, increasing from 2.5 to 4.1 deaths per 1000 cases.

We controlled for several biases when using a matched cohort approach. In particular, mortality is affected by how many patients require intensive care in a hospital setting<sup>14</sup>; increasing numbers of patients in the study period (1 October 2020 to 12 February 2021), compounded by staff absenteeism as a result of



**Fig 2 | Kaplan-Meier survival curve for S gene positive (previously circulating variants) and S gene negative (new variant VOC-202012/1) participants in the UK. The y axis has been truncated as mortality was low in both groups**

covid-19 infection or isolation because of contact with infected people, has placed intense strain on hospital services and a reduction in the staff to patient ratio. Staff absenteeism might have affected mortality and is a potential source of bias. We controlled for this by matching patients on administrative region and time of positive test result (within one day), which constrains pairs to receive care at the same place and time, and we suggest at a similar level of care. Although age related mortality is controlled for by matching on age (within five years), it is also controlled for by using the Cox proportional hazards model.

As this was a community based study, we do not have information on the S gene status of patients in hospitals. The community based testing (pillar 2) in this dataset covered a younger age group and hence represented less severe disease than patients detected through hospital based testing (pillar 1). Death remains a comparatively rare outcome in patients detected in the community compared with identified in-hospital deaths. Our study only includes about 8% of the total deaths that occurred during the study period. Of all coronavirus deaths, about 26% occurred in those who were identified in the community, and data on S gene status was only available for 30%.<sup>23</sup> Whether the increase in mortality from community based testing is also observed in elderly patients or in patients admitted to hospital remains to be seen.

We cannot exclude a selection bias. Community testing is largely self-selected, or driven by contact tracing. A potential bias remains if a higher proportion of patients with S gene negative infections without symptoms were undetected than patients with S gene positive infections. In this event, patients infected with VOC-202012/1 might be at a more advanced stage of disease when identified and have a higher apparent mortality. This could be consistent with the lower N gene cycle threshold values observed in S gene negative participants. Our analysis, or any retrospective study based on patients with symptoms, would not be able to detect this; however, early survey data suggest that people with S gene negative infections are, if anything, more likely to present for testing.<sup>22</sup> Dealing with this potential bias requires a study design capable of detecting asymptomatic infections in participants who are negative or positive for the S gene.

Some of the increased risk could be explained by comorbidities. Information was not available about comorbid conditions in the data we analysed, although this would be partly controlled for by matching on age, ethnicity, and index of multiple deprivation. Currently there is no evidence of a mechanistic reason why people with certain comorbidities would be infected with one variant and not another. It is possible, however, that people with certain comorbidities are at a higher risk of infection with VOC-202012/1 and have a higher mortality rate. This would tend to reduce the hazard ratio attributable to VOC-202012/1 alone.

Our preliminary estimate of the hazard ratio was 1.91 (95% confidence interval 1.35 to 2.71), which is marginally higher than the estimate presented here

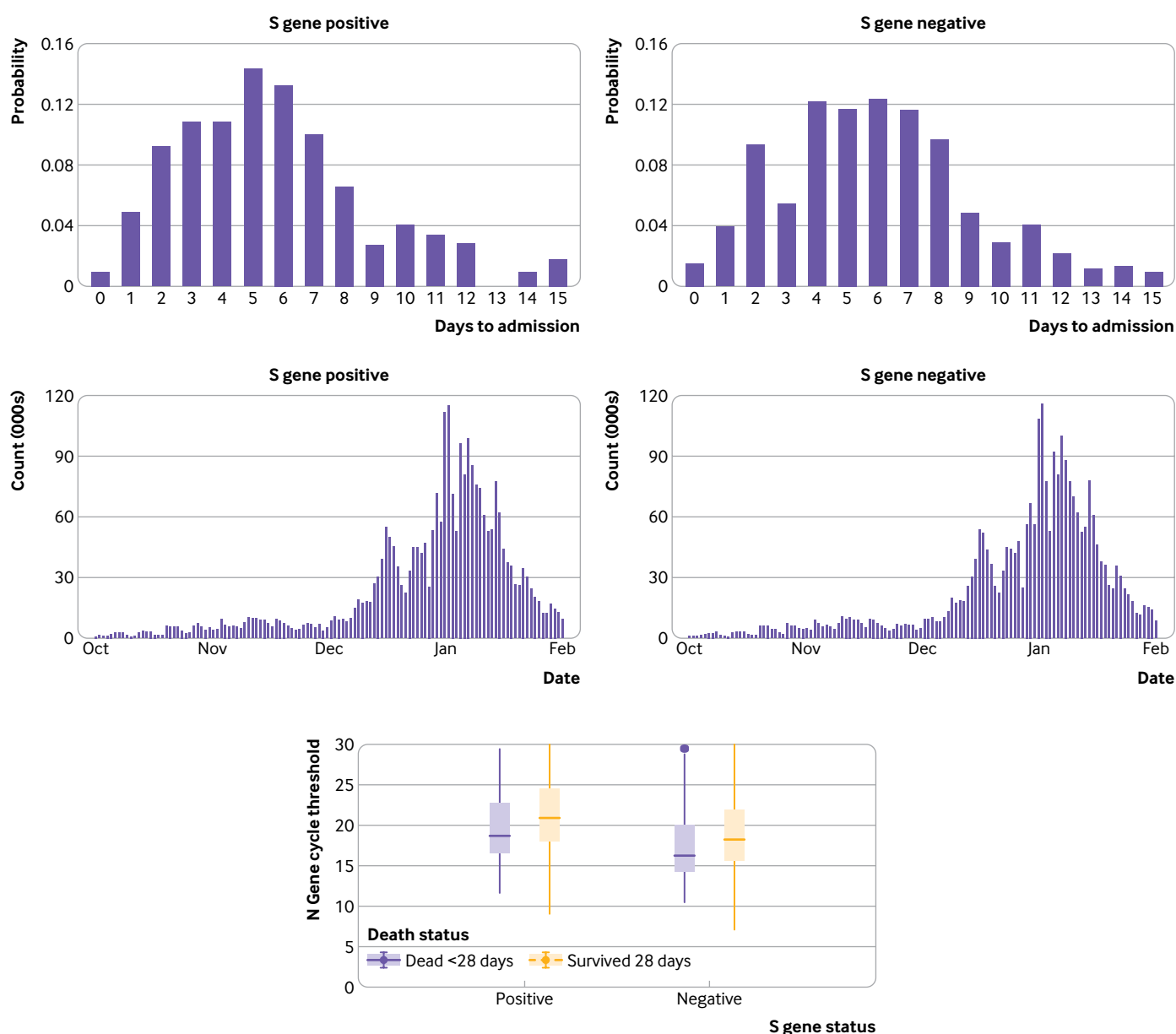


Fig 3 | Investigation of biases in S gene positive and S gene negative study arms

with compatible uncertainty.<sup>23,24</sup> This was based on 94 deaths in S gene negative patients and 49 deaths in S gene positive patients in 66 208 less strictly matched pairs, with a shorter study period, and limited follow-up. As the new variant outbreak has unfolded and more data have become available, we have been able to obtain more accurate central estimates by narrowing the tolerance for mismatches, extending the study period and increasing the proportion of patients with complete follow-up. The design of this study is well suited to determining, in an unbiased manner, whether the risk of death has increased, although we studied a comparatively small number of patients. Other study designs, involving the use of unpaired samples, might be better able to quantify the absolute increase in risk, albeit with more potential for bias.<sup>25</sup>

Other recent studies produced similar estimates of the increased hazard ratio. Although these studies use the same community based testing data, they had different study and analysis designs. The preliminary results of these studies were compatible point estimates of the mortality hazard ratio (1.3 to 1.65), and the confidence intervals of these studies overlap with those described here.<sup>23</sup> As with our work, these other estimates are being continuously re-evaluated as more data are acquired; and in subsequent updates some of these have been revised upwards.<sup>26</sup>

### Conclusions

The variant of concern, in addition to being more transmissible, seems to be more lethal. We expect this to be associated with changes in its phenotypic

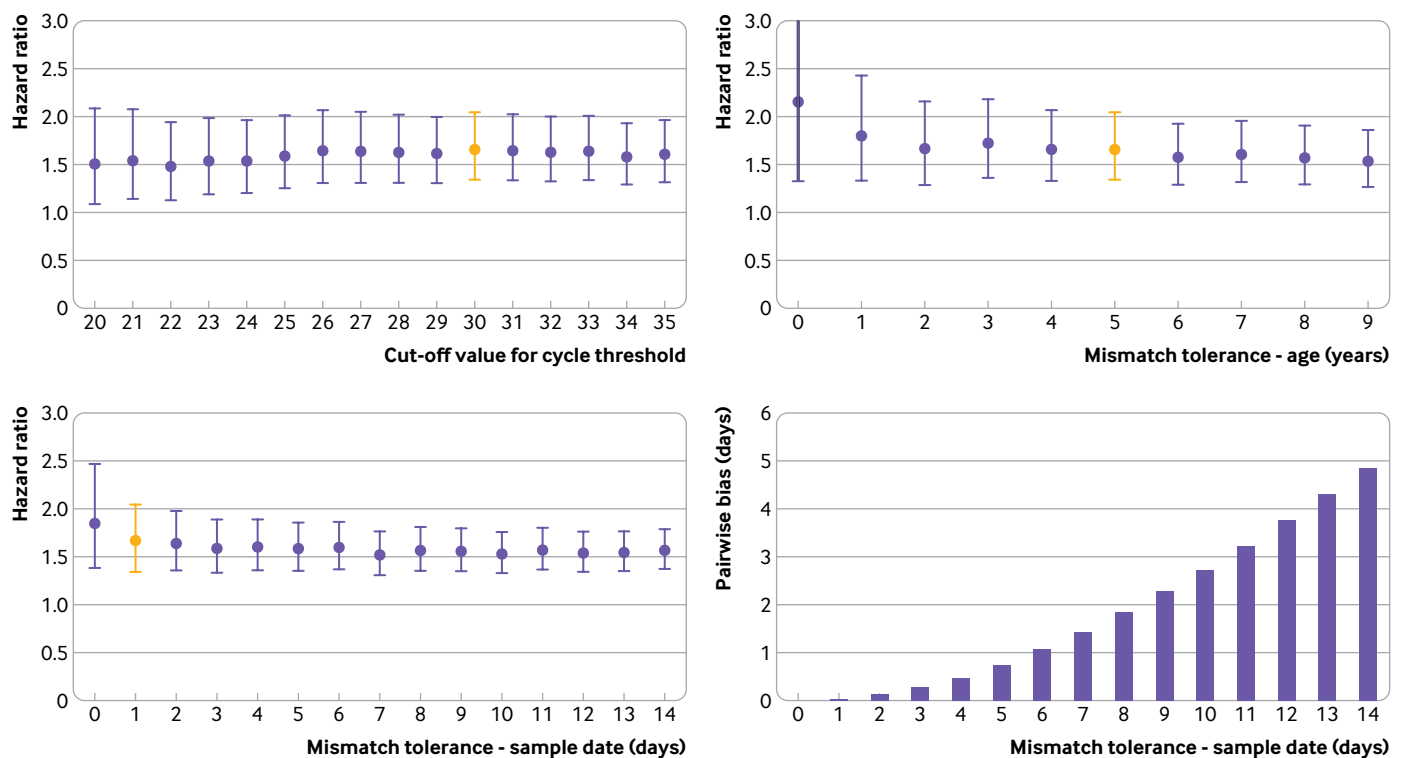


Fig 4 | Sensitivity analyses. Red bar indicates default assumptions (cycle threshold <30; age tolerance  $\pm 5$  years; sample date tolerance  $\pm 1$  day) from rest of study

properties because of multiple genetic mutations,<sup>27</sup> and we see no reason why this finding would be specific to the UK. This development, borne out in epidemiological analyses, implies that the rate of patients with serious infection requiring hospital attention will increase. At the time of writing (15 February 2021) the national lockdown appears to be effective at reducing the transmission rate of SARS-CoV-2 in the UK, but proliferation of the new variant has made it more difficult to control the covid-19 outbreak. The resulting number of deaths will scale linearly with the proportion of people infected with the new variant. Other analyses have indicated that the new variant is also associated with increased transmissibility, which would lead to a potentially exponential increase in the resulting number of deaths.<sup>12</sup> Clinicians at the front line should be aware that a higher mortality rate is likely even if quality of practice remains unchanged. This has broader implications for any vaccination allocation policy designed to reduce mortality in the late middle age groups, typical of the community identified patients in this dataset.

The question remains whether excess mortality due to VOC-202012/1 will be observed in other population groups, particularly elderly people, care home residents, and those with other comorbidities who generally present directly to hospital as an emergency. Hospital based studies require a mechanism to distinguish emerging variants from previously circulating variants, currently only done through genotyping. Owing to the effort involved, the proportion of genotyped samples representing patients

admitted to hospital remains low, and we recommend that PCR tests that specifically target VOC-202012/1 mutations should be more widely used.

Moreover, the emergence of VOC-202012/1 and its mutations (including E484K), combined with other variants of concern, including those identified in Brazil and South Africa,<sup>28</sup> highlights the capacity of SARS-CoV-2 to rapidly evolve new phenotypic variants, with mutants that evade vaccines being a real possibility.<sup>29</sup> Our study has helped to characterise the clinical presentation and outcome of one new variant, but given sufficient amounts of informative data our findings can be generalisable to other variants. Assessment of the clinical outcomes of multiple circulating phenotypic variants, however, requires scalable technology that is capable of identifying substantial numbers of patients infected with emerging variants (eg, broad PCR assay panels targeting variant foci<sup>30</sup>) and robust collection of outcome data.

In this study we controlled for the effect of time, geographical location, age, sex, ethnicity, and deprivation, but these are important factors to understand if future outcomes are to improve. Future work on the relative impact of these might allow for better targeting of resource allocation,<sup>31</sup> vaccine distribution strategies, and relaxation of restrictions.

#### AUTHOR AFFILIATIONS

<sup>1</sup>College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

<sup>2</sup>Somerset NHS Foundation Trust, Taunton, UK

<sup>3</sup>Joint Universities Pandemic and Epidemiological Research (JUNIPER consortium)

<sup>4</sup>University of Bristol, Bristol Veterinary School, Langford, Bristol, UK

<sup>5</sup>Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK

<sup>6</sup>Lancaster Medical School, Lancaster University, Bailrigg, Lancaster, UK

<sup>7</sup>The Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research, School of Life Sciences and Mathematics Institute, University of Warwick, Coventry, UK

<sup>8</sup>The Alan Turing Institute, British Library, London, UK

<sup>9</sup>Department of Engineering Mathematics, University of Bristol, Bristol, UK

We thank David Spiegelhalter for comments on an early draft and Nick Gent at Public Health England for providing access to the data.

Public Health England (PHE) collected data in a centralised database, which included details on the type of test and results. PHE provided anonymised data to contributors of the Scientific Pandemic Influenza Group on Modelling (SPI-M) as part of the covid-19 response under a data sharing agreement between PHE and the authors' institutions.

**Contributors:** RC and LDanon designed the study. All authors conceived the article. RC analysed the data and wrote the initial draft of the manuscript. KTA, EB-P, LDanon, JMR, and LDyson commented on, and revised, the manuscript. All authors read and approved the final manuscript. RC is the guarantor. The views presented here are those of the authors and should not be attributed to Somerset Foundation NHS Trust or Global Digital Exemplar. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

**Funding:** RC and KTA were supported by the Engineering and Physical Sciences Research Council (EPSRC; grants EP/N014391/1, EP/T017856/1) and NHS England, Global Digital Exemplar programme. LDanon and KTA were supported by The Alan Turing Institute (EPSRC grant EP/N510129/1). LDanon, RC, and EBP are supported by the Medical Research Council (MRC; MC/PC/19067). JMR is supported by EPSRC (EP/N014499/1) and MRC (MR/S004793/1, MR/V028456/1). EBP was partly supported by the National Institute for Health Research Health Protection Research Unit in Behavioural Science and Evaluation at University of Bristol, in partnership with Public Health England. LDanon, EBP, JMR, and LDyson are further supported by MRC (MR/V038613/1), and LDanon by EPSRC EP/V051555/1. LDyson also received support through the MRC through the covid-19 rapid response rolling call (grant No MR/V009761/1).

**Competing interests:** All authors have completed the ICMJE uniform disclosure form at [www.icmje.org/doi\\_disclosure.pdf](http://www.icmje.org/doi_disclosure.pdf) and declare: support from the Engineering and Physical Sciences Research Council, NHS England, Global Digital Exemplar programme, Alan Turing Institute, Medical Research Council, and National Institute for Health Research Health Protection Research Unit in Behavioural Science and Evaluation, in partnership with Public Health England; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

**Ethical approval:** The data were supplied from the Second Generation Surveillance System (SGSS) database and death reports after anonymisation under strict data protection protocols agreed between the University of Exeter and Public Health England. The ethics of the use of these data for these purposes was agreed by Public Health England with the UK government SPI-M(O)/SAGE committees. The research was assessed as not needing NHS research ethics committee review.

**Data sharing:** The data analysis code is available at doi:10.5281/zenodo.4543510. Data are available to members of universities who have data sharing agreements in place with Public Health England.

The lead author (RC) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

**Dissemination to participants and related patient and public communities:** Direct dissemination to study participants is not possible. Preprint versions of this analysis are available on *MedRxiv* ([www.medrxiv.org/content/10.1101/2021.02.09.21250937v2](http://www.medrxiv.org/content/10.1101/2021.02.09.21250937v2)).

**Provenance and peer review:** Not commissioned; externally peer reviewed.

This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which

permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: <http://creativecommons.org/licenses/by/4.0/>.

- 1 Wise J. Covid-19: New coronavirus variant is identified in UK. *BMJ* 2020;371:m4857. doi:10.1136/bmj.m4857
- 2 Investigation of novel SARS-CoV-2 variant: Variant of Concern 202012/01 - Technical briefing 4. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/959359/Variant\\_of\\_Concern\\_VOC\\_202012\\_01\\_Technical\\_Briefing\\_4.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959359/Variant_of_Concern_VOC_202012_01_Technical_Briefing_4.pdf)
- 3 Prime Minister announces Tier 4: 'Stay At Home' Alert Level in response to new COVID variant. GOV.UK. [www.gov.uk/government/news/prime-minister-announces-tier-4-stay-at-home-alert-level-in-response-to-new-covid-variant](http://www.gov.uk/government/news/prime-minister-announces-tier-4-stay-at-home-alert-level-in-response-to-new-covid-variant) (accessed 30 Jan 2021).
- 4 Risk related to the spread of new SARS-CoV-2 variants of concern in the EU/EEA – first update. *Eur. Cent. Dis. Prev. Control.* 2021. [www.ecdc.europa.eu/en/publications-data/covid-19-risk-assessment-spread-new-variants-concern-eueea-first-update](http://www.ecdc.europa.eu/en/publications-data/covid-19-risk-assessment-spread-new-variants-concern-eueea-first-update) (accessed 30 Jan 2021).
- 5 CDC. US COVID-19 Cases Caused by Variants. Centers for Disease Control and Prevention. 2020. [www.cdc.gov/coronavirus/2019-ncov/transmission/variant-cases.html](http://www.cdc.gov/coronavirus/2019-ncov/transmission/variant-cases.html) (accessed 30 Jan 2021).
- 6 The Helix COVID-19 Surveillance Dashboard. Helix. [www.helix.com/pages/helix-covid-19-surveillance-dashboard](http://www.helix.com/pages/helix-covid-19-surveillance-dashboard) (accessed 14 Feb 2021).
- 7 Investigation of novel SARS-CoV-2 variant: Variant of Concern 202012/01 - Technical briefing 1. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/947048/Technical\\_Briefing\\_VOC\\_SH\\_NJL2\\_SH2.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/947048/Technical_Briefing_VOC_SH_NJL2_SH2.pdf) (accessed 19 Jan 2021).
- 8 Investigation of novel SARS-CoV-2 variant: Variant of Concern 202012/01 - Technical briefing 2. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/949639/Technical\\_Briefing\\_VOC202012-2\\_Briefing\\_2\\_FINAL.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/949639/Technical_Briefing_VOC202012-2_Briefing_2_FINAL.pdf) (accessed 19 Jan 2021).
- 9 Investigation of novel SARS-CoV-2 variant: Variant of Concern 202012/01 - Technical briefing 3. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/959360/Variant\\_of\\_Concern\\_VOC\\_202012\\_01\\_Technical\\_Briefing\\_3.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959360/Variant_of_Concern_VOC_202012_01_Technical_Briefing_3.pdf)
- 10 Rambaut A. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology. *Virological*. 2020. <https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563> (accessed 30 Jan 2021).
- 11 Starr TN, Greaney AJ, Hilton SK, et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* 2020;182:1295-1310.e20. doi:10.1016/j.cell.2020.08.012
- 12 Davies NG, Barnard RC, Jarvis CI, et al. Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England. *medRxiv* 2020;12.24.20248822. doi:10.1101/2020.12.24.20248822
- 13 du Plessis L, McCrone JT, Zarebski AE, et al. COVID-19 Genomics UK (COG-UK) Consortium. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* 2021;371:708-12. doi:10.1126/science.abc2946
- 14 Wilde H, Mellan T, Hawryluk I, et al. The association between mechanical ventilator availability and mortality risk in intensive care patients with COVID-19: A national retrospective cohort study. *medRxiv* 2021;01.11.21249461. doi:10.1101/2021.01.11.21249461
- 15 PHE reporting of COVID-19 deaths: technical summary – 12 August 2020. 2020;15. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/916035/RA\\_Technical\\_Summary\\_-\\_PHE\\_Data\\_Series\\_COVID\\_19\\_Deaths\\_20200812.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/916035/RA_Technical_Summary_-_PHE_Data_Series_COVID_19_Deaths_20200812.pdf)
- 16 Seaman S, Samartisdid P, Kall M, et al. Nowcasting CoVID-19 Deaths in England by Age and Region. *medRxiv* 2020;09.15.20194209. doi:10.1101/2020.09.15.20194209
- 17 Cox DR, Oakes D. *Analysis of survival data*. CRC press, 1984.
- 18 R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2017.
- 19 Therneau TM. A package for survival analysis in r. 2020. <https://CRAN.R-project.org/package=survival>
- 20 Terry M, Therneau, Patricia M, Grambsch. *Modeling survival data: Extending the Cox model*. Springer, 2000.
- 21 Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* 2020;584:430-6. doi:10.1038/s41586-020-2521-4
- 22 Steel K, Davies B. Coronavirus (COVID-19) Infection Survey: characteristics of people testing positive for COVID-19 in England, 27

- January 2021. Office for National Statistics [www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsinthecommunityinengland/characteristicsofpeopletestingpositiveforcovid19inengland27january2021](http://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsinthecommunityinengland/characteristicsofpeopletestingpositiveforcovid19inengland27january2021)
- 23 Horby P. NERVTAG note on B.1.1.7 severity. NERVTAG [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/955239/NERVTAG\\_paper\\_on\\_variant\\_of\\_concern\\_\\_VOC\\_\\_B.1.1.7.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/955239/NERVTAG_paper_on_variant_of_concern__VOC__B.1.1.7.pdf) (accessed 31 Jan 2021).
- 24 Iacobucci G. Covid-19: New UK variant may be linked to increased death rate, early data indicate. *BMJ* 2021;372:n230. doi:10.1136/bmj.n230
- 25 Brazauskas R, Logan BR. Observational Studies: Matching or Regression? *Biol Blood Marrow Transplant* 2016;22:557-63. doi:10.1016/j.bbmt.2015.12.005
- 26 NERVTAG. Update note on B.1.1.7 severity, 11 February 2021. GOV.UK. [www.gov.uk/government/publications/nervtag-update-note-on-b117-severity-11-february-2021](http://www.gov.uk/government/publications/nervtag-update-note-on-b117-severity-11-february-2021) (accessed 14 Feb 2021).
- 27 Volz E, Mishra S, Chand M, et al. Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. *medRxiv* 2021;2020.12.30.20249034. doi:10.1101/2020.12.30.20249034
- 28 Mahase E. Covid-19: What new variants are emerging and how are they being investigated? *BMJ* 2021;372:n158. doi:10.1136/bmj.n158
- 29 Wibmer CK, Ayres F, Hermanus T, et al. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *bioRxiv* 2021;2021.01.18.427166. doi:10.1101/2021.01.18.427166
- 30 Harper H, Burrige AJ, Winfield M, et al. Detecting SARS-CoV-2 variants with SNP genotyping. *bioRxiv* 2020;2020.11.18.388140. doi:10.1101/2020.11.18.388140
- 31 Lacasa L, Challen R, Brooks-Pollock E, Danon L. A flexible method for optimising sharing of healthcare resources and demand in the context of the COVID-19 pandemic. *PLoS One* 2020;15:e0241027. doi:10.1371/journal.pone.0241027

**Supplementary information:** Additional material supplied by authors



# Supplementary material: Increased hazard of mortality in cases compatible with SARS-CoV-2 variant of concern 202012/1 - a matched cohort study

Robert Challen<sup>1,2,3</sup>; Ellen Brooks-Pollock<sup>3,4,5</sup>; Jonathan M Read<sup>3,6</sup>; Louise Dyson<sup>3,7</sup>; Krasimira Tsaneva-Atanasova<sup>1,8</sup>; Leon Danon<sup>3,5,8,9</sup>

- 1) College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, Devon, UK.
- 2) Somerset NHS Foundation Trust, Taunton, Somerset, UK.
- 3) Joint Universities Pandemic and Epidemiological Research (JUNIPER) consortium.
- 4) University of Bristol, Bristol Veterinary School, Langford, Bristol, UK.
- 5) Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK.
- 6) Lancaster Medical School, Lancaster University, Bailrigg, Lancaster, UK.
- 7) The Zeeman Institute for Systems Biology & Infectious Disease Epidemiology Research, School of Life Sciences and Mathematics Institute, University of Warwick, Coventry, UK.
- 8) The Alan Turing Institute, British Library, 96 Euston Rd, London, UK.
- 9) Department of Engineering Mathematics, University of Bristol, UK.

<b>Assessment of potential data biases</b>	<b>1</b>
<b>Case Matching</b>	<b>3</b>
<b>Combining estimates</b>	<b>4</b>
<b>Additional Proportional Hazards Models</b>	<b>5</b>
<b>Proportional hazards assumption</b>	<b>7</b>
<b>Supplementary material references</b>	<b>8</b>

## Assessment of potential data biases

In our analysis the percentage of COVID-19 cases that die within 28 days of a positive test appears lower than would be expected from a simple Infection Fatality Ratio (IFR) calculation of all COVID-19 cases. One reason for this is that the community-testing dataset we are using (Pillar 2) does not include testing in healthcare settings, and so represents a younger than average population. Figure S1 shows the distribution of S-gene positive, S-gene negative and equivocal S-gene cases in Pillar 2 by single year of age, compared to the distribution of people in the general population (grey). When we consider that elderly patients are less likely to be tested in the community, prior to admission to hospital, we can expect that mortality in the younger community cohort will be lower.

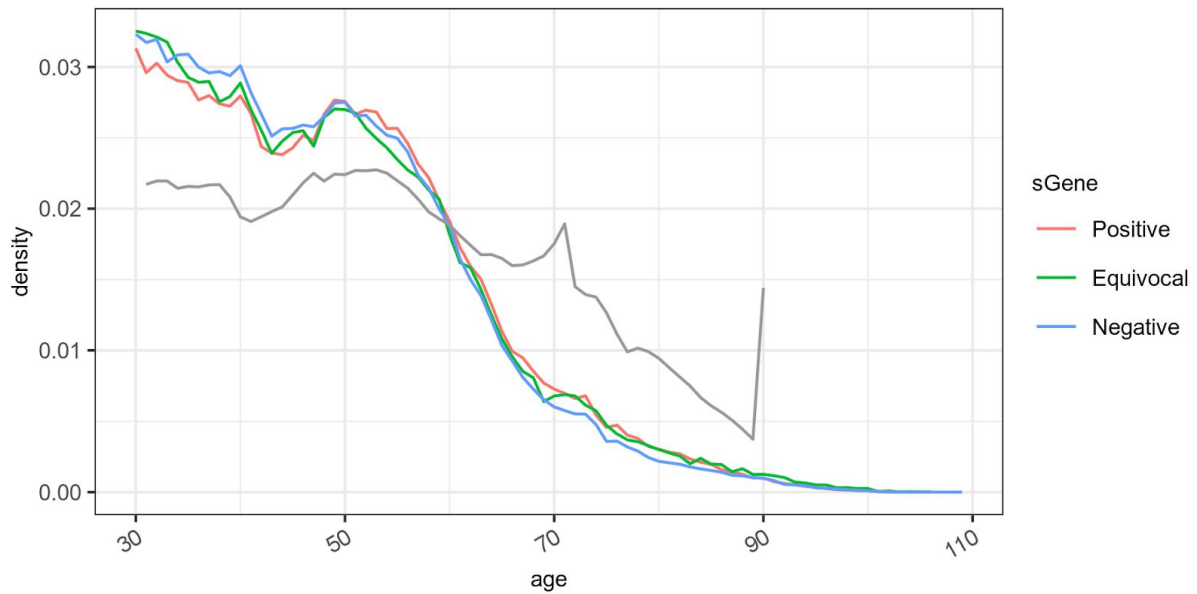


Figure S1 - Distribution of ages for S-gene positive (red), negative (blue) and equivocal (green) cases from Pillar 2 (community testing) data, compared to the general population (grey). In this figure we take unmatched COVID-19 cases for which S-gene data is available (941,518 cases) and compare the age distribution to that in the ONS 2019 mid year estimates for the general population (which has an upper age category of 90+, and now lags the testing data by 3 years).

A potential source of bias which could influence the estimation of the hazard ratio would occur if there was either a differential loss or a differential delay to reporting of outcome in either of the 2 arms of our matched cohort. In Figure S2 we investigate the delay in reporting of death and find the majority of all deaths are reported within 14 days, with minimal differences between S-gene positive and S-gene negative, and conclude that this source of bias is negligible.

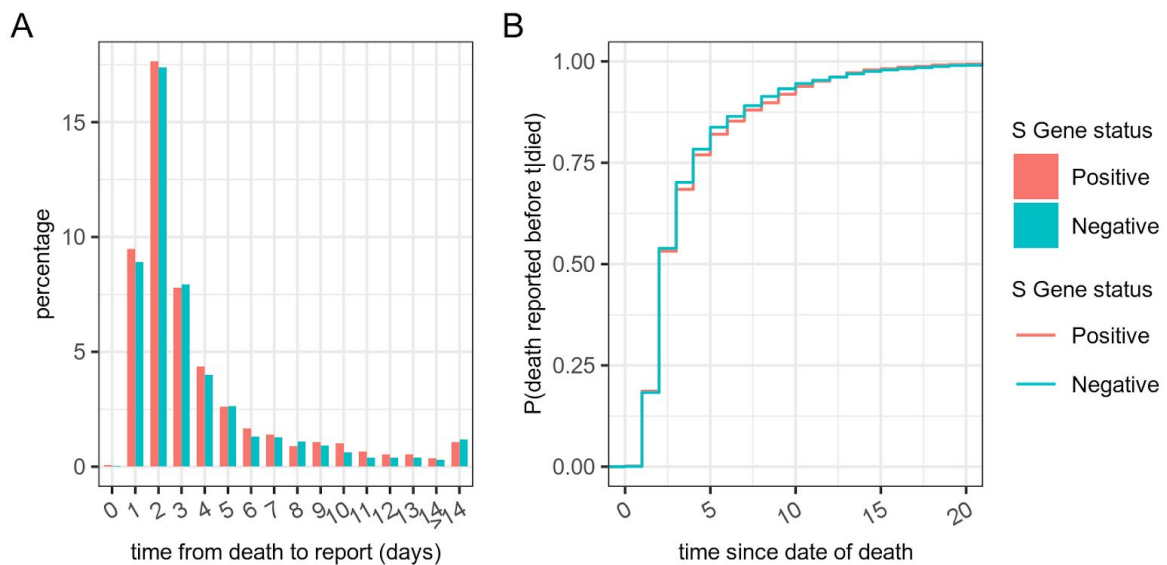
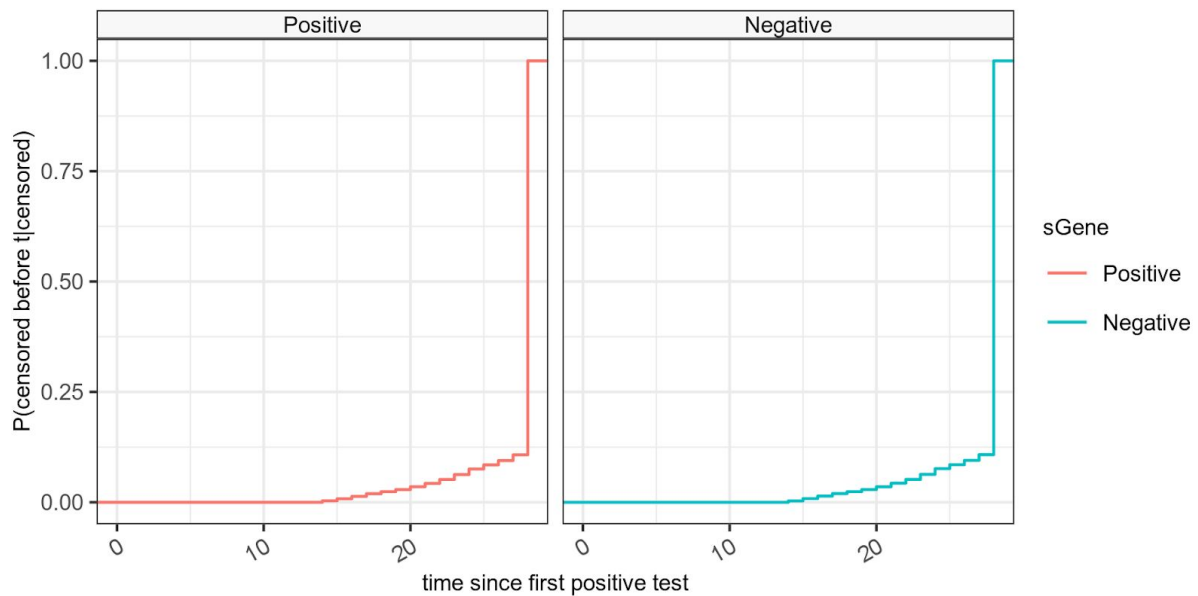


Figure S2 - Delays from date of death and reporting of death in the dataset studied. Panel A shows the distribution of times from date of death to report of death, with the cumulative distribution shown in panel B. We see no discernable difference between the distribution for S-gene positive (red) and S-gene negative (blue) individuals. Over 50% of deaths are reported within 3 days of the date of death.

We analyse data involving samples taken up to the 29th January 2021 and follow cases up for 28 days or the 12th February 2021 (whichever is earlier). This means that all cases have at least 14 days follow up and 85% of cases have the full 28 days as seen in Figure S3



*Figure S3 - Timing of censored cases shows the cumulative proportion of cases that are censored by time since their first positive test. Less than 15% of cases are censored in this analysis and no cases are censored before 14 days. The censoring of cases is indistinguishable between S-gene negative (blue) and positive (red) cases as we are matching cases nearly exactly by sample date.*

## Case Matching

Matching of S-gene negative and S-gene positive cases in this study is an unusual problem, as unlike most other case control studies the prevalence of both S-gene negative and S-gene positive cases was dynamic over the study period. This is in contrast to a rare effects matched cohort study where rare “exposed” cases can be matched to a number of “un-exposed” from the very large general population. In our case, as the prevalence of both S-gene negative and S-gene positives varies, the ease of pairing them also varies. This results in a potential asymmetry in the pairs matched which is time dependent as demonstrated in Figure S4.

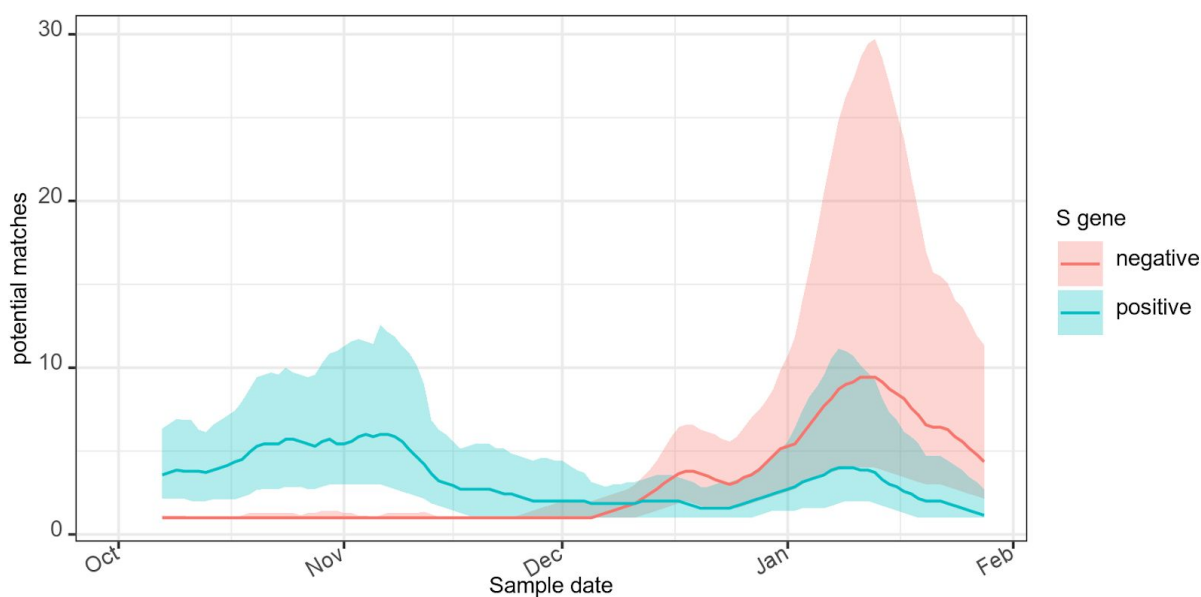


Figure S4 - The average number and interquartile ranges of the number of potential matches for individual S-gene negative (red) or S-gene positive (blue) cases as a function of time over the study period. Initially it was much easier to find S-gene positive matches for S-gene negatives which were relatively scarce. By the end of the time series the opposite is true.

Since individual S-gene negative cases can match multiple S-gene positive cases and vice versa it is necessary to have a strategy for selecting individuals or individual pairs from the range of possible options. We took three approaches to this. Firstly an unpaired selection strategy in which we take unique individuals out of the matched pairs into a single replicate. Secondly an edge sampling strategy in which we randomly sample pairs with replacement into multiple replicates, ensuring unique cases in each replicate, and thirdly a node sampling strategy in which we randomly sample first cases then associated paired matches into multiple replicates and ensure unique cases in each replicate.

The comparison of the unpaired cases, node and edge sampling strategies on the outcome of the hazard rate estimation is shown in Table S1. All other parameters in this analysis are the same as the central estimates in Table 2 of the main paper, which uses the node sampling strategy. All three matching strategies give results with similar estimates of hazard ratios. All models show a significantly elevated HR for S-gene negativity, and the confidence limits are very similar. The central estimates of the hazard ratio vary with different selection strategies but this variability is small compared to the confidence intervals.

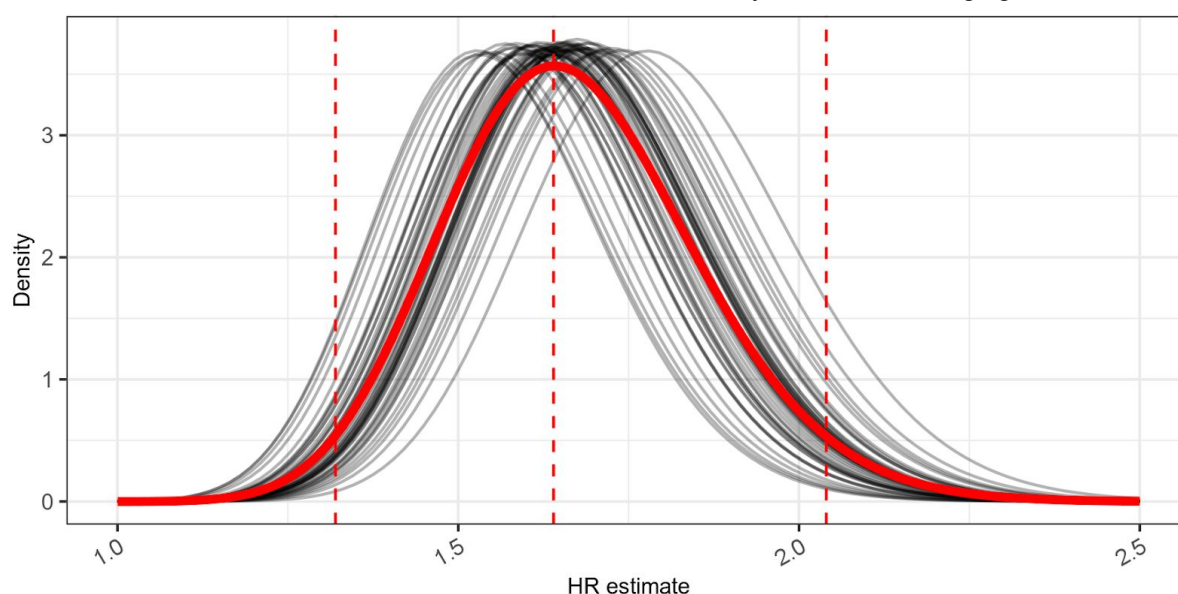
**Table S1. Comparison of three methods for resolving multiple matches in the data set, on the final hazard ratio estimates. The unpaired strategy selects all unique cases that match. The edge sampling strategy selects random replicates without replacement based on pairs, and the node sampling strategy selects random replicates without replacement based on the individuals.**

Strategy	Predictor	Value	Hazard ratio (95% CI)	p value
Edge sampling	S-gene status	Positive (ref)	—	—
		Negative	1.60 (1.26 – 2.03)	<0.001
	Age (per decade)		3.54 (3.25 – 3.86)	<0.001
Node sampling	S-gene status	Positive (ref)	—	—
		Negative	1.64 (1.32 – 2.04)	<0.001
	Age (per decade)		3.55 (3.28 – 3.84)	<0.001
Unpaired cases	S-gene status	Positive (ref)	—	—
		Negative	1.65 (1.37 – 1.98)	<0.001
	Age (per decade)		3.49 (3.29 – 3.71)	<0.001

## Combining estimates

Using random sampling generates 50 replicates of different case compositions, and for each replicate we fitted a set of Cox proportional hazards models, considering each replicate as an individual sample. This produces 50 differing estimates and confidence intervals for the hazard ratios and beta-coefficients of each component of the resulting mortality model. To produce a single combined estimate for the hazard of any given covariate, we assumed the estimates of beta-coefficients from the cox-models to be normally distributed and combined their probability density functions from each bootstrap replicate into a mixture distribution, as shown in red in figure S6. This mixture and its associated cumulative density function were used to determine the 95% confidence intervals, with a Newton-Raphson numerical approach (red dashed lines in figure S6).

Figure S6 shows the combined estimate of the hazard ratio of S-gene negative infection in the main model presented in our paper, and gives us a sense of the stability of our estimate in the face of the random variation introduced by the node sampling pair selection strategy. It shows our central estimate of the hazard ratio of 1.64 is the middle value of a family of estimates ranging from 1.5 to 1.8.



*Figure S6 - Combining point estimates from replicates was performed assuming the beta-coefficient of the cox model to be normally distributed, and summing the probability density functions to make a mixture distribution. In this figure the exponential of the beta-coefficients of the component and mixture distributions are shown as hazard ratio estimate distributions.*

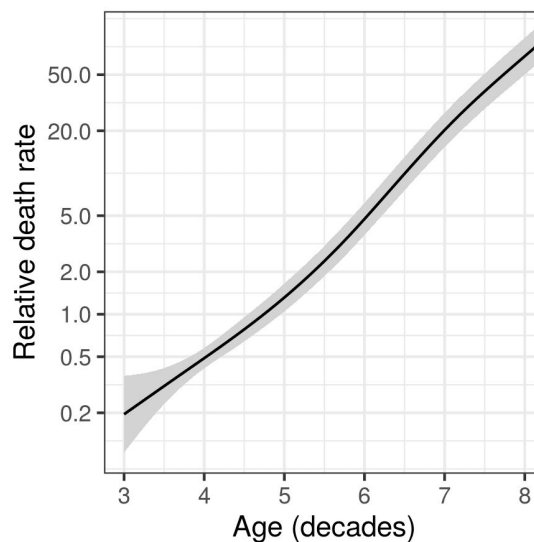
## Additional Proportional Hazards Models

To further assess the robustness of our findings, we fitted a model which included additional individual-level covariates: sex, ethnicity, and Index of Multiple Deprivation (IMD) of patient home location, on top of the S-gene status and age. We found the hazard ratio associated with negative S-gene status to be comparable to the model without the additional covariates as shown in Table S1. This is to be expected as the matched cohort approach ensures that these covariates are independently associated with mortality compared to the S-gene status. As may be expected from other studies of COVID-19 mortality, we find increasing age and male sex are significantly associated with a higher hazard of mortality. Because we exclude many cases in the pairing process we do not have sufficient power to determine any other associations.

**Table S2. Cox proportional hazard model with additional covariates.**

Model	Predictor	Value	Hazard ratio (95% CI)	p value
	S-gene status	Positive (ref)	—	—
		Negative	1.64 (1.31 – 2.04)	<0.001
	Age (per decade)		3.62 (3.33 – 3.94)	<0.001
	Ethnicity	White	—	—
		Afro-caribbean	1.03 (0.27 – 3.18)	0.455
		Asian	1.03 (0.71 – 1.49)	0.427
		Other	0.00 (1.00 – Inf)	0.301
		Unknown	0.00 (1.00 – Inf)	0.496
S-gene + covariates	IMD	1	1.53 (0.96 – 2.46)	0.037
		2	1.68 (1.04 – 2.73)	0.016
		3	1.16 (0.68 – 2.01)	0.292
		4	1.45 (0.85 – 2.48)	0.085
		5 (ref)	—	—
		6	1.38 (0.80 – 2.39)	0.125
		7	0.92 (0.51 – 1.66)	0.394
		8	0.96 (0.54 – 1.71)	0.440
		9	0.69 (0.37 – 1.29)	0.122
		10	0.67 (0.35 – 1.27)	0.110
	Gender	Female (ref)	—	—
		Male	2.13 (1.71 – 2.67)	<0.001

The relationship between age and hazard ratio is most simply described by a linear term in the proportional hazards model, however it is possible that this relationship is non-linear. To test this we constructed a model which involved S-gene status and age and a restricted polynomial spline, and visualised the resulting coefficient. We conducted this analysis on a single replicate, and on the combination of all replicates. Both showed the same pattern as in Figure S7 in which it is clear that any non-linearity in the contribution of age effects is very minor.



*Figure S7 - A restricted spline with 4 knots used to model the relationship between age as a continuous term and hazard of mortality in S-gene negative cases in an otherwise unconstrained model shows no obvious non-linearity.*

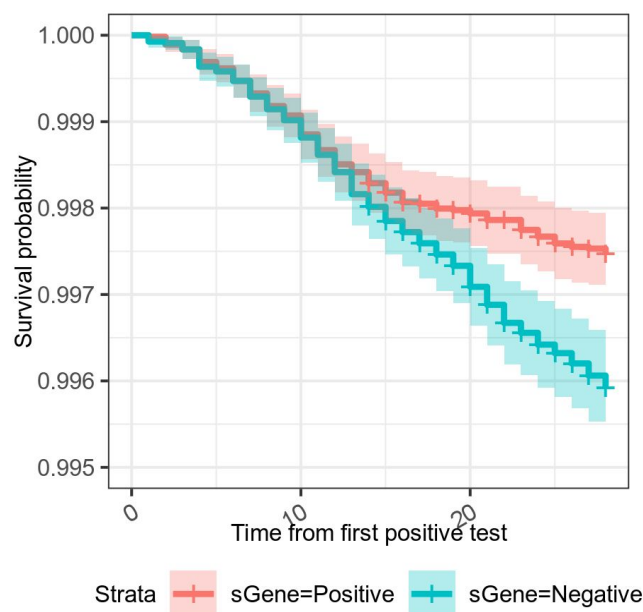
## Proportional hazards assumption

As described in the main text, we found that the assumption of constant hazard over time was violated; Figure S8 and Table S3. This is corroborated when we estimated the probability that the assumption was violated using the methods of Grambsch et al [1] and a further 3 methods for matched control studies, as presented in Xue et al [2]

**Table S3. Probability that the assumption of constant hazard over time was valid.**

method	P(assumption valid)
cox.zph	0.00459
Event time correlation	0.00452
KM Estimate correlation	0.00499
Rank event time correlation	0.00439

The proportional hazard assumption violation can be seen in the Kaplan Meier curve presented in the main paper ( and copied here for clarity). As the survival probabilities of S-gene positive and negative patients diverge around 14 days from the first positive test, we investigated whether we could resolve this violation of assumptions by fitting a model stratified by the time from first positive test result as a categorical variable of 0-14 days and 15-28 days.



*Figure S8 - A Kaplan Meier curve of mortality of S-gene positive and S-gene negative COVID-19 infections. There is a divergence noted at 14 days.*

The hazard ratios for this model is presented in Table S4. These can be interpreted as saying that during the course of a COVID-19 infection, once the patient has survived past 2 weeks there is a relative improvement in their chance of surviving the next 2 weeks (HR 0.5). However this improvement in survival is seen principally in S-gene positive cases and S-gene negative cases continue to be at high risk of death into weeks 3 and 4 of their clinical course.

**Table S4. Cox proportional hazard model with time dependent terms, considering the impact of S-gene status on an early period (days 0-14) and a late period (days 15-28).**

Model	Predictor	Value	Hazard ratio (95% CI)	p value
Early vs Late	Period	0-14 (ref)	—	—
		15-28	0.51 (0.35 – 0.73)	<0.001
	S-gene status	Positive (ref)	—	—
		Negative (Period 0-14)	1.23 (0.92 – 1.64)	0.079
		Negative (Period 15-28)	2.40 (1.66 – 3.47)	<0.001

For the time dependent model the proportional hazard assumption is not violated as demonstrated in table S5. Thus considering the follow up as 2 separate periods is enough to describe the increase in hazard ratio of mortality.

**Table S5. Probability that the assumption of constant hazard over time was valid with introduction of day 0-14 and day 15-28 periods.**

method	P(assumption valid)
cox.zph	0.756
Event time correlation	0.759
KM Estimate correlation	0.751
Rank event time correlation	0.758

Given this evidence that the hazard ratio is dependent on time, it is possible that the central estimate presented in the main paper will change as more data become available, and analysis which includes longer follow up becomes possible [3] depending on how the hazard ratio behaves past 28 days. This is in evolution and needs further investigation

The time dependence of mortality hazard has potential clinical significance. If S-gene negative COVID-19 infection is shown to be associated with later deaths it may be possible to modify this risk by earlier treatment or enhanced monitoring. This finding is something that can be validated in an unpaired retrospective observational cohort analysis, which would be powered to detect any subgroups for whom this may be a particular concern, and is a subject for future work.

## Supplementary material references

- 1 Grambsch PM, Therneau TM. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika* 1994;**81**:515–26. doi:10.2307/2337123
- 2 Xue X, Xie X, Gunter M, *et al.* Testing the proportional hazards assumption in case-cohort analysis. *BMC Med Res Methodol* 2013;**13**:88. doi:10.1186/1471-2288-13-88
- 3 Hernán MA. The Hazards of Hazard Ratios. *Epidemiol Camb Mass* 2010;**21**:13–5. doi:10.1097/EDE.0b013e3181c1ea43



# 8. Early epidemiological signatures of novel SARS-CoV-2 variants: establishment of B.1.617.2 in England

Status: submitted to Science

Date: 6th Dec 2021.

Candidate's contribution: This paper brings together research from a number of members of the Juniper Consortium. The geospatial analysis was conducted by me. The growth rate analysis was a synthesis of estimates performed by me and estimates performed by LG and CO. The analysis of age structure and community spread was initiated by me and developed further by CO, into the form present here. Other contributions from the wider team were made to the supplementary materials which I gratefully acknowledge.

Conceptualisation: RC, LDa, LDy, EBP, MJK;

Methodology: RC, LDa, LDy, EBP, MJK, DW, CO, FS, LP, WSH, RNT;

Investigation: all;

Visualisation: RC,CO,LG;

Funding acquisition: LDa, MJK, LDy, EBP, JRG, MJT;

Project administration: LDa, JRG, MJK, EBP;

Supervision: LDa, JRG, MJK, EBP;

Writing – original draft: LDa, RC, EBP, LDy, EMH, CO, HS;

Writing – review & editing: All

**Title:**

# Early epidemiological signatures of novel SARS-CoV-2 variants: establishment of Delta (B.1.617.2) in England

**One Sentence Summary:**

**Routine community testing data provided early warning of the emergence of the Delta variant of SARS-CoV-2 in England and the rapid expansion of local clusters into the wider community was subsequently confirmed by sequencing as the result of a selective advantage of Delta over Alpha.**

**Authors:**

Robert Challen<sup>1,2,3</sup>; Louise Dyson<sup>3,5</sup>; Christopher E. Overton<sup>3,8,9</sup>, Laura M. Guzman-Rincon<sup>3,5</sup>, Edward M. Hill<sup>3,5</sup>, Helena B. Stage<sup>3,8,10,11</sup>, Ellen Brooks-Pollock<sup>3,4</sup>, Lorenzo Pellis<sup>3,6,8</sup>, Francesca Scarabel<sup>3,8</sup>, David J. Pascall<sup>3,12</sup>, Paula Blomquist<sup>13</sup>, Michael Tildesley<sup>3,5</sup>, Daniel Williamson<sup>1,6</sup>, Trevelyan J. McKinley<sup>14</sup>, Stefan Siegert<sup>1</sup>, Xiaoyu Xiong<sup>1</sup>, Ben Youngman<sup>1</sup>, William S. Hart<sup>15</sup>, Robin N. Thompson<sup>5,16</sup>, JUNIPER<sup>3</sup>, Jonathan M. Read<sup>3,17</sup>, Julia R. Gog<sup>3,18</sup>, Matthew J. Keeling<sup>3,5</sup>, Leon Danon<sup>3,4,6,7\*</sup>

**Affiliations:**

1. College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, Devon, UK.
2. Somerset NHS Foundation Trust, Taunton, Somerset, UK.
3. Joint Universities Pandemic and Epidemiological Research (JUNIPER) consortium.
4. Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK.
5. The Zeeman Institute for Systems Biology & Infectious Disease Epidemiology Research, School of Life Sciences and Mathematics Institute, University of Warwick, Coventry, UK.
6. The Alan Turing Institute, British Library, 96 Euston Rd, London, UK.
7. Department of Engineering Mathematics, University of Bristol, UK.
8. Department of Mathematics, University of Manchester, UK.
9. Clinical Data Science Unit, Manchester University NHS Foundation Trust, UK.
10. Department of Physics and Astronomy, University of Potsdam, Germany.
11. Department of Physics, Humboldt University of Berlin, Germany.
12. MRC Biostatistics Unit, University of Cambridge, UK.
13. COVID Outbreak Surveillance Team, Public Health England, UK.
14. College of Medicine and Health, University of Exeter, Exeter, Devon, UK.
15. Mathematical Institute, University of Oxford, Oxford, UK.
16. Mathematics Institute, University of Warwick, Coventry, UK.
17. Lancaster Medical School, Lancaster University, UK.

18. Department of Applied Mathematics and Theoretical Physics (DAMTP), University of  
Cambridge, UK.  
\*Corresponding author.

**Abstract:**

The emergence of SARS-CoV-2 mutants with new phenotypic properties is a crucial challenge to the control of the ongoing pandemic. In April 2021 in England, against a background of near uniform Alpha variant (B.1.1.7) and associated S-gene negative test results, increasing numbers of S-gene positive cases were identified, in several small geographical clusters. These were later found to be Delta (B.1.617.2) infections. Evidence from the age profile of these cases shows transient importation followed by community transmission, with no signs of vaccine escape. Over May and June 2021, the Delta variant spread countrywide during a period when Alpha variant cases were declining, with an estimated selective advantage of Delta over Alpha of 53% (95% confidence interval 31-98%). This use of S-gene as a marker for variant cases is relevant to the current Omicron outbreak.

**Main text:**

The global COVID-19 pandemic, which started in 2019, has caused millions of SARS-CoV-2 cases and deaths worldwide. The high case burden has generated thousands of SARS-CoV-2 variants that are circulating globally (1). While most mutations show no concerning phenotypic change or selective advantage, others confer higher transmission, higher severity, or escape potential from antibodies generated from previous infection or vaccination (2). These are termed “variants-of-concern” (VOCs) (3, 4) and pose a serious threat to disease control.

As of July 2021, current VOCs include the Alpha variant (pango lineage B.1.1.7) (5), which emerged in southeast England in September 2020 (6). Its large number of accrued mutations conferred increased transmissibility compared to earlier variants (7, 8), and increased mortality (9, 10). The Beta variant (B.1.351), which emerged in South Africa in September 2020, shares some of the same mutations as Alpha, but in addition appears to show reduced sensitivity to immune responses acquired against the ‘wild-type’ Wuhan virus (11) or generated by current vaccines (12). The lineage behind the large number of cases in India in April-May 2021 (B.1.617), increased rapidly as a proportion of sequenced lineages in the UK against a background of Alpha sequences, suggesting a competitive advantage. This lineage was divided into three sublineages: B.1.617.1 (the Kappa variant), B.1.617.2 (the Delta variant) and B.1.617.3 (13). Two of the sublineages, B.1.617.1 and B.1.617.3 have the E484Q mutation, which may reduce viral neutralisation and could facilitate vaccine escape (14): these have been designated as “variants under investigation” (VUIs). The Delta variant was designated as a VOC by Public Health England on 6 May 2021 (15). There is conflicting evidence on sensitivity to vaccine-acquired immunity with early observational studies suggesting no clear change (16) but experimental studies showing reduced sensitivity (17).

Variants of SARS-CoV-2 that show increased transmissibility or demonstrate any ability to escape from vaccine-derived immunity, could generate large future waves of infection requiring further costly social distancing measures to prevent healthcare systems being overwhelmed (18, 19). The recent detection in South Africa of a novel variant (B.1.1.529), showing more than 30 mutations and concerning epidemiological patterns (20, 21) has prompted the World Health Organisation to label it a variant of concern, named Omicron. The emergence of Omicron highlights the critical importance of early identification and epidemiological understanding of VOCs .

Here we present our approach of combining genomic information with S-gene positivity status from testing data, linked to demographic and geographical data, that allowed us to:

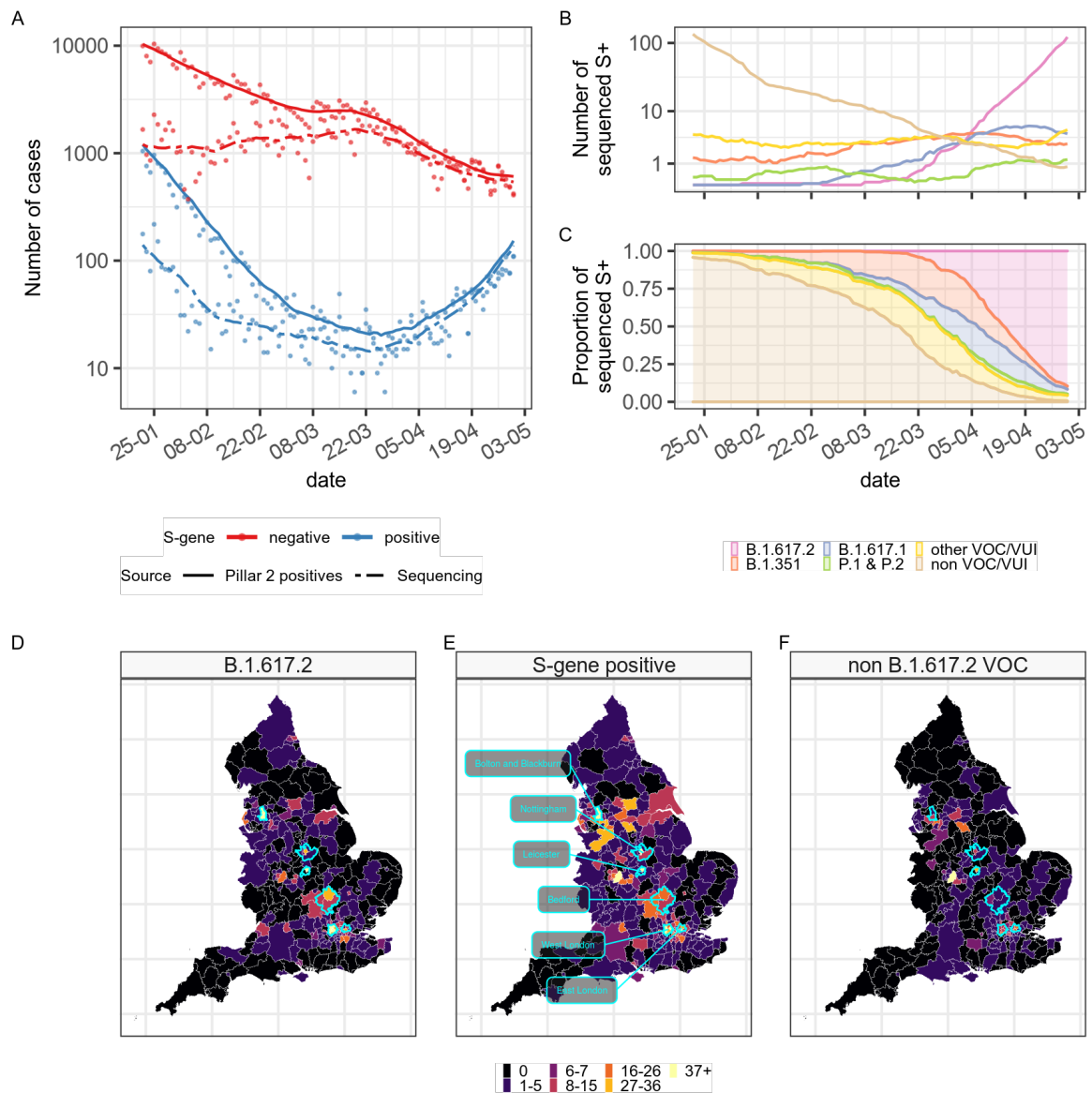
1. identify the early invasion of the Delta variant in England in routine surveillance data between mid April and early May 2021,
2. observe the regional variability in local growth of S-gene positive cases and relate this to the transition to dominance of the Delta variant,
3. differentiate between local importations and early limited outbreaks in specific regions and sustained community transmission, and
4. over the longer term (May - June 2021) and larger geographic scale, estimate the selective advantage that the Delta has over Alpha.

The emergence and progression of the Alpha variant in England in late 2020 could be observed because, in contrast to previously circulating ‘wild-type’ variants, Alpha exhibits a deletion in the SARS-CoV-2 genome at site 69-70 in the spike protein, leading to the ThermoFisher TaqPath quantitative PCR assay failing to amplify the S-gene target (22). During the emergence of the Alpha variant, the percentage of cases with an S-gene target failure, or SGTF, increased from 3% in October 2020 to 98% by March 2021 (23). In contrast to sequencing, which may take up to 3 weeks from a positive PCR test result, the presence or absence of an S-gene from the TaqPath assay is obtained without delay and requires no additional resources, providing a more immediate monitoring tool.

The high prevalence of Alpha in March 2021 and associated S-gene negative test results (13) allows us to use S-gene positive cases from the TaqPath assay in community testing (known as Pillar 2 in England and the UK) as a rapid signal for investigating community spread of S-gene positive variants, including the Beta and Delta variants. Social distancing restrictions in January 2021 in England meant the number of SARS-CoV-2 cases due to all variants fell to low levels. However, a steady rise in the absolute number of S-gene positive cases was observed from early April, while S-gene negative cases continued to fall (Fig. 1A). S-gene positive cases overtook S-gene negative cases by the end of May 2021.

Sequencing data revealed that this increase in S-gene positive case numbers was the result of the Delta variant, which increased while all other S-gene positive variants either remained stable or declined in number (Fig. 1B). From 1st May 2021, the Delta variant was the proven cause of more than 95% of all sequenced S-positive cases (Fig. 1C). The association between S-gene positive cases and Delta is supported by comparison of their early geographic

distributions (Figs 1D-1E), however during the early spread of Delta in March and April, S-gene positive cases were frequently due to other non-Alpha variants (Fig. 1F). Combining S-gene positive and emerging sequencing data enabled early identification of areas with high risk of Delta variant outbreaks (highlighted in cyan on Fig. 1 panels D-F), prompting further investigation. These areas (Fig. 1D) generally have a high asian population, with close ties to the Indian subcontinent, and may expect high rates of case importation.



*Fig. 1 - Spatial and temporal patterns in SARS-CoV-2 variants in England: (A) S-gene positive and negative case counts (solid lines) from 21st Jan 2021 to 1st May 2021 (21-day rolling average, logarithmic scale) and sequencing activity associated with those cases (dashed lines). (B) The number of cases of different variants, from sequencing samples of S-gene positive tests (21-day rolling average, log(1+y) scale). (C) Proportion of S-gene positive cases of different variants, showing recent dominance of the Delta variant (21-day rolling average). (D) The*

geographical distribution of sequence proven Delta (B.1.617.2) variant cases, (E) S-gene positive cases, and (F) other S-gene positive VOC/VUI between March and May 2021 show a number of areas with high levels of Delta variant and S-gene positive cases including Bolton & Blackburn, Leicester, Nottingham, Bedford, and parts of London (cyan). Other areas showed comparatively high levels of non B.1.617.2 S-gene positive variants (F).

The growth rate is a reliable measure of increasing cases that is based solely on observed data, and not reliant on any other assumptions. We use five methods to estimate the daily growth rate: three which use a maximum likelihood framework, a Generalised Additive Model, a piecewise Poisson generalised linear regression, a Poisson local polynomial regression; and two which use a Bayesian framework, a Gaussian Process model, and a Bayesian Poisson model (see supplementary methods for details).

A rapid rise of Delta in Bolton and Bedford was related to a sudden increase in the S-gene positive growth rate of 20% per day, corresponding to a minimum doubling time of 5 days (Fig. 2). Leicester & West London, which had higher rates of lineage B.1.617.1, showed a slower transition to Delta dominance and less marked elevation in S-gene positive growth rates up to the end of April 2021. The scale and timing of the increase in growth rates is linked to when Delta became the dominant S-gene positive strain (Fig. 2). Conversely, areas where a range of variants were sequenced, such as East London, initially showed lower growth rates. As a result of this developing picture, on May 3rd public health interventions were targeted at Bolton (24).

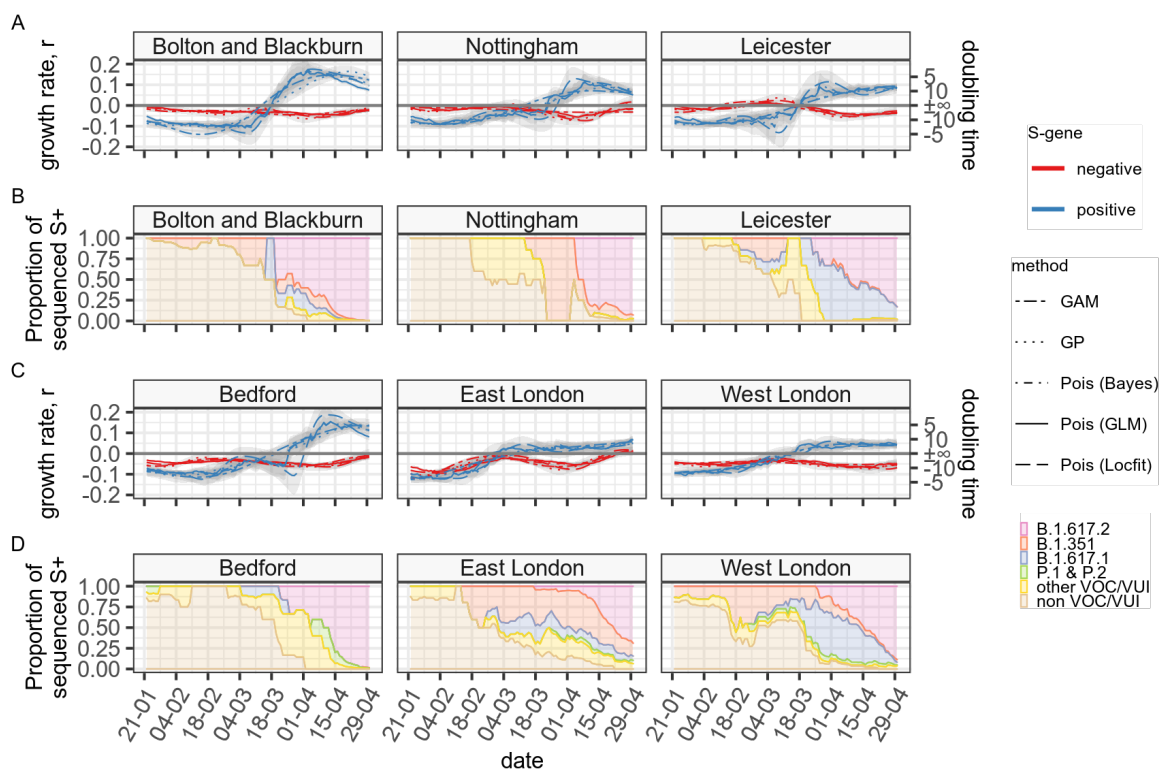


Fig. 2 - Growth rates and temporal patterns in SARS-CoV-2 variants in small areas where initial outbreaks of the Delta variant (B.1.617.2) were suspected as a result of increased incidence of S-gene positive cases between 21st Jan 2021 to 1st May 2021. Panels A and C

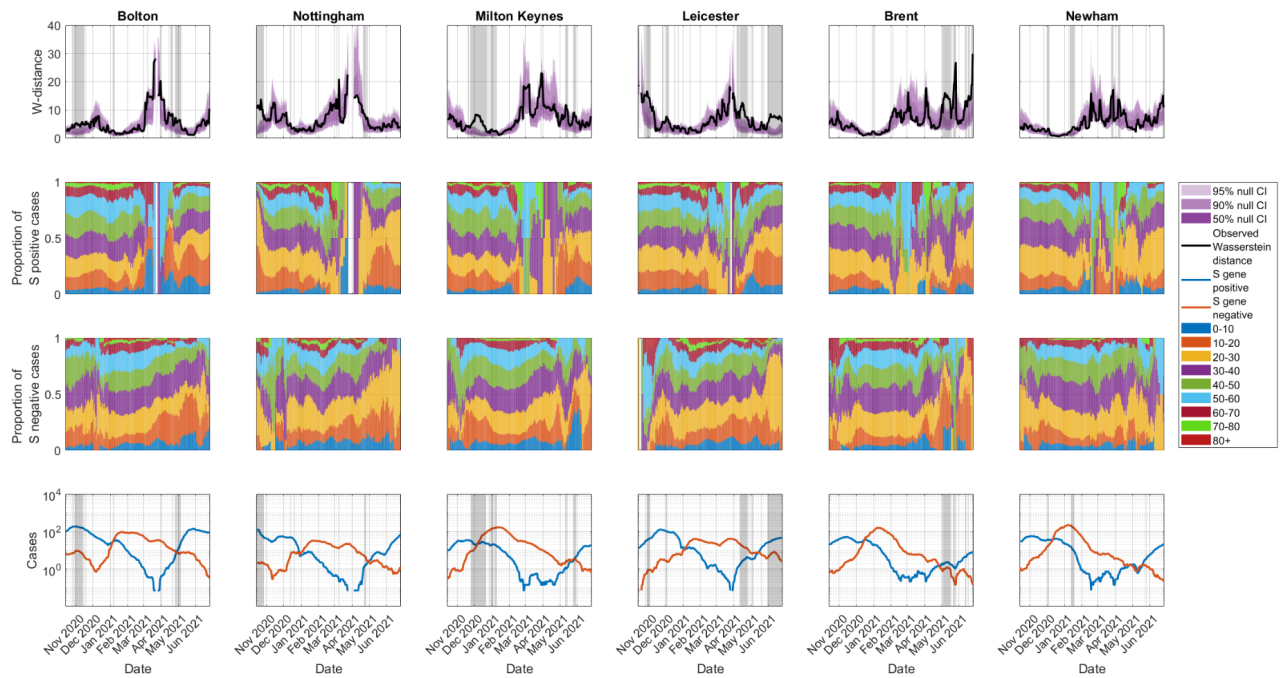
*show the exponential growth rate estimated using five methods in the small areas of England marked in Fig. 1D-F. Panels B and D show the relative proportion of sequencing results (over a rolling 21 day window).*

To determine when a localised outbreak in a specific subpopulation transitions to wider community spread we use the age distribution of cases. The case age distribution reflects population mixing patterns, age-stratified infectivity, susceptibility, severity, testing-seeking behaviour, and vaccine uptake. If most cases of a new variant are imported through travel, we would expect the observed age distributions to more closely reflect the age distribution of travellers than the general population (see supplementary text and Figs. S5-S6), before returning to a distribution reflecting community transmission.

We compute the Wasserstein distance (22, 23) between the age distribution of S-gene positive and negative cases and use a permutation test to generate a 95% confidence region (see supplementary text and Figs. S3-S4). In each region shown, the Wasserstein distance leaves the confidence region for short periods of time in late March and early April 2021 and subsequently declines in the more recent time points (Fig. 3), suggesting that introductions of Delta variant were quickly followed by sustained transmission in the community. This is consistent with the timing and opportunity of importation of the Delta variant (see Figs. S5-S6 for traveller status breakdown) as on 23rd April 2021, travel restrictions were imposed from India (21). In other LTLAs (see supplementary text and Figs S7-S9), the discrepancies between the age distributions are less abrupt and substantial, suggesting that fewer cases were imported directly.

Over the period studied, vaccination programmes were being rolled out across England in priority of decreasing age (25), resulting in a shift to cases to younger age groups. However vaccination should not cause a differential age distribution between S-gene positive and negative cases unless S-gene positive variants were vaccine-escaping, leading to more S-gene positive cases in older vaccinated people. We do not see any evidence of long term age structure divergence, suggesting that vaccine escape is not the sole driver of observed excess growth of S-gene positive variants.

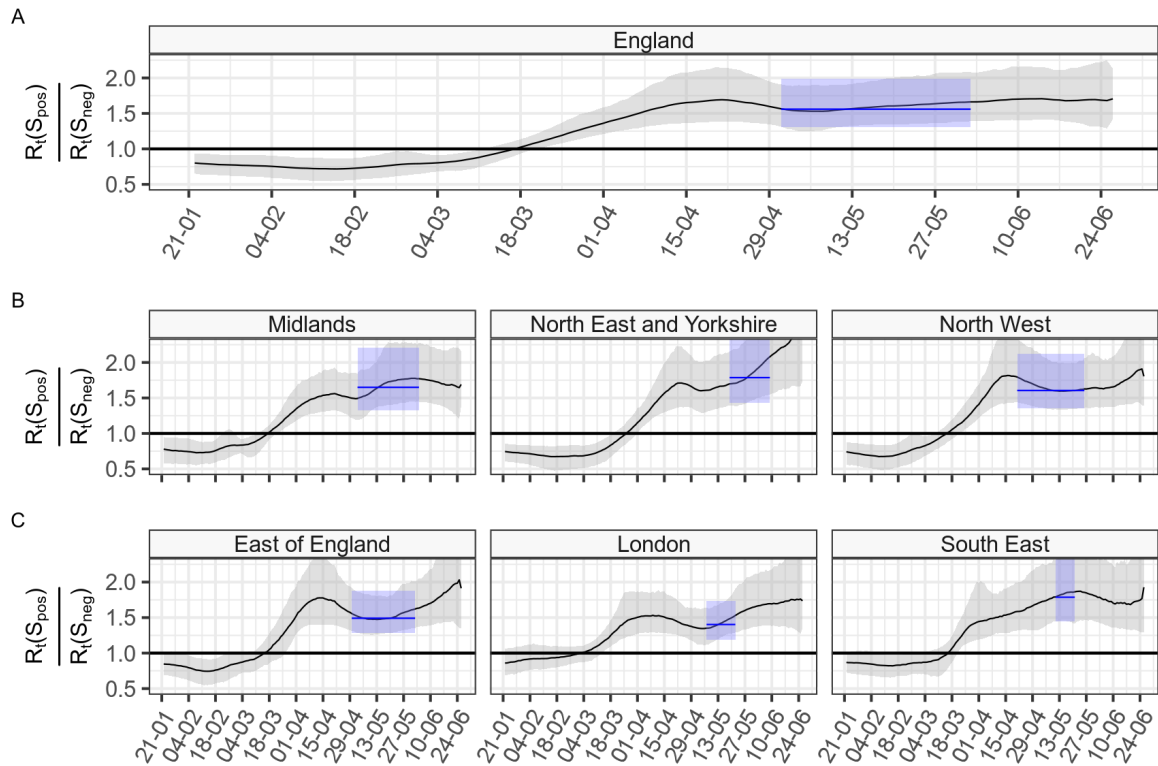




*Fig. 3 - Dynamics of age distributions of cases for each region. Top panels : Wasserstein distance between the age distribution of S-gene positive and S-gene negative cases (black), with 95% confidence intervals (purple); grey shaded regions indicate dates where the point estimate falls outside the confidence interval indicating a significant difference from noise. Middle panels: the age distribution of S-gene positive cases (second line), and S-gene negative cases (third line). Bottom panels: Number of cases S-gene negative (red) and S-gene positive (blue). All metrics are computed over a 14-day rolling window of cases by specimen date, plotted by first swab date. Bolton, Leicester and Nottingham correspond directly to the regions in Fig. 1, Milton Keynes is a sub region of the Bedford area, Brent is in West London, and Newham is in East London.*

The Delta variant spread broadly throughout May and June 2021 and was present in every region across England by July (26). This occurred despite continued restrictions, which continued to be effective in decreasing the number of Alpha variant cases. By the beginning of July, S-gene negative cases became less than 5% of all cases (supplementary text and Figs. S10-S12).

We measure the transmission advantage of the Delta variant over Alpha, by examining the reproduction number ratio of S-gene positives over S-gene negative cases in England and NHS regions (Fig. 4). In January 2021, this ratio was less than one indicating a transmission disadvantage of S-gene positive variants - largely representing residual non-VOC/VUI (mostly B.1.177), with the Delta variant yet to be detected. The reproduction number ratio began to increase at the beginning of March 2021 across all NHS regions and the whole of England, followed by a more dramatic increase, before stabilising when the Delta variant became dominant in all areas. Towards the end of the time series, as the number of Alpha cases fell to low numbers, instabilities in the reproduction number ratio are observed (Figure 4).



*Fig. 4 - Reproduction number ratios for S-gene positive and S-gene negative cases, estimated from the growth rate, combining estimates from four growth rate and seven generation time estimates (see supplementary information). The blue regions represent the period of time where the composition of S-gene positive cases was dominated by the Delta variant and there continued to be enough Alpha variant S-gene negative cases to allow stable growth rate estimates (see supplementary text and Figs. S10-S12). During this period we have the most stable estimates of the transmission advantage.*

For each region, the “stabilised period”, shown in blue on Fig. 4, occurs when the Delta variant accounts for the majority of S-gene positive cases, but there are still enough S-gene negative cases to make estimates of the reproduction number ratio (see supplementary text and Figs. S10-12). These periods vary from region to region, the longest corresponding to North West and East of England (Table 1). The reproduction number ratio is estimated to be between 31-98% for England, with a central estimate of 56%. These estimates align with findings from early household studies conducted by Public Health England (3), which estimated secondary attack rates of Delta to be 11.0% compared to 8.2% for Alpha. There are many caveats to these reproduction ratio estimates. The estimates vary with time (Fig. 4) and hence the point estimate depends on the precise window over which it is estimated (2). Variations in acquired immunity, restrictions and behaviours may influence the apparent growth advantage in a regionally specific way. The point reproduction number ratio estimates are significantly affected by both different generation time assumptions and growth rate estimates (see Fig. S16, Table S1 and Table S2, supplementary text and table S3). The rate of infections with unknown S-gene status is not fixed over time periods (supplementary text and Figs. S10-S12), possibly reflecting a redistribution of testing effort and acquisition biases, the impact of which is not quantified.

*Table 1 - Reproduction number ratio estimates for S-gene positive infections over S-gene negative infections over the time window in given regions where S-gene positive is most closely associated with the Delta variant and where S-gene negative infections (representative of Alpha) continue to be observed at significant numbers (See supplement text and Figs. S10-S12).*

Level	Name	Date range	Reproduction number ratio	Reproduction number ratio (percentage)
Country	England	01 May – 02 Jun	1.56 (1.31–1.98)	56% (31%–98%)
	East of England	30 Apr – 02 Jun	1.49 (1.28–1.88)	49% (28%–88%)
NHS region	London	07 May – 22 May	1.40 (1.19–1.74)	40% (19%–74%)
	Midlands	03 May – 04 Jun	1.65 (1.33–2.21)	65% (33%–121%)
	North East and Yorkshire	19 May – 09 Jun	1.79 (1.43–2.57)	79% (43%–157%)
	North West	21 Apr – 26 May	1.61 (1.36–2.12)	61% (36%–112%)
	South East	11 May – 21 May	1.79 (1.45–2.50)	79% (45%–150%)

## Discussion

We provide evidence that S-gene positive SARS-CoV-2 infections in England, although initially confined to travellers and their close contacts, quickly became established in the wider population. All indications are that these cases are due to the Delta variant. This analysis was only possible by combining widely collected routine testing data with the delayed signal from genomic sequencing. In the early part of 2021 the majority of positive SARS-CoV-2 tests in England were sequenced (74% in April 2021), but with processing delays: 21% of sequencing results were available within 1 week, 44% within 2 weeks, and 66% within 3 weeks, (as of 27th May 2021). We were able to draw timely conclusions about the spread of Delta from S-gene data because of the close relationship between S-gene positivity and the Delta variant in the limited geographical areas we identified in England during April. This allowed us to identify potential clusters of cases before sequencing was complete, prompting early public health health measures in those areas (24).

We estimate that S-gene positive cases were doubling every 7 days in England overall, and as rapidly as 5 days in some regions with identified outbreaks – for comparison, the shortest doubling time of SARS-CoV-2 infections in the UK was estimated to be 3 days (27). This growth was against a background of a shrinking epidemic of S-gene negative cases in the same areas. However, here we demonstrated that S-gene positive cases initially grew rapidly in distinct subpopulations, both regionally and in terms of age groups. This observation could be explained if behavioural patterns imply differences in contact rates (28, 29) or propensity for larger gatherings (30) that are not dependent on the biology of the Delta variant.

In the medium term we observe the convergence of the initial geographic and age related differences in cases to the background distribution, combined with a persistent elevation of the reproduction number ratio following convergence, which is inconsistent with a founder effect, or a decrease in vaccine effectiveness. The observation that age-related incidence is initially perturbed has the potential to affect vaccine effectiveness studies (16) if confounders are not completely accounted for. Analysis of Delta variant severity from Public Health England suggests that, when age is controlled for, the risk of hospitalisation is increased in infections caused by B.1.617.2 (3). Our analysis highlights the complex and dynamic relationship between space, time and age of cases in this outbreak, and there remain challenges in fully controlling for these variables when estimating phenotypic properties of emerging variants.

We estimate the transmission advantage of the Delta variant over Alpha to be 53% (95% CI 31-98%) at timescales and geographies carefully chosen to minimise founder effects and reduce model uncertainty. Transmission advantage is clear and persistent, regardless of the assumptions we make about generation times or growth rate estimation methodology. We find estimates that assume no change in the generation time between variants may overestimate transmission advantage, finding it 10% higher compared to estimates that take this into account (supplementary text and table S3). Mathematical modelling of the UK's proposed roadmap for easing social distancing predicts that variants with a large competitive advantage could generate a resurgence in cases and hospital admissions larger than experienced in January 2021 in the UK. The risks identified by this, and other, analyses prompted a reassessment of the roadmap to allow for further rollout of England's vaccination programme (31).

Despite remaining limitations (see supplementary text), our work provides a deeper understanding of the effect of SARS-CoV-2 variant dynamics that need to be accounted for when estimating transmissibility, severity (7, 9, 10) and vaccine escape potential (16). The United Kingdom has a well established disease surveillance programme, and the ability to undertake detailed epidemiological studies that establish phenotypic properties of globally circulating variants, which are often not possible in locations where variants emerge, or are first identified (32). Our study underscores the imperative to continue these efforts, with specific application to the recent identification of the Omicron variant, which can again be distinguished from Delta by S-gene target failure, the results of which have immediate global policy implications.

## References and Notes

1. T. Koyama, D. Platt, L. Parida, Variant analysis of SARS-CoV-2 genomes. *Bull. World Health Organ.* **98**, 495–504 (2020).
2. A. S. Luring, E. B. Hodcroft, Genetic Variants of SARS-CoV-2—What Do They Mean? *JAMA.* **325**, 529–531 (2021).
3. Public Health England, “SARS-CoV-2 variants of concern and variants under investigation: Technical briefing 14,” p. 66.
4. Tracking SARS-CoV-2 variants, (available at <https://www.who.int/emergencies/emergency-health-kits/trauma-emergency-surgery-kit-who-tesk-2019/tracking-SARS-CoV-2-variants>).

5. A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, O. G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
6. Public Health England, “Investigation of novel SARS-CoV-2 variant: Variant of Concern 202012/01 - Technical briefing 1,” (available at [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/947048/Technical\\_Briefing\\_VOC\\_SH\\_NJL2\\_SH2.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/947048/Technical_Briefing_VOC_SH_NJL2_SH2.pdf)).
7. N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. B. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne, T. Wenseleers, A. Gimma, W. Waites, K. L. M. Wong, K. van Zandvoort, J. D. Silverman, CMMID COVID-19 Working Group, COVID-19 Genomics UK (COG-UK) Consortium, K. Diaz-Ordaz, R. Keogh, R. M. Eggo, S. Funk, M. Jit, K. E. Atkins, W. J. Edmunds, Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*. **372** (2021), doi:10.1126/science.abg3055.
8. E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J. Laydon, G. Dabrera, Á. O'Toole, R. Amato, M. Ragonnet-Cronin, I. Harrison, B. Jackson, C. V. Ariani, O. Boyd, N. J. Loman, J. T. McCrone, S. Gonçalves, D. Jorgensen, R. Myers, V. Hill, D. K. Jackson, K. Gaythorpe, N. Groves, J. Sillitoe, D. P. Kwiatkowski, T. C.-19 G. U. (COG-U. Consortium, S. Flaxman, O. Ratmann, S. Bhatt, S. Hopkins, A. Gandy, A. Rambaut, N. M. Ferguson, Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data (2021), p. 2020.12.30.20249034, , doi:10.1101/2020.12.30.20249034.
9. N. G. Davies, C. I. Jarvis, W. J. Edmunds, N. P. Jewell, K. Diaz-Ordaz, R. H. Keogh, Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature*. **593**, 270–274 (2021).
10. R. Challen, E. Brooks-Pollock, J. M. Read, L. Dyson, K. Tsaneva-Atanasova, L. Danon, Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. *BMJ*. **372**, n579 (2021).
11. C. K. Wibmer, F. Ayres, T. Hermanus, M. Madzivhandila, P. Kgagudi, B. Oosthuysen, B. E. Lambson, T. de Oliveira, M. Vermeulen, K. van der Berg, T. Rossouw, M. Boswell, V. Ueckermann, S. Meiring, A. von Gottberg, C. Cohen, L. Morris, J. N. Bhiman, P. L. Moore, SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat. Med.* **27**, 622–625 (2021).
12. S. A. Madhi, V. Baillie, C. L. Cutland, M. Voysey, A. L. Koen, L. Fairlie, S. D. Padayachee, K. Dheda, S. L. Barnabas, Q. E. Bhorat, C. Briner, G. Kwatra, K. Ahmed, P. Aley, S. Bhikha, J. N. Bhiman, A. E. Bhorat, J. du Plessis, A. Esmail, M. Groenewald, E. Horne, S.-H. Hwa, A. Jose, T. Lambe, M. Laubscher, M. Malahleha, M. Masenya, M. Masilela, S. McKenzie, K. Molapo, A. Moultrie, S. Oelofse, F. Patel, S. Pillay, S. Rhead, H. Rodell, L. Rossouw, C. Taoushanis, H. Tegally, A. Thombrayil, S. van Eck, C. K. Wibmer, N. M. Durham, E. J. Kelly, T. L. Villafana, S. Gilbert, A. J. Pollard, T. de Oliveira, P. L. Moore, A. Sigal, A. Izu, Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. *N. Engl. J. Med.* **384**, 1885–1898 (2021).
13. Public Health England, “SARS-CoV-2 variants of concern and variants under investigation in England: Technical briefing 11,” (available at [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/986380/Variants\\_of\\_Concern\\_VOC\\_Technical\\_Briefing\\_11\\_England.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/986380/Variants_of_Concern_VOC_Technical_Briefing_11_England.pdf)).
14. A. J. Greaney, A. N. Loes, K. H. D. Crawford, T. N. Starr, K. D. Malone, H. Y. Chu, J. D. Bloom, Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host*

- Microbe*. **29**, 463-476.e6 (2021).
15. Public Health England, “SARS-CoV-2 variants of concern and variants under investigation: Technical Briefing 10,” p. 39.
  16. J. L. Bernal, N. Andrews, C. Gower, E. Gallagher, R. Simmons, S. Thelwall, J. Stowe, E. Tessier, N. Groves, G. Dabrera, R. Myers, C. Campbell, G. Amirthalingam, M. Edmunds, M. Zambon, K. Brown, S. Hopkins, M. Chand, M. Ramsay, “Effectiveness of COVID-19 vaccines against the B.1.617.2 variant” (2021), p. 2021.05.22.21257658, , doi:10.1101/2021.05.22.21257658.
  17. E. C. Wall, M. Wu, R. Harvey, G. Kelly, S. Warchal, C. Sawyer, R. Daniels, P. Hobson, E. Hatipoglu, Y. Ngai, S. Hussain, J. Nicod, R. Goldstone, K. Ambrose, S. Hindmarsh, R. Beale, A. Riddell, S. Gamblin, M. Howell, G. Kassiotis, V. Libri, B. Williams, C. Swanton, S. Gandhi, D. L. Bauer, Neutralising antibody activity against SARS-CoV-2 VOCs B.1.617.2 and B.1.351 by BNT162b2 vaccination. *The Lancet*. **397**, 2331–2333 (2021).
  18. L. Dyson, E. M. Hill, S. Moore, J. Curran-Sebastian, M. J. Tildesley, K. A. Lythgoe, T. House, L. Pellis, M. J. Keeling, Possible future waves of SARS-CoV-2 infection generated by variants of concern with a range of characteristics. *Nat. Commun.* **12**, 5730 (2021).
  19. SPI-M-O, “Summary of further modelling of easing restrictions – roadmap step 3, 5 May 2021,” (available at [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/984521/S1227\\_SPI-M-O\\_Summary\\_of\\_further\\_modelling\\_of\\_easing\\_restrictions\\_Roadmap\\_Step\\_3\\_2\\_.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/984521/S1227_SPI-M-O_Summary_of_further_modelling_of_easing_restrictions_Roadmap_Step_3_2_.pdf)).
  20. S. S. A. Karim, Q. A. Karim, Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. *The Lancet*. **0** (2021), doi:10.1016/S0140-6736(21)02758-6.
  21. E. B. Hodcroft, “CoVariants: SARS-CoV-2 Mutations and Variants of Interest.” (available at <https://covariants.org/>).
  22. A. Rambaut, N. Loman, O. Pybus, W. Barclay, J. Barrett, A. Carabelli, T. Connor, T. Peacock, D. Robertson, E. Volz, “Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology” (CoG-UK, 2020), (available at <https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563>).
  23. Public Health England, “SARS-CoV-2 variants of concern and variants under investigation in England: Technical briefing 7,” (available at [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/972247/Variants\\_of\\_Concern\\_VOC\\_Technical\\_Briefing\\_7\\_England.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/972247/Variants_of_Concern_VOC_Technical_Briefing_7_England.pdf)).
  24. Coping with the rise of the Delta variant in Bolton | Local Government Association, (available at <https://www.local.gov.uk/coping-rise-delta-variant-bolton>).
  25. Public Health England, “COVID-19 vaccine surveillance report - week 27,” p. 25.
  26. Public Health England, SARS-CoV-2 variants of concern and variants under investigation: Technical briefing 19, 55.
  27. L. Pellis, F. Scarabel, H. B. Stage, C. E. Overton, L. H. K. Chappell, E. Fearon, E. Bennett, K. A. Lythgoe, T. A. House, I. Hall, Challenges in control of COVID-19: short doubling time and long delay to effect of interventions. *Philos. Trans. R. Soc. B Biol. Sci.* **376**, 20200264 (2021).
  28. C. I. Jarvis, K. Van Zandvoort, A. Gimma, K. Prem, M. Auzenbergs, K. O’Reilly, G. Medley, J. C. Emery, R. M. G. J. Houben, N. Davies, E. S. Nightingale, S. Flasche, T. Jombart, J. Hellewell, S. Abbott, J. D. Munday, N. I. Bosse, S. Funk, F. Sun, A. Endo, A. Rosello, S. R. Procter, A. J. Kucharski, T. W. Russell, G. Knight, H. Gibbs, Q.

- Leclerc, B. J. Quilty, C. Diamond, Y. Liu, M. Jit, S. Clifford, C. A. B. Pearson, R. M. Eggo, A. K. Deol, P. Klepac, G. J. Rubin, W. J. Edmunds, CMMID COVID-19 working group, Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. *BMC Med.* **18**, 124 (2020).
29. L. Danon, J. M. Read, T. A. House, M. C. Vernon, M. J. Keeling, Social encounter networks: characterizing Great Britain. *Proc. R. Soc. B Biol. Sci.* **280**, 20131037 (2013).
  30. E. Brooks-Pollock, J. M. Read, T. House, G. F. Medley, M. J. Keeling, L. Danon, The population attributable fraction of cases due to gatherings and groups with relevance to COVID-19 mitigation strategies. *Philos. Trans. R. Soc. B Biol. Sci.* **376**, 20200273 (2021).
  31. Covid: Lockdown easing in England to be delayed by four weeks. *BBC News* (2021), (available at <https://www.bbc.com/news/uk-57464097>).
  32. M. Wadman, Scientists tracking coronavirus variants struggle with global blind spots | Science | AAAS, (available at <https://www.science.org/news/2021/05/scientists-tracking-coronavirus-variants-struggle-global-blind-spots>).
  33. R. M. Anderson, R. M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, Oxford, New York, 1992).

### **Acknowledgments:**

We thank the members of JUNIPER consortium for helpful comments on the manuscript.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.

### **Funding:**

- UKRI through the JUNIPER consortium (grant number MR/V038613/1), LDa LDy, JRG, MJK, TJM, CO, LG, FS, LP, DJP
- UKRI (grant number EP/V053507/1), RNT
- MRC (grant number MC/PC/19067), LDa, RC, EBP, MJK
- MRC (grant number MR/V009761/1) EMH, MJT, LDy, MJK
- MRC (unit programme number MC UU 00002/11), DJP
- EPSRC (EP/V051555/1), LDa, DW, TJM
- EPSRC through the The Alan Turing Institute, grant EP/N510129/1), LDa, DW
- EPSRC (EP/R513295/1), WSH
- National Institute of Health Research UK: EBP,
- Pfizer through investigator-led grants on respiratory tract infections. LDa
- Wellcome Trust and the Royal Society (grant 202562/Z/16/Z), LP, HS and CO
- Alexander von Humboldt Foundation, HS

### Author contributions:

- Conceptualization: RC, LDa, LDy, EBP, MJK
- Methodology: RC, LDa, LDy, EBP, MJK, DW, CO, FS, LP, WSH, RNT
- Investigation: all
- Visualization: RC,CO,LG,
- Funding acquisition: LDa, MJK, LDy, EBP, JRG, MJT
- Project administration: LDa, JRG, MJK, EBP
- Supervision: LDa, JRG, MJK, EBP
- Writing – original draft: LDa, RC, EBP, LDy, EMH, CO, HS
- Writing – review & editing: All

### Competing interests:

LDa is partly funded through an investigator-led grant funded by Pfizer.

### **List of Supplementary materials:**

#### Materials and Methods

- Materials: Description of data sources
- Materials: Processing I: case numbers for growth rate estimates
- Materials: Processing II: S-gene and genomic sequencing case data
- Methods: Instantaneous growth rate estimation
- Methods: Transmission advantage
- Methods: Comparison of the distribution of cases by age

#### Supplementary Text

- Age distribution of cases among travellers
- Comparing age distributions of S-gene positive and S-gene negative cases at LTLA level
- Growth rate in England and NHS Regions from late January to late June
- Selective advantage and generation time
- Limitations

Figs. S1 to S16

Tables S1 to S3

References (33-53):





# Science



## Supplementary Materials for Early epidemiological signatures of novel SARS-CoV-2 variants: establishment of Delta (B.1.617.2) in England

### Authors:

Robert Challen<sup>1,2,3</sup>; Louise Dyson<sup>3,5</sup>; Christopher E. Overton<sup>3,8,9</sup>; Laura M. Guzman-Rincon<sup>3,5</sup>; Edward M. Hill<sup>3,5</sup>; Helena B. Stage<sup>3,8,10,11</sup>; Ellen Brooks-Pollock<sup>3,4</sup>; Lorenzo Pellis<sup>3,6,8</sup>; Francesca Scarabel<sup>3,8</sup>; David J. Pascall<sup>3,12</sup>; Paula Blomquist<sup>13</sup>; Michael Tildesley<sup>3,5</sup>; Daniel Williamson<sup>1,6</sup>; Trevelyan J. McKinley<sup>14</sup>; Stefan Siegert<sup>1</sup>; Xiaoyu Xiong<sup>1</sup>; Ben Youngman<sup>1</sup>; William S. Hart<sup>15</sup>; Robin N. Thompson<sup>5</sup>; JUNIPER<sup>3</sup>; Jonathan M. Read<sup>3,17</sup>; Julia R. Gog<sup>3,18</sup>; Matthew J. Keeling<sup>3,5</sup>; Leon Danon<sup>3,4,6,7\*</sup>

### Affiliations:

1. College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, Devon, UK.
2. Somerset NHS Foundation Trust, Taunton, Somerset, UK.
3. Joint Universities Pandemic and Epidemiological Research (JUNIPER) consortium.
4. Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK.
5. The Zeeman Institute for Systems Biology & Infectious Disease Epidemiology Research, School of Life Sciences and Mathematics Institute, University of Warwick, Coventry, UK.
6. The Alan Turing Institute, British Library, 96 Euston Rd, London, UK.
7. Department of Engineering Mathematics, University of Bristol, UK.
8. Department of Mathematics, University of Manchester, UK.
9. Clinical Data Science Unit, Manchester University NHS Foundation Trust, UK.
10. Department of Physics and Astronomy, University of Potsdam, Germany.
11. Department of Physics, Humboldt University of Berlin, Germany.
12. MRC Biostatistics Unit, University of Cambridge, UK.
13. COVID Outbreak Surveillance Team, Public Health England, UK.
14. College of Medicine and Health, University of Exeter, Exeter, Devon, UK.
15. Mathematical Institute, University of Oxford, Oxford, UK.
16. Mathematics Institute, University of Warwick, Coventry, CV4 7AL, UK.
17. Lancaster Medical School, Lancaster University, UK.
18. Department of Applied Mathematics and Theoretical Physics (DAMTP), University of Cambridge, UK.  
\*Corresponding author.

## Materials and Methods

### Materials: Description of data sources

We estimated lineage-specific growth rates, age distributions and geographical spread using data provided by Public Health England (PHE) collected as part of the pandemic monitoring effort in England and provided to the Scientific Pandemic Influenza group on Modelling (SPI-M).

We use information on variants of concern (VOCs) or variants under investigation (VUIs) in the UK from four data streams: the positive SARS-CoV-2 cases line list; the S-gene line list detailing TaqPath test results; the VAM line list detailing variants of concern identified by genomic sequencing from COG-UK; and the CTAS (Contact Tracing Advisory Service) line list, detailing genomic sequencing results that are not variants of concern. Descriptive analyses of the frequency of variants of concern among patients who have a S-gene positive TaqPath test result were performed using all four data sources to create a combined genomic and S-gene data set, de-duplicated into unique episodes of infection. For the analysis of growth rates, and associated age distributions, we used only the positive cases line list and the S-gene line list, de-duplicated using 4 different methods, as detailed below, and excluding the last 4 days of case counts, which are subject to reporting delays, and may bias recent estimates of growth rates. In our main results we present findings based on unique episodes of infection, and provide further sensitivity analyses in this section.

Here, we describe each dataset in turn and the data fields that are applicable to the sequencing and/or determining S-gene status of specimens. We later describe the pipeline to combine the multiple data sources.

#### *Dataset 1: Variant line list (VAM):*

This contains a list of specimens sequenced by the Covid Genomics UK Consortium (COG-UK), provided by Public Health England (PHE), that were genomically confirmed as VOCs or VUIs; as of 20 May 2021 in the UK there were 5 VOCs and 7 VUIs (Table 1 of PHE SARS-CoV-2 variants of concern and variants under investigation in England: Technical Briefing 10; (15)). Additionally, each record included the traveller status of the individual (Traveller, Contact of Traveller, Not travel-associated, Refused or Uncontactable, Awaiting information). The VAM is deduplicated to one VOC/VUI call per person. If multiple VOCs per person are identified (rare), non-B.1.1.7 (e.g. B.1.617.2) is prioritised over B.1.1.7 which is prioritised over unclassified.

#### *Dataset 2: CTAS line list:*

This contains genomic information collected through the Contact Tracing Advisory Service (CTAS) and also included traveller status. It lists all sequences that could be linked to the contact tracing system, including sequences that were not VOCs or VUIs. CTAS line list was the only source to contain sequences that were neither VOCs or VUIs.

#### *Dataset 3: "Pillar 1 and 2 line lists":*

These are lists detailing the first case for an individual (i.e. they are deduplicated) within Pillars 1 and 2 of the UK SARS-CoV-2 mass testing programme. Pillar 1 encompasses virus testing in PHE laboratories and NHS hospitals for those with a clinical need, and health and care workers. Pillar 2 contains records of virus tests for the wider population. It records the ethnicity, age and location (to coarse grained geographic level) for each positive case. In addition, each line list record has a categorical value for “Asymptomatic\_indicator” which details the symptomatic status of the individual on the date of testing for community (Pillar 2) tests. A value of “N” indicates the person declared symptoms at the time of testing; and “Y” can be interpreted as tests conducted for screening purposes (i.e. asymptomatic testing). Hospital based testing (Pillar 1) testing is always “U” for unknown, however the majority of hospital testing can be assumed to be due to symptoms, with a small minority due to testing of staff.

#### *Dataset 4: “S-gene line list”:*

For Pillar 2 tests performed using the ThermoFisher TaqPath system, these supplied RT-PCR cycle threshold (Ct) values and their classification according to whether the S-gene target of the TaqPath assay failed to amplify. The TaqPath assay is a multiplex test designed to target three distinct regions of the SARS-CoV-2 genome (N,ORF1ab,S). Each record in the S-gene line list was classified as either S-gene negative, S-gene positive or equivocal, according to these Ct criteria: S-gene negative -  $N \leq 30$  Ct; S undetected;  $ORF \leq 30$  Ct; (also referred to as S-gene dropout) and S-gene positive -  $N \leq 30$  Ct;  $S \leq 30$  Ct;  $ORF \leq 30$  Ct; (also known as “triple-positive”), and equivocal - other combinations. Tests taken during recovery are frequently equivocal when CT values rise.

#### Materials: Processing I: case numbers for growth rate estimates

The S-gene line list is provided on a per-test basis, and some people have multiple S-gene test results. This can be the result of repeated testing during a single episode of infection or multiple episodes of repeated infection, and there is some ambiguity in this. We ensured each infection episode was counted no more than once in the following four ways:

*1) Selecting only cases for whom the S-gene test result is within 4 days of their first ever positive result for that patient (“first infection”)*

This is simple but potentially biases the data, under-representing people who have had multiple infections or for whom the first test in an infection episode was done in a lab that does not use TaqPath tests, and followed up with a TaqPath test later in the infection.

*2) Selecting only cases with a first positive TaqPath test result for an individual (“first taqpath”)*

By taking the first ever TaqPath test result for an individual we maximise the number of individuals we identify but potentially bias the data because the date of the sample may not be completely representative of the date of onset of disease, if the patient was initially diagnosed by a laboratory that does not use TaqPath, or by a lateral flow device. Also cases with repeated infections, separated by a long time would be excluded from analysis.

### *3) Selecting only the last positive TaqPath result for an individual (“last taqpath”)*

By taking the last ever TaqPath test result we also potentially bias the sample by making the time point of infection more recent, in the subset of patients who have had multiple TaqPath test results. However, if these test results are widely separated in time, this strategy may be appropriate as it will pick up repeated infection episodes.

### *4) Inferring continuous infection episodes based on delays between tests (“infection episodes”)*

This strategy involves grouping all known test results, both TaqPath and non TaqPath, together into continuous episodes, containing sequences of positive test results separated by 28 days or fewer (or 56 days in the event of an equivocal S-gene test result). Each episode may contain multiple test results, and if there are TaqPath results within the episode that are positive, and there are no negative S-gene results, the episode is deemed to be caused by a S-gene positive infection. Conversely if there are negative TaqPath test results within an episode and no positive results the episode is deemed to be caused by a S-gene negative infection. It is categorised as equivocal if TaqPath test results are all equivocal, and unknown if there are no TaqPath test results for that episode.

This strategy has the benefit that we correctly identify the onset date of the episode regardless of when the TaqPath testing was done in an episode, and potential episodes of re-infection are detected. Compared to other methods it may appropriately result in earlier infection dates if they have had a few tests done in laboratories that do not use TaqPath tests.

For sensitivity, we tested different datasets obtained with the four alternative deduplication processes (1-4) described here and these are presented in Fig. S1. Consistent results across different methods and data cleaning processes give us confidence on the general trends of S-gene positive and negative cases. Different datasets are also generated for both all cases and those described as symptomatic. Comparative analysis of these datasets are presented in Fig. S2.

## Materials: Processing II: S-gene and genomic sequencing case data

Assembling a linked data set for genomic variant and TaqPath results informs our analysis of case numbers. The data processing pipeline has five main steps:

### *Step 1: Create a joint confirmed sequence case list from the CTAS and VAM line lists*

- From the VAM line list we took proven VOC and VUI cases (i.e. confirmed through sequencing).
- From the CTAS line list we took proven non-VOC/VUI cases.
- We used the following groups: B.1.351, P.1 & P.2, B.1.617.1, B.1.617.2, other VOC/VUI.
- We included the B.1.1.7 and B.1.525 variants within the “other VOC/VUI” category.

### *Step 2: Gather unsequenced S-gene positive cases from the S-gene line list*

- We added entries in the sequenced case list from Step 1 with a FINALID not present to the group of unsequenced S-gene positive cases from the S-gene positive line list, deduplicated using the infection episodes strategy.

*Step 3: Construct a combined list of sequenced VOCs and VUIs (from Step 1) and unsequenced S-gene positive cases (from Step 2)*

- We include both confirmed and suspected cases with VOCs/VUIs of interest.

*Step 4: Using Pillar 1 and 2 line lists, determine if a case was asymptomatic undertaking a test as part of a screening process:*

- Check for match of FINALID and specimen\_date fields in the Pillar 1 and 2 line lists and the Step 3 combined list.
- Extract value from the Asymptomatic\_indicator field: U for Pillar 1; N for Pillar 2 and symptomatic; Y for Pillar 2 and test conducted for screening purposes (i.e. asymptomatic testing)
- Non-matches could occur if it was not the first sample that was sent for sequencing and we could not match the date. This may happen either when a case is a reinfection or if multiple specimens were taken. We recorded these instances as “Unknown” for asymptomatic indicator.

*Step 5: Using Pillar 1 and 2 line lists, determine ethnicity, age and location of a case:*

- Check for a match of the FINALID field between the Pillar 1 and 2 line lists and the Step 3 combined list.
- If a match is found, extracted data from the fields for ethnicity, age and patient location (LTLA level).
- We assume these remain unchanged throughout the study period, acknowledging this will not account for people aging or moving residence.

## Methods: Instantaneous growth rate estimation

We measured the growth rate of the S-gene positive and S-gene negative cases to assess any evidence of differential growth of VOCs. The growth rate describes the exponential rate at which cases are growing or declining in a given area. As opposed to the widely used reproduction number (33), the growth rate can be estimated directly from data and provides a direct measure of the speed of growth of cases regardless of whether they derive from direct transmission or other sources such as importation. It is therefore a more reliable measure to investigate trends when prevalence is low and importation may be significant. From a given growth rate, classical methods allow us to compute the corresponding reproduction number using estimates of the generation time distribution (34–36), but this implicitly makes the assumption that the observed growth is driven entirely by local transmission. Additionally, the generation time distribution of a new variant is often hard to infer from the scarce available data, and may in general be different from that of the currently circulating one.

Especially in situations of low prevalence, as at the time of writing (May 2021), outbreaks can be very heterogeneous across the country. For this reason, in addition to looking at aggregated national figures, which could average out areas seeing rapid spread with others still in decline, we estimated the growth rates in different smaller-scale geographies independently. These local outbreaks can be indicative of the speed of growth of a national epidemic should a variant become widespread across the country, although conditions at the local scale may not translate to larger geographies. However, estimating reliable trends in the growth rate when looking at small scale geographies is challenging due to the very low number of cases involved and the fact that the instantaneous growth rate is undefined when no cases are observed in a given time period.

To identify this problem and associated uncertainty introduced by model assumptions we used five different methods to estimate the instantaneous growth rate, each of them with different assumptions, as described below.

### *Generalised Additive Model method. (Maximum likelihood)*

To estimate growth rates we adapt a generalised additive model (GAM) where the number of cases on day  $t$ ,  $I(t)$ , is assumed to be given by  $I(t) \propto \exp(s(t))$  for some smoother function  $s(t)$ . We use a log link and a penalised spline as implemented in the R package *mgcv* (37). The over-dispersed noise inherent in both disease dynamics and surveillance data motivates the use of a negative binomial error structure, and a day-of-the-week fixed effect is added to capture daily variability within a 1-week period. The number of knots used by the spline is fixed as one twentieth the length of the time-series (for time-series shorter than 200 days the default number of knots is used) to avoid over smoothing the data or losing signal in the noise. The instantaneous local growth rate is then the time derivative of the smoother. The GAM can lead to boundary effects from the choice of smoother, so the most recent central estimates may not be reliable. The growth rate is assumed independent for each geographical area and case definition considered, where case definitions include S-gene positive and S-gene negative. The model used is an extension of the model developed by Pellis et al. 2021 (27).

### *Bayesian Gaussian Process method.*

The growth rate is estimated independently for each geographical area. For a given area  $i$ , the method assumes that the daily count of cases is distributed as a negative binomial function with risk parameter  $\mu_{ti}$ , where  $t$  is a day index. We decompose the log-relative-risk parameter into a Gaussian process (GP) and a weekday random effect:  $\log(\mu_{ti}) = gt + wt$ , where  $w$  has a log-gamma distribution with shape 0 and rate 0.01, and  $g$  is a GP with a Matern covariance function with  $\nu=3/2$ , length scale 1 and precision  $\tau$ . The hyperparameter  $l$  is assigned a log-normal prior with mean 1 and precision 1, while  $\tau$  has a log-normal distribution with mean -3.5 and precision 100. The growth rate is calculated as the first derivative of the GP. To remove fine-scale fluctuations, the derivative is approximated using a centered difference approximation. The model is implemented using the R-package INLA, where the GP is obtained as the weak solution of a stochastic partial differential equation (38, 39).

### *Poisson regression generalised linear model method. (Maximum likelihood )*

The growth rate is estimated for every day using a generalised linear model, where the number of cases on day  $t$ ,  $I(t)$ , is assumed to be given by  $I(t) \propto \exp(rt)$ . A value of  $r$  for a given time is obtained by fitting a quasi-Poisson model to a time slice of case counts, assuming the growth rate is constant over that time period, as a direct estimate of  $r$  as the growth rate.. The model is refitted at each time point using a sliding time window of the data covering the 4 weeks before and 4 weeks after each time point. When considering dates within the most recent 4 weeks there is less data to estimate the growth rate on, so the most recent estimates are both less reliable and more representative of the growth rate in the past. As quite a large window of data is used, this method is slow to respond to step changes in the instantaneous growth rate, as the model assumptions are invalid and the resulting fits can be demonstrated to be low quality at areas of discontinuity.

### *Poisson regression local likelihood method.*

The growth rate is estimated independently in each geographical area. The number of cases on day  $t$ ,  $I(t)$ , is assumed to be given by  $I(t) \sim \text{Poiss}(\lambda_t)$  where  $\lambda_t$  is defined as an order 1 polynomial determined by maximising the local likelihood of a quasi-poisson model, to account for overdispersion, using the nearest 28 data points, and a logarithmic link function, as implemented in the R package locfit (40). The local maximum likelihood estimation is conducted on data weighted by day of week, with Saturday, Sunday and Monday down-weighted compared to Tuesdays to Fridays to account for fluctuation in testing rates at the weekend. The time derivative of the locally fitted polynomial is a direct estimate of  $r$  as the growth rate, and is evaluated at every time point. At the end of the time-series there is less future data to estimate the growth rate on, so the most recent estimates are based on more past data. This method is similar to the one above, but adjusts to step changes more quickly. However it may overshoot in these situations.

### *Bayesian Poisson method.*



The same approach as the poisson regression was also implemented in Bayesian framework using an observation level random effect, rather than quasi-poisson error distribution to account for the overdispersion. This was fitted in the R package brms v 2.15.0 (41). The priors used were a normal (mean 0, standard deviation 5) prior on the intercept of the model, a standard normal prior on the growth rate ( ), and an exponential (lambda=1) prior on the standard deviation of the observation random effect distribution. As in the poisson regression method, the surrounding 8 weeks of case counts were used, and the caveats about the most recent 4 weeks of data apply. This method was used to validate the main estimates in Fig. 2 of the paper, and is not being used in the sensitivity analysis.

Figs. 2 show estimates of the instantaneous growth rate obtained using the five different methods.

### Methods: Reproduction number ratio

The reproduction number ratio is an estimate of the transmission advantage of S-gene positive over S-gene negative cases ( $T_{1>2}$ ). It is estimated from the growth rate, where the relationship between reproduction numbers and growth rates is determined by the shape of the generation interval distributions as described by Wallinga et al. (2007) (42).

$$T_{1>2} = \frac{R_1}{R_2}$$

More precisely, if  $R$  is the reproduction number of infections caused by a given variant,  $r$  is the growth rate of cases of that variant and  $M$  is the moment generating function of the generation interval distribution, then the following relationship holds

$$R = \frac{1}{M(-r)}.$$

We assume that the generation interval  $G$  is Gamma distributed with shape  $\alpha$ , rate  $\beta$ , then the mean ( $\bar{G}$ ) and variance ( $\bar{S}^2$ ) are given by:

$$\bar{G} = E(G) = \alpha/\beta$$

$$\bar{S}^2 = V(G) = \alpha/\beta^2$$

Assuming no change in the generation interval between variants,  $R$  simplifies to:

$$R = \left(1 + \frac{r}{\beta}\right)^\alpha$$

And hence the reproduction ratio advantage resolves to:

$$T_{1>2} = \left(\frac{\beta+r_1}{\beta+r_2}\right)^\alpha$$

Solutions only exist when  $r > -\beta$ , or equivalently  $r > -\bar{G}/\bar{S}^2$ . Small growth rates are not possible to obtain for distributions of  $G$  with long heavy tails, since a rapid decay cannot be explained by large generation times. To estimate  $T_{1>2}$  for low growth rates, the distribution of  $G$  must be truncated.

We employed an empirical form for  $G$  as proposed in (42), where the range of the generation time ( $a$ ) is divided into 21 intervals with bounds  $a_0, \dots, a_n = 0, 0.5, 1.5, \dots, 19.5, 25$ . This does not require that we assume a gamma distribution and allows us to compare parametric with empirical distributions for  $G$ . If  $y_1, \dots, y_n$  are the associated probabilities in the periods ending at  $a_1, \dots, a_n$ , we obtain:

$$R = \frac{r}{\sum_{i=1}^n \frac{y_i}{a_i - a_{i-1}} (e^{-ra_{i-1}} - e^{-ra_i})}.$$

The estimates from the continuous and the discrete versions of these equations differ by a small amount (mean continuous = 1.61; mean discrete = 1.59; RMSE: 0.022) which reflects the fact

that some combinations of  $r$  and generation time cannot be calculated with the continuous equations, and there are some differences due to the inclusion of empirical generation time estimates. The discrete estimates are regarded as the better estimate and continuous estimates discarded.

Estimations in Table 1 are obtained using the generation interval assumptions from posterior samples from Challen et al. (2020) (43), Ganyani et al. (2020) (44), Hart et al. (2021) (45) independent and mechanistic models (45). We also used additional empirical, household and intrinsic estimates from Hart et al. (personal communication) which are generated using the same methodology as their previous work, but applied to UK household data containing differential recording of Alpha versus Delta cases. These contain different generation time estimates for the 2 variants (see Fig. S16). Where the generation time estimates were non-parametric and given as a mean and standard deviation, we assumed a Gamma distribution with the same mean and standard deviation. For each of these 7 sources, 100 sample estimates of the generation time distribution are selected at random in pairs for both S-gene positive and S-gene negative groups. For the empirical, household and intrinsic Hart estimates the pairs are different distributions for S-gene positive (Delta) and S-gene negative (Alpha) groups, for the other sources they are assumed to be unchanged between S-gene positive and negative groups.

Each of these 700 paired generation time distribution estimates were used to estimate the reproduction number for S-gene positive and negative cases on each day, for each of the first four growth estimation methods described above (the bayesian poisson method was excluded as we did not have growth rate estimates for all geographies). We restricted transmission advantage calculations to the “symptomatic” cases and the “infection episodes” dataset.

This results in 2800 estimates of the reproduction number ratio for every day, with a varying combination of growth rate model, and generation time distribution source. Summary statistics and confidence limits were empirically estimates from these ungrouped (table 1 main paper); grouped by day over the “stabilised period” (Fig. 4 - see later in supplementary text for definition of stabilised period), and grouped by growth rate model and generation time source (see Fig. S16, table S1, and table S2).

#### Methods: Comparison of the distribution of cases by age

Individuals’ transmission patterns typically depend on age, and interactions between different age groups can drive epidemics. Using empirically-derived age mixing matrices describing those interactions, next generation matrices can be calculated (46), which map the age distribution of infected individuals in a population to the age distribution after one generation of infections. As the epidemic progresses, the age distribution of infected individuals converges to the dominant eigenvector of the next generation matrix (47), provided interaction mixing and transmission patterns remain constant, and assuming the same generation time distribution for all ages. Under the assumption that mixing and transmission patterns (i.e. relative susceptibility and infectivity by age) are similar for all variants of a pathogen up to a multiplicative constant describing an overall increase or decrease in transmissibility, we would expect cases across different variants to have the same age distribution, regardless of whether

the epidemic is growing or declining. However, when a new variant emerges, it may emerge preferentially within and between certain age groups, though as the variant-level epidemic progresses the age distribution should converge to the same dominant eigenvector. In the context of the Delta variant, many cases arrived in the UK through travel, possibly resulting in an age bias in Delta variant cases aligned with the age distribution of travellers. Following these seeding events, we may observe three primary outcomes: continued growth through importation of cases, continued growth through community transmission, or local extinction of the variant. If we observe continued growth through importation, the age distribution of cases will reflect those of the imported cases. If continued growth occurs through community transmission, the age distribution of cases should gradually shift towards the dominant eigenvector, which describes the mixing and transmission patterns in the community. If local extinction occurs, the cases should die out, and we again expect a return to the dominant eigenvector, but without a corresponding growth in cases. Therefore, by studying the age distribution of different variants in tandem with the growth rate, we can gain insight into whether there is community transmission.

We consider cases at a relatively small spatial aggregation level of a Lower Tier Local Authority (LTLA), since the age-mix of cases may vary across wider spatial regions. For the LTLA that experiences the most substantial discrepancy between the two age distributions in the six areas considered in Fig. 3, the dynamics of perturbation and equilibration are summarised with the Wasserstein (or Kantorovich) distance metric (48, 49), which captures discrepancies between the age distribution of S-gene negative and positive samples, (see supplementary text and Figs. S3-S4 for additional details, all other LTLAs within these regions are shown Figs. S7-S9). These metrics are sensitive to sample size, so, to assess uncertainty, we generate a 95% significance level confidence region for the null hypothesis that S-gene positive and S-gene negative have the same age distribution, based on a permutation test. If the distance metric falls outside this confidence region, the null hypothesis is rejected, and we conclude that the distributions are different.

Given a metric space  $M$  provided with a metric  $d$ , the 1st Wasserstein distance, or ‘earth mover’s distance’, or Mallow’s distance, intuitively describes the minimum cost of transforming one probability distribution to another. Let  $X$  and  $Y$  be random variables with distributions  $P$  and  $Q$  in  $R^d$ , respectively. The Wasserstein distance between  $P$  and  $Q$ ,  $M_p(P, Q)$ , can be defined as the minimum of the expected difference between  $X$  and  $Y$ , taken over all joint probability distributions  $F$  for  $(X, Y)$  such that the marginal distribution of  $X$  is  $P$  and the marginal distribution of  $Y$  is  $Q$ , i.e.,

$$M_p(P, Q) = \min_F \{ (E_F \|X - Y\|^p)^{1/p} : (X, Y) \sim F, X \sim P, Y \sim Q \},$$

We consider the 1st Wasserstein distance, defined for  $p=1$ . Some of the advantages of the Wasserstein distance rather than other measures, e.g., the Kullback–Leibler divergence, in the context of age distributions, are: 1) the Wasserstein distance satisfies the properties of a metric (e.g., it is symmetric); 2) it is well defined for probability distributions with different support;

3) it accounts for the relative position of points in the parameter space, providing a measure of distance not only in a probability sense, but also in a metric sense. For discrete distributions, the latter translates into the fact that non-adjacent bins are considered being further away than adjacent bins.

To ensure conclusions are not driven by the choice of distance metric, we also use the Kolmogorov-Smirnov distance measure. This is defined as the largest absolute difference between two cumulative distribution functions, i.e.,

$$D = \sup_x |F(x) - G(x)|,$$

where  $F$  and  $G$  denote the cumulative distribution functions of the probability distributions under study. The Kolmogorov-Smirnov test is extremely easy to compute, but suffers from some drawbacks including that it focuses on the point of maximal difference rather than measuring more global differences between distributions. We here use this measure only as an additional validation of the tests performed using the Wasserstein distance.

To compare two empirical distributions (e.g., the observed age distributions of S- gene positive versus S-gene negative cases), we consider the null hypothesis that the two samples are drawn from the same probability distribution. We do this by using a permutation test: we combine the two samples into a single distribution, from which we randomly draw 1000 permutations that split this single distribution into two samples, with the respective sample sizes of the original two samples. For each pair of permutation samples, we calculate their (Wasserstein or Kolmogorov-Smirnov) distance. These distances represent possible values that could be obtained if the two samples were drawn from the same distribution. From the 1000 permutation distances, we can calculate percentiles in order to obtain a confidence region. We use the latter to determine whether the observed samples are likely to have been drawn from the same distribution: if the distance calculated from the original samples lies within the confidence region, then there is insufficient evidence to reject the null hypothesis that the two samples were drawn from the same distribution, at the corresponding confidence level.

Fig. S3, focuses on Leicester as an example. The bottom three panels show a snapshot of the age distributions at given points in time. Panel (a) shows the age distribution of cases between 23rd March and 5th April 2021. Visually comparing the S+ and S- age distributions, there does not seem to be a significant difference between the two, at least within the binomially distributed uncertainty bounds. This is in agreement with the fact that the Wasserstein distance for the corresponding date range falls well within the confidence region. The second panel shows the age distributions between 5th-18th April 2021. There is now a more evident difference in the observed age distributions, with S+ cases occurring in much older individuals. The corresponding Wasserstein distance is outside of the confidence region, confirming the significant difference between the two age distributions. The final panel shows the distributions for cases between 23rd April and 6th May 2021. Visually the two age distributions appear similar again, and the Wasserstein distance is back within the confidence region, so there is no evidence to suggest a significant difference between the distributions. Therefore, the relative position of the computed Wasserstein distance and the confidence region obtained by a

permutation test allows quantification of the visual relationship between the age distributions, and indicates whether such distance is significant based on the sample sizes involved. This is particularly important when either one or both sample sizes are low and visual inspection may be difficult. In this case, the Wasserstein distance can be very high even when the two samples are drawn from the same distribution, leading to large confidence regions and therefore allowing to quantify the confidence in the difference between distributions.

To investigate temporal changes in the relationship between the two age distributions, we considered the age distributions over a rolling time window. We opted for a 14-day time window, as this ensured that there were sufficient cases to get some insight into the age distribution, whilst still having a fine-grained temporal resolution. Using a longer window would result in a temporal correlation of long duration, so short-term, but substantial perturbations to the age distributions could be missed. This method appears well calibrated at LTLA level, with the results shown in Figs. S7 to S9 and Fig. 3 in the main text. Statistically significant Wasserstein distances are only observed when there is a clear difference between the age distributions. Looking back to October 2020, there are very few significant differences, with most arriving during periods of change, such as the easing of lockdown restrictions in December 2020 and implementation of lockdown in January 2021, which will have perturbed mixing patterns in the community. We only consider LTLA level, as across wider spatial regions we might not observe the same age distributions, due to differences in local demographics, though they are likely to be relatively similar. However, these small discrepancies could lead to the model suggesting significant differences in the age distributions, whereas in reality S-gene positive and S-gene negative cases could be spreading in separate subregions. Performing this for the regions of interest, for the Wasserstein distance, gives the results shown in Fig. 3 in the main paper. To verify that our conclusions regarding comparisons between the data streams are not driven by the assumptions of the Wasserstein distance, in Fig. S4, we show the same analysis for the Kolmogorov-Smirnov distance alongside the Wasserstein distance. Comparing the results, both metrics suggest similar conclusions. Therefore, our conclusions are not driven by the choice of model.

## **Supplementary Text**

### Age distribution of cases among travellers

Cases in travellers may not reflect the age distribution of community acquired cases. To investigate this, we compare the subset of PCR positive cases in confirmed travellers to all PCR positive cases (Fig. S5). Both samples consider cases between 28/02/2021 and 24/05/2021, to ensure the time frames are comparable. The age distribution of traveller cases are skewed to older individuals, suggesting that the perturbations seen in the age distribution of S gene positive cases are likely to be caused by imported cases.

The age distribution of travellers arriving in the UK is assumed to be comparable to the ages of those leaving the country. To this end, we use the ages of travellers participating in the Civil Aviation Authority's Departing Passenger Survey from 2019 (50) (note the 2020 survey was cancelled), though we acknowledge that the distribution may be biased by variable survey

engagement rates. We consider the six English airports with the highest passenger counts in 2020: Birmingham International (BHX), London Gatwick (LGW), London Heathrow (LHR), London Stansted (STN), Luton (LTN), and Manchester (MAN) (51). In particular, we highlight LHR and MAN as these are presumed to be the most likely final destination airports for long-haul international flights from India, and cover both the London and North West regions where many Delta cases were initially detected.

Overall, age of travellers varies with the flight type and reason for travel, but is generally distinct from the age distribution of the general population of England as per the ONS 2019 estimates (52) (see Fig. S6). There are higher proportions of travellers aged 20-64 than expected from the English population, a picture which is consistent with individuals who also travel for work and/or have greater levels of disposable income for leisure travel. Overall, the mean age of travellers is seen to be higher than the median age of the population.

In early summer 2021 a similar picture was emerging in other European countries. Denmark, Germany and the Netherlands, were seeing Delta variant cases growing on a background of predominantly Alpha infections (over 90% of cases in each country). In Denmark, where up to 91% of cases are sequenced, 36% of Delta variant cases had recently travelled, yet the rapid growth in cases across all regions of Denmark suggests that Delta variant and its sub-lineages has shifted to spreading beyond the household and close contacts of identified travellers. As of the 30th of May, Delta was identified in 0.5% of cases (53). Germany and the Netherlands are sequencing a smaller proportion of cases (10% and 3% respectively), and report an increase in Delta variant cases in early May. In Germany, the Delta variant accounts for 2% of cases in late May, with less than 0.5% in the Netherlands. This is consistent with travel-related importation of Delta and a delay in reporting.

#### Comparing age distributions of S-gene positive and S-gene negative cases at LTLA level

In the main text, we compare the age distribution across six LTLAs where we have identified significant perturbations to the S-gene positive age distribution. In this section, we consider the remaining LTLAs within the six areas of interest (Bolton and Blackburn, Nottingham region, Leicester region, Bedford region, East London, and West London). The results for the resultant 18 LTLAs are shown in Fig. S7-S9. We observe that during April and May the only significant perturbation, outside of the six regions shown in Fig. 3 in the main text, occurs in Blaby. However, this appears to be driven by very few S-gene positive cases rather than a genuine change to the age distribution. In the LTLAs shown, many regions observed perturbations prior to April. These generally seem to correspond to B.1.351 spikes in the areas, so could correspond to importation of these cases.

#### Growth rate in England and NHS Regions from late January to late June

In England, the estimated instantaneous growth rate of S-gene negative cases appears relatively stable and below zero, from the beginning of February to mid May 2021, indicating that cases were consistently declining (Fig. 2A). Conversely, a clear increase in S-gene positive cases since the beginning of April (Fig. 1A), is confirmed with growth rate estimates remaining

above zero (Fig. 2A), with doubling times as short as 7 days in early May 2021. The transition from comparatively lower growth rates of S-gene positive cases at the beginning of the time series, to comparatively higher growth rates of S-gene positive cases at the end of the time series mirrors the transition seen in Fig. 1B from primarily “non VOC/VUI” cases at the beginning to Delta (B.1.617.2) cases towards the end of the observed period. This pattern is seen earliest at the local level in Fig. 2 but subsequently was observed at longer time-scales and at broader geographic scales, England wide (Fig. S10) and at the level of NHS regions (Figs. S11-S12), which extend out to late June.

In panel B of Figs. S10-S12 we see the proportion of variants found when sequencing S-gene positive cases. Solid vertical lines are marked that indicate the date from which in each region we can state with confidence that the vast majority of S-gene positive cases were subsequently found to be Delta (B.1.617.2) on genomic sequencing. To be precise this marks the time when both the upper binomial confidence interval for the probability of the Delta variant being detected given an S-gene positive case is greater than 0.95 and the lower confidence interval for the probability of the Delta variant being detected given an S-gene positive case is greater than 0.7.

In panel C of Figs. S10-S12 we see the proportion of cases that are either S-gene positive, negative, equivocal or unknown. At this point in time in England, S-gene negative cases were the Alpha variant (B.1.1.7) almost without exception. The dashed vertical line on Figs. S10-S12 therefore represents the date within each region after which we can say with certainty that the Alpha variant becomes uncommon. To be precise this marks the time when both the upper binomial confidence interval for the probability of a S-gene negative case being detected given a positive SARS-CoV-2 PCR test is less than 0.1 and the lower confidence interval for the probability of the S-gene negative case being detected given a positive SARS-CoV-2 PCR test is greater than 0.05.

The period between the solid vertical line and the dashed vertical line are described in the main text as the “stabilised period” during which we can make comparisons between the growth rate of S-gene positive and S-gene negative cases and infer that they relate to Delta and Alpha infections respectively. These are the time periods referred to in table 1 in the main text, and tables S1-S2 in this supplement.

The “unknown” S-gene test results are a result of 2 processes. Either infections which are first detected in hospital, or community cases for which TaqPath testing was not performed. It has to be assumed that the cases that have an “unknown” S-gene status are composed of a similar proportion of Alpha and Delta cases, as those for whom an S-gene status is known, but this is not able to be proven. The degree of TaqPath testing is not uniform across England over the time period studied, as is seen in Fig. S14. The areas with the highest TaqPath testing coverage and hence least uncertainty due to tests with “unknown” S-gene status, include the North West of England, the North East and Yorkshire and the Midlands.



### Selective advantage and generation time

The ability for a novel variant to spread more rapidly than previous variants is a selective advantage. This may be due to enhanced viral replication within a host, enhanced transmission between hosts, or a novel ability to evade immune responses. A transient selective advantage can also be the result of founder effects when a novel variant emerges in a community which is particularly favourable for rapid spread (2). However a new variant may also appear to be spreading more rapidly, if the same number of cases are infected, but the speed at which they are infected is quicker, due to a reduction in the time between sequential infections (the generation time).

In the assessment of the potential advantage of the Alpha variant over previously circulating variants (7) there was the possibility that the increased growth rate of the Alpha variant was solely due to a reduction in generation time, as at the time of identification both previously circulating and Alpha variants were increasing. As soon as the previously circulating strain started to decline during November 2020, whilst Alpha cases continued to rise, it became apparent that the selective advantage Alpha had over previous strains could not be solely due to a reduction in generation time.

The Delta outbreak we studied happened during a phase where the previously dominant Alpha variant was already in steep decline due to social distancing measures put in place in early January. As a result there was no question that the selective advantage shown by Delta could be solely due to a shorter generation time. However in the early part of the outbreak it has been equally difficult to tease apart the contribution of generation time, transient founder effects, and biological transmission advantage to the overall picture of the observed growth advantage enjoyed by Delta. Due to the rapid decline in the number of cases of the Alpha variant the time period over which we can get further information to study this phenomenon in England is now over.

The existence of the estimates of generation time selective for Alpha and Delta variant from Hart et al. (personal communication), and an assessment of the selective advantage in community spread at large scale after initial outbreaks have passed, allows us to establish the relative importance of generation time, founder effects, and evolutionary transmission advantages. In Fig. S16 and table S1, the empirical, household, and intrinsic estimates of the generation time are variant specific (see also Fig. S15). These generally produce lower reproduction number ratios than are obtained from the other 4 generation time sources. Similarly the estimates of the reproduction number ratio over time (Fig. 4 main text) show a fairly consistent pattern between regions of an initial hump followed by a subsequent “stabilised period” during which we estimate reproduction number ratios. These patterns are clearest for England, The North West and the East of England. Taken together we could assert that the best estimates for the transmission advantage of Delta over Alpha would be those estimates that use the variant specific generation time estimates, during the stabilised periods in England, the North West and the East of England. The estimates are shown in table S1, have central estimates that range from 40% (East of England / household) to 72% (North West / intrinsic) and they conveniently bracket the central estimate presented in the main text.

In table S3 we compare an estimate of the generation time using Hart et al. (personal communication) intrinsic estimates, which differentiate between Alpha and Delta, by both calculating the reproduction number ratio assuming there are variant specific differences in generation time and then assuming there are no differences in generation time between the variants and they are either fixed at the value used for Alpha (Alpha only) or fixed at an average value between the two (Combined). This gives us three estimates: 1.76 (using Alpha generation time for both variants), 1.69 (using average generation time for both variants) and 1.66 (using variant specific estimates).

The gap between the estimates based on an average generation time and the variant specific generation time is smaller than we anticipated, at 0.03. This might be explained by the fact that the reproduction number for Alpha, which is under 1 over this timeframe is pushed further away from 1 by the larger Alpha specific generation time compared to the average generation time, whereas the reproduction number for Delta, which is above 1 over this timeframe is pulled towards 1 by its shorter Delta specific generation time compared to the average generation time, thus the ratio between the two reproduction numbers is less affected than if the generation time for Delta were larger than that for Alpha.

More relevant possibly, is the comparison between the variant specific estimate (1.66) and the estimate made using only the Alpha variant generation time, applied to both variants (1.76). This gives us an indication of the degree of overestimation of the transmission advantage resulting from the assumption that the variants have the same generation time when in reality Delta's generation time may be shorter. On this evidence, the assumption that there is no change in the generation time between the two variants might account for 10% of the resulting estimated transmission advantage (8% - 14%).

The observed biological transmission advantage represented as the ratio between effective reproduction numbers for the two variants, can be due to a difference in the basic reproduction number, or a difference in the reduction of the effective reproduction number conferred by immunity. In the second case this is either escape from naturally derived immunity or vaccine derived immunity. Our age analysis shows that there is no evidence of a persistent change in age structure in the S-gene positive and S-gene negative cases as might have been expected if vaccine escape is a significant driver for infection during our observation period. It remains possible though that the time scales on which we observe this are too small and vaccine escape is a longer term driver of Delta growth.

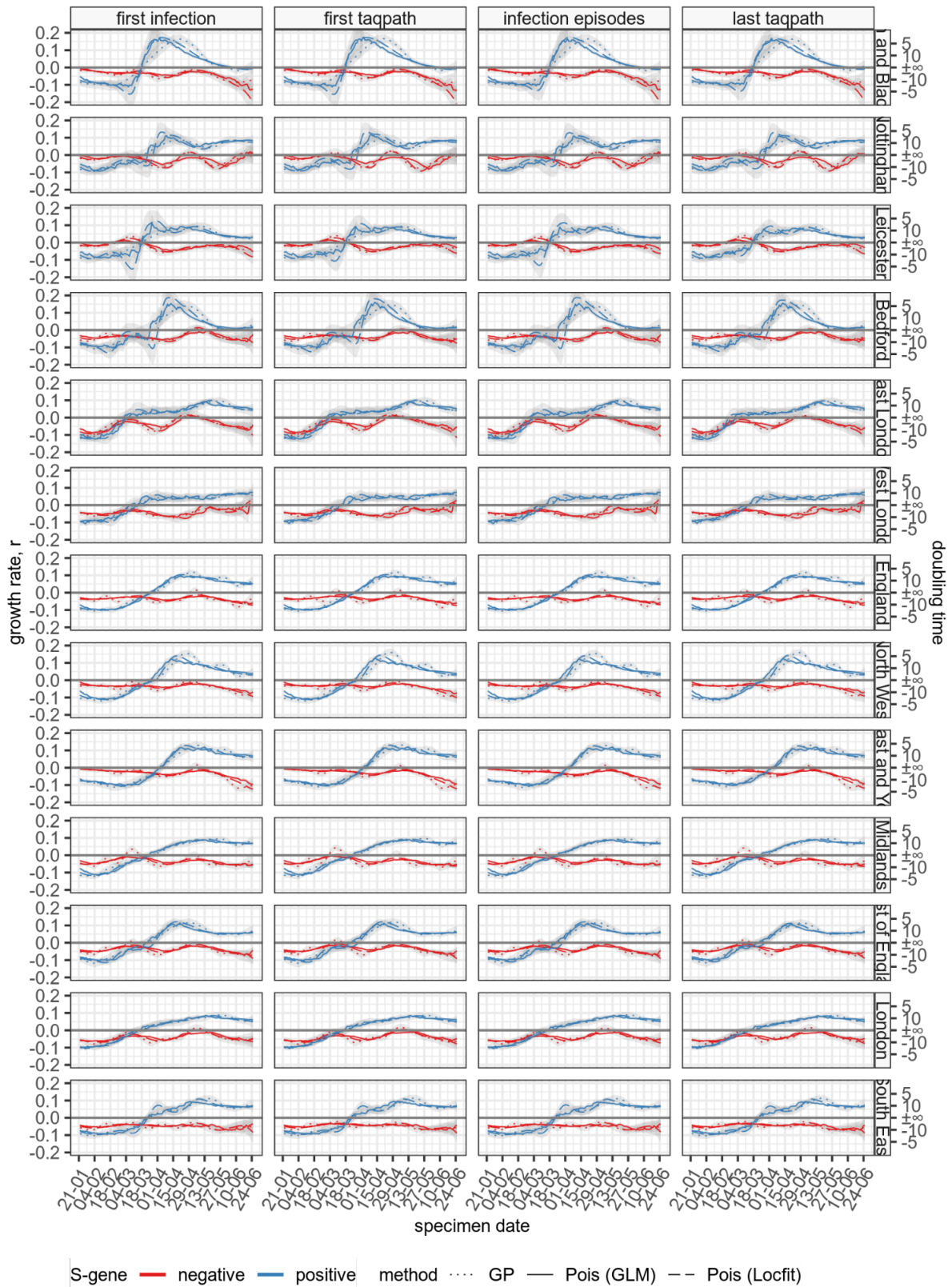
### Limitations

Our study has remaining limitations. Pre-B.1.1.7 wild-type SARS-CoV-2, and other variants such as B.1.351 and P.1, are also detected in the S-gene positive signal, particularly earlier in the time series, so it remains possible that increased transmissibility inferred from the S-gene positive signal is misattributed to B.1.617.2.

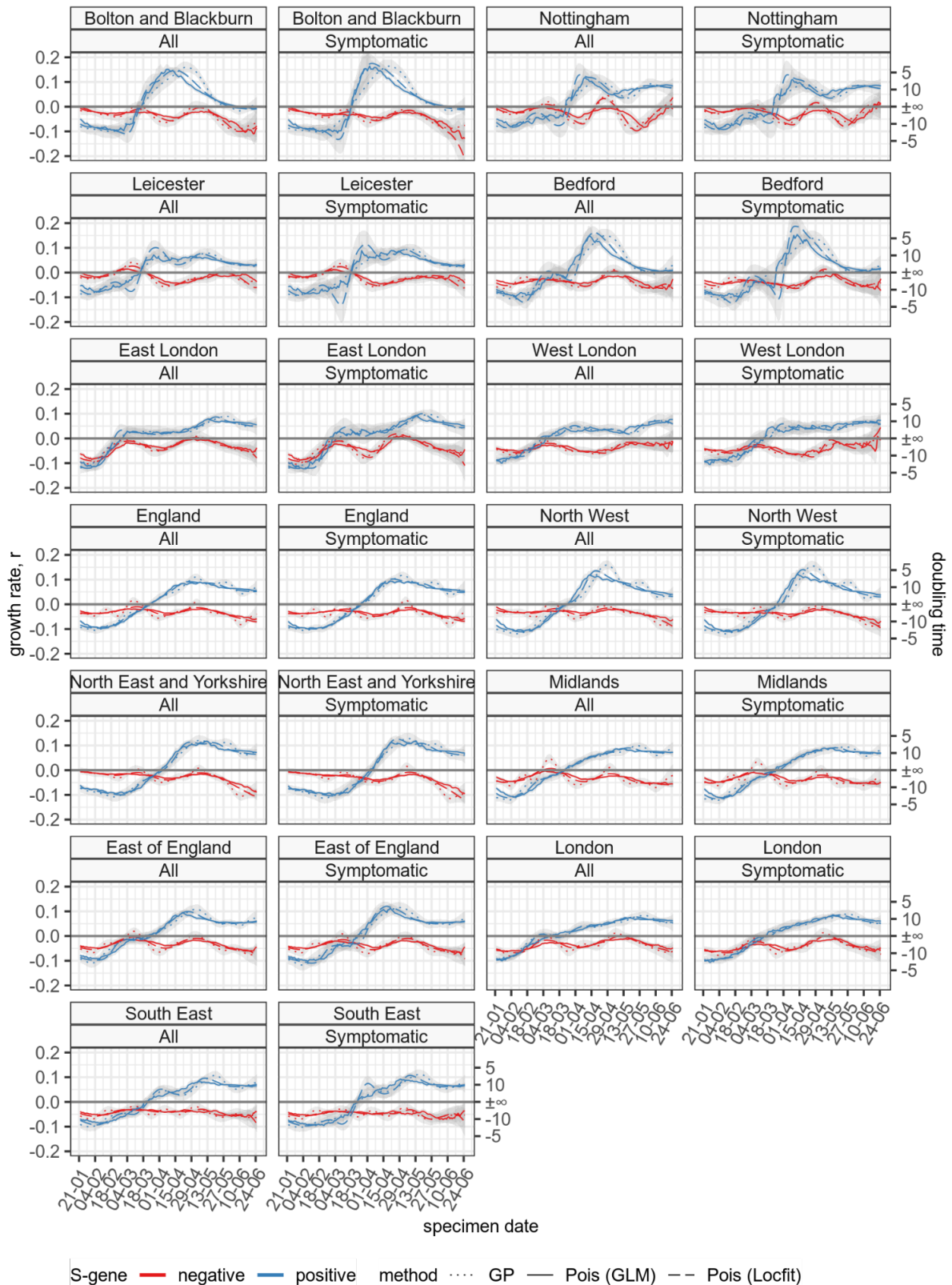
TaqPath testing coverage is not complete or uniform, and we may be underestimating the number of S-gene positive or negative cases. The proportion of Pillar 2 cases that have a S-gene result is higher in the North West and Midlands than in the South West and South East of

England (see Fig. S14). There is also the possibility of differential case acquisition rates between S-gene positive and S-gene negative cases which may be exacerbated in the shorter time frames and small geographic regions by surge testing conducted by public health.

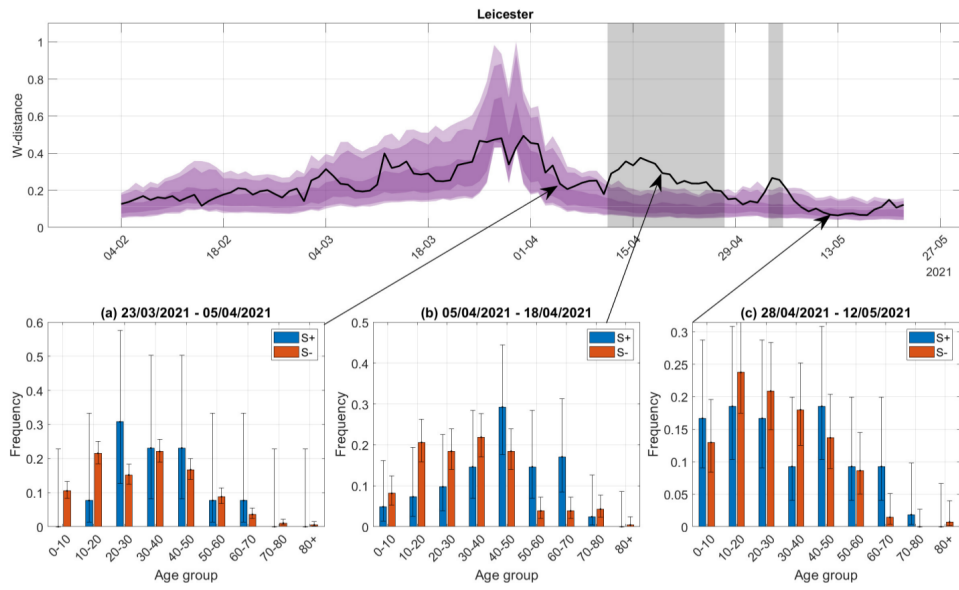
We do not have traveller status data linked, unless the case has been sequenced, so there is a remaining bias from imported cases in the S-gene positive data, which could inflate initial estimates of growth, although in the latter part of the period studied travel restrictions and quarantine measures were in place.



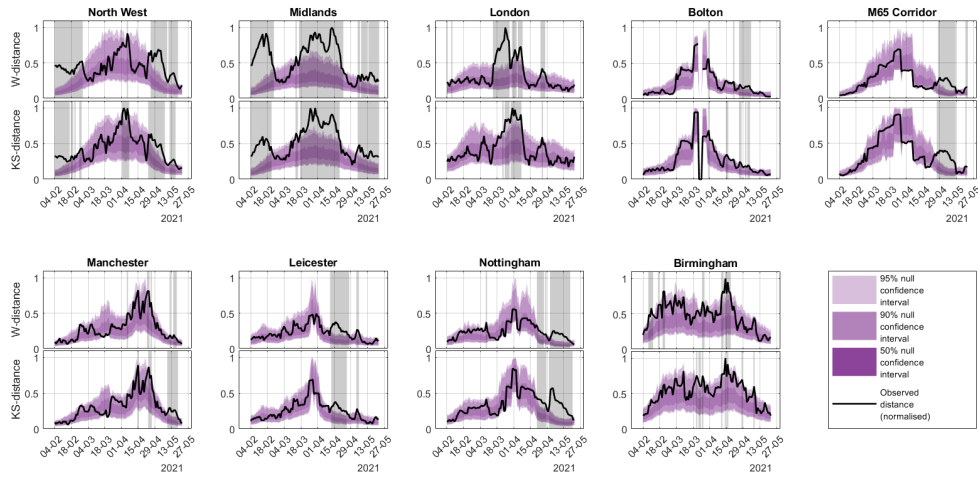
*Fig. S1 - Deduplication strategy sensitivity analysis: Instantaneous growth rates and doubling times for all regions estimated using three different methods and data obtained using the four different deduplication strategies described above.*



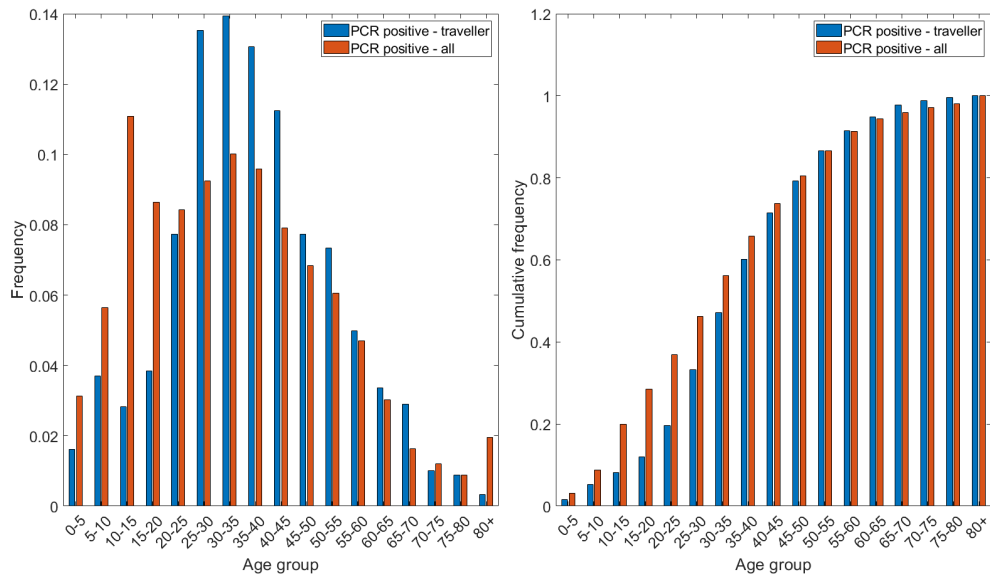
*Fig. S2 - Case subtype selection sensitivity analysis: Instantaneous growth rates and doubling times for all regions estimated using three different methods using all identified cases, versus the subset of cases described as “symptomatic”.*



*Fig. S3: Comparing the Wasserstein time-series to snapshots of the observed age distributions for S-gene positive and S-gene negative cases in Leicester. This figure illustrates how the Wasserstein distance relates to the observed distributions, and how we should interpret the time-series and confidence region.*

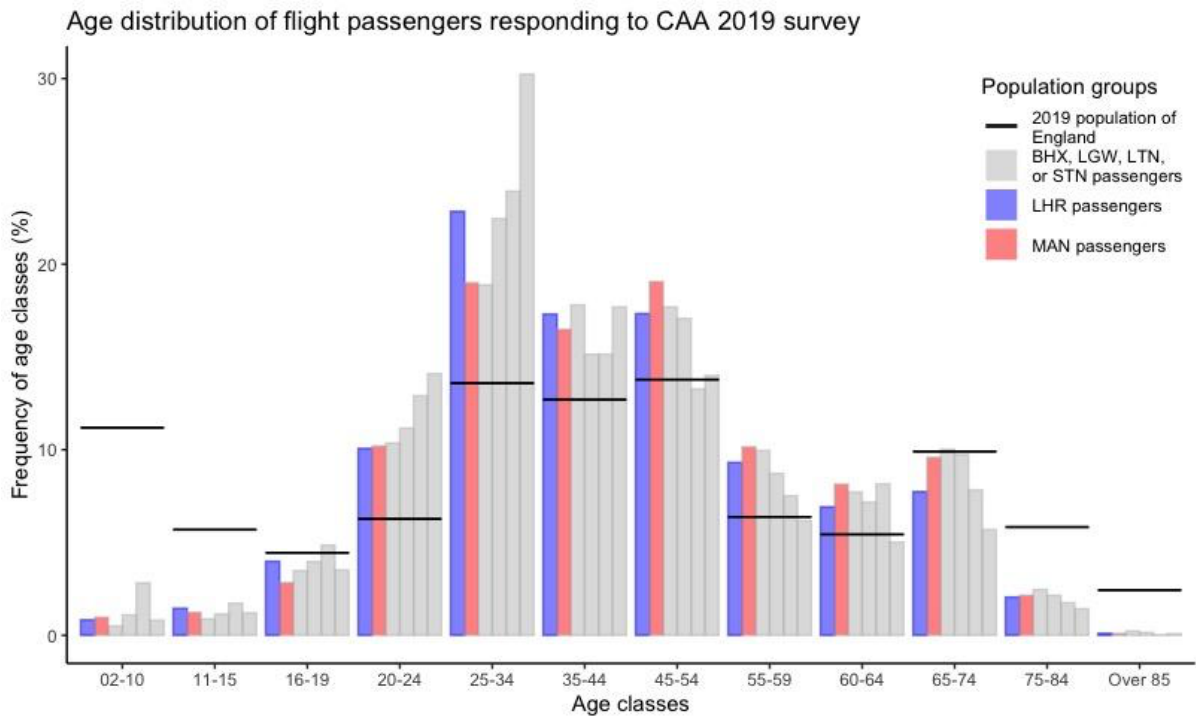


*Fig. S4: Sensitivity analysis: Comparison of the Kolmogorov-Smirnov distance between the S+ and S- age distributions and the Wasserstein distance, in the areas of concern. This considers the age distribution among a two-week rolling aggregation of cases, plotted by first swab data in the two-week window. The black curve indicates the distance between the two age distributions. The blue shaded region is generated through 1000 Monte Carlo samples of a permutation test, and indicates the confidence region for distance metrics if the two samples were drawn from the same distribution. The grey shaded regions indicate dates where the distance metric falls outside of the confidence region, denoting a significant difference between the age distribution of S-gene positive and S-gene negative cases.*

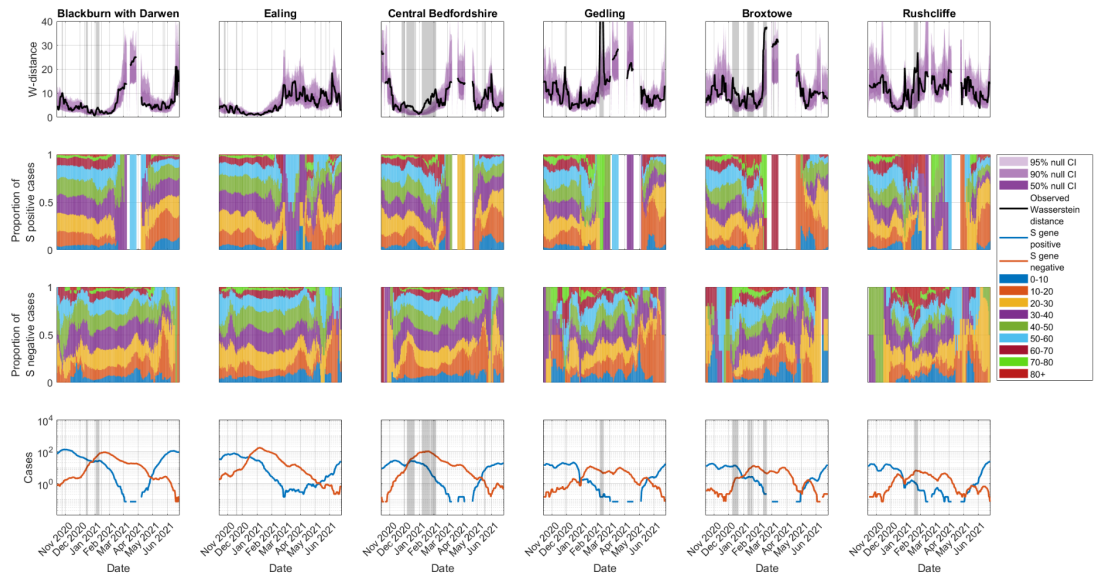


*Fig. S5: Age distribution of cases in travellers versus all cases. The left plot shows the proportion of cases in each age group, and the right plot shows the cumulative frequency up to each age group. From the left plot, the traveller age distribution appears skewed towards older individuals. The right plot clarifies this, with the age distribution of all cases having a substantially higher cumulative frequency for lower ages.*





*Fig. S6: Age distribution of flight passengers from the six busiest English airports, as estimated by the CAA 2019 passenger survey. We highlight London Heathrow (LHR) and Manchester (MAN) airports as we presumed those to welcome the largest volume of long-haul passengers from India, but note that the qualitative features of the age distribution are preserved across other airports. The horizontal bars indicate the proportions of each age group in the 2019 mid-year ONS population estimates (52). People aged 20-64 are more likely to travel than other age groups.*



*Fig. S7 - Dynamics of age distributions of cases for Blackwith with Darwen, Ealing, Central Bedfordshire, Gedling, Broxtowe, and Rushcliffe. For each region, we show the Wasserstein distance between the age distribution of S-gene positive and S-gene negative cases (top panels), the age distribution of S-gene positive cases (second panels), the age distribution of S-gene negative cases (third panels) and the number of cases (bottom panels). All metrics are computed over a 14-day rolling window of cases by specimen date, plotted by first swab date. The black curve indicates the Wasserstein distance between the two age distributions. The purple shaded region indicates the confidence region for the Wasserstein distance if the two samples were drawn from the same distribution, generated through 1000 Monte Carlo samples of a permutation test. The grey shaded regions indicate dates where the Wasserstein distance falls outside of the 95% confidence region, denoting a significant difference between the age distribution of S-gene positive and S-gene negative cases. The red curve indicates S-gene negative cases and the blue curve indicates S-gene positive cases.*

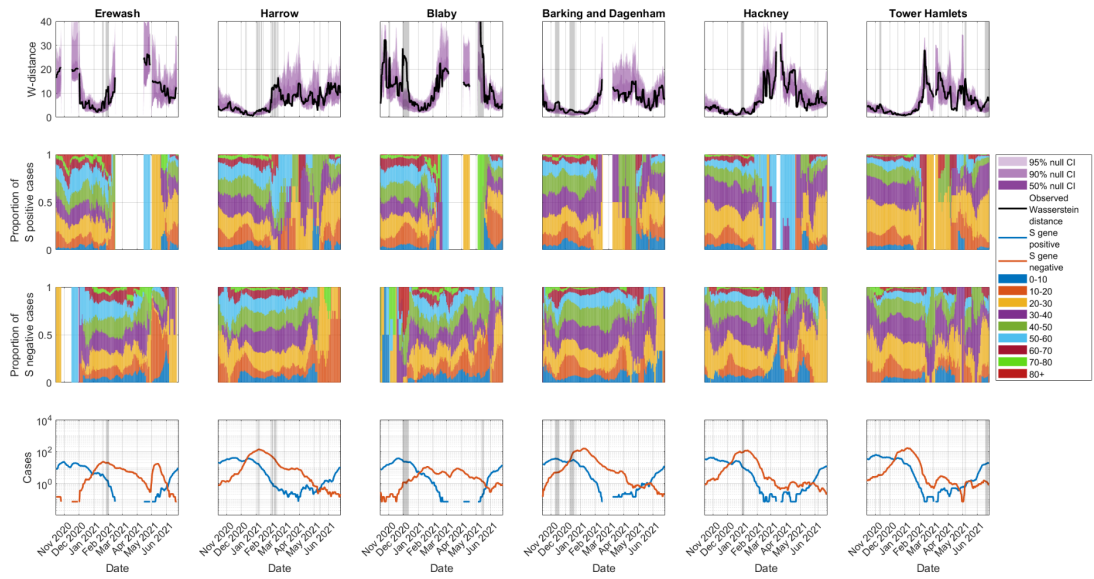
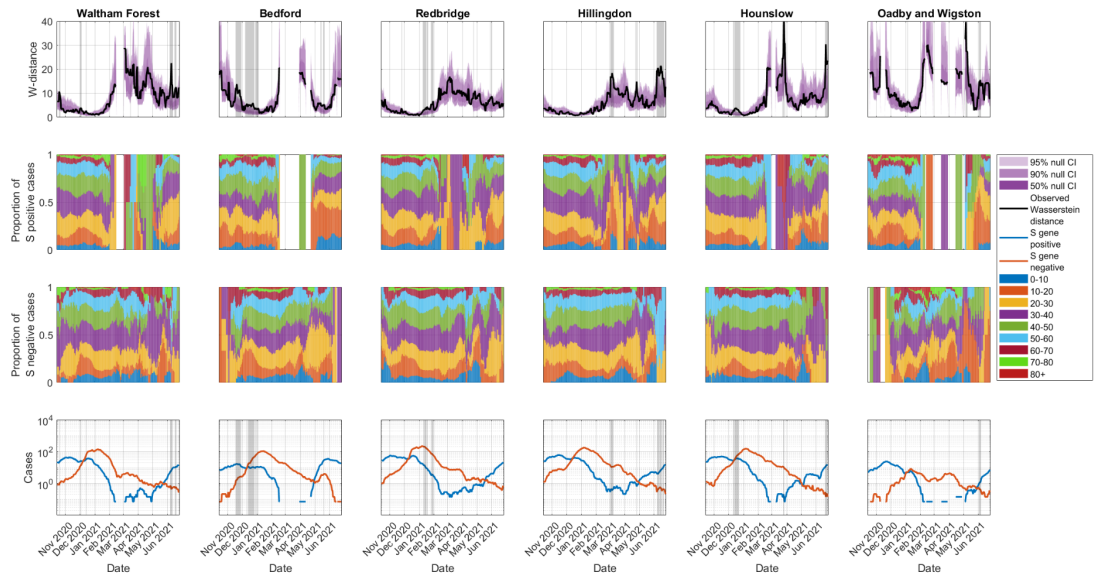


Fig. S8 - Dynamics of age distributions of cases for Erewash, Harrow, Blaby, Barking and Dagenham, Hackney, and Tower Hamlets. For each region, we show the Wasserstein distance between the age distribution of *S*-gene positive and *S*-gene negative cases (top panels), the age distribution of *S*-gene positive cases (second panels), the age distribution of *S*-gene negative cases (third panels) and the number of cases (bottom panels). All metrics are computed over a 14-day rolling window of cases by specimen date, plotted by first swab date. The black curve indicates the Wasserstein distance between the two age distributions. The purple shaded region indicates the confidence region for the Wasserstein distance if the two samples were drawn from the same distribution, generated through 1000 Monte Carlo samples of a permutation test. The grey shaded regions indicate dates where the Wasserstein distance falls outside of the 95% confidence region, denoting a significant difference between the age distribution of *S*-gene positive and *S*-gene negative cases. The red curve indicates *S*-gene negative cases and the blue curve indicates *S*-gene positive cases.



*Fig. S9 - Dynamics of age distributions of cases for Waltham Forest, Bedford, Redbridge, Hillingdon, Hounslow, Oadby and Wigston. For each region, we show the Wasserstein distance between the age distribution of S-gene positive and S-gene negative cases (top panels), the age distribution of S-gene positive cases (second panels), the age distribution of S-gene negative cases (third panels) and the number of cases (bottom panels). All metrics are computed over a 14-day rolling window of cases by specimen date, plotted by first swab date. The black curve indicates the Wasserstein distance between the two age distributions. The purple shaded region indicates the confidence region for the Wasserstein distance if the two samples were drawn from the same distribution, generated through 1000 Monte Carlo samples of a permutation test. The grey shaded regions indicate dates where the Wasserstein distance falls outside of the 95% confidence region, denoting a significant difference between the age distribution of S-gene positive and S-gene negative cases. The red curve indicates S-gene negative cases and the blue curve indicates S-gene positive cases.*

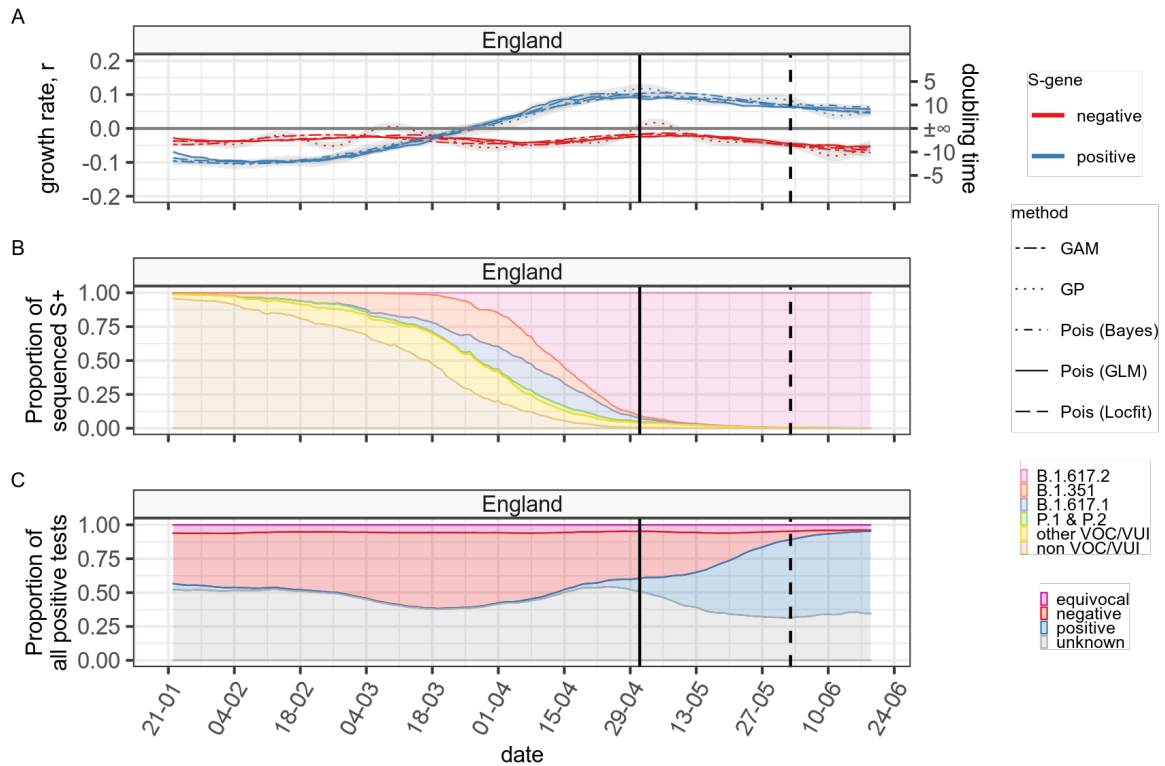
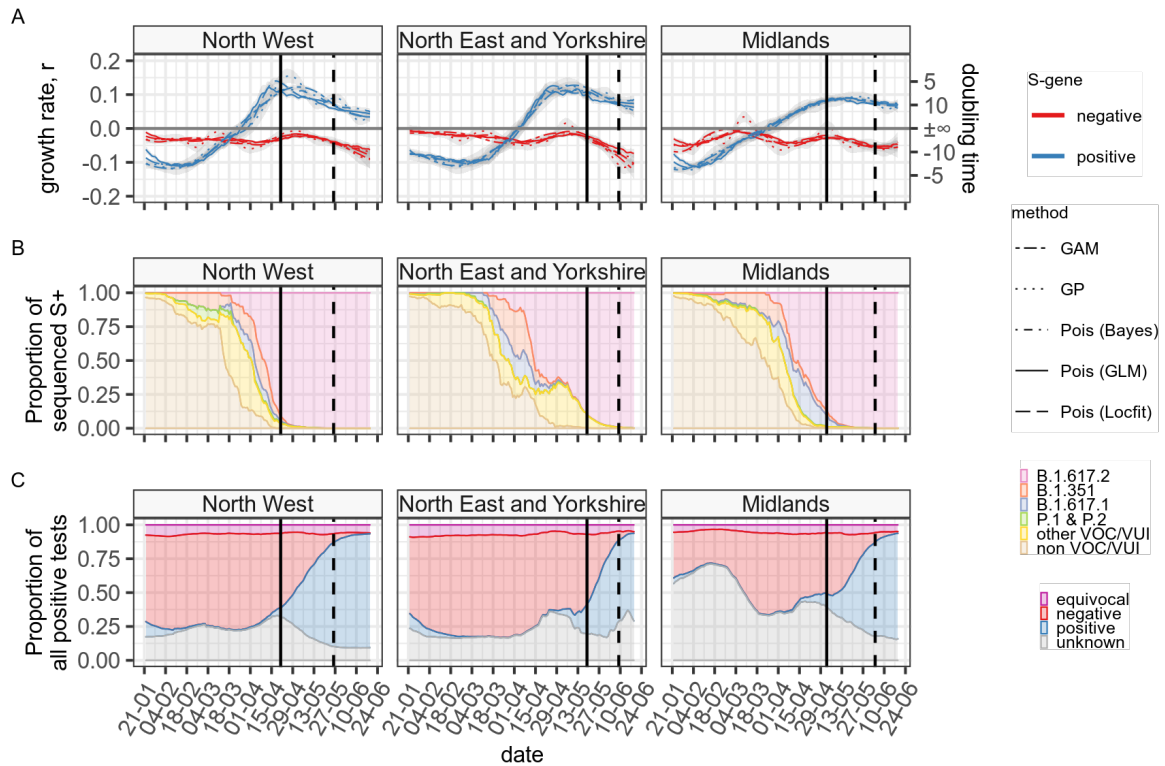
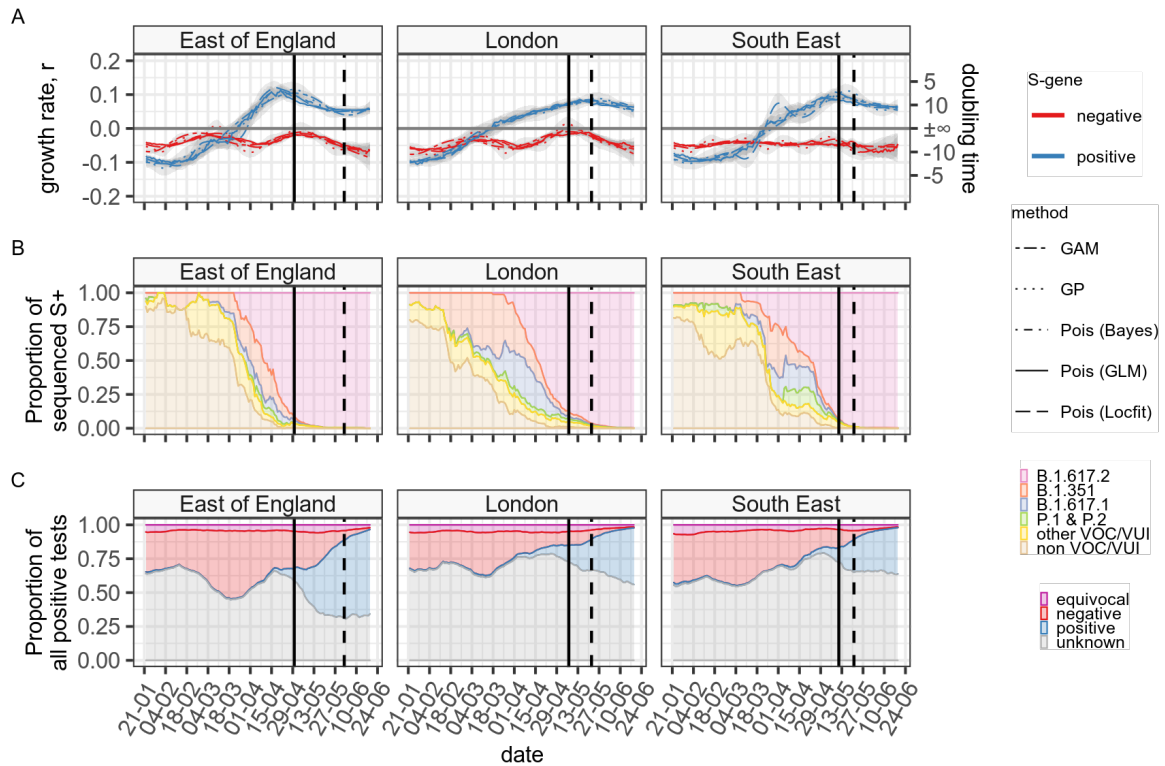


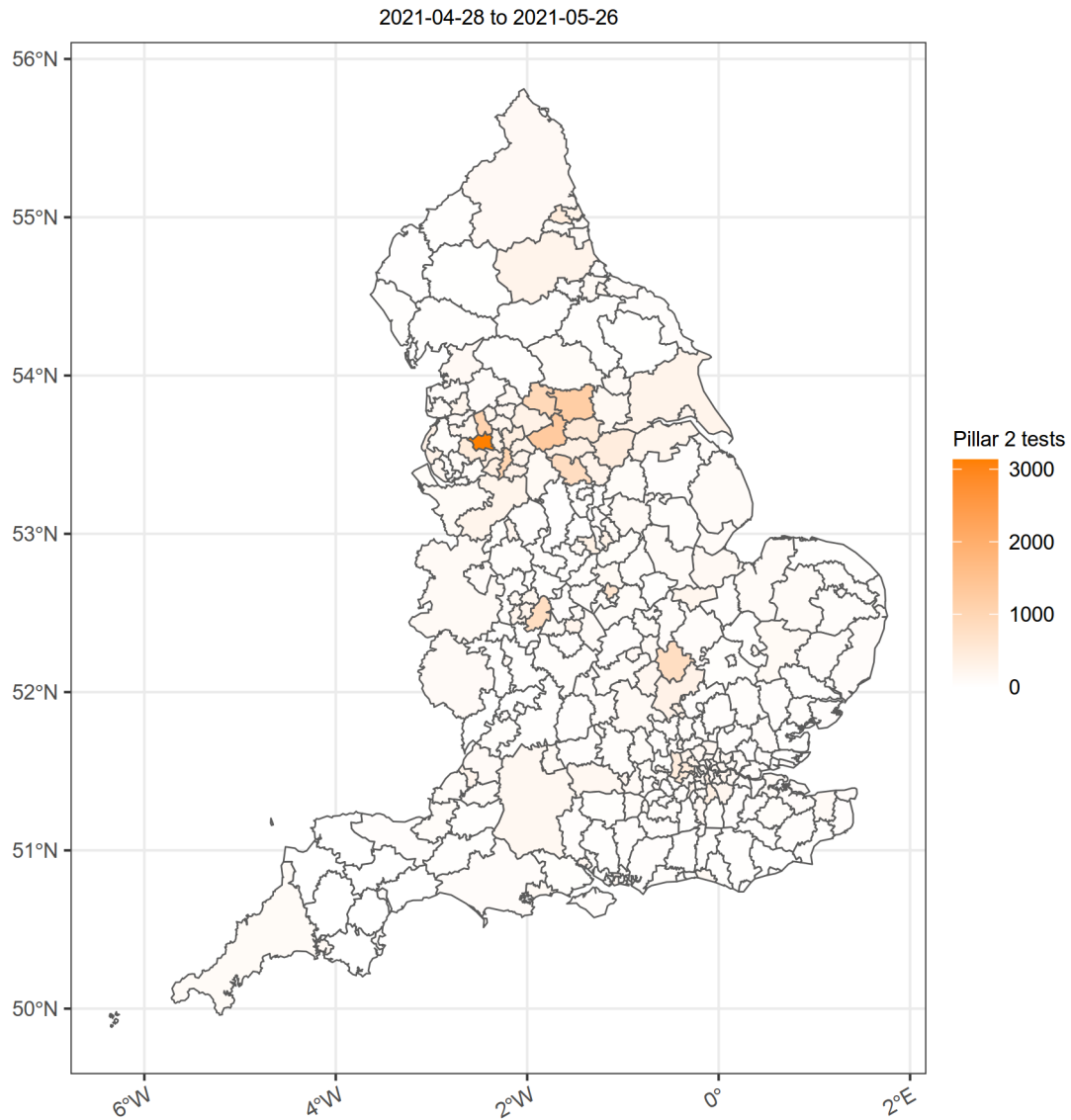
Fig. S10: Growth rate (A), S+ variant proportions (B) and S-gene statuses (C) for positive test results in all test episodes in England. The solid vertical line represents the time after which the vast majority of S-gene positive cases can be regarded as Delta variant (B.1.617.2) and the dashed vertical line, the time before which there are sufficient B.1.1.7 S-gene negative cases to allow a stable estimate of the growth rate of the Alpha variant. Transmission advantage estimates in Table 1 in the main paper are based on the period between the two vertical lines. Over this period the proportion of cases for which there was no known S-gene status decreased.



*Fig. S11: Growth rate (A), S+ variant proportions (B) and S-gene statuses (C) for positive test results in all test episodes in Northern NHS regions. The solid vertical line represents the time after which the vast majority of S-gene positive cases can be regarded as Delta variant (B.1.617.2) and the dashed vertical line, the time before which there are sufficient B.1.1.7 S-gene negative cases to allow a stable estimate of the growth rate of the Alpha variant. Transmission advantage estimates in Table 1 in the main paper are based on the period between the two vertical lines. Over this period the proportion of cases for which there was no known S-gene status was low and decreased in the North West and Midlands regions but no decrease was observed in the North East and Yorkshire.*

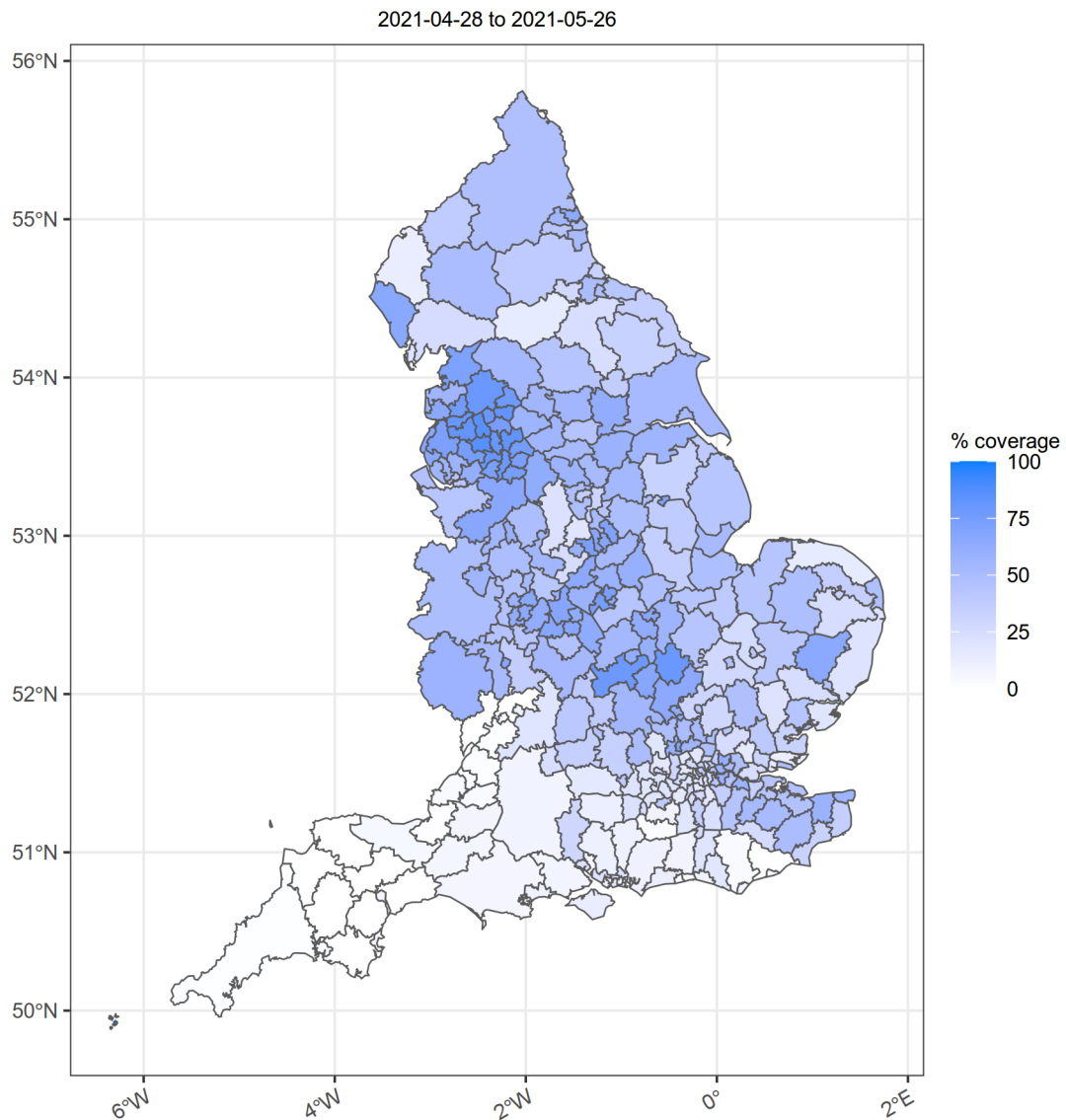


*Fig. S12: Growth rate (A), S+ variant proportions (B) and S-gene statuses (C) for positive test results in all test episodes in Southern NHS regions (excluding the South West where TaqPath testing was not routinely performed). The solid vertical line represents the time after which the vast majority of S-gene positive cases can be regarded as Delta variant (B.1.617.2) and the dashed vertical line, the time before which there are sufficient B.1.1.7 S-gene negative cases to allow a stable estimate of the growth rate of the Alpha variant. Transmission advantage estimates in Table 1 in the main paper are based on the period between the two vertical lines. Over this period the proportion of cases for which there was no known S-gene status was initially over 50% but decreased in all three regions, but most significantly in the East of England.*



*Fig. S13 - Pillar 2 positive cases by LTLA between 17th April 2021 and 15th May 2021. The degree of testing using the TaqPath assay varies from lab to lab. Since 1st March 2021 coverage of the S-gene test has been more extensive in the regions which we identify as problematic. This may be the result of an acquisition bias. Since 1st March 2021 the number of pillar 2 positive cases varies substantially from region to region reflecting areas where the second wave of the epidemic in England, driven by B.1.1.7 has taken more time to die down.*





*Fig. S14 - TaqPath test coverage in Pillar 2 by LTLA between 17th April 2021 and 15th May 2021. The proportion of tests that are performed using the TaqPath testing system and therefore for which we will have S-gene results generally covers those areas which have had a lot of Pillar 2 testing. However, regions with low case numbers also tend to have low TaqPath coverage and we must regard the S-gene signal in these areas as unreliable. Overall the TaqPath coverage in London for example is only about 30%*

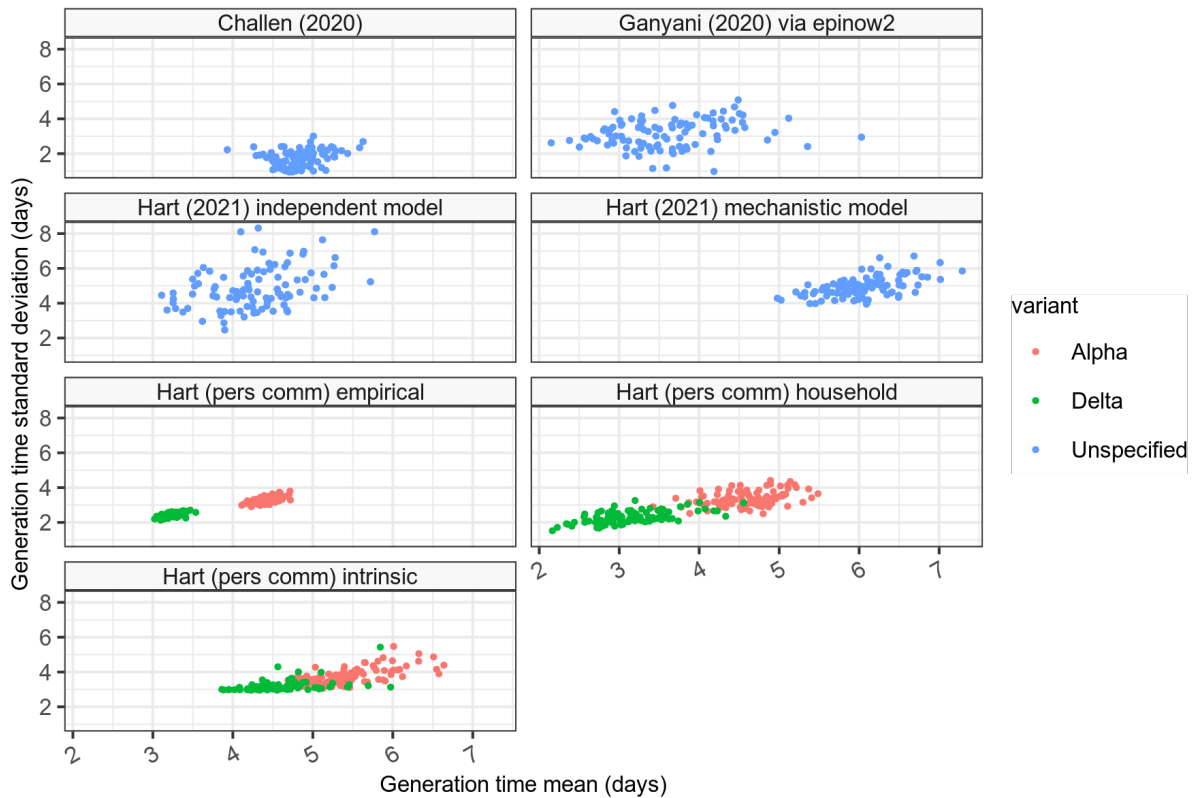


Fig. S15 - Generation time mean and standard deviation estimates from Challen *et al.* (2020) (43), Ganyani *et al.* (2020) (44), and Hart *et al.* (2021) (45) which are based on data from the early outbreak before Alpha and Delta variants were identified. Also included are estimates from Hart *et al.* (personal communication, pending publication) which apply the same methodology as previously published to UK household data containing differential recording of Alpha versus Delta cases.

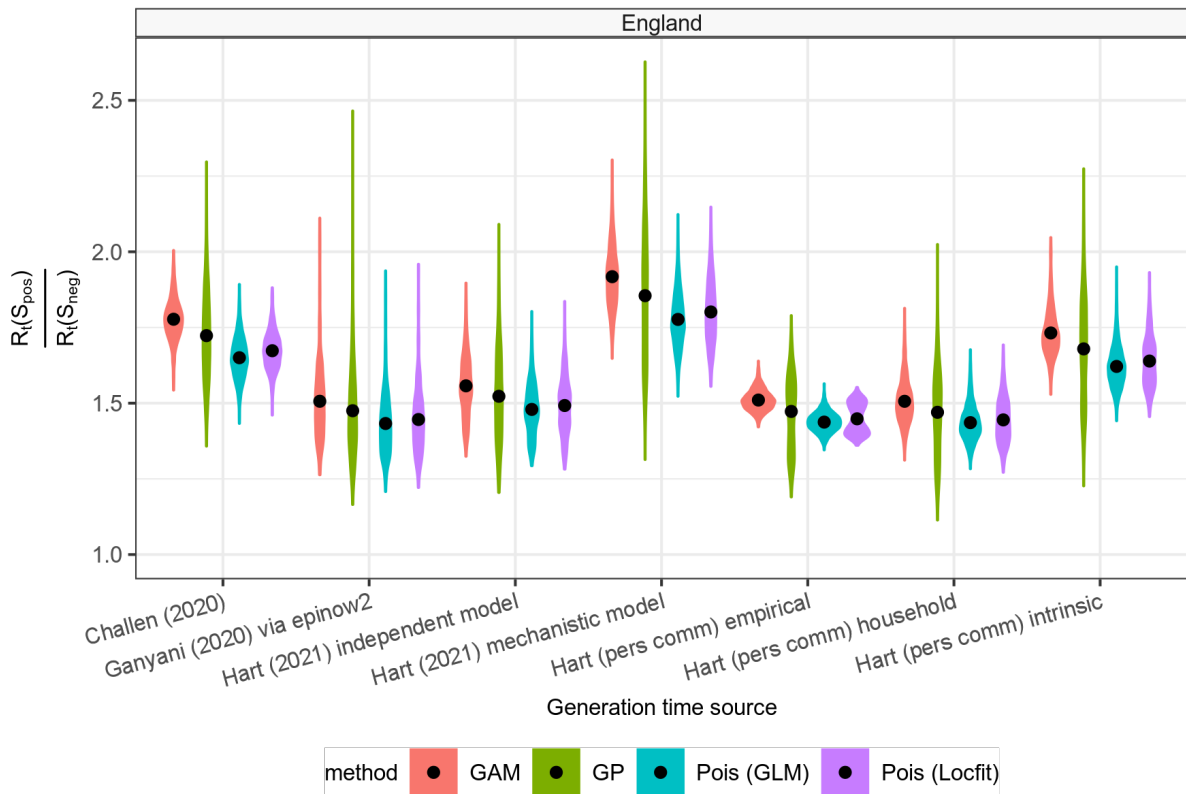


Fig. S16 - Reproduction number ratios for *S*-gene positive versus *S*-gene negative infections in England estimated between 1st May 2021 and 2nd June 2021. During this period *S*-gene positive cases were closely associated with Delta infections and *S*-gene negative cases were closely associated with Alpha. The estimates are made independently for each day, each of four growth rate estimation methods and using samples from each of seven generation time estimates. Summary estimates presented here are combined by day, showing the sensitivity to generation time, and choice of growth rate model. Generation time estimates have an impact on the central estimate of the reproduction number ratio, whereas the confidence limits are more sensitive to choice of growth rate model. It should be noted that the right most three groups of estimates using the empirical, household and intrinsic estimates from Hart et al. (person communication) have different generation time distributions for Alpha versus Delta, with Delta generation times being smaller than Alpha (see Fig. S15). This results in lower estimates for the reproduction number ratio for Delta (and by implication a lower transmission advantage) as some of the observed growth rate increase results from faster cycling through generations. However, it is clear that regardless of this effect all models and all estimates are estimating a reproduction number ratio significantly above 1.

*Table S1: Reproduction number ratio estimates for S-gene positive infections over S-gene negative infections in given regions over a time window described in Figs. S10-S12. Estimates are made independently for each day, each of four growth rate estimation methods and using samples from each of seven generation time estimates. Summary estimates presented here are combined by day and growth rate method, showing the sensitivity to generation time. All confidence intervals are above 1.*

Level	Name	Date range	Generation time	R number ratio	R number ratio (%age)
Country	England	01 May – 02 Jun	Challen (2020)	1.70 (1.52–1.93)	70% (52%–93%)
			Ganyani (2020) via epinow2	1.45 (1.27–1.77)	45% (27%–77%)
			Hart (2021) independent model	1.51 (1.32–1.75)	51% (32%–75%)
			Hart (2021) mechanistic model	1.83 (1.57–2.15)	83% (57%–115%)
			Hart (pers comm) empirical	1.47 (1.30–1.62)	47% (30%–62%)
			Hart (pers comm) household	1.46 (1.27–1.67)	46% (27%–67%)
			Hart (pers comm) intrinsic	1.66 (1.45–1.91)	66% (45%–91%)
Country	England	30 Apr – 02 Jun	Challen (2020)	1.59 (1.44–1.83)	59% (44%–83%)
			Ganyani (2020) via epinow2	1.39 (1.24–1.68)	39% (24%–68%)
			Hart (2021) independent model	1.45 (1.28–1.70)	45% (28%–70%)
			Hart (2021) mechanistic model	1.72 (1.51–2.07)	72% (51%–107%)
			Hart (pers comm) empirical	1.41 (1.30–1.60)	41% (30%–60%)
			Hart (pers comm) household	1.40 (1.27–1.64)	40% (27%–64%)
			Hart (pers comm) intrinsic	1.58 (1.41–1.86)	58% (41%–86%)
NHS region	London	07 May – 22 May	Challen (2020)	1.53 (1.30–1.74)	53% (30%–74%)
			Ganyani (2020) via epinow2	1.34 (1.17–1.58)	34% (17%–58%)
			Hart (2021) independent model	1.37 (1.19–1.57)	37% (19%–57%)
			Hart (2021) mechanistic model	1.60 (1.33–1.87)	60% (33%–87%)
			Hart (pers comm) empirical	1.34 (1.16–1.47)	34% (16%–47%)
			Hart (pers comm) household	1.32 (1.15–1.50)	32% (15%–50%)
			Hart (pers comm) intrinsic	1.48 (1.24–1.70)	48% (24%–70%)
NHS region	Midlands	03 May – 04 Jun	Challen (2020)	1.79 (1.49–2.14)	79% (49%–114%)
			Ganyani (2020) via epinow2	1.51 (1.28–1.90)	51% (28%–90%)
			Hart (2021) independent model	1.58 (1.32–1.93)	58% (32%–93%)
			Hart (2021) mechanistic model	1.96 (1.55–2.48)	96% (55%–148%)
			Hart (pers comm) empirical	1.54 (1.30–1.78)	54% (30%–78%)
			Hart (pers comm) household	1.54 (1.29–1.82)	54% (29%–82%)
			Hart (pers comm) intrinsic	1.77 (1.43–2.15)	77% (43%–115%)

North East and Yorkshire	19 May – 09 Jun	Challen (2020)	1.93 (1.66–2.48)	93% (66%–148%)
		Ganyani (2020) via epinow2	1.61 (1.35–2.17)	61% (35%–117%)
		Hart (2021) independent model	1.70 (1.42–2.29)	70% (42%–129%)
		Hart (2021) mechanistic model	2.16 (1.77–3.11)	116% (77%–211%)
		Hart (pers comm) empirical	1.64 (1.45–2.11)	64% (45%–111%)
		Hart (pers comm) household	1.64 (1.42–2.14)	64% (42%–114%)
		Hart (pers comm) intrinsic	1.92 (1.62–2.63)	92% (62%–163%)
North West	21 Apr – 26 May	Challen (2020)	1.77 (1.55–2.19)	77% (55%–119%)
		Ganyani (2020) via epinow2	1.49 (1.30–1.86)	49% (30%–86%)
		Hart (2021) independent model	1.54 (1.35–1.83)	54% (35%–83%)
		Hart (2021) mechanistic model	1.90 (1.66–2.29)	90% (66%–129%)
		Hart (pers comm) empirical	1.50 (1.39–1.66)	50% (39%–66%)
		Hart (pers comm) household	1.49 (1.35–1.73)	49% (35%–73%)
		Hart (pers comm) intrinsic	1.72 (1.53–2.04)	72% (53%–104%)
South East	11 May – 21 May	Challen (2020)	1.95 (1.67–2.45)	95% (67%–145%)
		Ganyani (2020) via epinow2	1.61 (1.36–2.12)	61% (36%–112%)
		Hart (2021) independent model	1.70 (1.44–2.16)	70% (44%–116%)
		Hart (2021) mechanistic model	2.18 (1.82–2.88)	118% (82%–188%)
		Hart (pers comm) empirical	1.65 (1.48–1.98)	65% (48%–98%)
		Hart (pers comm) household	1.65 (1.45–2.04)	65% (45%–104%)
		Hart (pers comm) intrinsic	1.93 (1.66–2.50)	93% (66%–150%)

*Table S2: Transmission advantage estimates for S-gene positive infections over S-gene negative infections in given regions over a time window described in Figs. S10-S12. Estimates are made independently for each day, each of 4 growth rate estimation methods and using samples from each of seven generation time estimates. Summary estimates presented here are combined by day and generation time estimate source, showing the sensitivity to the growth rate model. All confidence intervals are above 1.*

Level	Name	Date range	Model	R number ratio	R number ratio (%age)
Country	England	01 May – 02 Jun	GAM	1.61 (1.38–2.01)	61% (38%–101%)
			GP	1.57 (1.26–2.08)	57% (26%–108%)
			Pois (GLM)	1.52 (1.33–1.86)	52% (33%–86%)
			Pois (Locfit)	1.54 (1.33–1.91)	54% (33%–91%)
NHS region	East of England	30 Apr – 02 Jun	GAM	1.53 (1.33–1.88)	53% (33%–88%)
			GP	1.53 (1.27–1.99)	53% (27%–99%)
			Pois (GLM)	1.46 (1.28–1.77)	46% (28%–77%)
			Pois (Locfit)	1.47 (1.27–1.81)	47% (27%–81%)
	London	07 May – 22 May	GAM	1.42 (1.25–1.70)	42% (25%–70%)
			GP	1.37 (1.13–1.81)	37% (13%–81%)
			Pois (GLM)	1.43 (1.27–1.72)	43% (27%–72%)
			Pois (Locfit)	1.40 (1.21–1.73)	40% (21%–73%)
	Midlands	03 May – 04 Jun	GAM	1.68 (1.39–2.19)	68% (39%–119%)
			GP	1.70 (1.24–2.39)	70% (24%–139%)
			Pois (GLM)	1.61 (1.37–2.01)	61% (37%–101%)
			Pois (Locfit)	1.63 (1.36–2.11)	63% (36%–111%)
North East and Yorkshire	19 May – 09 Jun	GAM	1.86 (1.50–2.50)	86% (50%–150%)	
		GP	1.93 (1.46–2.97)	93% (46%–197%)	
		Pois (GLM)	1.66 (1.38–2.15)	66% (38%–115%)	
		Pois (Locfit)	1.75 (1.45–2.28)	75% (45%–128%)	
North West	21 Apr – 26 May	GAM	1.66 (1.40–2.11)	66% (40%–111%)	
		GP	1.63 (1.35–2.25)	63% (35%–125%)	
		Pois (GLM)	1.56 (1.34–1.96)	56% (34%–96%)	
		Pois (Locfit)	1.59 (1.35–2.04)	59% (35%–104%)	
South East	11 May – 21 May	GAM	1.78 (1.48–2.34)	78% (48%–134%)	
		GP	1.96 (1.52–2.80)	96% (52%–180%)	
		Pois (GLM)	1.66 (1.40–2.11)	66% (40%–111%)	

Pois (Locfit)

1.77 (1.46–2.33)

77% (46%–133%)

---

Table S3: Transmission advantage estimates for S-gene positive infections over S-gene negative infections in England between the 1st May and 2nd June 2021. Estimates are made independently for each day, each of 4 growth rate estimation methods and using samples from Hart et al. (personal communication) intrinsic estimates, and calculated either assuming different generation times for each variant (Alpha vs Delta), or assuming the same average generation time for both variants (Combined), or assuming the same generation time as would have existed prior to Delta emergence (Alpha only). Summary estimates presented here are combined by day and show the magnitude of the effect of the assumption that generation times of Alpha and Delta are the same.

Generation time	R number ratio	R number ratio (%age)
Alpha only	1.76 (1.53–2.07)	76% (53%–107%)
Alpha vs Delta	1.66 (1.45–1.93)	66% (45%–93%)
Combined	1.69 (1.48–1.95)	69% (48%–95%)

## References

33. R. M. Anderson, R. M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, Oxford, New York, 1992).
34. R. N. Thompson, J. E. Stockwin, R. D. van Gaalen, J. A. Polonsky, Z. N. Kamvar, P. A. Demarsh, E. Dahlgvist, S. Li, E. Miguel, T. Jombart, J. Lessler, S. Cauchemez, A. Cori, Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics*. **29**, 100356 (2019).
35. J. Wallinga, P. Teunis, Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. *Am. J. Epidemiol.* **160**, 509–516 (2004).
36. A. Cori, N. M. Ferguson, C. Fraser, S. Cauchemez, A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *Am. J. Epidemiol.* **178**, 1505–1512 (2013).
37. S. N. Wood, *Generalized Additive Models: An Introduction with R* (Chapman and Hall/CRC, Boca Raton, ed. 2, 2017).
38. H. Rue, S. Martino, N. Chopin, Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71**, 319–392 (2009).
39. F. Lindgren, H. Rue, J. Lindström, An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**, 423–498 (2011).
40. C. Loader, J. Sun, L. Technologies, A. Liaw, *locfit: Local Regression, Likelihood and Density Estimation* (2020; <https://CRAN.R-project.org/package=locfit>).
41. P.-C. Bürkner, Advanced Bayesian Multilevel Modeling with the R Package brms. *R J.* **10**, 395–411 (2018).
42. J. Wallinga, M. Lipsitch, How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B Biol. Sci.* **274**, 599–604 (2007).



43. R. Challen, E. Brooks-Pollock, K. Tsaneva-Atanasova, L. Danon, *medRxiv*, doi:10.1101/2020.11.17.20231548.
44. T. Ganyani, C. Kremer, D. Chen, A. Torneri, C. Faes, J. Wallinga, N. Hens, Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance*. **25**, 2000257 (2020).
45. W. S. Hart, S. Abbott, A. Endo, J. Hellewell, E. Miller, N. Andrews, P. K. Maini, S. Funk, R. N. Thompson, *medRxiv*, doi:10.1101/2021.05.27.21257936.
46. C. Castillo-Chavez, H. W. Hethcote, V. Andreasen, S. A. Levin, W. M. Liu, Epidemiological models with age structure, proportionate mixing, and cross-immunity. *J. Math. Biol.* **27**, 233–258 (1989).
47. O. Diekmann, J. A. P. Heesterbeek, J. A. J. Metz, On the definition and the computation of the basic reproduction ratio  $R_0$  in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* **28**, 365–382 (1990).
48. E. Levina, P. J. Bickel, The Earth Mover’s distance is the Mallows distance: some insights from statistics. *Proc. Eighth IEEE Int. Conf. Comput. Vis. ICCV 2001*. **2**, 251–256 vol.2 (2001).
49. L. V. Kantorovich, Mathematical Methods of Organizing and Planning Production. *Manag. Sci.* **6**, 366–422 (1960).
50. Departing passenger survey | UK Civil Aviation Authority, (available at <https://www.caa.co.uk/Data-and-analysis/UK-aviation-market/Consumer-research/Departing-passenger-survey/Departing-passenger-survey/>).
51. Airport data 2021 01 | UK Civil Aviation Authority, (available at <https://www.caa.co.uk/Data-and-analysis/UK-aviation-market/Airports/Datasets/UK-Airport-data/Airport-data-2021-01/>).
52. Population estimates - Office for National Statistics, (available at <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates>).
53. Danish Covid-19 Genome Consortium, Genomic overview of SARS-CoV-2 in Denmark, (available at <https://www.covid19genomics.dk/statistics>).

# A. Reproduction number validation methodology

## A.1 Introduction

The reproduction number represents the ratio between the number of secondary cases resulting from each primary case. At the beginning of an outbreak assuming no prior immunity and a freely mixing population, this is described as the basic reproduction number,  $R_0$  (Vegvari et al., 2021).

The effective reproduction number  $R_t$ , is a time varying quantity, which may be defined in terms of the basic reproduction number  $R_0$ , the fraction of contacts that people are making at a given time  $C_t$ , compared to a freely mixing population, and the fraction of the population that is still susceptible to infection  $S_t$ .

$$R_t = S_t C_t R_0$$

The infectivity profile is another probability distribution. Most often represented in discrete form  $\omega_1, \omega_2, \dots, \omega_s$ , that defines the likelihood that a case infected at time  $t$  resulted from a case infected between the times  $t - s$  and  $t - s + 1$ . This definition implies that  $\omega_{s \leq 0} = 0$  as that would apply to secondary infections resulting from

primary infections in the future, and that the discrete time measure  $s$  here represents the upper bound of the equivalent continuous unit time interval, rather than, for example, the middle of the interval.

Connecting to the instantaneous reproduction number and the infectivity profile is the quantity  $\beta$  which is the “transmissibility” of an infection in an average individual, infected at a given time,  $t$ , at a given number of days post infection,  $\tau$ .  $\beta$  is related to the in host viral load of an infection, and the number of contacts that infected individuals make with susceptible individuals.

$$\beta_{t,s} = R_t \omega_s$$

There are two basic types of effective reproduction number to consider. The simplest conceptually is the forward-looking definition, in which the reproduction number is the number of secondary infections generated by a single primary infection which occurs at time  $t$ , this is known as the case reproduction number  $R_t^{case}$  (Fraser, 2007). The case reproduction number reflects the state of the epidemic at a specific point in time but is limited by the fact that it can only be determined after the event. If we assume  $I_0, I_1, \dots, I_t$  is a time series of infection counts, assumed to be drawn from some discrete probability distribution with expected value  $\bar{I}_t$  then  $R_t^c$  is given by:

$$R_t^{case} = \frac{\sum_{s=1}^{\infty} \bar{I}_{t+s} \omega_s}{\bar{I}_t}$$

We can alternatively define the backward-looking effective reproduction number as the inverse ratio of the number of primary infections that cause the secondary infections observed at time  $t$ , this is known as the instantaneous reproduction number,  $R_t^{inst}$ . In an evolving epidemic the instantaneous reproduction number is able to be calculated using data that has already been observed and is hence the more

useful quantity. The rest of this summary considers the instantaneous version of the effective reproduction number, which we refer to as the reproduction number or  $R_t$ .

$$R_t^{inst} = \frac{\bar{I}_t}{\sum_{s=1}^{\infty} \bar{I}_{t-s} \omega_s}$$

Alternative methods exist to derive the reproduction number from growth rate (Wallinga and Lipsitch, 2007) which we describe in more detail later. These produce a reproduction number that does not have a neat definition in terms of the infection case counts and produces an estimate that is between the two flavours of reproduction number presented here.

The renewal equation method can be used to calculate the instantaneous reproduction number and has a reference implementation in the R package EpiEstim (Cori et al., 2021; Cori et al., 2013; Thompson et al., 2019). This allows for a variety of configuration options to suit different use cases. We concentrate on the estimation of a time series of the reproduction number, for which the main parameters are a single fixed mean and standard deviation for the  $R_t$  prior, the window over which the estimate will be performed, and an infectivity profile. The infectivity profile can be specified in a number of ways, but for our purposes we concentrate on the discrete empirical version (“non\_parametric\_si” option). Figure A.1 shows the behaviour of EpiEstim on an outbreak in the Flu2009 data set included in EpiEstim for a range of different estimation windows, and using the given infectivity profile.

This simple example demonstrates some variability in the estimates that we wished to be able to quantify, in that there is a clear trade-off between bias and variance in the window selection. This is determined in part by the case counts within the time window of the estimate. With smaller estimate windows and less data at either end of the time series the estimates revert to the prior  $R_t$  distribution, whereas with longer windows, at least in this scale and duration of outbreak, the detail is lost.

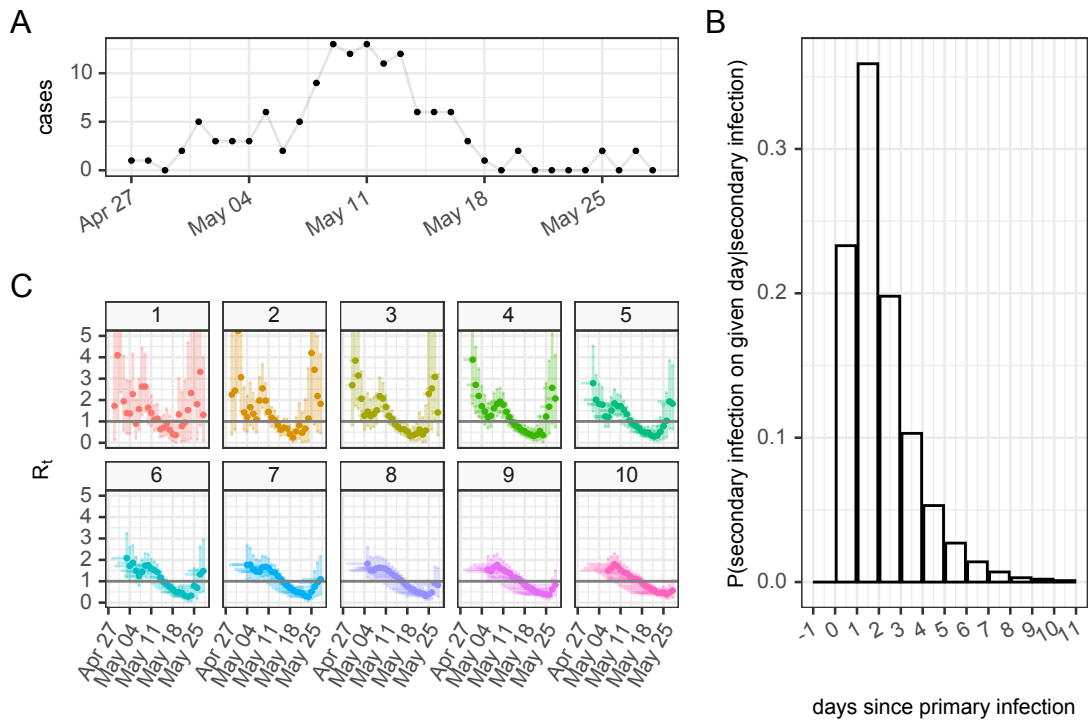


Figure A.1: Panel A shows a case count for the Flu2009 outbreak data set, Panel B shows the discrete infectivity profile, and panel C shows the  $R_t$  time series estimates for a range of different estimation windows. Vertical error bars show the confidence in  $R_t$  whereas horizontal error bars show the date range for which the assumption of constant  $R_t$  for each estimate is applied

There is a need to be able to assess the performance of an individual estimation method and parameterisation, against a standard and to be able to compare against each other. The purpose of this paper is to describe the validation procedure and define the associated quality metrics we employ to compare estimates. This will be described in terms of  $R_t$  estimation using three parameterisations of EpiEstim (14 day window, 7 day window and 4 day window) but is extensible to comparisons between other methods and to other observations we may wish to estimate such as case incidence and exponential growth rates as well.

## A.2 Validation methodology

We construct synthetic data sets with known values of expected incidence,  $R_t$  and exponential growth rate. This synthetic data set is generated using an initial case

count, and a time series of exponential growth rate values. The growth rate time series are either a step function with 6 different predefined levels or a smooth cubic spline passing through 6 predefined control points, over the course of a theoretical year. The growth rate time series is accumulated and applied to the initial incidence to generate a time series of expected incidence. Theoretical values of  $R_t$  are calculated from the growth rate using the methods of Wallinga and Lipsitch, 2007 and assuming a synthetic generation time which is gamma distributed with mean ( $\mu$ ) of 5 days and standard deviation ( $\sigma$ ) of 4. This uses the following relationship, where  $M$  is the moment generating function of the gamma distribution with shape parameter  $\alpha = \mu^2/\sigma^2$  and rate parameter  $\beta = \mu/\sigma^2$ .

$$\begin{aligned} R_t &= \frac{1}{M(-r_t)} \\ &= \left(1 - \frac{r_t}{\beta}\right)^{-\alpha} \\ &= \left(1 - \frac{r_t\sigma^2}{\mu}\right)^{-\mu^2/\sigma^2} \end{aligned}$$

From each simulation of case incidence random bootstrap samples are drawn from a Poisson distribution whose rate is the expected incidence. Optionally a weekend effect is simulated, a second Poisson sample is made on Saturday, Sunday and Mondays in the time series with a rate given as a fraction of the expected incidence (0%, 3% or 10%). This second random sample is subtracted from the first on Saturdays and Sundays, and added on Mondays, giving a weekly cycle to case counts similar to that seen in reality. Any resulting negative values are set to zero.

We simulate using one smooth and one stepped growth rate time series in combination with three levels of weekend effect mentioned above, and with two initial case counts (100 and 10,000), giving an overall 12 different configurations. The synthetic gamma distributed generation time is discretised on whole day intervals as required by the estimation methods, imposing that the generation time is known precisely and does not vary for this validation analysis.

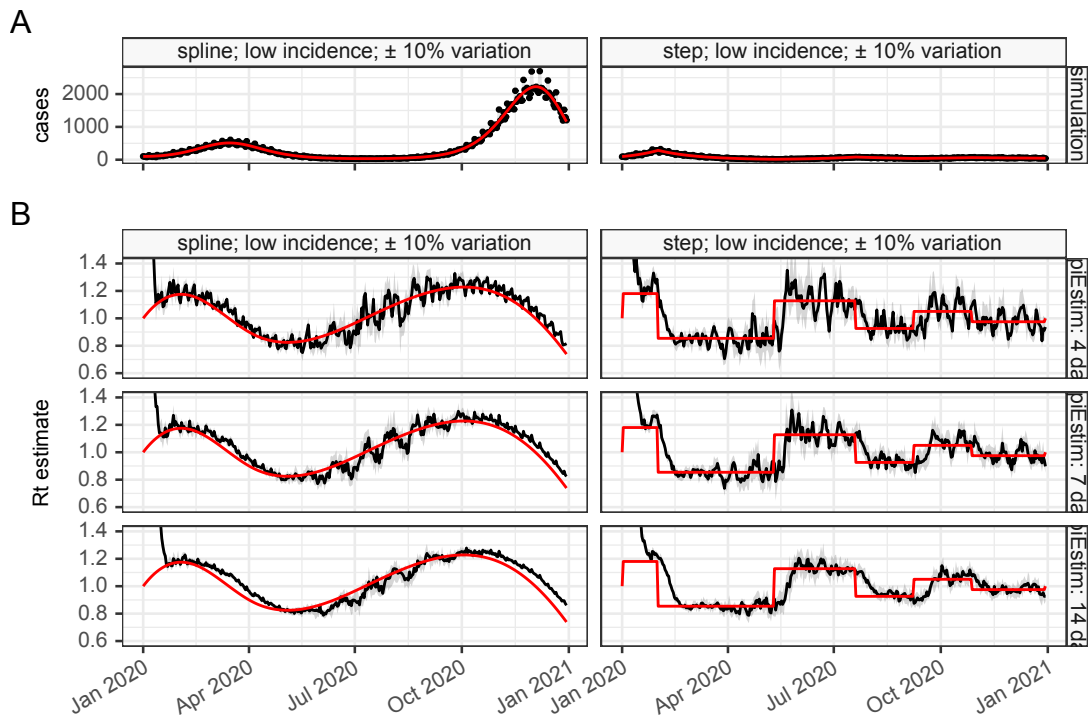


Figure A.2: Panel A shows a case count, for a synthetic data sets generated from a spline function for growth rate, with low (100) initial case counts, and a 10% weekly variability. Expected incidence curves are shown in red and black points represent one set of data samples (out of 10 generated). Panel B shows the estimated reproduction number using the renewal equation method with a gamma distributed infectivity profile (mean 5 days, sd 4), and a fixed window of 14 days. The red line shows the theoretical  $R_t$  value.

In Figure A.2 we show 1 sample from the smooth test configuration, with 10% weekly variation and 100 initial cases. The reproduction number is estimated with EpiEstim in its default configuration with a fixed window of 14 days. Estimate uncertainty is larger when case numbers are low. The estimates (black lines in panel B) are clearly lagged compared to the theoretical  $R_t$  value, particularly noticeably when case counts are high. At the beginning of the time series the estimate is consistently inappropriately high reflecting the prior distribution of the  $R_t$  estimate, and a boundary effect at the beginning of the time series. Although only one time series from each configuration are being shown here, 10 are generated, to simulate data variability. Similarly we are showing two simulation configurations here as examples, but overall we have 12 configurations which describe different combinations of smoothness, weekend effect, and initial incidence.

### A.3 Quantifying estimate delay

Before we can assess the quality of the estimates we need to determine how much delay there is in the estimates. To do this we use another synthetic data set which is generated from a triangular wave for growth rate with period of 91 days. For a range of different time lags (0 days to 28 days) we calculate the root mean squared error between the theoretical  $R_t$  and median of the estimate shifted backwards in time. In Figure A.3 panel B we quantify the delay between the theoretical and the estimated  $R_t$ , for a range of estimation methods. The delay is the result of the combination of estimate being a backward looking instantaneous reproduction number, which integrates information from the past according to the generation time distribution, and the method's estimation window, which in this case is 4, 7, or 14 days. From visual inspection of the time series in panel A there is no compelling evidence that the lag is significantly different from one time period to another, although there is a hint that the delay is shorter when case numbers are high and rising. For subsequent analysis each model's estimates of  $R_t$  are adjusted backwards by the nearest integer number of days derived from this lag analysis. It must be pointed out that the true value of the reproduction number is calculated using the methods of Wallinga and Lipsitch, 2007 which produces a different type of reproduction number to the instantaneous reproduction number produced by the renewal equation methods presented here.

### A.4 Quantification of accuracy, bias and calibration

To investigate the bias and calibration of the estimation method we compare four statistics derived from the  $R_t$  estimates and associated theoretical  $R_t$  values, using the lag adjusted estimates. Estimates of  $R_t$  include a mean, standard deviation and



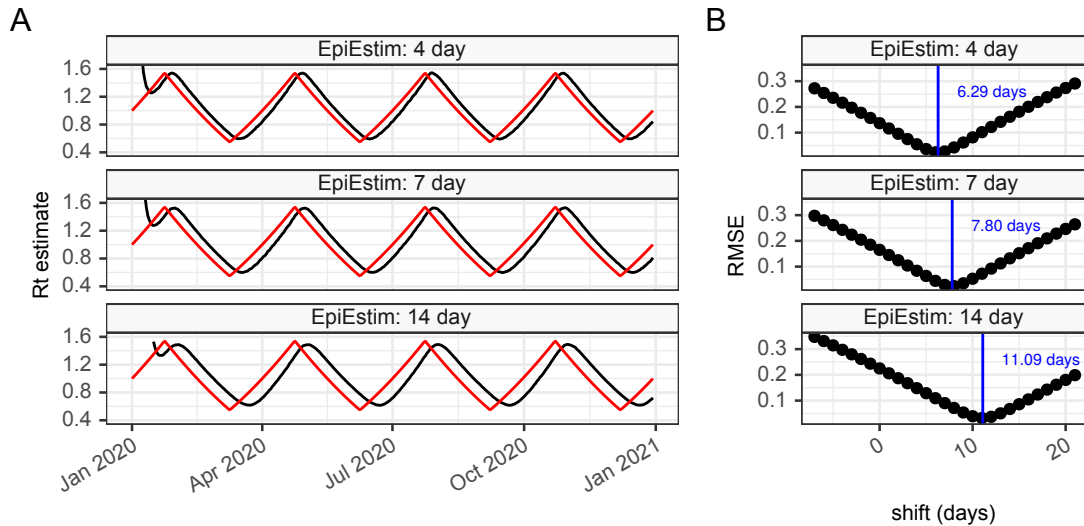


Figure A.3: Analysis of  $R_t$  estimate time delays. In panel A as estimate of  $R_t$  based on a periodic growth rate clearly shows estimate lag in all methods of estimating  $R_t$ . In panel B the root mean squared error for a range of values of lag are calculated. The minimum RMSE is depends on the estimation model employed, however there is not an obvious linear relationship.

2.5%, 5%, 25%, 50%, 75%, 95% and 97.5% quantiles. To identify bias, for each point we calculate the difference between median estimate and theoretical  $R_t$  values as the residual ( $\xi_{bias}$ ). On this measure a value of 0 represents an unbiased estimate, and bias is presented on the same scale as the  $R_t$  estimate.

To assess precision we measure the calibration of the estimate ( $\xi_{cal}$ ). This is defined as 1 if the actual value is within the confidence intervals of the estimate, and 0 if the actual value is outside. A well calibrated set of estimates should have an average close to 1, and poorly calibrated estimates will be closer to zero.

We examine the confidence of the estimate using a continuous data analogy to the verification rank histogram (Anderson, 1996; Bröcker, 2008; Siegert et al., 2020). The verification rank histogram examines the distribution of the ranks of each of the true values among the collection of associated point estimates. The shape of the distribution of these ranks determines the appropriateness of the confidence limits (Hamill, 2001). Appropriate variation in point estimates lead to a flat histogram with the true values falling evenly over the point estimates. Not enough variability

in the point estimates results in a U-shaped histogram, too much in a dome shaped histogram. For our purposes we have individual estimates that describe the estimate as a a posterior distribution through a set of quantiles. This estimate is associated with a single true  $R_t$  value. We can derive a cumulative density function (CDF) of each estimate ( $F(x)$ ) from the quantiles, by filling a monotonic Hermite spline to them, and use this CDF to estimate the quantile of the theoretical  $R_t$  value with respect to the estimate distribution [ $F(x = R_t)$ ]. The quantile of the actual  $R_t$  value ( $\xi_{quant}$ ) with regards to the posterior distribution of the  $R_t$  estimate is equivalent to the rank in the rank histogram, and we define this as the “actual value quantile” for all the paired estimate distributions and true  $R_t$  values. The density plot of the actual value quantile, for a range of estimates, we define as the quantile density plot and this has the same interpretation as the rank histogram.

The continuous ranked probability score (CRPS) is a measure of performance for probabilistic estimates of a scalar observation. It is a quadratic measure of the difference between the estimate cumulative distribution function (CDF) and the empirical CDF of the observation (Zamo and Naveau, 2018). Loosely speaking, it describes the mass of the probability density function that is closer to the mean of the estimate, than the true value is. For each estimate it can be calculated directly from the paired CDF of an estimate ( $F(x)$ ) and a scalar true value ( $y$ ) as the following, where  $\mathbb{I}$  is the indicator function. This estimate-by-estimate value is referred to as the instantaneous CRPS ( $\xi_{cprs}$ ), and low values imply higher quality estimates:

$$CRPS_{inst}(F, y) = \int_{-\infty}^{\infty} \left( F(x) - \mathbb{I}(x \geq y) \right)^2 dx$$

In summary empirical distributions for the 4 metrics are derived from each individual estimate:

bias :

$$\xi_{bias} = \left[ F_n^{-1}(0.5) - R_{t,n} \right]$$

calibration :

$$\xi_{cal} = \left[ \mathbb{I} \left( F_n^{-1}(0.025) \leq R_{t,n} \leq F_n^{-1}(0.975) \right) \right]$$

actual value quantile :

$$\xi_{quant} = \left[ F_n(x = R_{t,n}) \right]$$

instantaneous continuous rank probability score :

$$\xi_{cprs} = \left[ CPRS_{inst}(F_n, R_{t,n}) \right]$$

In our case, of particular interest in assessing estimates of  $R_t$  is the critical threshold of  $R_t = 1$  when an epidemic transitions from growth to decline or vice-versa. A specific measure for this scenario detects if the confidence limits of an estimate are both the opposite side of 1 to the true value. An ideal situation is that this never happens. This “critical threshold” metric ( $\xi_{crit}$ ) is defined as follows assuming the sign function is defined as  $sgn(x) = x/|x|$ :

critical threshold :

$$\xi_{crit} = \left[ \mathbb{I} \left( sgn(F_n^{-1}(0.025) - 1) = sgn(R_{t,n} - 1) \right. \right. \\ \left. \left. || sgn(R_{t,n} - 1) = sgn(F_n^{-1}(0.975) - 1) \right) \right]$$

When we wish to compare the performance of a given method against another we need aggregate the estimate level scores achieved against a particular validation data set. The median and inter-quartile ranges of the bias and instantaneous continuous rank probability score of all estimates can be simply calculated to give an overall metric for the performance of the estimation method. Likewise the calibration and critical threshold of each estimate can be aggregated into a single percentage for the method.

The actual value quantiles ( $\xi_{quant}$ ), are inspected visually and have the same interpretation as the verification rank histogram, with a U-shaped density curve representing over-precision and a dome shaped density curve representing excessive uncertainty, and an ideal result being a uniform distribution. The average value of  $\xi_{quant}$  should be close to 0.5. To quantify the shape of the quantile density plot, define the deviation of the values  $\xi_{quant}$  for each of  $N$  estimates from the expected value of 0.5, and compare this to the ideal value, which is the standard deviation of a uniform distribution with support between 0 and 1 ( $\sqrt{\frac{1}{12}} = 0.289$ ). This gives us the following “quantile deviation” metric, which ranges between 0.211 and -0.289 with positive values representing more deviation than expected, a U-shaped quantile density plot and over precise estimation methods, and negative values representing less deviation than expected, a dome shaped quantile density plot and excessive uncertainty in the estimation method:

$$Q_{dev} = \sqrt{\frac{\sum(\xi_{quant} - 0.5)^2}{N - 1}} - \sqrt{\frac{1}{12}}$$

quantile deviation :

These metrics can be used to summarise the performance of each individual estimation method and compare it to others. At a top level we can aggregate the individual estimates over time and over the different simulations, to produce a single set of 5 metrics for each estimation method, with a focus on the overall performance. The quality metrics are visualised as box plots for the bias and CRPS, a simple proportion for the calibration, and critical threshold measure, and the distribution shape for the quantile density in the form of a violin plot. In this case the U shaped verification rank histogram becomes an I shaped quantile density violin plot, representing over-precision, and a dome shaped verification rank histogram, becomes an O shaped quantile density violin plot, representing excessive uncertainty.

The top level comparison in Figure A.4 shows similar performance of the different

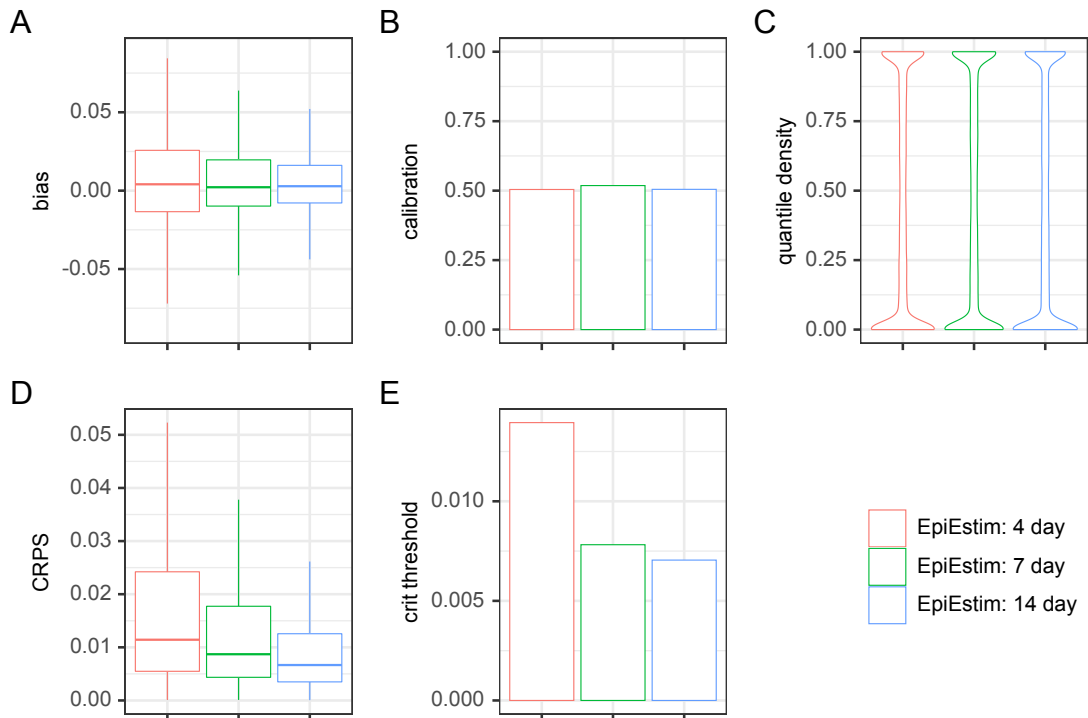


Figure A.4: Estimate quality metric summaries for multiple estimation methods. In this instance we compare the performance of the renewal equation method with 14, 7, and 4 days as the windowing. Panel A show summary statistics for the bias of  $R_t$  estimates, panel B shows the calibration. Panel C shows the quantile density, panel D the CRPS and panel E the critical threshold calibration.

methods on most metrics. Unsurprisingly there is more spread in the error of estimates using a 4 day window seen in Panel A. All methods are similarly calibrated (panel B - higher is better) and show similar levels of over precision (Panel C). The CRPS (lower is better) is clearly higher for the 7 day and 4 day estimates, with the 14 day estimate performing best overall, reflecting the decrease in noise seen in panel A. 14 day model is also best at predicting growth when the epidemic is in decline and vice-versa, with less than 1% critical threshold error, although this is relatively good across all methods. The visual comparison is backed up by the quantitative metrics presented in Table A.1, which also gives us a quantitative comparison for the quantile divergence which confirms the over precision is worst (highest quantile deviation) for the 14 day estimates.

To further assess the details of why the individual methods differ we investigate an intermediate level of detail, in which the metrics are broken down by differences in

Table A.1: Estimate quality metric summaries for the three estimation methods compared here.

metric	EpiEstim: 4 day	EpiEstim: 7 day	EpiEstim: 14 day
average bias	0.0041 [IQR -0.013 – 0.026]	0.0022 [IQR -0.0098 – 0.02]	0.0028 [IQR -0.0078 – 0.016]
average calibration	50.4% [95% CI 49.9% – 50.9%]	51.8% [95% CI 51.4% – 52.3%]	50.5% [95% CI 50.0% – 50.9%]
average critical threshold	1.4% [95% CI 1.3% – 1.5%]	0.8% [95% CI 0.7% – 0.9%]	0.7% [95% CI 0.6% – 0.8%]
CRPS	0.011 [IQR 0.0055 – 0.024]	0.0087 [IQR 0.0044 – 0.018]	0.0067 [IQR 0.0035 – 0.013]
quantile deviation	0.115	0.111	0.112
estimate delay	6.29 days	7.80 days	11.09 days

the simulated data as described below.

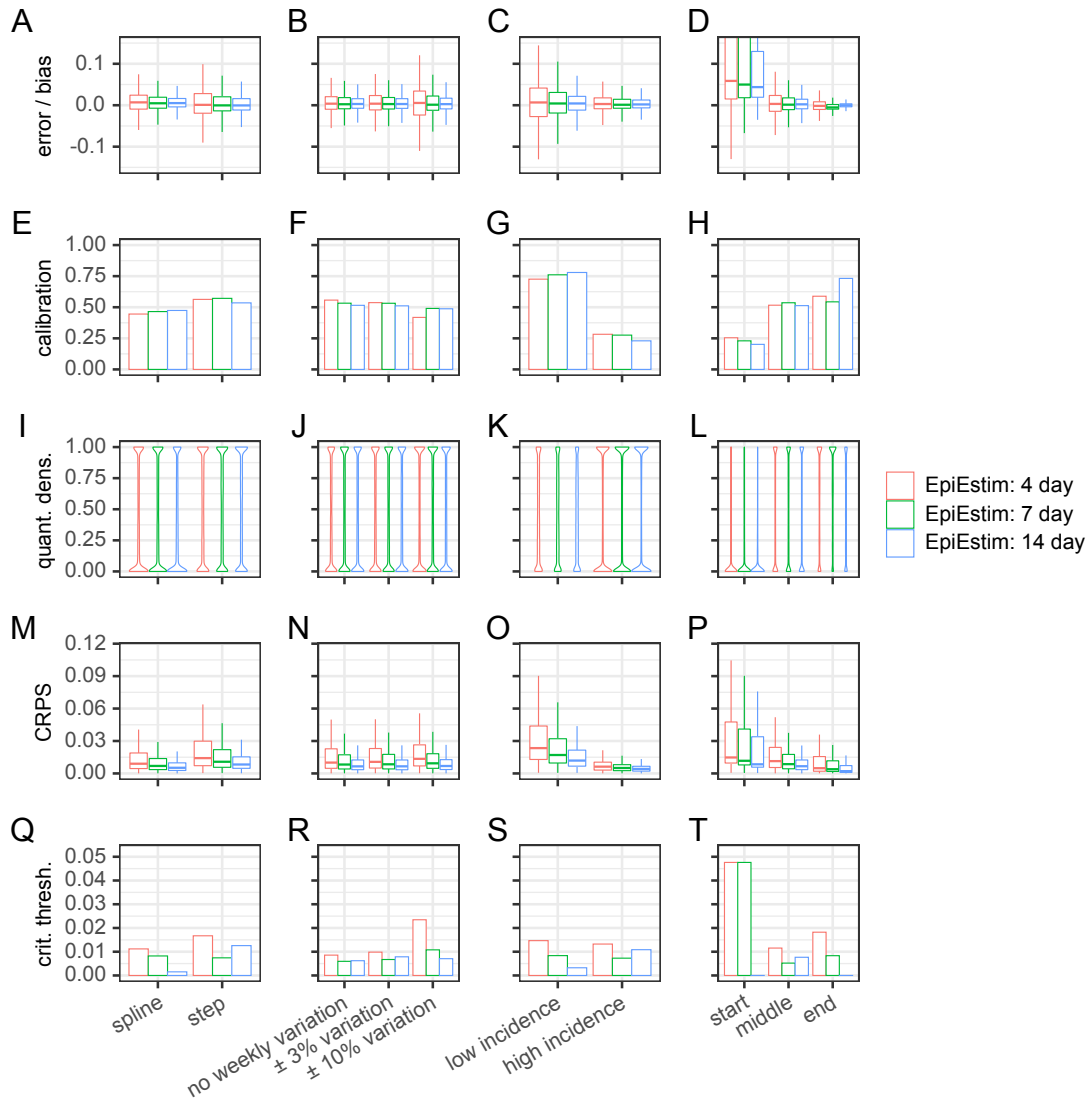


Figure A.5: Estimate quality metric intermediate detail and model comparison. Panel A-D show summary statistics for the bias of  $R_t$  estimates broken down by the simulation smoothness, weekly variability, initial incidence, and time series boundary status. Panels E-F shows the calibration for the same subdivisions. Panels I-L shows the quantile density, panel M-P the CRPS, and panel Q-T the critical thresholds for the same subdivisions.

At this intermediate detail the aggregation of the quality metrics are faceted by

the salient features of the simulation, which inform us of where the estimators strengths and weaknesses are. These including the the smoothness of the growth rate function (first column), the degree of weekend effect (second column), initial size (third column), and also the proximity of the estimate to either end of the time series (final column). This last measure is defined as whether the estimate is within 21 days of the start or end of the time series. The latter is important because boundary effects particularly on most recent end of the time series may be important for real time estimates of  $R_t$ . This intermediate view of the quality of the estimates expands on the findings from Figure A.4, for example, in Panel G and K in Figure A.5, we note the low calibration and excessive precision of all the estimates is due to the high incidence simulations. In panel O however we note that the CPRS is lower (better) in the higher incidence scenarios suggesting that the over-precise estimates do come with improved accuracy (also seen in panel C).

The boundary effect analysis in panels D,H,L,P demonstrates that all the methods investigated here are biased high in the first 3 weeks, and poorly calibrated. The critical threshold metric (panels Q,R,S & T) shows that the 14 day estimate method is able to detecting transitions between growth to decline well. The simulations with high weekly variation is seen to the accuracy of the 4 day estimate on this measure (panel R). The impact of the weekly variation on the 4 day estimate is also seen in panel B where increase spread of error occurs with increasing variation.

On the face of this comparison there is evidence to favour the 14 day estimate methodology.

To further understand the performance of the 14 day model, we can also visualise these metrics over time and between selected simulation configurations. For the single estimation methodology (14 days window) we can compare low and high incidence and smoothness of the simulation. We present in detail the 4 simulations with no weekend effects, 2 of which are based on smoothly varying growth rates, and 2 on

stepped growth rates, with different numbers of initial cases (low incidence: 100 and high incidence: 10000).

Figure A.6 shows variations over time of the quality metrics for simulations with no weekly variability. This allows us to detect which parts of the input time series are causing poorer estimate performance and helps us identify where improvements are required. Panel A shows the modelled case count, and one sample from the simulation for 4 synthetic data simulation configurations. Panel B shows the simulation reproduction number (red) and one estimate (black) based on one sample from the simulation, using the renewal equation method with a gamma distributed infectivity profile (mean 5 days, standard deviation 4), and a fixed window of 14 days with the time corrected to account for estimate lags. Panel C shows the bias of each individual  $R_t$  estimate and rolling quantiles over a 28 days period. Panel D shows the calibration of individual estimates and the rolling mean of the calibration (14 day window). Panel E shows the quantile density over time (with density as the shade), the rolling 10%, 30% 50% 70% and 90% quantiles are shown as grey lines calculated over a 28 day window. In panel F the instantaneous CRPS for individual estimates and the rolling quantiles are shown. In Panel G the critical threshold measure shows where the estimates confidence limits are the other side of the  $R_t$  critical threshold of 1 to the actual value. In all panels quantiles shown in red are 2.5% and 97.5% if dotted, 25% and 75% if dashed and the median is a solid red line. Blue horizontal lines represent the median for the statistic for the whole time series.

For this selection of simulation configurations and estimation method we again find that there is no evidence of strong systematic bias. Over all simulations in Figure A.6 the bias is 0.0028 [IQR -0.0078 – 0.016] (reproduction number is unit-less). There is a suggestion in Panel C that this varies over time. The calibration of the estimates is very variable, and overall 50.5% [95% CI 50.0% – 50.9%] of the estimates include the actual value in their confidence intervals. Panel D demonstrates that this is far worse when case numbers are high and during step changes in the reproduction



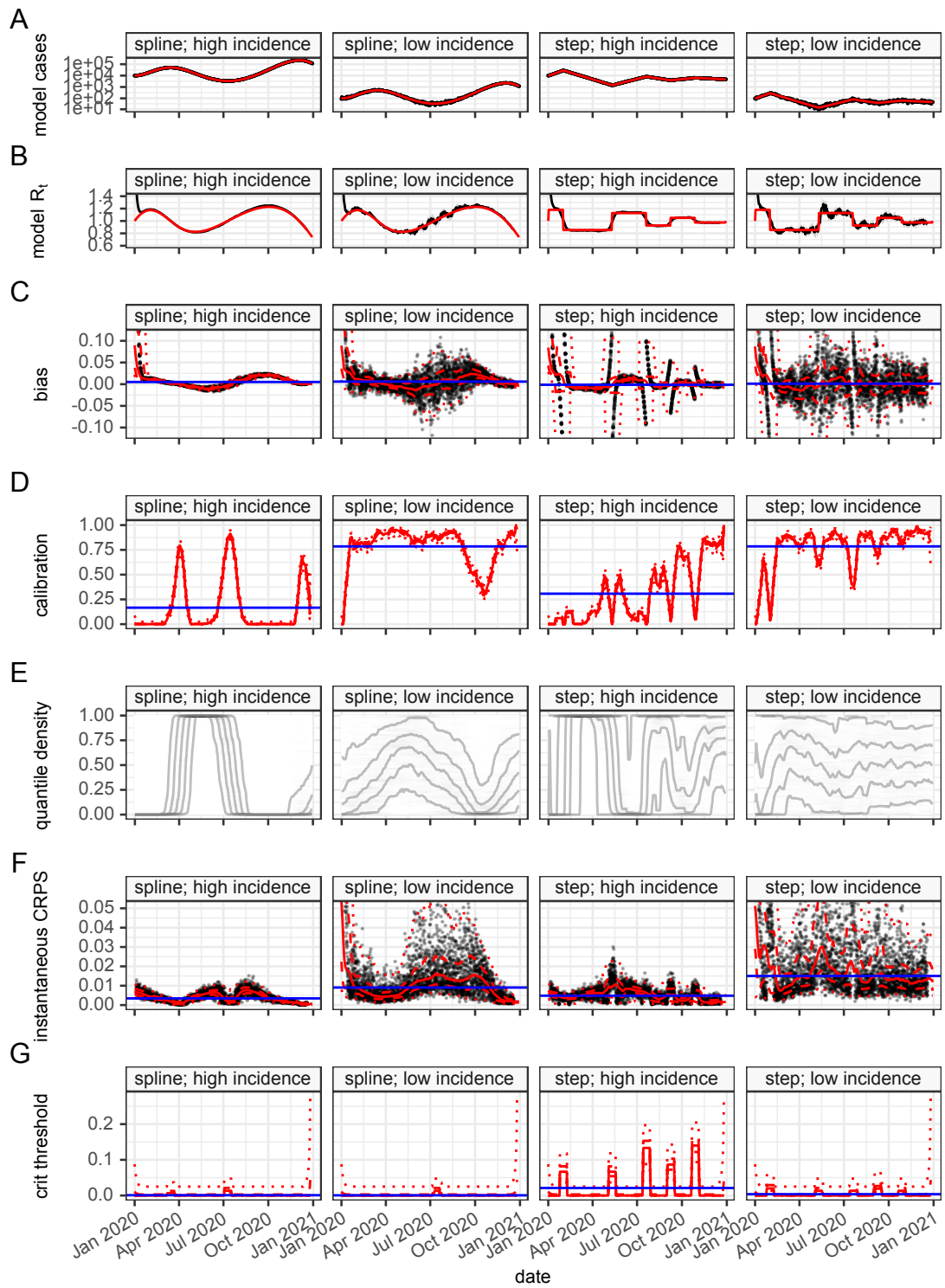


Figure A.6: Time series of estimate quality metrics for simulations with no weekly variability show which parts of the time-series pose the greatest problem for the estimation method. Panel A shows the modelled case count. Panel B shows estimate of the reproduction number. Panel C shows the bias of each individual  $R_t$  estimate. Panel D shows the calibration of individual estimates. Panel E shows the quantile density over time as 10%, 30% 50% 70% and 90% quantiles (rolling 28 day window). In panel F the instantaneous CRPS for individual estimates. In Panel G the critical threshold measure over time.

number, which is due to over precision in the estimates. The quantile density (panel E) backs this up with the distribution swinging from one extreme to the other in all the parts of the simulations with high incidence, implying an over-precise estimate. However although the calibration is poor and the quantile density is far from a flat distribution, in the higher incidence simulations the bias of the estimates is smaller. When these observations are taken together, despite the over precision, we see in panel E that the CRPS is in fact lowest for the higher incidence simulations. Overall we can conclude that that this estimation methodology is most accurate with high case numbers but produces slightly biased estimates with excessive confidence. It performs less well when there is an abrupt change in  $R_t$  (panel G), and this means that 0.7% [95% CI 0.6% – 0.8%] of the estimates using this method incorrectly predict growth when in fact the epidemic is in decline, or vice-versa.

## A.5 Summary

This appendix describes a methodology for verifying the estimation of  $R_t$  using synthetic data sets designed to highlight particular issues, by comparing the posterior distributions of estimates to the known  $R_t$  values used to generate the data sets. The different methods compared here have different degrees of delay between the actual  $R_t$  using in creating the data set and the estimate ranging from 7.75 to 10.85 days. The performance of the estimation methods can be summarised with 5 metrics, the bias, calibration, quantile deviation, continuous ranked probability score, and the critical threshold measure. At a summary level the combination of these metrics allows performance of different estimate methods to be quantitatively compared, and demonstrate that despite a degree of over precision, and resulting lower calibration, the 14 day estimator performs best overall. Furthermore by comparing performance of different estimation methods against different simulations with specific features, we can qualitatively assess situations in which individual estimation methodologies perform better or worse. A detailed analysis of the time series of the metrics gives us

further insight into situations where each given estimate method may under-perform.

This validation methodology is extensible to other metrics other than  $R_t$  where a set of confidence intervals or quantiles are available, particularly estimates of growth rate and incidence.

# **B. Renewal equation reproduction number estimation and Jepidemic implementation validation**

This appendix provides a detailed review of the renewal equation methodology for estimating the effective reproduction number as presented by Cori et al. (Cori et al., 2021; Cori et al., 2013; Thompson et al., 2019). This supported much of the analysis contained in this thesis, and was applied to data from the SARS-CoV-2 outbreak in the UK to support the work of the Scientific Pandemic Influenza - Modelling subgroup (SPI-M) and track the progress of the epidemic in the UK. The implementation of this method changed over the course of the pandemic to address specific issues that arose. These changes have been implemented in a Java library with bindings to the R language called “jepidemic”, which is open source and available on GitHub (Challen, 2022). The implementation of this library diverges from the reference implementation, EpiEstim, and this appendix summarises those changes and quantifies the benefits expected in terms of estimation accuracy, using the validation methodology described in appendix A.

## B.1 Review of the renewal equation method for estimating the effective reproduction number.

We assume  $I_0, I_1, \dots, I_t$  is a time series of infection counts, assumed to be a single sample drawn from some time varying discrete probability distribution with expected value  $\bar{I}_t$ . The infectivity profile is another probability distribution, that defines the likelihood that a case infected at time  $t$  resulted from a case infected between the times  $t - s$  and  $t - s + 1$ , and as defined here most often represented in discrete form  $\omega_1, \omega_2, \dots, \omega_s$ . This definition implies that  $\omega_{s \leq 0} = 0$  as that would apply to secondary infections resulting from primary infections in the future. The discrete time measure  $s$  here represents the upper bound of the equivalent continuous unit time interval, rather than, for example, the middle of the interval.

As in appendix A, we can define the backward-looking effective reproduction number as the inverse ratio of the number of primary infections that cause the secondary infections observed at time  $t$ , this is known as the instantaneous reproduction number,  $R_t^i$ . In an evolving epidemic the instantaneous reproduction number is able to be calculated using data that has already been observed, and is hence a more useful quantity than the other forms of the reproduction number presented in Appendix A. The rest of this summary considers only the instantaneous version of the effective reproduction number, which we refer to as the reproduction number or  $R_t$ .

$$R_t^i = \frac{\bar{I}_t}{\sum_{s=1}^t \bar{I}_{t-s} \omega_s} \quad (1)$$

With the definitions above, we consider the number of new cases on a day  $I_t$ , to be the number of cases observed in previous time points convolved by the infectivity profile, and scaled by the reproduction number. From this we define the quantity  $\Lambda_t$  as the number of primary cases that resulted in a secondary case at time  $t$ , and which

is the denominator in (1). This definition disregards the possible role of co-infection, of an individual infectee by multiple infectors and assumes secondary cases result from one, and only one, primary infection.

$$\Lambda_t = \sum_{s=1}^t I_{t-s} \omega_s \quad (2)$$

$$E[I_t | I_0, \dots, I_{t-1}, \omega, R_t] = R_t \Lambda_t$$

We also assume that as a count of infections, the case incidence can be modelled as a Poisson distributed quantity,  $I_t \sim Pois(\lambda_t)$  and therefore  $\lambda_t = \bar{I}_t$ . Given the infectivity profile distribution  $\omega$ , the number of cases we expect to see on a given day, is given by the Poisson distribution probability density function:

$$P(I_t) = \frac{\lambda_t^{I_t} e^{-\lambda_t}}{I_t!}$$

$$\lambda_t \approx E[I_t | I_0, \dots, I_{t-1}, \omega, R_t] \quad (3)$$

$$P(I_t | I_0, \dots, I_{t-1}, \omega, R_t) = \frac{(R_t \Lambda_t)^{I_t} e^{-R_t \Lambda_t}}{I_t!}$$

We are interested in producing estimates of the reproduction number that are conditioned on the data we have available. To do this we assume  $R_t$  is constant over a short time period of  $\tau$  days prior to and including the date of the estimate  $[t - \tau + 1; t]$  (and defined as  $R_{t,\tau}$ ).

As an aside, we use the following relationship:

$$\begin{aligned}
P(A_3|A_2, A_1) \times P(A_2|A_1) &= \\
&= \frac{P(A_3, A_2, A_1)}{P(A_2, A_1)} \times \frac{P(A_2, A_1)}{P(A_1)} \\
&= \frac{P(A_3, A_2, A_1)}{P(A_1)} \\
&= P(A_3, A_2|A_1)
\end{aligned}$$

Consider the combined probability of observing  $I_{t-\tau+1} \dots I_t$ , given the other information available to us, we can express this as the product:

$$\begin{aligned}
&P(I_t|I_0, \dots, I_{t-1}, \omega, R_{t,\tau}) \times \\
&P(I_{t-1}|I_0, \dots, I_{t-2}, \omega, R_{t,\tau}) \times \\
&P(I_{t-2}|I_0, \dots, I_{t-3}, \omega, R_{t,\tau}) \times \\
&\quad \dots \\
&P(I_{t-\tau+1}|I_0, \dots, I_{t-\tau-1}, \omega, R_{t,\tau}) \\
&= P(I_{t-\tau+1}, \dots, I_t|I_0, \dots, I_{t-\tau}, \omega, R_{t,\tau})
\end{aligned}$$

Furthermore substituting (3) for the left hand side we derive the following expression for the combined probability of observing  $I_{t-\tau+1} \dots I_t$ :

$$\begin{aligned}
P(I_{t-\tau+1}, \dots, I_t|I_0, \dots, I_{t-\tau}, \omega, R_{t,\tau}) &= \prod_{s=t-\tau+1}^t \frac{(R_{t,\tau}\Lambda_s)^{I_s} e^{-R_{t,\tau}\Lambda_s}}{I_s!} \\
&= R_{t,\tau}^{\sum_{s=t-\tau+1}^t I_s} e^{-R_{t,\tau}(\sum_{s=t-\tau+1}^t \Lambda_s)} \prod_{s=t-\tau+1}^t \frac{\Lambda_s^{I_s}}{I_s!}
\end{aligned} \tag{4}$$

To make use of the mathematical property of the conjugate prior of the Poisson distribution, we further assume a prior belief that  $R_{t,\tau}$  is Gamma distributed with shape parameter  $\alpha$  and rate parameter  $\beta$  and hence by definition:

$$P(R_{t,\tau}) = \frac{\beta^\alpha}{\Gamma(\alpha)} R_{t,\tau}^{\alpha-1} e^{-\beta R_{t,\tau}} \quad (5)$$

We wish to determine the posterior probability of  $R_t$  given the evidence  $I_0, \dots, I_t, \omega$ , i.e. we wish to identify  $P(R_{t,\tau}|I_0, \dots, I_t, \omega)$ . We have a prior probability  $P(R_{t,\tau})$ , and an expression for the likelihood of  $I_t - \tau + 1, \dots, I_t$  given  $I_0, \dots, I_{t-1}$ , the infectivity profile  $\omega$  and  $R_{t,\tau}$ . To do this we use Bayes theorem to restate the posterior probability of the relationship  $P(A \cup B) = P(A|B)P(B)$  in two stages, firstly:

$$P(R_{t,\tau}|I_0, \dots, I_t, \omega) = \frac{P(R_{t,\tau}, I_0, \dots, I_t, \omega)}{P(I_0, \dots, I_t, \omega)}$$

And secondly:

$$P(R_{t,\tau}, I_0, \dots, I_t, \omega) = P(R_{t,\tau}, I_{t-\tau+1}, \dots, I_t | I_0, \dots, I_{t-\tau}, \omega) P(I_0, \dots, I_{t-\tau}, \omega)$$

Substituting and re-organising:

$$P(R_{t,\tau}|I_0, \dots, I_t, \omega) \frac{P(I_0, \dots, I_t, \omega)}{P(I_0, \dots, I_{t-\tau}, \omega)} = P(R_{t,\tau}, I_{t-\tau+1}, \dots, I_t | I_0, \dots, I_{t-\tau}, \omega)$$

We have expressions for both the evidence  $P(I_{t-\tau+1}, \dots, I_t | I_0, \dots, I_{t-\tau}, \omega, R_{t,\tau})$  and prior belief  $P(R_{t,\tau})$ , and we can relate these to the right hand side of the previous expression:



$$P(R_{t,\tau}, I_{t-\tau+1}, \dots, I_t | I_0, \dots, I_{t-\tau}, \omega) = \frac{P(I_{t-\tau+1}, \dots, I_t | I_0, \dots, I_{t-\tau}, \omega, R_{t,\tau})P(R_{t,\tau})}{P(I_{t-\tau+1}, \dots, I_t | I_0, \dots, I_{t-\tau}, \omega)}$$

And combining these last two expressions gives us:

$$P(R_{t,\tau} | I_0, \dots, I_t, \omega) \frac{P(I_0, \dots, I_t, \omega)}{P(I_0, \dots, I_{t-\tau}, \omega)} = \frac{P(I_{t-\tau+1}, \dots, I_t | I_0, \dots, I_{t-\tau}, \omega, R_{t,\tau})P(R_{t,\tau})}{P(I_{t-\tau+1}, \dots, I_t | I_0, \dots, I_{t-\tau}, \omega)}$$

This includes components which are not conditional in any way on  $R_t$ . Given  $I_0, \dots, I_t, \omega$  these components are constant:

$$\frac{P(I_0, \dots, I_{t-\tau}, \omega)}{P(I_0, \dots, I_t, \omega)P(I_{t-\tau+1}, \dots, I_t | I_0, \dots, I_{t-\tau}, \omega)} = K$$

Which gives us the following expression for the posterior of  $R_t$  given our prior belief and the evidence:

$$P(R_{t,\tau} | I_0, \dots, I_t, \omega) = KP(I_{t-\tau+1}, \dots, I_t | I_0, \dots, I_{t-\tau}, \omega, R_{t,\tau})P(R_{t,\tau})$$

Using the expressions for the likelihood (4), and the prior probability (5) derived above we can express the posterior as:

$$\begin{aligned}
P(R_{t,\tau}|I_0, \dots, I_t, \omega, \alpha, \beta) &= K \left( \prod_{s=t-\tau+1}^t \frac{(R_{t,\tau} \Lambda_t)^{I_s} e^{-R_{t,\tau} \Lambda_s}}{I_s!} \right) \left( \frac{\beta^\alpha}{\Gamma(\alpha)} R_{t,\tau}^{\alpha-1} e^{-\beta R_{t,\tau}} \right) \\
&= K R_{t,\tau}^{\alpha + \sum_{s=t-\tau+1}^t I_s - 1} e^{-R_{t,\tau} (\beta + \sum_{s=t-\tau+1}^t \Lambda_s)} \left( \prod_{s=t-\tau+1}^t \frac{\Lambda_s^{I_s}}{I_s!} \right) \left( \frac{\beta^\alpha}{\Gamma(\alpha)} \right)
\end{aligned}$$

Which we noting has a form similar to a Gamma distribution with shape  $\alpha'$  and scale  $\beta'$ :

$$\begin{aligned}
\alpha' &= \alpha + \sum_{s=t-\tau+1}^t I_s \\
\beta' &= \beta + \sum_{s=t-\tau+1}^t \Lambda_s
\end{aligned} \tag{6}$$

This leads us to the conclusion that the posterior distribution of  $R_t$  is also Gamma distributed, with shape ( $\alpha'$ ) and rate ( $\beta'$ ) with a constant normalising factor which can be ignored:

$$\begin{aligned}
P(R_{t,\tau}|I_0, \dots, I_t, \omega) &= \frac{\beta'^{\alpha'}}{\Gamma(\alpha')} R_{t,\tau}^{\alpha'-1} e^{-R_{t,\tau} \beta'} \left[ K \left( \prod_{s=t-\tau+1}^t \frac{\Lambda_s^{I_s}}{I_s!} \right) \left( \frac{\Gamma(\alpha') \beta'^{\alpha'}}{\Gamma(\alpha) \beta'^{\alpha'}} \right) \right] \\
R_{t,\tau}|I_0, \dots, I_t, \omega &\sim \text{Gamma} \left( \alpha + \sum_{s=t-\tau+1}^t I_s, \beta + \sum_{s=t-\tau+1}^t \Lambda_s \right)
\end{aligned}$$

This final expression for the reproduction number explicitly integrates information from the last  $t \dots t - \tau + 1$  time points. However within the  $\Lambda_s$  term from (2) there is also information stretching back further into the past, depending on the nature of  $\omega$ . In reality the duration from an infector and an infectee in practice is limited, and for SARS-CoV-2 we think secondary infections are rare after 10 days. In this case if we consider  $\omega$  to have a finite  $N_\omega$  terms, then knowledge of the time series between  $I_{t-\tau-N_\omega} \dots I_t$  is sufficient to make an estimate of  $R_{t,\tau}$ .

## B.2 Limitations

Considering again the difference between the case based and the instantaneous reproduction number, it is a property of the instantaneous reproduction number that, in the face of a step change in the case based reproduction number, the instantaneous reproduction number will only fully account for that change when  $N_\omega$  days have elapsed. With the method presented here the additional delay introduced by the windowing must be accounted for when relating the estimates of  $R_t$  to exact points in time, and relating them to the case based reproduction number. It is also of note that it is difficult to incorporate anomalous or missing data into the method presented here as a single missing value invalidates the estimates over the next  $N_\omega + \tau$  time points.

As a final observation, the coefficient of variation ( $\kappa$ ) of a gamma distribution is the reciprocal of the square root of the shape parameter, which for the  $R_t$  estimate is give by (6):

$$\begin{aligned}\kappa &= \frac{\text{sd}}{\text{mean}} \\ \kappa &= \frac{\sqrt{\frac{\alpha}{\beta^2}}}{\frac{\alpha}{\beta}} \\ \kappa &= \frac{1}{\sqrt{\alpha}} \\ \kappa_{R_t} &= \frac{1}{\sqrt{\alpha + \sum_{s=t-\tau+1}^t I_s}}\end{aligned}$$

The form of  $\kappa_{R_t}$  is highly influenced by the count of infections. When infection numbers are in the thousands per day, the coefficient of variation becomes small regardless of the prior parameterisation. This is independent of the infectivity profile and leads to very certain estimates of  $R_t$  particularly when infection numbers are large for a sustained period of time. It is not clear whether this certainty is always

appropriate. An inherent assumption that the observed infections ( $I_t$ ) are a true representation of the expected value of the infections ( $\bar{I}_t$ ), and that these are Poisson distributed is core to this method, and over-dispersion of the observed case counts is not adjusted for.

The method presented here represents an estimate over a time window where  $R_t$  is assumed to be constant. When this is in fact not the case, and  $R_t$  is changing rapidly, the violation of this assumption leads to a certain but rapidly changing estimate that does not reflect reality. Shortening the time window over which the estimate is made in this situation may help, but this may in turn lead to excessive variation in central estimates particularly in the case where there is a weekly cycle to observations.

### **B.3 Implementation considerations**

The reference implementation of this method is provided by the R package EpiEstim. This has a range of features and configuration. The main element of this are various ways to configure the infectivity profile, either as a parameterised probability distribution, which is then discretised, or directly as an empirical set of weights ( $\omega$ ). There is also the option to provide uncertainty around the infectivity profile either as uncertainty bounds on the distribution parameters, which are then sampled to produce a set of parameterised distributions, which are then in turn discretised, or to provide that infectivity profile uncertainty directly as a sequence of empirical distributions ( $\omega_a, \omega_b, \omega_c, \dots$ ). In either event the algorithm progresses using such a sequence of empirical distributions, each one of which representing one possible infectivity profile. The estimate of  $R_t$  for all profiles ( $R^{profiles}$ ) is then calculated as a combination of all the possible estimates of  $R_t$  given each of the infectivity profiles, and a size of window  $\tau$ .

$$R^{profiles}(t, \tau) = \{R_{t,\tau,\omega} : \omega \in \omega_a, \omega_b, \omega_c, \dots\}$$

### B.3.1 Prior selection

The default implementation uses the same fixed prior gamma distribution for  $R_t$  for all points in the time series. This is configurable but recommended to be set to a value (e.g. 5). Reversion to the prior  $R_t$  when case incidence is low, for example at the start of the time series, means  $R$  at very low incidence may become biased towards the set prior value.

An alternative to this fixed prior, is to use an “informed” prior that assumes that estimates of  $R_t$  are likely to be continuous in time, and uses previous posterior  $R_t$  estimates to calculate priors for the next time point. The previous time point posteriors are combined with a scale factor  $k$  that increases the standard deviation of the prior distribution compared to the posterior of the previous time step, whilst keeping the mean constant. This essentially allows the prior at time  $t$  to be the posterior at time  $t - 1$  plus a random walk, the variation of which is controlled by  $k$ . By enforcing the continuity in time the aim is to stabilise noisy estimates of  $R_t$  when incidence is low, however this strategy may worsen the over-precise estimates seen when case numbers are high.

$$\begin{aligned}
E(R_{t,prior}) &= E(R_{t-1,posterior}) \\
V(R_{t,prior}) &= k^2 V(R_{t-1,posterior}) \\
\alpha_{t,prior} &= \frac{E(R_{t-1,posterior})^2}{k^2 V(R_{t-1,posterior})} \\
\alpha_{t,prior} &= \frac{\alpha_{t-1,posterior}^2}{\beta_{t-1,posterior}^2} \frac{\beta_{t-1,posterior}^2}{k^2 \alpha_{t-1,posterior}} \\
\alpha_{t,prior} &= \frac{\alpha_{t-1,posterior}}{k^2} \\
\beta_{t,prior} &= \frac{E(R_{t-1,posterior})}{k^2 V(R_{t-1,posterior})} \\
\beta_{t,prior} &= \frac{\alpha_{t-1,posterior}}{\beta_{t-1,posterior}^2} \frac{\beta_{t-1,posterior}^2}{k^2 \alpha_{t-1,posterior}} \\
\beta_{t,prior} &= \frac{\beta_{t-1,posterior}}{k^2} \\
R_{t,prior} &\sim \text{Gamma}\left(\frac{\alpha_{t-1,posterior}}{k^2}, \frac{\beta_{t-1,posterior}}{k^2}\right)
\end{aligned}$$

### B.3.2 Windowing strategy and posterior estimate selection

In appendix A we observed the selection of a windowing parameter may have a significant effect on the bias variance trade-off, and selecting a single window may produce estimates that may either be too precise or too noisy. Picking the right value is also constrained by the observation that in the face of weekly periodicity of case incidence,  $R_t$  estimates may over-fit the data when windows are too short.

It is computationally efficient to calculate an array of windows at the same time, and with all windows available we open up some options to select the best posterior in a different way. The first possibility is to adopt an adaptive strategy allows window selection to be determined by the case incidence within the window ( $\sum I_t$ ). By selecting longer windows where case numbers are small we both improve the certainty of the estimate, and reduce its noise, and when case numbers are high we reduce the window size to allow more rapid adaptation to changing  $R_t$ , with less risk of over-fitting. This strategy is summarised as follows where  $R^{adaptive}$  is the set of estimates where the window size  $\tau$  is the smallest possible value that encompasses enough data.

$$R^{adaptive}(t, \tau_{min}, \tau_{max}, I_{min}) = \{R^{profiles}(t, \tau) : \tau = \min(\tau_{max} \mid \tau_{min} \leq \tau; I_{min} \leq \sum_{s=t-\tau+1}^t I_s)\}$$

In the validation study we observed that the EpiEstim reference implementation with 14 day window produces a lagged, over-precise estimate. The degree of lag determines the overall accuracy of the method when  $R_t$  is rapidly changing. This happens because the assumption that  $R_t$  is constant over a window period  $\tau$  is violated. The precision of the  $R_t$  estimate is unwarranted in this situation. Once we have a range of windows calculated we may address this problem. Firstly, with a range of different time windows for any given time point, there are a set of estimates of  $R_t$  that are equally valid, and which capture different assumptions about the length of time over which  $R_t$  is constant. Secondly, if the time point in consideration is  $s$  days in the past, there are also estimates from later time points, where  $s \leq t + \tau$  that are also relevant to time  $t$ . In this case the  $R_s, \tau$  estimate assumes the reproduction number is constant over the time period  $s - \tau - 1 \dots s$ . If we consider allowing  $\tau$  to vary between two limits ( $\tau_{min} \leq \tau \leq \tau_{max}$ ) then we can describe the set of all estimates of effective  $R_t$  that are relevant to a single time point as  $R_{t,all}$ :

$$R^{all}(t, \tau_{min}, \tau_{max}) = \{R^{profiles}(s, \tau) : \tau_{min} \leq \tau \leq \tau_{max}; t \leq s \leq t + \tau\}$$

This is better explained visually and Figure B.1 demonstrates at time  $s = 10$  that estimates where  $10 \leq s + \tau \leq 16$  are all of relevance to the time point  $s$ . All possible combinations of window and future time point  $R_t$  estimates that are relevant to a specific point in time are shown. Each red point represents an  $R_t$  estimate that is based on some assumption about the reproduction number on day 10, and all of these estimates may be combined to produce a final  $R_t$  estimate for day 10. This set of estimates provides a broader set of assumptions than a single window can

provide and therefore may reduce the unwanted over-precision of estimates when  $R_t$  is changing rapidly.

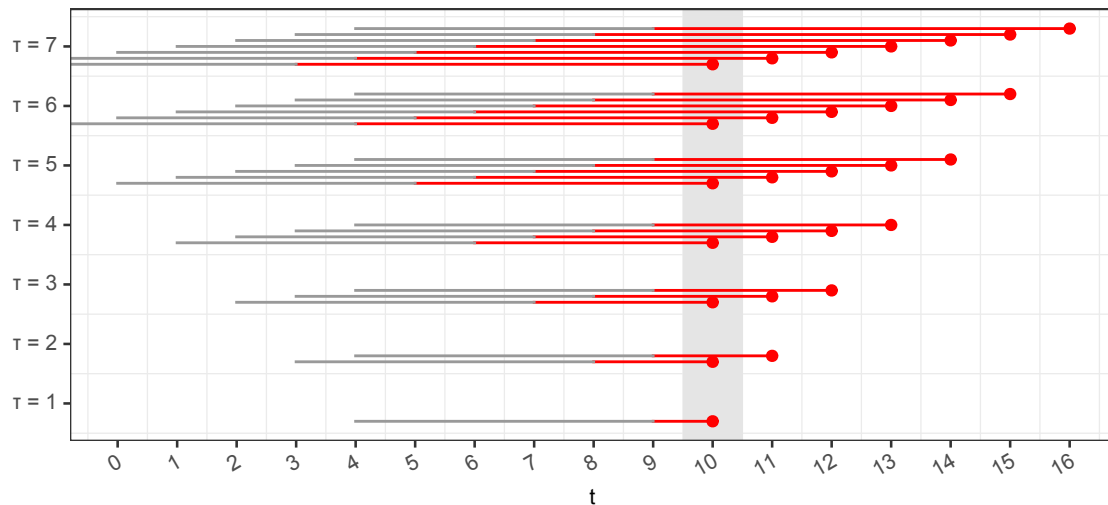


Figure B.1: A graphical representation of the information involved in estimates of  $R_t$  using the renewal equation method with different length windows (y-axis). The highlighted estimates (red dots) all use estimation windows that span the 10th day and during each of these windows there is an assumption of constant  $R_t$ . In situations where the true  $R_t$  is dynamic, combining these estimates may reflect the overall uncertainty.

### B.3.3 Combining posterior estimates

In all the estimates thus far ( $R^{profiles}$ ,  $R^{adaptive}$ , and  $R^{all}$ ) the estimate of  $R_t$  take the form of a set of posterior gamma distributions for each time point. This is a result of the fact that there are multiple infection profiles, expressing uncertainty, or there are multiple windows over which the estimate is calculated, or there are multiple days over which the estimates are collected. These estimates must be combined, and there are different possible strategies for doing this.

In the reference implementation, multiple estimates resulting from multiple infection profiles, are combined by constructing an empirical distribution from Monte-Carlo random sampling of the posterior Gamma distributions of all estimates. Quantiles are estimated from this empirical distribution. This is not deterministic, and for a reasonable degree of accuracy is computationally expensive.



When we consider that the only information we need to get from the mixture of posterior distributions is a set of quantiles, an alternative strategy therefore is to construct a mixture distribution from the set of posteriors and solve it numerically for the quantiles. Given random sampling does the same process in an un-targeted way this is actually less overall effort and provides a deterministic result.

The posterior estimates should be similar to one another. A reasonable approximation therefore is to consider the mixture distribution as another Gamma distribution with first and second moments matching those of the mixture distribution. This is quick to calculate and allows us to rapidly combine the potentially large number of estimates that arise from the multiplicative combinations of infection profiles, variable window length and variable day of estimate. The quality of this estimate will depend on how different the distributions are from each other, but this again produces a deterministic result. The parameterisation of the estimated gamma distribution is expressed below in terms of shape ( $\alpha$ ) and rate ( $\beta$ ) parameters, composed of a mixture of gamma distributions ( $\sim \text{Gamma}(\alpha_i, \beta_i)$ ) from the posterior estimates.

$$\begin{aligned}
E[X] = \mu &= \frac{\alpha}{\beta} = \frac{1}{n} \sum_1^n \mu_i = \frac{1}{n} \sum_1^n \frac{\alpha_i}{\beta_i} \\
E[(X - \mu)^2] = \sigma^2 &= \frac{\alpha}{\beta^2} = E[X^2] - \mu^2 \\
&= \frac{1}{n} \left( \sum_1^n E[X_i^2] \right) - \mu^2 \\
&= \frac{1}{n} \left( \sum_1^n \sigma_i^2 + \mu_i^2 \right) - \mu^2 \\
&= \frac{1}{n} \left( \sum_1^n \frac{\alpha_i}{\beta_i^2} + \left( \frac{\alpha_i}{\beta_i} \right)^2 \right) - \left( \frac{1}{n} \sum_1^n \frac{\alpha_i}{\beta_i} \right)^2 \\
&= \frac{1}{n} \left( \sum_1^n \frac{\alpha_i}{\beta_i^2} + \frac{\alpha_i^2}{\beta_i^2} - \frac{1}{n} \frac{\alpha_i^2}{\beta_i^2} \right) \\
&= \frac{1}{n} \left( \sum_1^n \frac{\alpha_i + \frac{n-1}{n} \alpha_i^2}{\beta_i^2} \right) \\
\alpha = \frac{\mu^2}{\sigma^2} &= \frac{1}{n} \frac{\left( \sum_1^n \frac{\alpha_i}{\beta_i} \right)^2}{\left( \sum_1^n \frac{\alpha_i + \frac{n-1}{n} \alpha_i^2}{\beta_i^2} \right)} \\
\beta = \frac{\mu}{\sigma^2} &= \frac{\left( \sum_1^n \frac{\alpha_i}{\beta_i} \right)}{\left( \sum_1^n \frac{\alpha_i + \frac{n-1}{n} \alpha_i^2}{\beta_i^2} \right)}
\end{aligned}$$

In all the strategies described above there is the potential to weight specific estimates more than others. As each infectivity profile is taken to be equally likely this only make sense when combining estimates made over many time windows, and such a weighing may be based on a function of the size of the window, or potentially a function of the number of cases observed within the window. This is beyond the scope of this description.

## B.4 Validation and comparison

The implementation of the renewal equation method has options for informed prior selection, broader posterior selection, and different methods for combining the posteriors. To evaluate the impact of these different implementation strategies we

compare them using the methods described in appendix A in the next section. In all comparisons we use the validation data set described in appendix A with a fixed infectivity profile distribution based on a discretised Gamma distribution with mean of 5 and standard deviation of 4.

### **B.4.1 Informed prior selection**

In this comparison we are looking at the performance change resulting from changing the prior selection strategy from a fixed prior to that of an informed prior. Our baseline reference methodology is the default EpiEstim configuration including a fixed prior  $R_t$  with mean of 1.2 and standard deviation of 4, and with  $R_t$  estimates calculated over a 7 days period. We vary the configuration by adopting an informed prior strategy with a step size factor  $\kappa = 1.25$  and the same 7 day window, and secondly with  $\kappa = 1.125$  and a smaller 4 day window.

The qualitative result of these changes is shown in Figure B.2, which demonstrates a modest reduction in high frequency noise, particularly in time periods where the incidence is low, for the same time period, and that this is retained when we shorten the time window from 7 to 4 days, if we also reduce the step size factor.

As the use of an informed prior constrains the rate of change of the  $R_t$  estimate in the face of sudden changes in the true value, it is expected that smaller values of  $k$ , the random walk step factor, result in estimates with more stiffness in time, and hence an increase in the delay. This is borne out by Figure B.3 in which smaller step size values result in more estimate delay. This partly defeats the purpose of introducing the informed prior as a way to shorten the window needed and hence improve responsiveness to step changes in  $R_t$ .

A quantitative comparison of the 3 methods in Figure B.4 and Table B.1 reveals

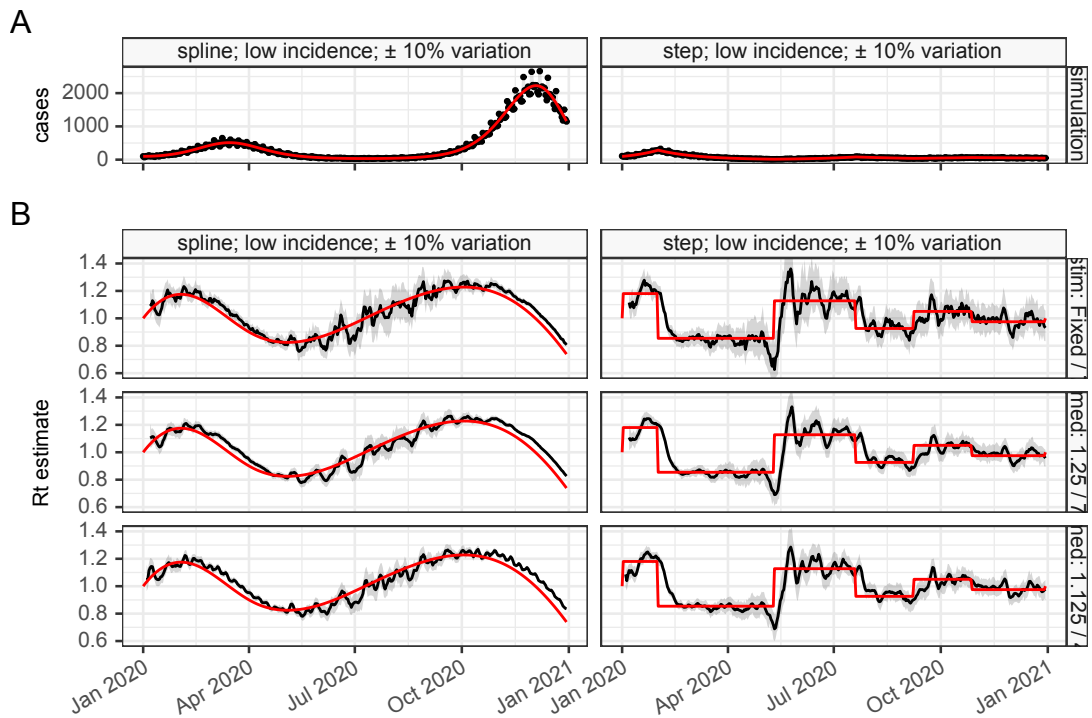


Figure B.2: Qualitative estimates of  $R_t$  (black) against simulated (red) comparing 3 methods that vary in their prior selection process

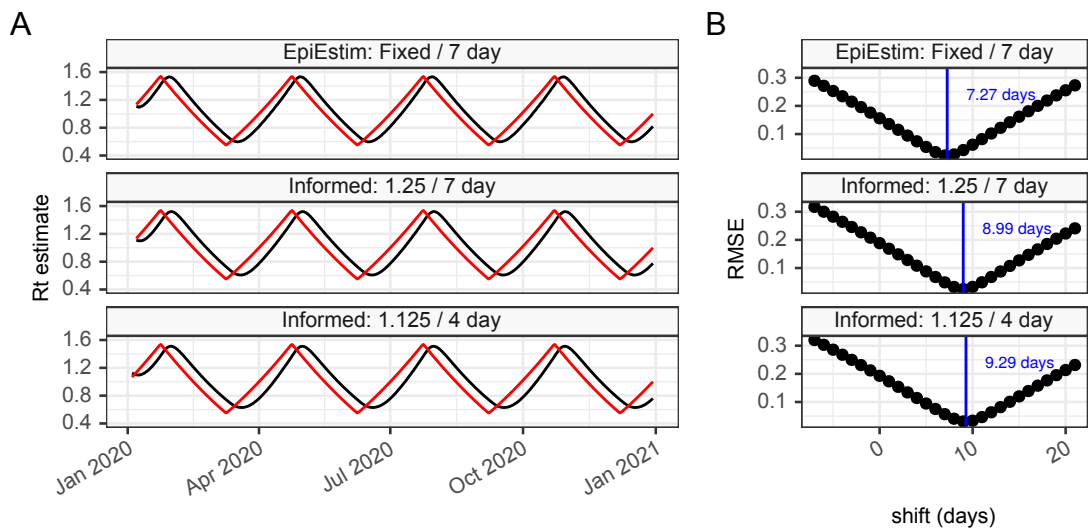


Figure B.3: Time delay analysis of  $R_t$  (black) against simulated (red) comparing 3 methods that vary in their prior selection process

that the informed prior does improve the estimate against the validation set when assessed by the continuous rank probability score (CRPS), but as predicted the informed prior also worsens the over-precision of the estimator, with I shaped violin plots, and increasing quantile deviation scores. Although not demonstrated here we expect the improvement in CRPS to be most noticeable for the smoothly varying

validation scenarios rather than step changes which we think it will perform less well on.

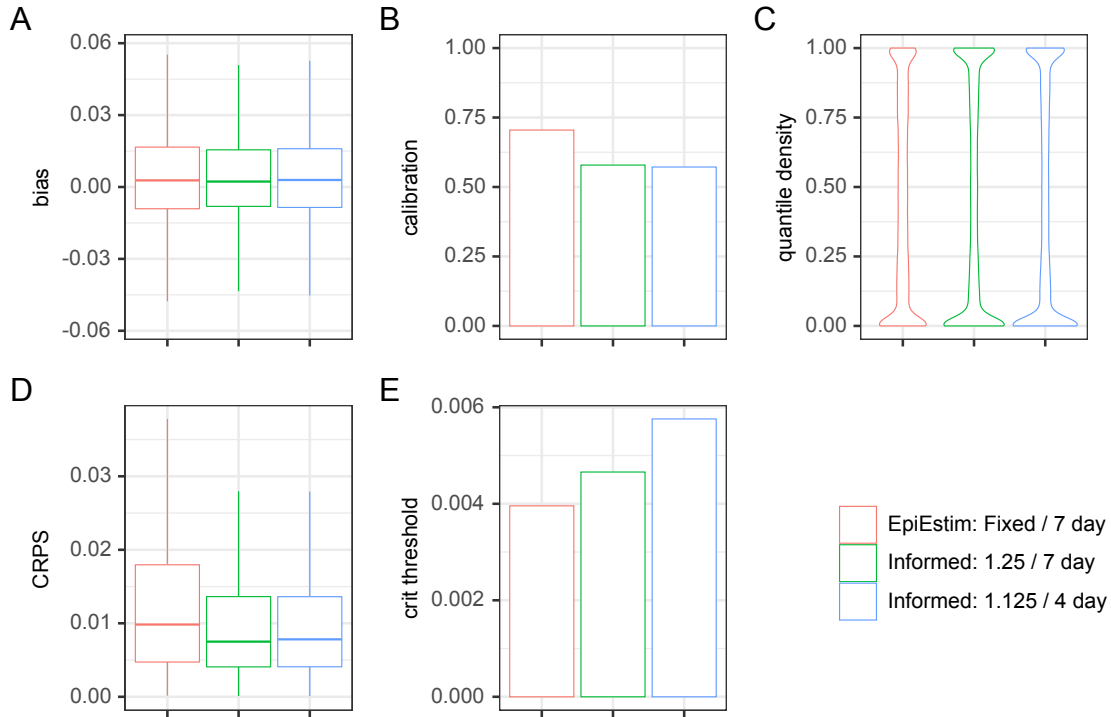


Figure B.4: Quantitative analysis of  $R_t$  (black) against simulated (red) comparing 3 methods that vary in their prior selection process

Table B.1: Summary of quantitative analysis of  $R_t$  (black) against simulated (red) comparing 3 methods that vary in their prior selection process

metric	EpiEstim: Fixed / 7 day	Informed: 1.25 / 7 day	Informed: 1.125 / 4 day
average bias	0.0028 [IQR -0.0091 – 0.017]	0.0023 [IQR -0.0081 – 0.016]	0.0029 [IQR -0.0085 – 0.016]
average calibration	70.5% [95% CI 70.1% – 70.9%]	57.9% [95% CI 57.4% – 58.4%]	57.2% [95% CI 56.7% – 57.7%]
average critical threshold	0.4% [95% CI 0.3% – 0.5%]	0.5% [95% CI 0.4% – 0.5%]	0.6% [95% CI 0.5% – 0.7%]
CRPS	0.0098 [IQR 0.0047 – 0.018]	0.0075 [IQR 0.0041 – 0.014]	0.0078 [IQR 0.0041 – 0.014]
quantile deviation	0.0744	0.109	0.109
estimate delay	7.27 days	8.99 days	9.29 days

## B.4.2 Posterior selection

In this comparison we are looking at the performance change resulting from changing the posterior selection strategy. Our baseline reference methodology is the default EpiEstim configuration including a fixed prior  $R_t$  with mean of 1.2 and standard deviation of 4. We select posterior  $R_t$  estimates that are calculated over a 7 days period. We vary the configuration by firstly allowing the posterior to be automatically

selected to ensure a minimum number of cases (in this instance 100) in the estimation window (down to a minimum of 4 days). This is the adaptive window described above.

The second comparison is against the combined posterior estimate of all windows that span a given time point, and described above. The multiple estimates of  $R_t$  for the different window sizes, and start and end dates, obtained this way are combined by empirical re-sampling as in the original EpiEstim implementation.

The qualitative result of these changes is shown in Figure B.5, which demonstrates the adaptive window results in excess high frequency noise in the scenario with weekend variation shown, compared to the reference implementation. This is an indication of over-fitting as a result of the selection of short (<7 day) windows. The “All windows” strategy on the other hand produces a stable estimate with broader confidence intervals which aligns closely to the true value.

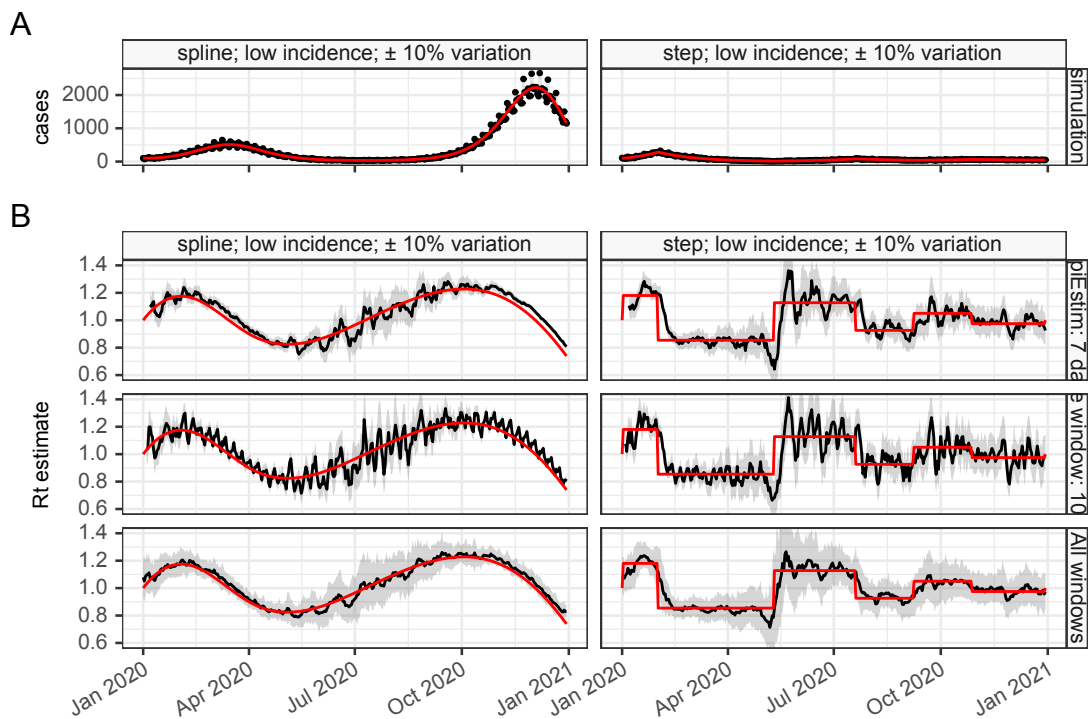


Figure B.5: Qualitative estimates of  $R_t$  (black) against simulated (red) comparing 3 methods that vary in their posterior selection process

The different posterior selection strategies involve integrating information from

different sized windows, this in turn affects the degree of time delay within the estimate as shown in Figure B.6, with the reference implementation having the most delay, whereas the “all windows” strategy includes estimates from short and long windows and therefore responds relatively quickly to change. It must be pointed out that the true value of the reproduction number is calculated using the methods of Wallinga and Lipsitch, 2007 which produces a different type of reproduction number to the instantaneous reproduction number produced by the renewal equation methods presented here.

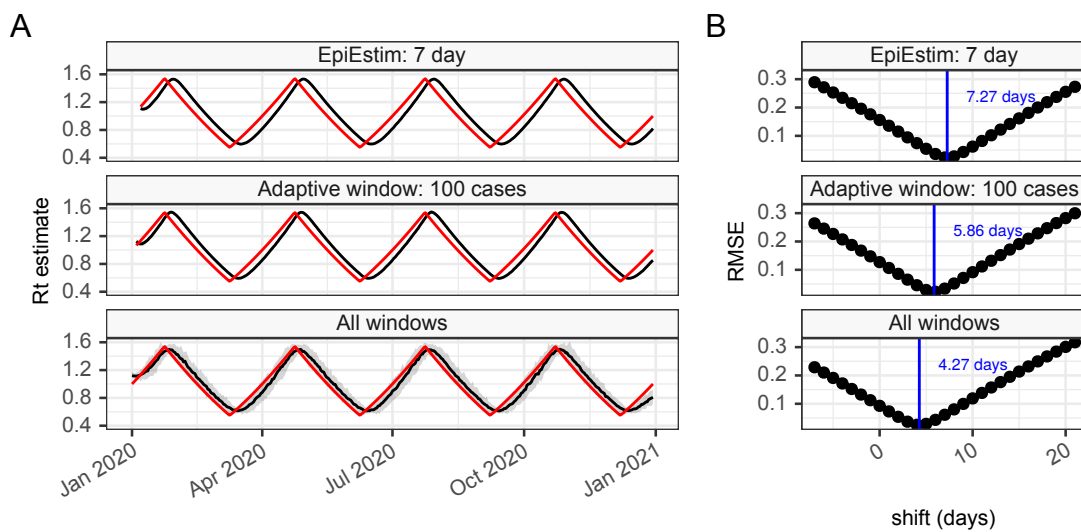


Figure B.6: Time delay analysis of  $R_t$  (black) against simulated (red) comparing 3 methods that vary in their posterior selection process

The purpose of investigating the posterior selection was to reduce the over-precision, and improve calibration of the estimates without compromising the overall performance. In Figure B.7 and Table B.2 we can see that the “All windows” strategy is well calibrated, successful in reducing over precision, with a tendency towards excessive uncertainty, seen in the O shaped quantile density plot, and negative quantile deviation score, and that overall it performs equally well as the reference implementation in terms of CRPS.

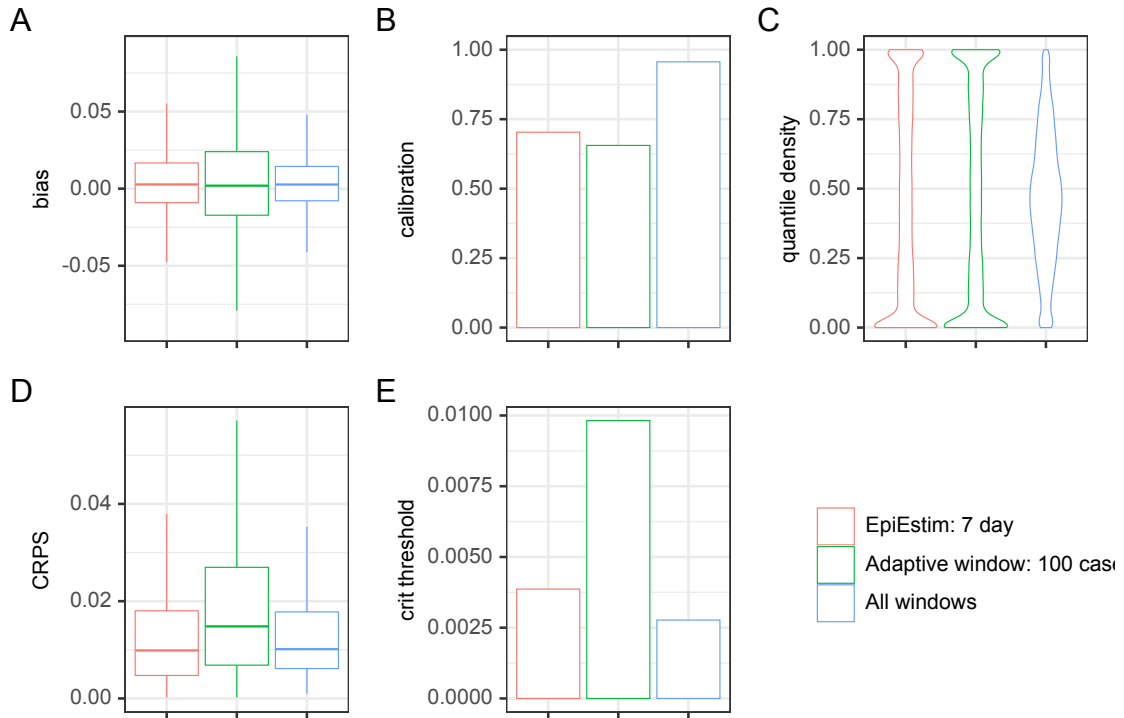


Figure B.7: Quantitative analysis of  $R_t$  (black) against simulated (red) comparing 3 methods that vary in their posterior selection process

Table B.2: Summary of quantitative analysis of  $R_t$  (black) against simulated (red) comparing 3 methods that vary in their posterior selection process

metric	EpiEstim: 7 day	Adaptive window: 100 cases	All windows
average bias	0.0027 [IQR -0.0091 – 0.017]	0.0019 [IQR -0.017 – 0.024]	0.0027 [IQR -0.0079 – 0.014]
average calibration	70.3% [95% CI 69.9% – 70.7%]	65.6% [95% CI 65.1% – 66.0%]	95.6% [95% CI 95.4% – 95.8%]
average critical threshold	0.4% [95% CI 0.3% – 0.4%]	1.0% [95% CI 0.9% – 1.1%]	0.3% [95% CI 0.2% – 0.3%]
CRPS	0.0099 [IQR 0.0047 – 0.018]	0.015 [IQR 0.0068 – 0.027]	0.01 [IQR 0.0061 – 0.018]
quantile deviation	0.0743	0.0883	-0.0481
estimate delay	7.27 days	5.86 days	4.27 days

### B.4.3 Combining posteriors

For this comparison we further examine the “All windows” strategy from above, which combines a number of different  $R_t$  posterior estimates, by varying the method in which these are combined. The original EpiEstim implementation uses random sampling to combine posteriors, and we compare this to both a formal estimation of the quantiles of a mixture of posterior estimates, and an approximation of the mixture using the method of matching moments described above. The qualitative analysis in Figure B.8 suggests this change has very limited effect on the estimate quality, and in Table B.3 this is confirmed with the quantitative metrics. As the



mixture approximation is far less computationally expensive than the other methods  
it seems this is best approach.

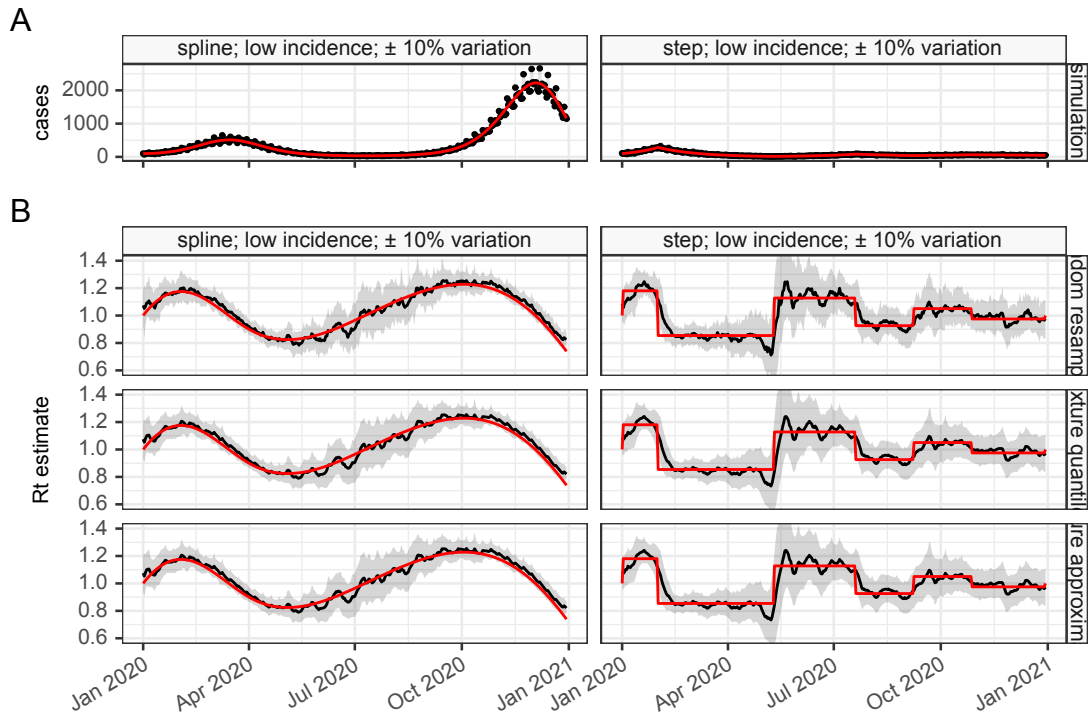


Figure B.8: Qualitative estimates of  $R_t$  (black) against simulated (red) comparing 3 methods that vary in their posterior combination process

Table B.3: Summary of quantitative analysis of  $R_t$  (black) against simulated (red) comparing 3 methods that vary in their posterior combination process

metric	Random resampling	Mixture quantiles	Mixture approximation
average bias	0.0027 [IQR -0.0078 – 0.014]	0.0027 [IQR -0.0077 – 0.014]	0.0027 [IQR -0.0077 – 0.014]
average calibration	95.6% [95% CI 95.4% – 95.8%]	95.5% [95% CI 95.3% – 95.7%]	95.5% [95% CI 95.3% – 95.7%]
average critical threshold	0.3% [95% CI 0.2% – 0.3%]	0.3% [95% CI 0.2% – 0.3%]	0.3% [95% CI 0.2% – 0.3%]
CRPS	0.01 [IQR 0.0062 – 0.018]	0.01 [IQR 0.0061 – 0.017]	0.01 [IQR 0.0061 – 0.017]
quantile deviation	-0.0482	-0.0496	-0.0496
estimate delay	4.23 days	4.24 days	4.24 days

## B.5 Overall combination

With the step-by-step analysis above we can make informed decision about how to best go about estimating  $R_t$  using the renewal equation method to address some of its limitations. Estimating  $R_t$  by combining all the estimation windows available, into a single Gamma distribution approximating the mixture of posteriors, when combined with an informed prior based on a posterior estimates from previous time-steps,

with a moderate choice for the step size factor, has potential to perform well as an estimator. This is tested below, with the step size factor of 1.25. The qualitative results of this comparison are in Figure B.9, which shows the improved method producing stable, and seemingly accurate estimates of  $R_t$ .

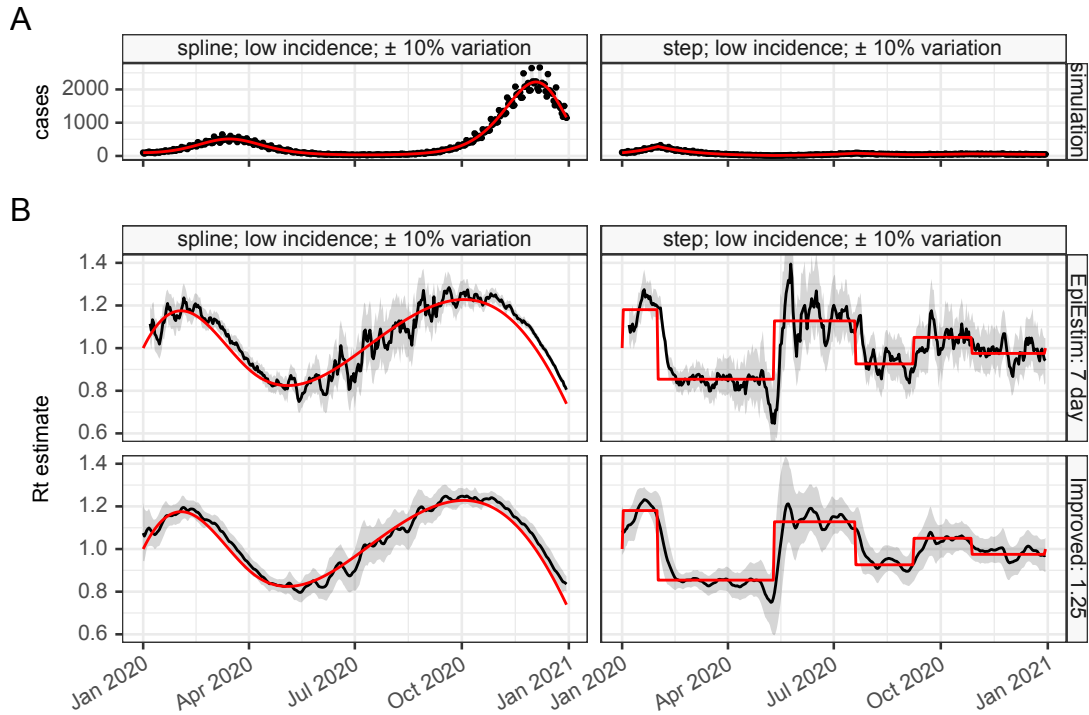


Figure B.9: Qualitative estimates of  $R_t$  (black) against simulated (red) comparing the reference implementation to a proposed improved method

As before the introduction of a informed prior introduces hysteresis and delay into the estimates compared with the fixed prior (shown in Figure B.10), but this delay is still less than the reference implementation.

A qualitative analysis in Figure B.11 and Table B.4 demonstrates the improved estimators are indeed better calibrated than the reference implementation, and perform better on the CRPS score. They have a more uniform quantile density plot and a quantile deviation score very close to 0. This suggests we have been successful in improving the estimate quality.

For completeness we examine what drives the improved estimator's performance, and in Figure B.12 we see notable improvements in calibration occur particularly in

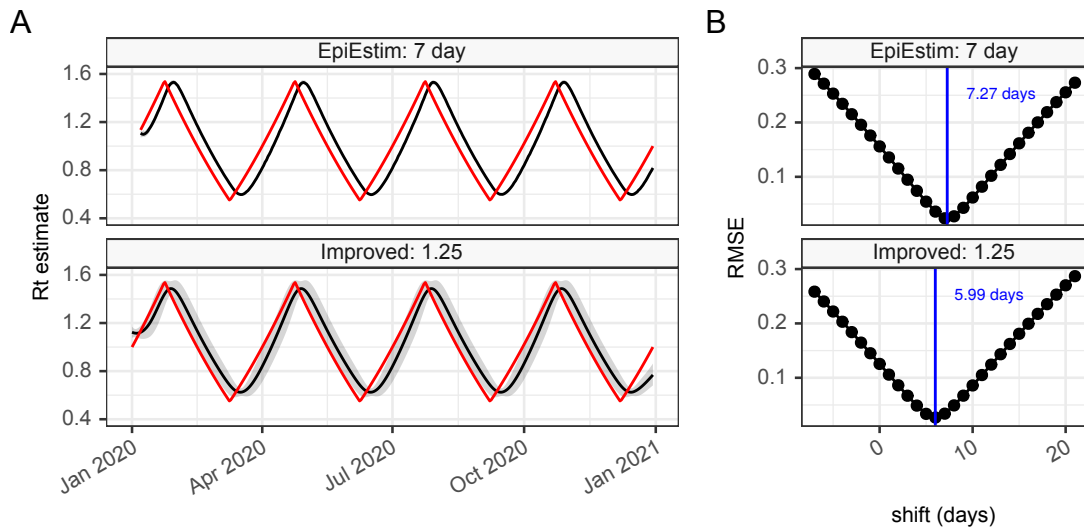


Figure B.10: Time delay analysis of  $R_t$  (black) against simulated (red) comparing the reference implementation to a proposed improved method

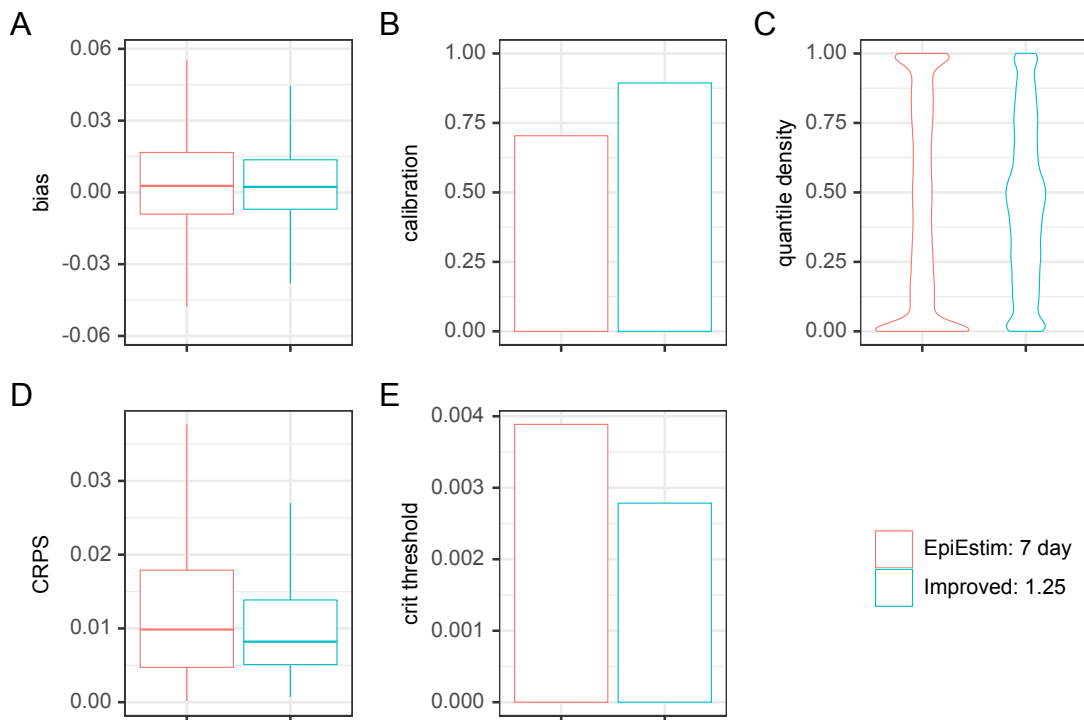


Figure B.11: Quantitative analysis of  $R_t$  (black) against simulated (red) comparing the reference implementation to a proposed improved method

the high incidence scenarios (panel G) and the reduction in over-precision is again most marked in the high incidence scenarios (panel K).

Table B.4: Summary of quantitative analysis of  $R_t$  (black) against simulated (red) comparing the reference implementation to a proposed improved method

metric	EpiEstim: 7 day	Improved: 1.25
average bias	0.0027 [IQR -0.0091 – 0.017]	0.0023 [IQR -0.007 – 0.014]
average calibration	70.4% [95% CI 70.0% – 70.8%]	89.4% [95% CI 89.1% – 89.7%]
average critical threshold	0.4% [95% CI 0.3% – 0.5%]	0.3% [95% CI 0.2% – 0.3%]
CRPS	0.0099 [IQR 0.0047 – 0.018]	0.0082 [IQR 0.0051 – 0.014]
quantile deviation	0.0744	0.00181
estimate delay	7.27 days	5.99 days

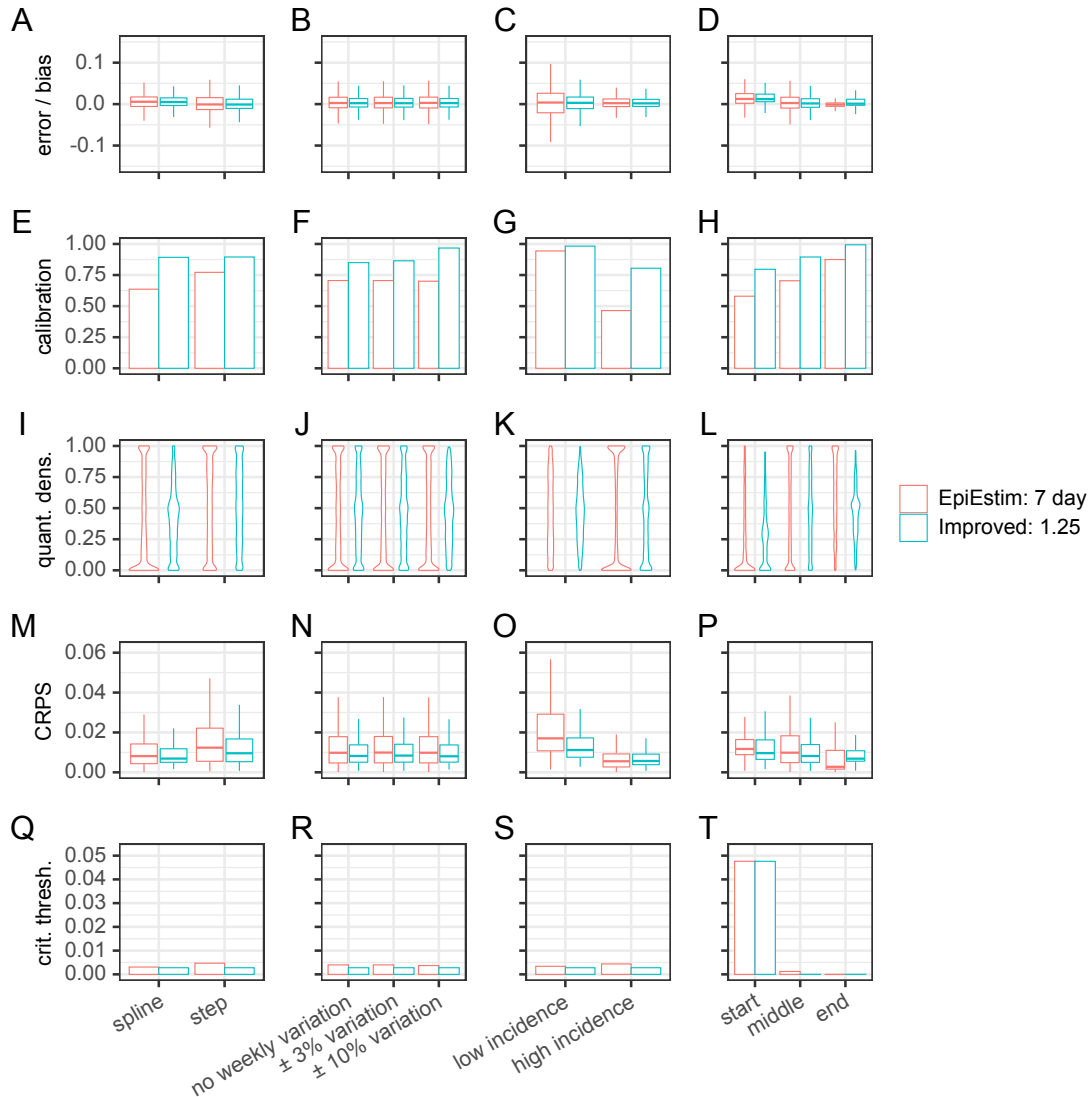


Figure B.12: Detailed breakdown of quantitative analysis of  $R_t$  (black) against simulated (red) comparing the reference implementation to a proposed improved method

## B.6 Summary

In this appendix we have described the detail of the renewal equation method for estimating  $R_t$ . We find a few limitations of the method, and particularly the over-

precise estimates that can occur when case incidence is high, and the limitations resulting from the assumption that  $R_t$  is constant over a window of time. We undertook a re-implementation of the algorithm in Java and used this to explore the possibilities of combining estimates with different window size assumptions and with different start and end dates, which provides us with a broader set of estimates that better reflects uncertainty. Our implementation also allows for the imposition of a constraint of continuity on the time-series of the estimates, which allows for more precise estimates when case numbers are small. Combined together these alterations produce a novel method for estimating the reproduction number that is better calibrated, and performs well under detailed validation.

# C. Incidence and growth rate estimation

## C.1 Introduction

The effective reproduction number ( $R_t$ ) is a useful measure of the state of the epidemic but by itself does not inform us of the likely impact that the epidemic will have. To do that we need an unbiased estimator of case counts. As the reproduction number is conditioned on the infectivity profile  $\omega$  the timing of the impact of rapidly growing case counts is also not described by the reproduction number. For this a more useful measure is the real-time exponential growth rate ( $r_t$ ), which it has been argued is more informative for decision making (Pellis et al., 2021), as combined with case counts it can inform how quickly cases will reach a particular threshold when intervention is required. The reproduction number implicitly describes a closed population, which when describing sub-populations may be unhelpful. The growth rate does not make such assumptions, can be inferred directly from data and this does not require sophisticated models.

In this appendix we describe the approach to estimation the real-time growth rate adopted throughout the majority of this thesis, and we validate that against synthetic epidemics using the validation methodology described in Appendix A.

## C.2 Estimating incidence and growth rate.

As before, we assume  $I_0, I_1, \dots, I_t$  is a time series of infection counts, drawn from some discrete probability distribution with expected value  $\bar{I}_t$ . The epidemic growth process is exponential by nature. The real-time growth rate is defined as the rate of change of the logarithm of infections:

$$\begin{aligned}\overline{I_{t+\Delta t}} &= \bar{I}_t e^{r_t \Delta t} \\ r_t &= \frac{\log(\overline{I_{t+\Delta t}}) - \log(\bar{I}_t)}{\Delta t} \\ r_t &= \frac{d\log(\bar{I}_t)}{dt}\end{aligned}\tag{7}$$

We assume  $I_t$  is a Poisson distributed quantity, with a rate parameter which is a function of time:

$$\begin{aligned}I_t &\sim \text{Poisson}(\lambda_t) \\ \bar{I}_t &= \lambda_t \\ r_t &= \frac{d\log(\lambda_t)}{dt}\end{aligned}$$

Due to the probability density function underlying the Poisson distribution estimating both  $\bar{I}_t$  and  $r_t$  from data is straightforward, in a range of frameworks. Following the methods of Loader, 1999 we represent  $\lambda_t$  as a polynomial function of time, and adopt a local maximum likelihood estimator with a logarithmic link function such that:

$$I = g(y) = e^y$$

$$y = g^{-1}(I) = \log(I)$$

$$y_t(t_i) = \langle a, A(t_i - t) \rangle = a_0 + a_1(t_i - t) + \dots + a_n(t_i - t)^n$$

$$l(y, \lambda) = \log(P(y|\lambda))$$

The global log-likelihood of a set of  $\lambda$  estimates  $\lambda_0, \lambda_1, \dots, \lambda_n$  is then given by:

$$\mathcal{L}(\lambda) = \sum_{i=1}^n l(Y_i, \lambda_i)$$

A corresponding localised log-likelihood function evaluated at time  $t$  can be created using a local kernel function  $w$ .

$$\mathcal{L}_t(\lambda) = \sum_{i=1}^n w_i(t) \cdot l(Y_i, \langle a, A(t_i - t) \rangle)$$

Which is maximised over the parameter set  $a$ , to generate a local estimate of the polynomial. From this direct estimates of the local  $\bar{I}_t$  and  $r_t$  are given by:

$$\bar{I}_t = e^{a_0} r_t = \frac{dA}{dt}(t) = a_1$$

The local kernel function is, by default, chosen to be a tri-cube function truncated between -1 and 1 and applied to a window of the data defined by a bandwidth  $n$ . Bandwidth selection can be done both as a range of  $t - n/2 \dots t + n/2$  entries within the time series or the nearest  $n$  data points:

$$w_i(u) = \left(1 - \left|\frac{i - u}{n}\right|^3\right)^3$$



Using this method, the incidence and growth rate estimation therefore depend on the bandwidth and the polynomial degree. Given that the most interesting part of the data is the right hand boundary, the choice of bandwidth selection method also is important, as this influences behaviour at this boundary.

The confidence of the estimates of  $a$ , given the logarithmic transformation are described in detail in Loader, 1999 and less straightforward. The confidence estimates of predictions rely on a variance stabilising link function, and a local regression weight diagram, and an assumption of normality in the transformed space, however the derivation of the confidence intervals on the parameters themselves is not fully detailed. As such an empirical assessment of calibration and precision of growth rate estimates is warranted.

### C.3 Estimating $R_t$ from the growth rate.

With an estimate of the growth rate available we would like to be able to make an estimate of the reproduction number, derived from this growth rate. We adopt the method of Wallinga and Lipsitch, 2007. This describes the generation interval distribution  $g(a)$  as a continuous function which describes the probability of a secondary infection related to the “age”  $a$  of a primary infection, i.e. the time from primary to secondary infection. This is combined with the Lotka-Euler equation (Anderson and May, 1992; Coale, 1972; Dublin and Lotka, 1925; Feller, 1941) to generate the following relationship between  $R$  and  $r$

$$\frac{1}{R} = \int_{a=0}^{\infty} e^{-ra} g(a) da$$

Which is identifiable as the Laplace transformation of  $g(a)$  and hence the moment

generating function  $M$  of the distribution  $g()$ . This leads to the following simplified relationship:

$$R = \frac{1}{M(-r)} \quad (8)$$

This relationship can be used to connect  $R$  with  $r$  for any given generation interval distribution assuming a moment generating function exists. In practice, as infection duration is measured in whole days, and in the implementation of the renewal equation method for estimating  $R_t$  (Cori et al., 2021; Thompson et al., 2019) the discrete version of the generation interval distribution is used and referred to as the infectivity profile  $(\omega_1, \omega_2, \dots, \omega_a)$ .

The discrete version of the relationship in (8) is derived in Wallinga and Lipsitch, 2007 and given as:

$$R = \frac{r}{\sum_{i=1}^n \omega_i \frac{e^{-ra_{i-1}} - e^{-ra_i}}{a_i - a_{i-1}}} \quad (9)$$

Where  $a_i$  is the upper boundary of the discretisation and  $a_{i-1}$  is the lower bound such that:

$$\omega_i = \int_{a_{i-1}}^{a_i} g(a) da$$

This allows for arbitrary discretisation intervals. In the renewal equation method, however, the discretisation of  $\omega$  is enforced to be on whole day boundaries such that  $a_0 = 0, a_1 = 1, \dots$ . This simplifies (9) to:

$$R = \frac{r}{\sum_{i=1}^n \omega_i (e^{-r(i-1)} - e^{-ri})}$$

With our estimates of the growth rate from the first section, and their associated quantiles, we can combine this method with a sampling approach to generate an alternative estimate for the quantiles of the reproduction number.

## C.4 Implementation

The growth rate algorithm described is simply implemented using the R library `LocFit` (Loader et al., 2020) which provides built in local likelihood estimators for Poisson distributed count variables, using a logarithmic link function, and can directly estimate incidence and the derivatives of the log-transformed incidence data, which provides us with a direct estimate of the growth rate. This can also be used to calculate a first derivative of growth. By fitting a Quasi-Poisson model rather than a direct Poisson model we can account for over-dispersion of the real life incidence data, and as an aside we can obtain a time varying estimate of the degree of dispersion.

The key parameterisation of the local likelihood estimate depend on a polynomial degree, and the kernel bandwidth. As we are applying to a time-series we express the band-width in terms of a time window size, in the unit of days. The algorithm is implemented in a R package called “`jepidemic`”, which is open source and available on GitHub (Challen, 2022).

## C.5 Growth rate validation

In this section we compare the performance of the growth rate estimator with a selection of parameterisations, involving linear or quadratic local polynomials fitted to either data included within a bandwidth of 14 or 21 days, against synthetic data with known growth rates. This follows the methodology described in Appendix A, with minor configuration changes to adjust for the fact the estimates of growth rate have a threshold value of 0, where the epidemic transitions from growth to decline, rather than the threshold value of 1 for the reproduction number.

In Figure C.1 we see qualitative estimates of the growth rate estimates against simulated for a range of configurations. In general the growth rate estimates appear robust, but as may be expected more accurate if the growth rate is a smoothly varying quantity rather than a discontinuous function. Polynomial degree does not have an obvious effect except at the boundaries of the time series, which the quadratics remain accurate up to the end of the time-series, whereas the linear fit appear less so.

From Figure C.1 there is no obvious delay to the growth rate estimates, and this is borne out quantitatively by the lag analysis in Figure C.2 which shows all methods to have the same degree of delay at 0.5 days. This specific amount of delay is possibly introduced as a result of treating discrete daily count data as if it is continuous in time.

A quantitative analysis of the quality of the growth rate estimates is shown in Figure C.3 and Table C.1. These show the estimators are all unbiased. Those with shorter windows are better calibrated (Panel B), and less frequently predict growth when the epidemic is declining than those with longer windows (Panel E) although the frequency of this is very low in all estimators. Overall the estimators with the longer windows appear to marginally over precise (with a positive quantile deviation in

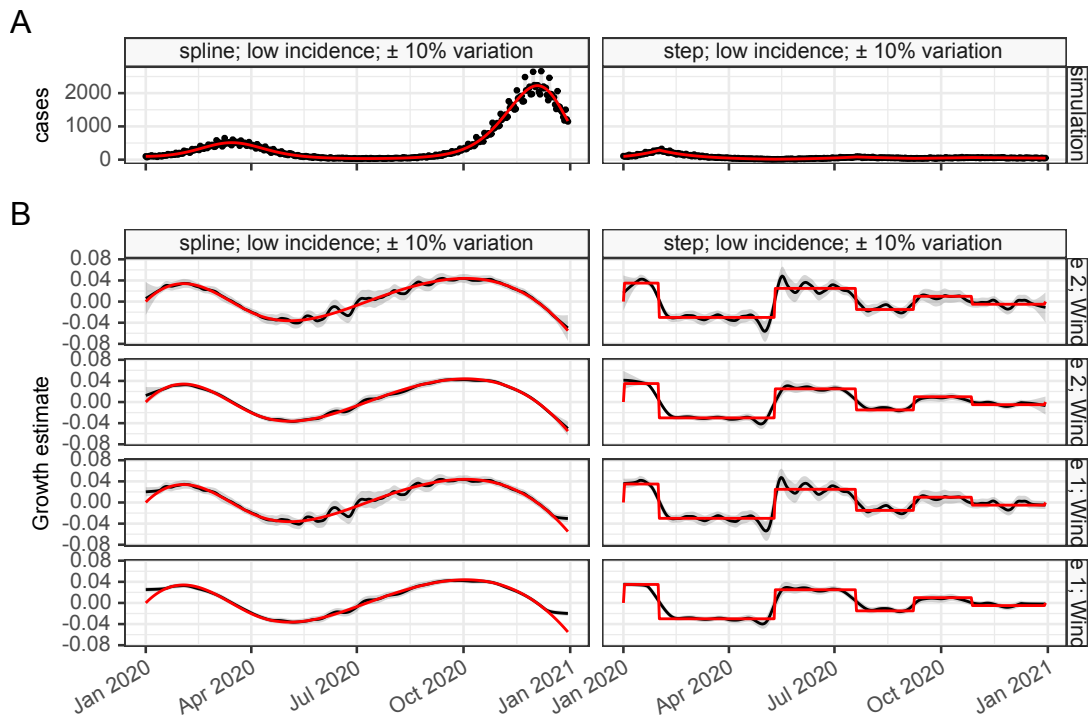


Figure C.1: Qualitative estimates of growth rate (black) against simulated (red) comparing 4 configurations of the locally fitted polynomial method that vary in the parameterisation of the polynomial degree and the bandwidth

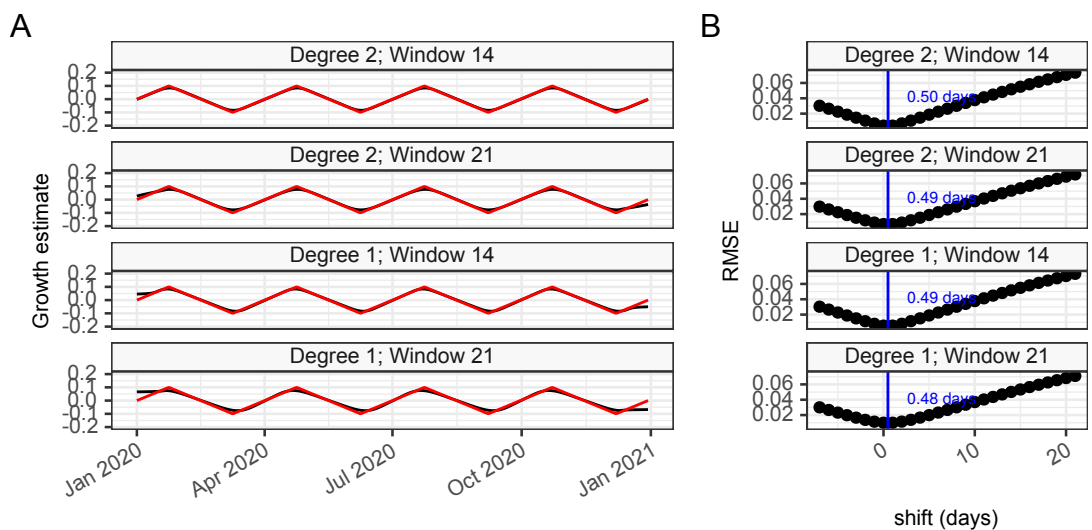


Figure C.2: Time delay analysis of growth rate (black) against simulated (red) comparing 4 configurations of the locally fitted polynomial method that vary in the parameterisation of the polynomial degree and the bandwidth

Table C.1 and I shaped quantile density plots in Panel C). However the CPRS score, as an overall measure combining accuracy and precision, is best (lowest) for the quadratic polynomial fitted over a longer time window.

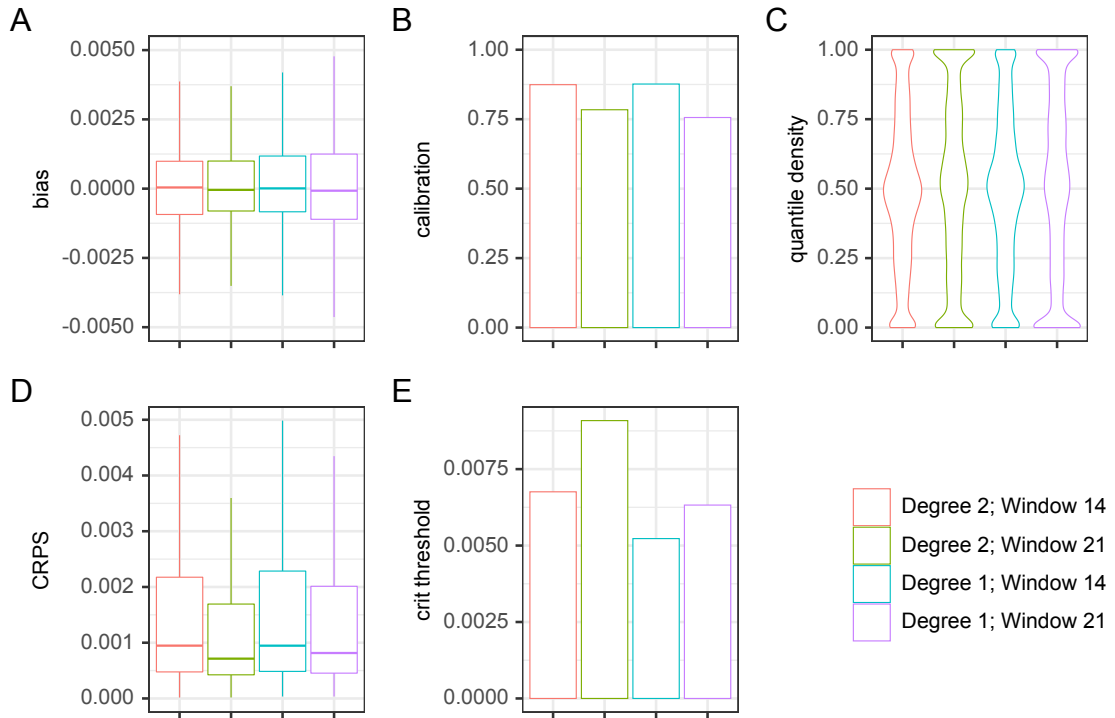


Figure C.3: Estimate quality metric summaries for multiple estimation methods. In this instance we compare the performance of the growth rate estimators that vary in the parameterisation of the polynomial degree and the bandwidth. Panel A show summary statistics for the bias of  $R_t$  estimates, panel B shows the calibration. Panel C shows the quantile density, panel D the CRPS and panel E the critical threshold calibration.

Table C.1: Summary of quantitative analysis of growth rate (black) against simulated (red) comparing 4 configurations of the locally fitted polynomial method that vary in the parameterisation of the polynomial degree and the bandwidth

metric	Degree 2; Window 14	Degree 2; Window 21	Degree 1; Window 14	Degree 1; Window 21
average bias	4.3e-05 [IQR -0.00093 – 0.00099]	-4.3e-05 [IQR -0.00081 – 0.001]	9.5e-06 [IQR -0.00083 – 0.0012]	-7.4e-05 [IQR -0.0011 – 0.0012]
average calibration	87.4% [95% CI 87.1% – 87.7%]	78.4% [95% CI 78.0% – 78.8%]	87.6% [95% CI 87.3% – 88.0%]	75.6% [95% CI 75.2% – 76.0%]
average critical threshold	0.7% [95% CI 0.6% – 0.8%]	0.9% [95% CI 0.8% – 1.0%]	0.5% [95% CI 0.5% – 0.6%]	0.6% [95% CI 0.6% – 0.7%]
CRPS	0.00095 [IQR 0.00048 – 0.0022]	0.00071 [IQR 0.00042 – 0.0017]	0.00095 [IQR 0.00049 – 0.0023]	0.00082 [IQR 0.00045 – 0.002]
quantile deviation	-0.0084	0.0331	-0.00753	0.0443
estimate delay	0.50 days	0.49 days	0.49 days	0.48 days

There is a hint in Figure C.2 that the behaviour at the boundaries may be different and therefore a full breakdown of the validation metrics by the scenario is shown in Figure C.4. The boundary effect, which is most obvious for the estimators relying on degree 1 polynomials is observed clearly in panels D, and P.

Overall all the estimators perform well. There is weak evidence that the estimators with polynomial degree 2 are better than the others, and particularly if the window

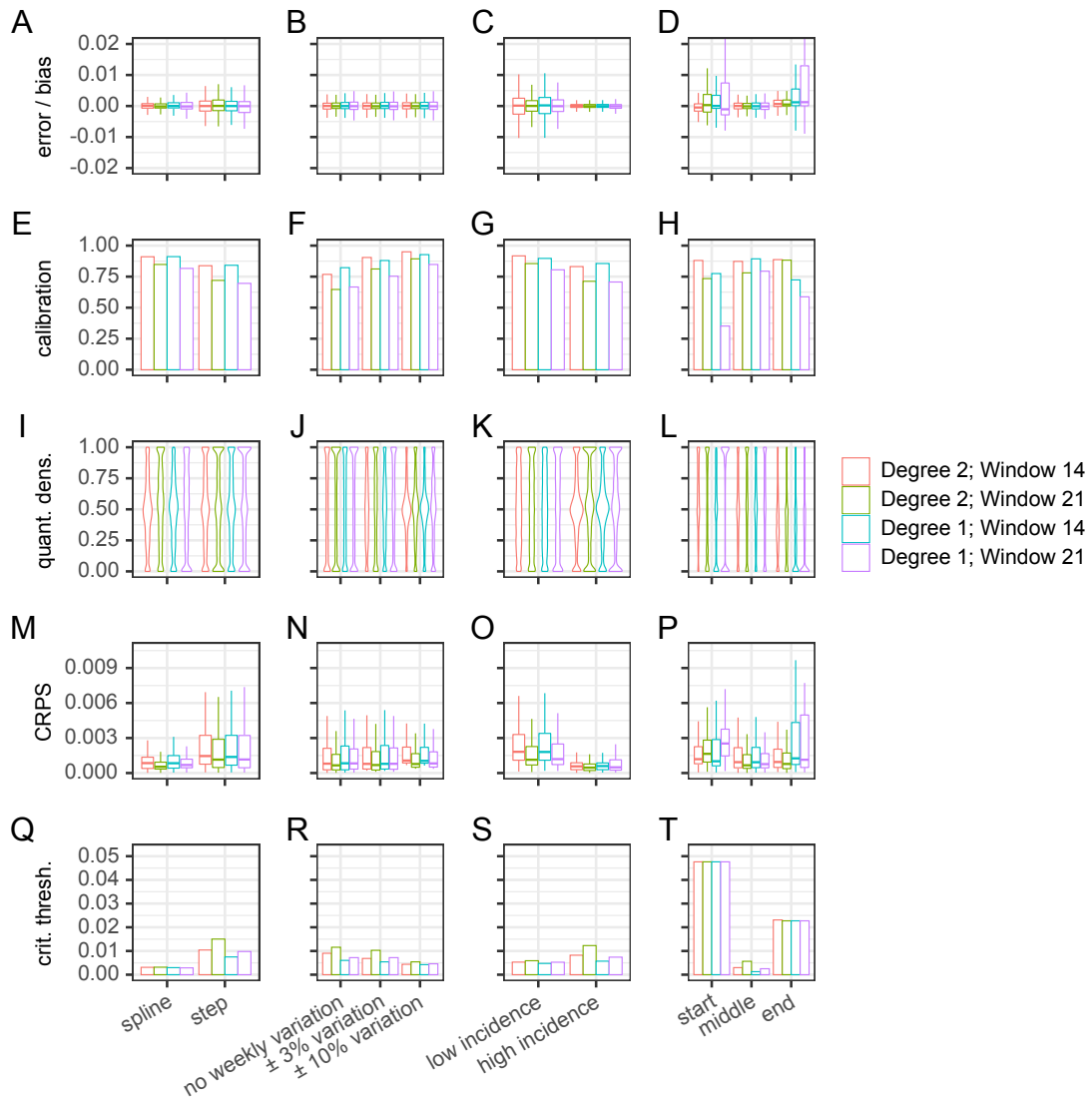


Figure C.4: Estimate quality metric intermediate detail and model comparison. Panel A-D show summary statistics for the bias of growth rate estimates broken down by the simulation smoothness, weekly variability, initial incidence, and time series boundary status. Panels E-F shows the calibration for the same subdivisions. Panels I-L shows the quantile density, panel M-P the CRPS, and panel Q-T the critical thresholds for the same subdivisions.

is longer, however the longer windows do tend to reduce the calibration and result in over-precise estimates. As a balanced default we therefore generally select a polynomial degree of 2 with a 14 day window.

## C.6 $R_t$ validation

The method for conversion of growth rate to  $R_t$  described above can be combined with these growth rate estimates. Uncertainty is propagated by sampling values from the growth rate estimate distribution and as a result is somewhat slow and non-deterministic. It however forms a useful comparison to the  $R_t$  estimation methods presented in Appendix B.

Qualitatively the derived  $R_t$  estimates are shown in Figure C.5. The growth rate based methods seem comparable but with less uncertainty than the renewal equation based method.

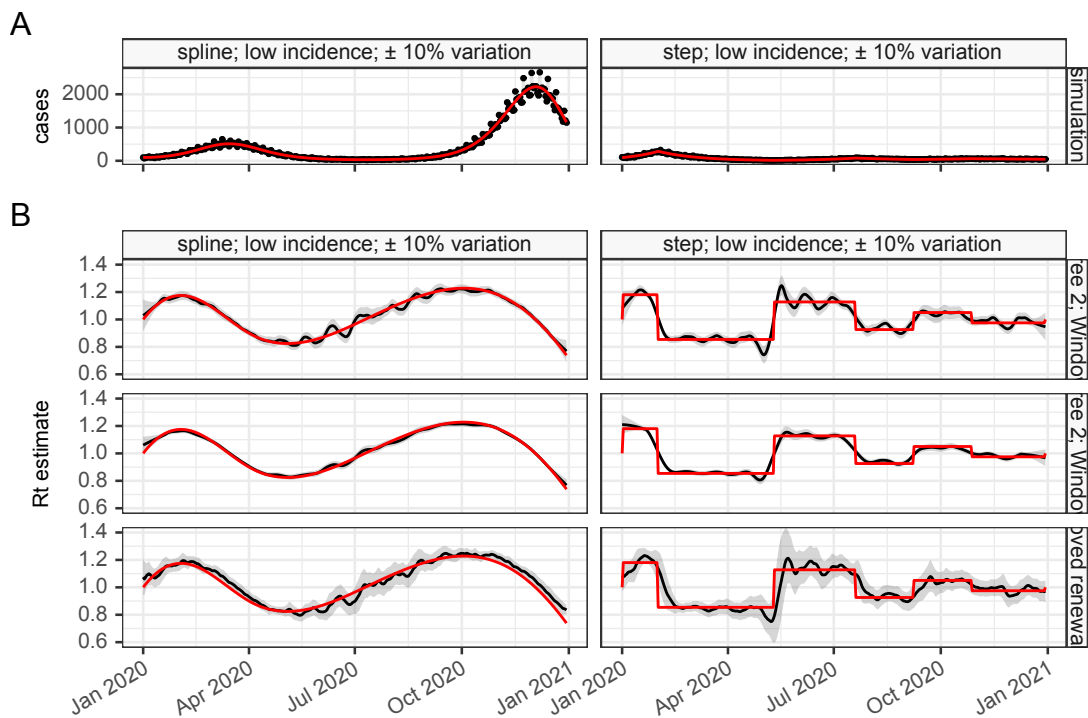


Figure C.5: Qualitative estimates of  $R_t$  (black) against simulated (red) comparing 2 configurations of the growth rate based  $R_t$  estimators using methods presented here and varying by time window (14 or 21 days) with the best performing renewal equation method from Appendix B (labelled “Improved renewal eqn”).

As the Wallinga and Lipsitch, 2007 method is also used to generate the simulated values of  $R_t$  from simulated growth rates, it is unsurprising that the growth rate based methods produce estimates with minimal lag, compared to those based on



the renewal equation method, as Figure C.6 shows, because the true value of the reproduction number is a different type of reproduction number to the instantaneous reproduction number produced by the renewal equation methods presented here.

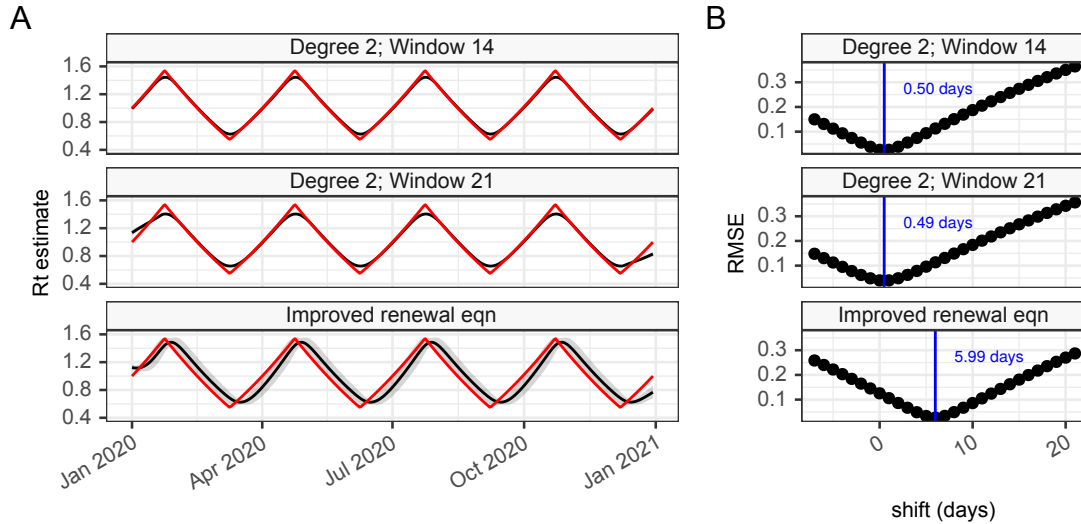


Figure C.6: Time delay analysis of  $R_t$  (black) against simulated (red) comparing 2 configurations of the growth rate based  $R_t$  estimators using methods presented here and varying by time window (14 or 21 days) with the best performing renewal equation method from Appendix B (labelled “Improved renewal eqn”).

As previously the quantitative comparison shown in Figure C.7 and Table C2 the growth rate based methods. This shows that the growth rate based methods are accurate but somewhat over precise. The overall CRPS metric shows the best performing estimate is the growth rate based estimator however this is offset by their generally lower calibration, and over-precision.

Table C.2: Summary of quantitative analysis of  $R_t$  (black) against simulated (red) comparing 3 methods that vary in their prior selection process

metric	Degree 2; Window 14	Degree 2; Window 21	Improved renewal eqn
average bias	0.00017 [IQR -0.0065 – 0.0066]	0.00016 [IQR -0.0065 – 0.0071]	0.0023 [IQR -0.007 – 0.014]
average calibration	75.2% [95% CI 74.8% – 75.6%]	60.1% [95% CI 59.7% – 60.6%]	89.4% [95% CI 89.1% – 89.7%]
average critical threshold	0.9% [95% CI 0.8% – 1.0%]	1.1% [95% CI 1.0% – 1.2%]	0.3% [95% CI 0.2% – 0.3%]
CRPS	0.0049 [IQR 0.0032 – 0.011]	0.0042 [IQR 0.0024 – 0.0089]	0.0082 [IQR 0.0051 – 0.014]
quantile deviation	0.0412	0.0899	0.00181
estimate delay	0.50 days	0.49 days	5.99 days

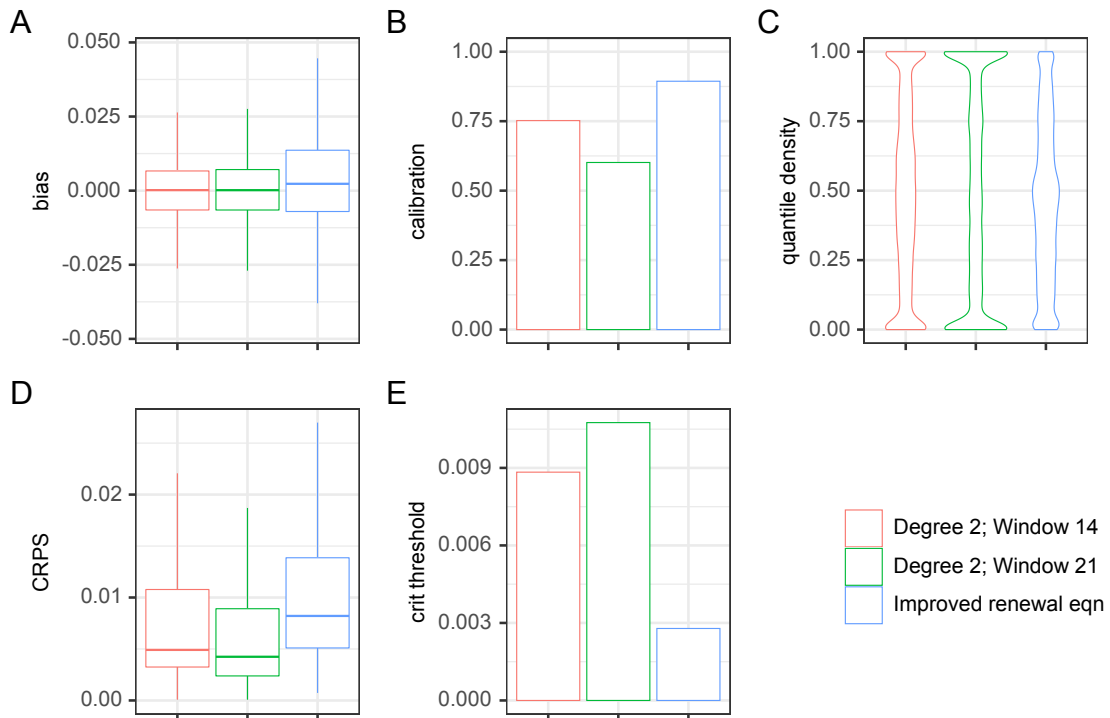


Figure C.7: Estimate quality metric summaries for multiple estimation methods. In this instance we compare the performance of the growth rate estimators that vary in the parameterisation of the polynomial degree and the bandwidth. Panel A show summary statistics for the bias of  $R_t$  estimates, panel B shows the calibration. Panel C shows the quantile density, panel D the CRPS and panel E the critical threshold calibration.

## C.7 Summary

In this appendix we present our chosen method for estimating the real-time growth rate of an exponentially growing epidemic, by local likelihood estimation of a Poisson model with a time varying polynomial as the rate parameter. This is shown to be robust to a range of different synthetic scenarios with known growth rate. Configured to use a quadratic and a local window of 14 days produces a estimate that balances general performance with calibration. Translating this growth rate estimate into a reproduction number produces an accurate estimate of the reproduction number, although the same methodology is used for calculating the true value of the reproduction number as the estimate. The resulting estimate of the reproduction number is also over precise and comparatively slow to calculate so is not felt to be a significant improvement over the best performing renewal equation method described in Appendix B. One advantage to note though is that due to the fact that the growth

rate estimate can be made right up to the end of the time series without any delay being introduced, the derived estimates of the reproduction number are potentially more up to date than those derived from the renewal equation method.

# D. Bayesian growth rate and reproduction number estimation

## D.1 Introduction

In previous appendices we demonstrated approaches to estimating the growth rate using a maximum likelihood estimator, and the effective reproduction number ( $R_t$ ) using a Bayesian framework with specific assumptions about the exponential growth resulting from the reproduction number. In this section we derive another simpler method for deriving both quantities within a single Bayesian framework, using minimal assumptions about the underlying processes, as a counterpoint for comparison. This method is implemented in a Java library with bindings to the R language called “jepidemic”, which is open source and available on GitHub (Challen, 2022).

## D.2 Estimation of the incidence of infection

As in previous appendices we assume  $I_0, I_1, \dots, I_t$  is a time series of infection counts, assumed to be drawn from some discrete probability distribution with expected value  $\bar{I}_t$ . We assume  $I_t$  is a Poisson distributed quantity, with a rate parameter which is a function of time:

$$E[I_t] = \lambda_t$$

If we assume  $\lambda_t$  is constant over a short time period  $[t - \tau; t + \tau]$  then by the definition of the Poisson distribution:

$$P(I_{t-\tau}, \dots, I_{t+\tau} | \lambda_t) = \prod_{s=t-\tau}^{t+\tau} \frac{e^{-\lambda_t} \lambda_t^{I_s}}{I_s!}$$

Using our knowledge of the conjugate prior of the Poisson distribution we assume the rate parameter,  $\lambda_t$ , to be gamma distributed with shape parameter  $\alpha$  and rate parameter  $\beta$ , and if  $n = 2\tau + 1$  then we can use a Bayesian framework to derive posterior estimates of the distribution of the Poisson rate, labelled  $\alpha'$  and  $\beta'$  conditioned on the data we observe over  $[t - \tau; t + \tau]$ :

$$\begin{aligned}
P(I_{t-\tau}, \dots, I_{t+\tau} | \lambda_t) &= \frac{e^{-n\lambda_t} \lambda_t^{\sum I_s}}{\prod_{s=t-\tau}^{t+\tau} I_s!} \\
P(\lambda_t) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_t^{\alpha-1} e^{-\beta\lambda_t} \\
P(\lambda_t | I_{t-\tau}, \dots, I_{t+\tau}) &= \frac{P(I_{t-\tau}, \dots, I_{t+\tau} | \lambda_t) P(\lambda_t)}{P(I_{t-\tau}, \dots, I_{t+\tau})} \\
P(\lambda_t | I_{t-\tau}, \dots, I_{t+\tau}) &= \frac{e^{-n\lambda_t} \lambda_t^{\sum I_s}}{\prod_{s=t-\tau}^{t+\tau} I_s!} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_t^{\alpha-1} e^{-\beta\lambda_t} \\
P(\lambda_t | I_{t-\tau}, \dots, I_{t+\tau}) &\propto \lambda_t^{\sum I_s + \alpha - 1} e^{-(2\tau + 1 + \beta)\lambda_t} \\
P(\lambda_t | I_{t-\tau}, \dots, I_{t+\tau}) &\sim \text{Gamma}\left(\sum_{s=t-\tau}^{t+\tau} I_s + \alpha, 2\tau + 1 + \beta\right) \\
\alpha' &= \alpha + \sum_{s=t-\tau}^{t+\tau} I_s \\
\beta' &= 2\tau + 1 + \beta
\end{aligned} \tag{10}$$

The posterior estimate of the Poisson rate  $\lambda$  is gamma distributed by definition. An

estimate of the likely value of  $I_t$  ( $\bar{I}_t$ ) given the observed data is given by the posterior predictive distribution:

$$\begin{aligned}\bar{I}_t &\sim \text{NegBin}\left(\alpha', \frac{1}{\beta' + 1}\right) \\ E(\bar{I}_t | I_{t-\tau}, \dots, I_{t+\tau}) &= \frac{\alpha'}{\beta'} \\ V(\bar{I}_t | I_{t-\tau}, \dots, I_{t+\tau}) &= \alpha' \left( \frac{\beta' + 1}{\beta'^2} \right)\end{aligned}$$

### D.3 Estimation of the growth rate

The exponential growth rate  $r_t$  is the gradient of the logarithm of  $I$  ( $\frac{d}{dt} \log(\bar{I}_t)$ ) with respect to time. We approximate this as the difference between two estimates of the true value of  $I$  separated by a small time period  $2m$ :

$$\begin{aligned}r_t &\approx \frac{1}{2m} (\log(E(I_{t+m})) - \log(E(I_{t-m}))) \\ r_t &= \frac{1}{2m} \log \frac{\lambda_{t+m}}{\lambda_{t-m}}\end{aligned}$$

Given that  $I$  is assumed to be Poisson distributed the expected value is  $\lambda_t$ . If we further define  $y = g(r)$  such that:

$$\begin{aligned}y &= g(r_t) = e^{2\tau r_t} \\ r_t &= g^{-1}(y) = \frac{1}{2\tau} \log(y)\end{aligned}$$

Then we can express the distribution of  $g(r)$  in terms of a ratio of Gamma distributed quantities. For simplification we have also assumed that  $m = \tau$ . This means that in estimating the growth rate we use two posterior estimates of the Poisson rate based on  $\tau + 1$  data points. The growth rate estimate is therefore based on  $2\tau + 1$

data points, as the central point overlaps. This is a choice made to simplify the mathematics but we could have used any 2 estimates which are close together. This approach also has the benefit of not re-using information.

$$\begin{aligned}
r_t &= \frac{1}{2\tau} \log \frac{\lambda_{t+\tau}}{\lambda_{t-\tau}} \\
g(r_t) &\sim \left( \frac{\text{Gamma}(\alpha + \sum_{s=t}^{t+2\tau} I_s, \beta')}{\text{Gamma}(\alpha + \sum_{r=t-2\tau}^t I_r, \beta')} \right) \\
\alpha'_{\tau+} &= \alpha + \sum_t^{t+2\tau} I_s \\
\alpha'_{\tau-} &= \alpha + \sum_{t-2\tau}^t I_s \\
g(r_t) &\sim \left( \frac{\text{Gamma}(\alpha'_{\tau+}, \beta')}{\text{Gamma}(\alpha'_{\tau-}, \beta')} \right)
\end{aligned}$$

The ratio of 2 gammas with same rate parameter is a Beta Prime distributed quantity (Pham-Gia and Turkkan, 2011; Leemis and McQueston, 2008), which describes  $g(r)$ :

$$\begin{aligned}
r_t \sim Y &= \frac{1}{2\tau} \log \left( \text{BetaPrime}(\alpha'_{\tau+}, \alpha'_{\tau-}) \right) \\
g(r_t) \sim X &= \text{BetaPrime} \left( \alpha + \sum_t^{t+2\tau} I, \alpha + \sum_{t-2\tau}^t I \right) \\
\text{BetaPrime}(\alpha_1, \alpha_2) &: \tag{11} \\
f(x) &= \frac{1}{B(\alpha_1, \alpha_2)} x^{\alpha_1-1} (1+x)^{-\alpha_1-\alpha_2} \\
F(x) &= I_{\frac{x}{1+x}}(\alpha_1, \alpha_2)
\end{aligned}$$

Where  $I$  is the regularised beta function and  $B$  is the complete beta function. The support for  $g(r_t)$  is  $[0 \dots \infty]$  and hence the support for  $r_t$  is  $[-\infty \dots \infty]$ . Given that  $g(r_t)$  can be differentiated and is a strictly increasing function we can apply the following transformation (Pishro-Nik, 2014; Taboga, 2021; Taboga, 2017):

$$\begin{aligned}
Y &= \frac{1}{2\tau} \log(X) \\
g(x) &= \frac{1}{2\tau} \log(x) \\
g^{-1}(y) &= e^{2\tau y} \\
\frac{dg^{-1}}{dy} &= 2\tau e^{2\tau y} \\
F_Y(y) &= F_X(g^{-1}(y)) \\
f_Y(y) &= f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}
\end{aligned}$$

Which gives us a probability density function for an estimate of  $r_t$  based on the Bayesian posterior of  $\lambda_t$ :

$$\begin{aligned}
f_Y(r_t) &= f_X(g^{-1}(r_t)) \frac{dg^{-1}(r_t)}{dr_t} \\
f_Y(r_t) &= 2\tau e^{2\tau r_t} f_X(e^{2\tau r_t}) \\
f_Y(r_t) &= 2\tau e^{2\tau r_t} \frac{\left( e^{r_t(\alpha'_{\tau+}-1)} (1 + e^{r_t})^{(-\alpha'_{\tau+}-\alpha'_{\tau-})} \right)}{B(\alpha'_{\tau+}, \alpha'_{\tau-})} \\
f_Y(r_t) &= \frac{2\tau}{B(\alpha'_{\tau+}, \alpha'_{\tau-})} \left( e^{r_t(\alpha'_{\tau+}+2\tau-1)} (1 + e^{r_t})^{(-\alpha'_{\tau+}-\alpha'_{\tau-})} \right)
\end{aligned}$$

and which has the following cumulative probability function:

$$\begin{aligned}
F_Y(r_t) &= F_Y(g^{-1}(r_t)) \\
F_X(r_t) &= F_Y(e^{2\tau r_t}) \\
F_X(r_t) &= I_{\frac{e^{2\tau r_t}}{1+e^{2\tau r_t}}}(\alpha_{\tau+}, \alpha_{\tau-})
\end{aligned}$$

where  $I_x(a, b)$  is the regularised Beta function (Weisstein, 2010).

The quantile function for  $g(r_t)$  is similarly transformed backwards to give a quantile estimate for  $r_t$ , and we can sample from  $Y$  and transform using  $g^{-1}(y)$  to get unbiased



samples of  $r_t$ .

## D.4 Reproduction number estimates from Poisson rate

If we continue to assume  $I_t \sim \text{Poisson}(\lambda_t)$  and an estimate of  $\lambda_t$  is available. As before  $\omega_1, \omega_2, \dots, \omega_s$  is another probability distribution, the infectivity profile, that defines the likelihood that a case infected at time  $t$  resulted from a case infected between the times  $t - s$  and  $t - s + 1$ . This definition implies that  $\omega_{s \leq 0} = 0$  as that applies to infections in the future, and that the discrete time measure  $s$  here represents the upper bound of the equivalent continuous unit time interval, rather than, for example, the middle of the interval.

As before we define the backward-looking effective reproduction number  $R_t$  as the inverse ratio of the number of primary infections associated with secondary infections observed at time  $t$ , this is known as the instantaneous reproduction number.

$$R_t = \frac{\bar{I}_t}{\sum_{s=1}^t \bar{I}_{t-s} \omega_s}$$

The posterior distribution of  $\lambda_t$  is an estimate of the  $\bar{I}_t$  distribution and therefore:

$$R_t = \frac{\lambda_t}{\sum_{s=1}^t \lambda_{t-s} \omega_s}$$

Assuming the posterior distribution of  $\lambda_t \sim \text{Gamma}(\alpha', \beta')$  as described above, we consider the denominator as the sum of scaled gamma distributions, and we can say

the distribution of the denominator is of the form:

$$\lambda_{t-s}\omega_s \sim \text{Gamma}\left(\alpha'_{t-s}, \frac{\beta'_{t-s}}{\omega_s}\right)$$

A commonly used approximation for a sum of gamma distributed variables is as another gamma distribution with the same first and second moments as described by Covo and Elalouf, 2014. In the case of a set of gamma distributions, with parameters  $\alpha_i$  and  $\beta_i$ , an approximation of the sum is another gamma distribution with parameters  $\alpha_{sum}$  and  $\beta_{sum}$ :

$$\begin{aligned} E(X_i) &= \frac{\alpha_i}{\beta_i} \\ V(X_i) &= \frac{\alpha_i}{\beta_i^2} \\ E(X_{sum}) &= \sum_i E(X_i) = \sum_i \frac{\alpha_i}{\beta_i} \\ V(X_{sum}) &= \sum_i V(X_i) = \sum_i \frac{\alpha_i}{\beta_i^2} \\ \alpha_{sum} &= \frac{E(X_{sum})^2}{V(X_{sum})} = \frac{\left(\sum_i \frac{\alpha_i}{\beta_i}\right)^2}{\sum_i \frac{\alpha_i}{\beta_i^2}} \\ \beta_{sum} &= \frac{E(X_{sum})}{V(X_{sum})} = \frac{\sum_i \frac{\alpha_i}{\beta_i}}{\sum_i \frac{\alpha_i}{\beta_i^2}} \end{aligned}$$

The moment matching approximation for the sum of gamma distributions can be empirically tested. In Figure D.1 we demonstrate its use on a random set of five gamma distributions show in panel A and the sum of a million random draws from each of these probability distributions in panel B. In panel B the red line is the predicted Gamma distribution based on the estimator. Further validation of the estimator is warranted but in simple cases the estimator is observed to perform well.

Using this approximation we estimate  $R_t$  to be distributed as the ratio of 2 Gamma

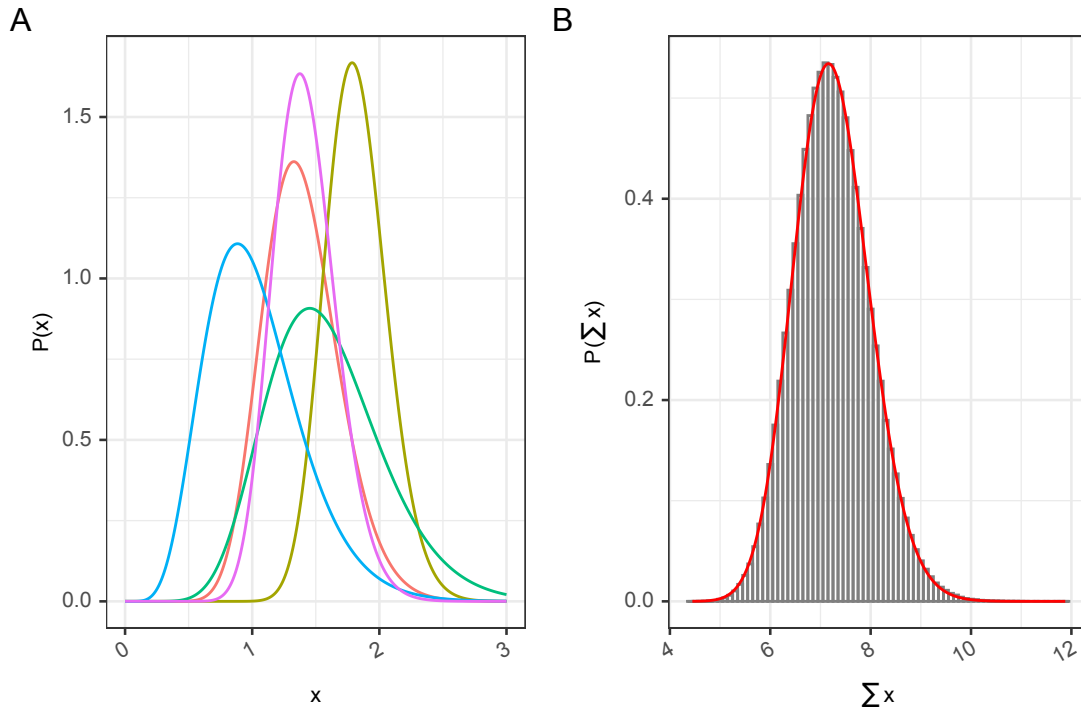


Figure D.1: A test of an approximation of a sum of Gamma distributed quantities (panel A) as a single Gamma distribution based on matching moments (panel B). In this particular example the sum of Gamma distributions is well approximated by another single Gamma distribution.

distributions: the numerator with shape and rate parameters  $\alpha'$  and  $\beta'$ , and the denominator with shape and rate parameters  $\alpha''$  and  $\beta''$  as described:

$$R_t \sim \frac{\beta'_t \text{Gamma}(\alpha'_t, 1)}{\beta''_t \text{Gamma}(\alpha''_t, 1)}$$

$$R_t \sim \text{BetaPrime}(\alpha'_t, \alpha''_t, 1, \frac{\beta''_t}{\beta'_t})$$

Where given:  $s \in (1..t)$

$$\alpha''_t = \frac{\left( \sum_s \frac{\alpha'_{t-s} \omega_s}{\beta'_{t-s}} \right)^2}{\sum_s \frac{\alpha'_{t-s} \omega_s^2}{(\beta'_{t-s})^2}} \tag{12}$$

$$\beta''_t = \frac{\sum_s \frac{\alpha'_{t-s} \omega_s}{\beta'_{t-s}}}{\sum_s \frac{\alpha'_{t-s} \omega_s^2}{(\beta'_{t-s})^2}}$$

This distributional form of  $R_t$  as a generalised Beta Prime distribution (Research, 2010; Kotz, 1995) with three shape parameters  $(\alpha_1, \alpha_2, \alpha_3)$  and one scale parameter

( $\beta$ ) and with the probability density and cumulative probability functions as given below:

Where  $BetaPrime(\alpha_1, \alpha_2, \alpha_3, \beta)$  :

$$f(x) = \frac{\alpha_3 \left(1 + \left(\frac{x}{\beta}\right)^{\alpha_3}\right)^{-\alpha_1 - \alpha_2} \left(\frac{x}{\beta}\right)^{\alpha_1 \alpha_3 - 1}}{\beta Beta(\alpha_1, \alpha_2)}$$

$$F(x) = I_{\frac{x^{\alpha_3}}{x^{\alpha_3} + \beta^{\alpha_3}}}(\alpha_1, \alpha_2)$$

and where  $I_x$  is the regularised Beta function. This form for the  $R_t$  estimate assumes only the Gamma posterior estimate for the Poisson rate, and so can be readily calculated directly from the analytic form of the posteriors identified in the first part of this method. This could be a biased if the denominator of the ratio is not a good estimate for the weighted sum in reality, however the moment matching estimate works well when the distributions are not completely different. In our case the fact that the  $\lambda_t$  estimates are part of the same time series and scaled by the infectivity profile, does make it quite likely the distributions will have similar characteristics.

This constitutes a method for estimating the expected incidence, the growth rate, and the instantaneous reproduction number from a time series of observed case incidence making only a minimal set of assumptions.

## D.5 Implementation considerations

In (10) we derived the standard posterior Gamma distribution of a Poisson rate parameter based on observed cases and a prior belief about that rate ( $\lambda_t$ ). The prior belief about the rate at time  $t$  can be informed by the value of the rate at earlier times, and optionally by the growth rate such that  $I_{t+1} = I_t e^{r t}$ . The full distributional forms of the posteriors at previous time steps could be used to generate

a prior, but this was found to result in a prior that is far too strong and reflects the assumption that the Poisson rate is constant over a small time period. Instead of this a better option is a prior that has a very low coefficient of variation, and we found a value of 1 to be appropriate, resulting in the distributional form for the prior,  $\lambda_{t+1,prior}$  in terms of the posterior of incidence ( $\lambda_{t,posterior}$ ) and growth rate ( $r_{t,posterior}$ ).

$$\lambda_{t+1,prior} \sim \text{Gamma}\left(1, \frac{1}{\lambda_{t,posterior} e^{\bar{r}_{t,posterior}}}\right)$$

This prior will influence the posterior estimates of incidence most strongly when incidence is low, and this may indeed have a desirable stabilising effect on the estimates in this situation. Increasing the coefficient of variation from 1 to a higher number may provide additional stability and is a topic for future exploration.

As stated the method does not account for uncertainty in the infectivity profile  $\omega_s$ . This can be included by sampling the infectivity profile, and estimating an  $R_t$  distribution for each infectivity profile. The estimator for  $R_t$  can be made using any posterior estimate of  $\lambda_t$ , for which we have a range of possibilities based on the desired data window ( $\tau$ ) we wish to use. Much as in Appendix B there are potentially  $R_t$  estimates for each of the data window sizes selected for  $\lambda_{t,\tau}$  and each of the infectivity profiles. These multiple estimates are combined as a weighted mixture distribution of the multiple  $R_t$  estimates, to generate an overall summary estimate while retaining uncertainty. By default a uniform weighting is applied but non uniform weighting based on the window size is also a topic for future exploration.

As described in appendix A the case counts may have weekly periodicity, or have anomalous values for other reasons. Within this framework it is possible to adjust the weighting of each data point on a case-by-case basis, by appropriately adjusting the posteriors in (10). This opens the door for increasing the uncertainty when

estimates span weekends, for example, and warrants more investigation.

## D.6 Growth rate validation

The growth rate estimates using this method can be compared to those from the maximum likelihood local polynomial approach described in Appendix C, with a polynomial degree of 2 and window of 14 days. This uses the same synthetic data and methodology as described in Appendix A.

The result of the comparison is shown qualitatively in Figure D.2, and shows the Bayesian estimator seems to perform well with comparable levels of uncertainty and close matching to the true value

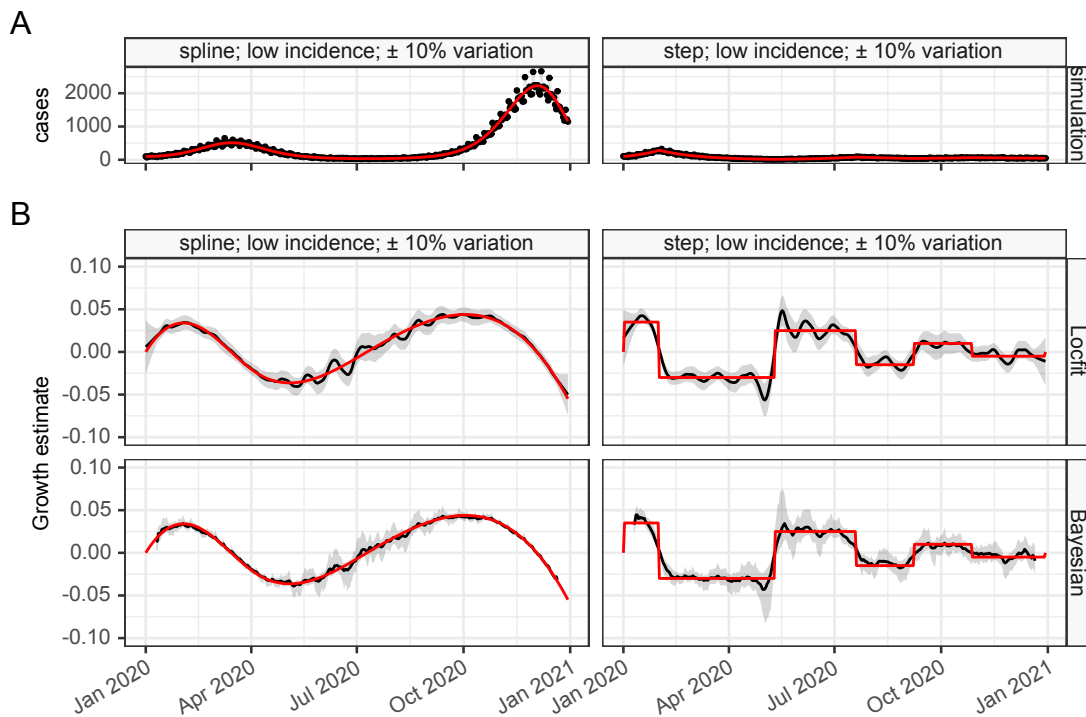


Figure D.2: Qualitative estimates of growth rate (black) against simulated (red) comparing the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C

As the growth rate estimates use two estimated of the Poisson rate centered around the time point in question we anticipate there to be no delay, similar to the locfit

method. This is confirmed in Figure D.3.

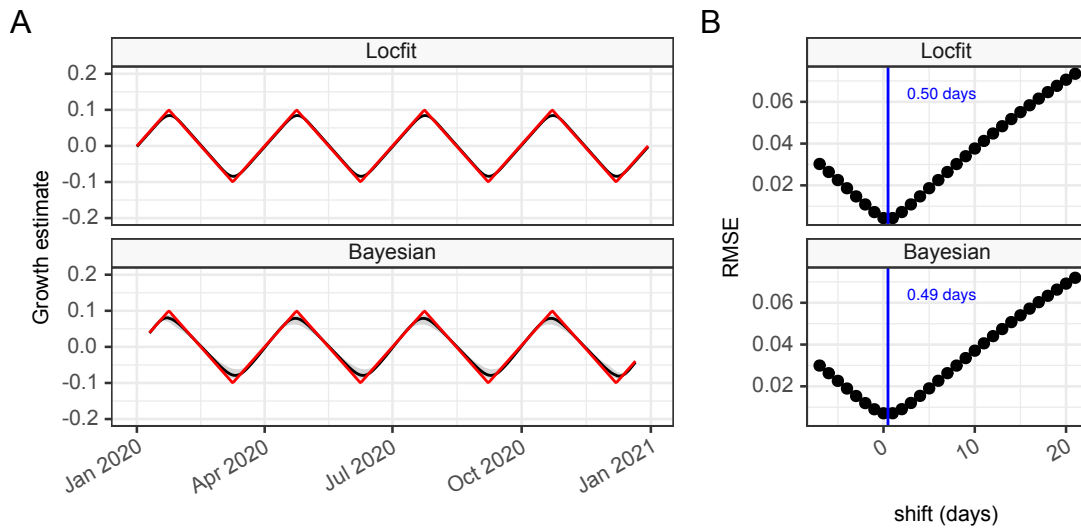


Figure D.3: Time delay analysis of growth rate (black) against simulated (red) comparing the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C

Quantitative comparison of the methods shown in Figure D.4 and Table D.1 show the Bayesian method performs slightly better overall than the LocFit method as judged by the continuous rank probability score, but with reassuringly little difference between the 2 methods.

Table D.1: Summary of quantitative analysis of  $R_t$  (black) against simulated (red) comparing the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C

metric	Locfit	Bayesian
average bias	4.3e-05 [IQR -0.00093 – 0.00099]	-0.00011 [IQR -0.001 – 0.0011]
average calibration	87.4% [95% CI 87.1% – 87.7%]	85.6% [95% CI 85.3% – 85.9%]
average critical threshold	0.7% [95% CI 0.6% – 0.8%]	0.4% [95% CI 0.4% – 0.5%]
CRPS	0.00095 [IQR 0.00048 – 0.0022]	0.00089 [IQR 0.00036 – 0.0018]
quantile deviation	-0.0084	0.0302
estimate delay	0.50 days	0.49 days

More detailed analysis did not demonstrate any obvious difference between the 2 methods although the Bayesian estimator is unable to produce a growth rate estimate right up to the end of the input time series. This is due to the symmetric nature of the estimation window, which is not a requirement for the LocFit estimator.

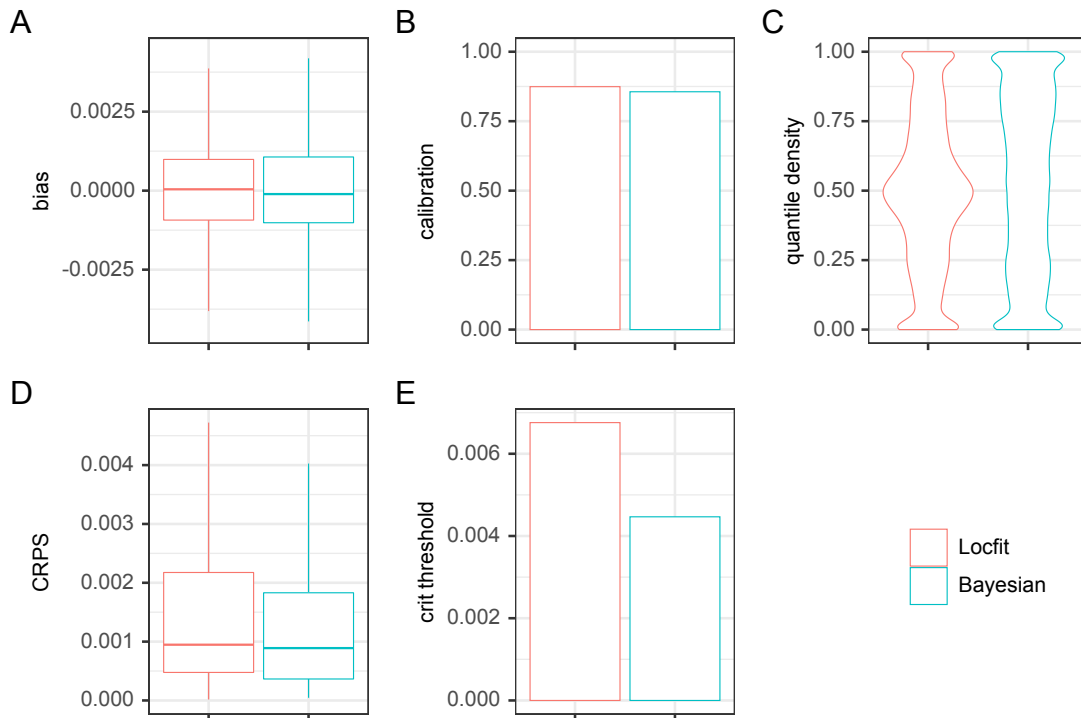


Figure D.4: Estimate quality metric summaries for multiple estimation methods. In this instance we compare the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C. Panel A show summary statistics for the bias of  $R_t$  estimates, panel B shows the calibration. Panel C shows the quantile density, panel D the CRPS and panel E the critical threshold calibration.

## D.7 $R_t$ validation

The derivation of the reproduction number within the Bayesian estimation framework can be compared to estimates from both the locally fitted polynomial LocFit approach from Appendix C and the renewal equation approach from Appendix B.

Qualitatively as seen in Figure D.5 the 3 methods are similar, in that they all faithfully reproduce the true values. As noted in appendix C the LocFit method is over-precise compared to the others, and there is little to choose between the Bayesian and renewal equation methods.

The timing of the estimates shown in Figure D.6 are also similar between the Bayesian and renewal equation methods although we note again that the Bayesian method cannot produce estimates up to the end of the time-series, however this is



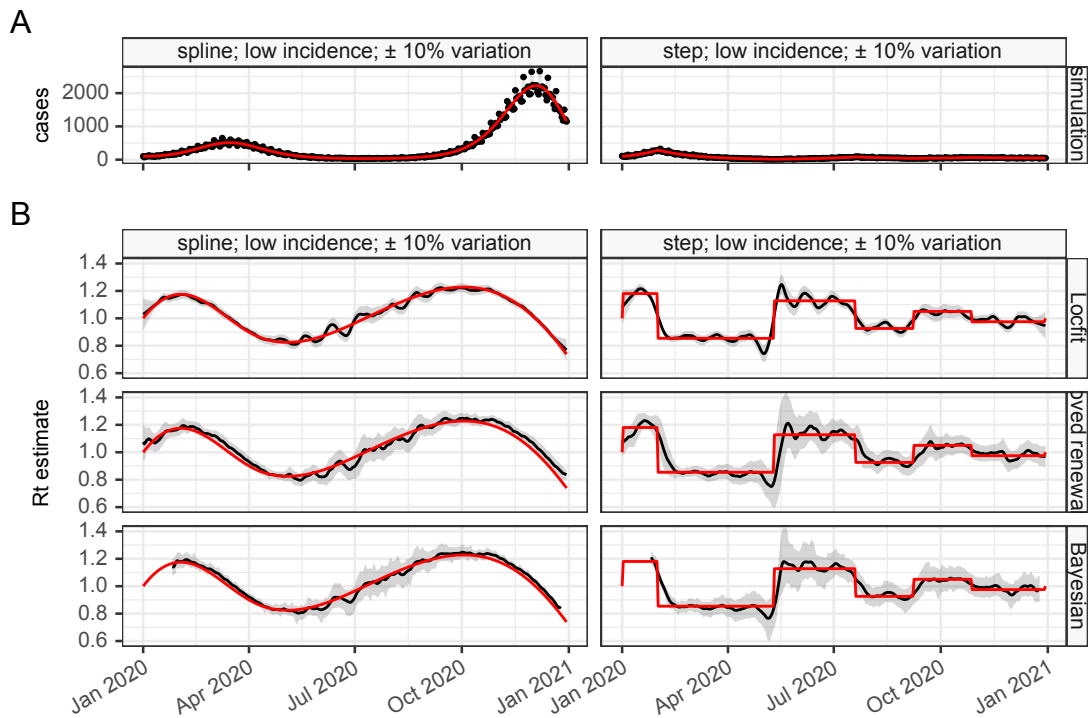


Figure D.5: Qualitative estimates of  $R_t$  (black) against simulated (red) comparing the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C and the renewal equation method from Appendix B

partly offset by those estimates having less lag.

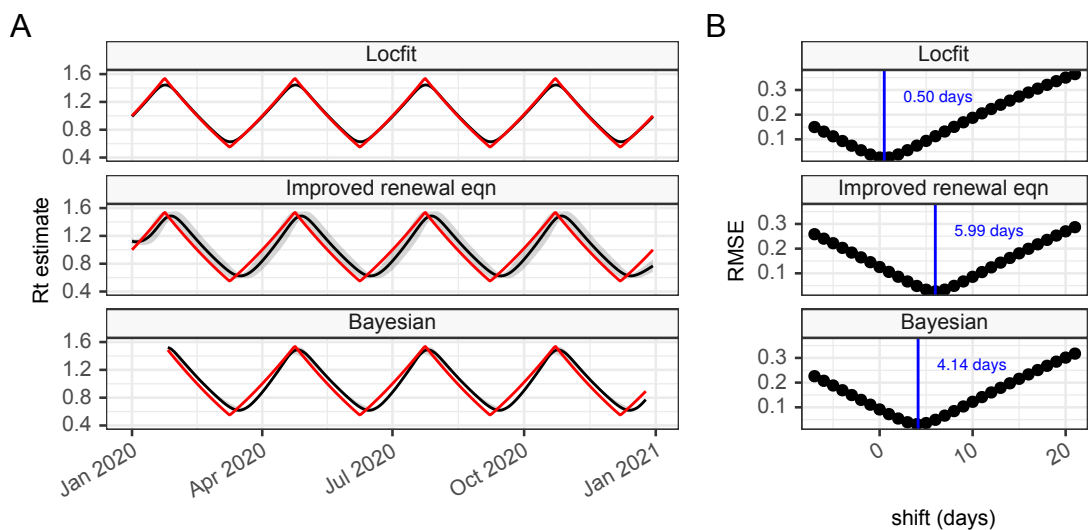


Figure D.6: Time delay analysis of  $R_t$  (black) against simulated (red) comparing the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C and the renewal equation method from Appendix B

Quantitative analysis in Figure D.7 and Table D.2 shows little to separate the renewal equation and this Bayesian method, with the renewal equation method being slightly

better calibrated.

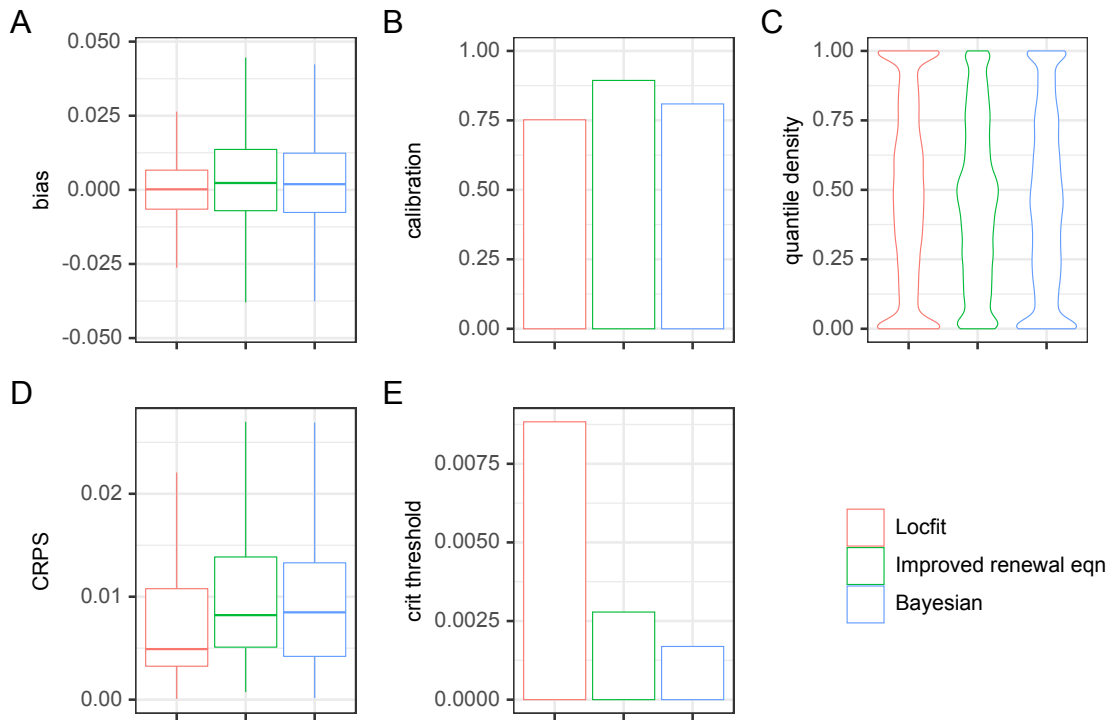


Figure D.7: Estimate quality metric summaries for multiple estimation methods. In this instance we compare the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C and the renewal equation method from Appendix B. Panel A show summary statistics for the bias of  $R_t$  estimates, panel B shows the calibration. Panel C shows the quantile density, panel D the CRPS and panel E the critical threshold calibration.

Table D.2: Summary of quantitative analysis of  $R_t$  (black) against simulated (red) comparing the Bayesian estimator described here with the maximum likelihood Poisson model implementation from Appendix C and the renewal equation method from Appendix B

metric	Locfit	Improved renewal eqn	Bayesian
average bias	0.00017 [IQR -0.0065 – 0.0066]	0.0023 [IQR -0.007 – 0.014]	0.0019 [IQR -0.0076 – 0.012]
average calibration	75.2% [95% CI 74.8% – 75.6%]	89.4% [95% CI 89.1% – 89.7%]	80.9% [95% CI 80.5% – 81.3%]
average critical threshold	0.9% [95% CI 0.8% – 1.0%]	0.3% [95% CI 0.2% – 0.3%]	0.2% [95% CI 0.1% – 0.2%]
CRPS	0.0049 [IQR 0.0032 – 0.011]	0.0082 [IQR 0.0051 – 0.014]	0.0085 [IQR 0.0042 – 0.013]
quantile deviation	0.0412	0.00181	0.0278
estimate delay	0.50 days	5.99 days	4.14 days

Detailed breakdowns did not provide any further insight into the differences between the methods.

## D.8 Summary

We introduced an alternative Bayesian framework for estimating case incidence, growth rate and reproduction number at the same time. This is robust to a range of different synthetic scenarios and performs in a very similar manner to the locally fitted polynomial for growth rate and to the best performing renewal equation method. The estimates are slightly reduced in value by the fact they cannot estimate right up to the end of the time series, unlike the other methods, but provide welcome validation by which to benchmark the other methods.

# Bibliography

- J. L. Anderson (1st July 1996). ‘A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations’. In: *Journal of Climate* 9.7, pp. 1518–1530. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2 (Cited on page 207).
- R. M. Anderson and R. M. May (27th Aug. 1992). *Infectious Diseases of Humans: Dynamics and Control*. Oxford, New York: Oxford University Press. 766 pp. ISBN: 978-0-19-854040-3 (Cited on page 247).
- Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen and M. Wang (14th Apr. 2020). ‘Presumed Asymptomatic Carrier Transmission of COVID-19’. In: *JAMA* 323.14, pp. 1406–1407. ISSN: 0098-7484. DOI: 10.1001/jama.2020.2565 (Cited on page 9).
- BBC News (3rd Apr. 2020a). ‘Coronavirus: Nightingale Hospital Opens at London’s ExCel Centre’. In: *BBC News. UK*. URL: <https://www.bbc.com/news/uk-52150598> (visited on 17/12/2020) (Cited on page 12).
- BBC News (21st Mar. 2020b). ‘Coronavirus: Thousands of Extra Hospital Beds and Staff’. In: *BBC News. UK*. URL: <https://www.bbc.com/news/uk-51989183> (visited on 16/12/2020) (Cited on page 12).
- BBC News (27th Feb. 2021a). ‘Covid-19: India in a ‘delicate Phase’ of Its Coronavirus Battle as Cases Surge’. In: *BBC News. India*. URL: <https://www.bbc.com/news/world-asia-india-56206004> (visited on 11/10/2021) (Cited on page 27).
- BBC News (14th June 2021b). ‘Covid: Lockdown Easing in England to Be Delayed by Four Weeks’. In: *BBC News. UK*. URL: <https://www.bbc.com/news/uk-57464097> (visited on 04/10/2021) (Cited on page 30).

- BBC News (5th Jan. 2021c). ‘Covid: New Lockdowns for England and Scotland Ahead of ‘Hardest Weeks’’. In: *BBC News. UK*. URL: <https://www.bbc.com/news/uk-55538937> (visited on 11/10/2021) (Cited on page 27).
- BBC News (22nd Feb. 2021d). ‘Lockdown: Boris Johnson Unveils Plan to End England Restrictions by 21 June’. In: *BBC News. UK*. URL: <https://www.bbc.com/news/uk-56158405> (visited on 07/10/2021) (Cited on page 27).
- J. Bröcker (6th Aug. 2008). ‘On Reliability Analysis of Multi-Categorical Forecasts’. In: *Nonlinear Processes in Geophysics* 15.4, pp. 661–673. ISSN: 1023-5809. DOI: 10.5194/npg-15-661-2008 (Cited on page 207).
- F. Caldicott (3rd Apr. 2020). *Data Sharing during This Public Health Emergency*. GOV.UK. URL: <https://www.gov.uk/government/speeches/data-sharing-during-this-public-health-emergency> (visited on 11/01/2022) (Cited on page 35).
- I. Carpenter, M. B. Ram, G. P. Croft and J. G. Williams (Aug. 2007). ‘Medical Records and Record-Keeping Standards’. In: *Clinical Medicine* 7.4, pp. 328–331 (Cited on page 5).
- D. Cereda, M. Manica, M. Tirani, F. Rovida, V. Demicheli, M. Ajelli, P. Poletti, F. Trentini, G. Guzzetta, V. Marziano et al. (1st Dec. 2021). ‘The Early Phase of the COVID-19 Epidemic in Lombardy, Italy’. In: *Epidemics* 37, p. 100528. ISSN: 1755-4365. DOI: 10.1016/j.epidem.2021.100528. arXiv: 2003.09320 (Cited on pages 9 and 12).
- R. Challen (Jan. 2022). *Jepidemic: R and Java Based Tools for Epidemic Modelling*. Version all versions. Zenodo. DOI: 10.5281/zenodo.5820292 (Cited on pages 32, 218, 249 and 258).
- R. Challen, G. Griffith, L. Lacasa and K. Tsaneva-Atanasova (2021a). ‘Algorithmic Hospital Catchment Area Estimation Using Label Propagation’. Manuscript submitted for publication. Manuscript submitted for publication. University of Exeter. URL: <https://terminological.github.io/arear/articles/catchment-areas.pdf> (Cited on page 15).

- R. Challen (2019). ‘Emerging Safety Issues in Artificial Intelligence’. Perspective. Perspective. Patient safety network, Agency for Healthcare Research and Quality. URL: <http://psnet.ahrq.gov/perspective/emerging-safety-issues-artificial-intelligence> (visited on 23/11/2020) (Cited on page 5).
- R. Challen (11th May 2021). *Potential Community Transmission of B.1.617.2 Inferred by S-gene Positivity - Briefing Note, 11 May 2021*. SAGE. URL: <https://www.gov.uk/government/publications/juniper-potential-community-transmission-of-b16172-inferred-by-s-gene-positivity-briefing-note-11-may-2021> (visited on 07/10/2021) (Cited on page 30).
- R. Challen, E. Brooks-Pollock, K. Tsaneva-Atanasova and L. Danon (21st Dec. 2021b). ‘Meta-Analysis of the Severe Acute Respiratory Syndrome Coronavirus 2 Serial Intervals and the Impact of Parameter Uncertainty on the Coronavirus Disease 2019 Reproduction Number’. In: *Statistical Methods in Medical Research*, p. 09622802211065159. ISSN: 0962-2802. DOI: 10.1177/09622802211065159 (Cited on page 17).
- R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards and K. Tsaneva-Atanasova (1st Mar. 2019). ‘Artificial Intelligence, Bias and Clinical Safety’. In: *BMJ Quality & Safety* 28.3, pp. 231–237. ISSN: 2044-5415, 2044-5423. DOI: 10.1136/bmjqs-2018-008370. pmid: 30636200 (Cited on pages 5 and 6).
- R. Challen, L. Dyson, C. E. Overton, L. M. Guzman-Rincon, E. M. Hill, H. B. Stage, E. Brooks-Pollock, L. Pellis, F. Scarabel, D. J. Pascall et al. (7th June 2021c). ‘Early Epidemiological Signatures of Novel SARS-CoV-2 Variants: Establishment of B.1.617.2 in England’. Manuscript submitted for publication. Manuscript submitted for publication. University of Exeter. URL: <https://www.medrxiv.org/content/10.1101/2021.06.05.21258365v1> (visited on 20/06/2021) (Cited on page 30).
- R. Challen, K. Tsaneva-Atanasova, T. Edwards, L. Gompels, M. Dayer, M. Pitt and L. Danon (1st July 2020). ‘Factors Influencing Digital Review of Pathology Test Results in an Inpatient Setting: A Cross-Sectional Study’. In: *JAMIA Open* 3.2, pp. 290–298. DOI: 10.1093/jamiaopen/ooaa003 (Cited on pages 7 and 19).

- A. J. Coale (1972). *The Growth and Structure of Human Populations: A Mathematical Investigation*. Princeton, N.J: Princeton University Press. xvii, 227. ISBN: 978-0-691-09357-4 (Cited on page 247).
- A. Cori, S. Cauchemez, N. M. Ferguson, C. Fraser, E. Dahlgvist, P. A. Demarsh, T. Jombart, Z. N. Kamvar, J. Lessler, S. Li et al. (7th Jan. 2021). *EpiEstim: Estimate Time Varying Reproduction Numbers from Epidemic Curves*. Version 2.2-4. URL: <https://CRAN.R-project.org/package=EpiEstim> (visited on 07/07/2021) (Cited on pages 17, 31, 202, 218 and 248).
- A. Cori, N. M. Ferguson, C. Fraser and S. Cauchemez (Nov. 2013). ‘A New Framework and Software to Estimate Time-Varying Reproduction Numbers during Epidemics’. In: *American Journal of Epidemiology* 178.9, pp. 1505–1512 (Cited on pages 17, 31, 202 and 218).
- S. Covo and A. Elalouf (Jan. 2014). ‘A Novel Single-Gamma Approximation to the Sum of Independent Gamma Variables, and a Generalization to Infinitely Divisible Distributions’. In: *Electronic Journal of Statistics* 8.1, pp. 894–926. ISSN: 1935-7524, 1935-7524. DOI: 10.1214/14-EJS914 (Cited on page 264).
- L. Danon, E. Brooks-Pollock, M. Bailey and M. Keeling (19th July 2021). ‘A Spatial Model of COVID-19 Transmission in England and Wales: Early Spread, Peak Timing and the Impact of Seasonality’. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 376.1829, p. 20200272. DOI: 10.1098/rstb.2020.0272 (Cited on pages 12 and 13).
- L. Danon, T. House and M. J. Keeling (Dec. 2009). ‘The Role of Routine versus Random Movements on the Spread of Disease in Great Britain’. In: *Epidemics* 1.4, pp. 250–258. ISSN: 1755-4365. DOI: 10.1016/j.epidem.2009.11.002. pmid: 21352771 (Cited on page 13).
- N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. B. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne et al. (9th Apr. 2021a). ‘Estimated Transmissibility and Impact of SARS-CoV-2 Lineage B.1.1.7 in England’. In: *Science* 372.6538. DOI: 10.1126/science.abg3055. pmid: 33658326 (Cited on page 24).

- N. G. Davies, C. I. Jarvis, W. J. Edmunds, N. P. Jewell, K. Diaz-Ordaz and R. H. Keogh (May 2021b). ‘Increased Mortality in Community-Tested Cases of SARS-CoV-2 Lineage B.1.1.7’. In: *Nature* 593.7858 (7858), pp. 270–274. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03426-1 (Cited on page 25).
- N. G. Davies, A. J. Kucharski, R. M. Eggo, A. Gimma, W. J. Edmunds, T. Jombart, K. O’Reilly, A. Endo, J. Hellewell, E. S. Nightingale et al. (1st July 2020). ‘Effects of Non-Pharmaceutical Interventions on COVID-19 Cases, Deaths, and Demand for Hospital Services in the UK: A Modelling Study’. In: *The Lancet Public Health* 5.7, e375–e385. ISSN: 2468-2667. DOI: 10.1016/S2468-2667(20)30133-X (Cited on pages 12 and 13).
- Deepmind (21st June 2017). *Enhancing patient safety at Taunton and Somerset NHS Foundation Trust*. Deepmind. URL: <https://deepmind.com/blog/article/taunton-and-somerset-nhs-foundation-trust-partnership> (visited on 13/09/2021) (Cited on page 5).
- Z. Du, X. Xu, Y. Wu, L. Wang, B. J. Cowling and L. A. Meyers (June 2020). ‘Serial Interval of COVID-19 among Publicly Reported Confirmed Cases’. In: *Emerging Infectious Diseases* 26.6, pp. 1341–1343. ISSN: 1080-6040. DOI: 10.3201/eid2606.200357. pmid: 32191173 (Cited on page 11).
- L. I. Dublin and A. J. Lotka (1st Sept. 1925). ‘On the True Rate of Natural Increase: As Exemplified by the Population of the United States, 1920’. In: *Journal of the American Statistical Association* 20.151, pp. 305–339. ISSN: 0162-1459. DOI: 10.1080/01621459.1925.10503498 (Cited on page 247).
- W. Feller (1941). ‘On the Integral Equation of Renewal Theory’. In: *The Annals of Mathematical Statistics* 12.3, pp. 243–267. ISSN: 0003-4851. DOI: 10.1214/aoms/1177731708. JSTOR: 2235855 (Cited on page 247).
- N Ferguson, D Laydon, G Nedjati Gilani, N Imai, K Ainslie, M Baguelin, S Bhatia, A Boonyasiri, Z. Cucunuba Perez, G Cuomo-Dannenburg et al. (9th Mar. 2020). *Report 9: Impact of Non-Pharmaceutical Interventions (NPIs) to Reduce COVID19 Mortality and Healthcare Demand*. MRC Centre for Global Infectious Disease Analysis: Imperial College London. URL: <https://www.imperial.ac>.



- uk/medicine/departments/school-public-health/infectious-disease-epidemiology/mrc-global-infectious-disease-analysis/covid-19/report-7-repatriation-flights-covid-19/ (Cited on pages 12 and 13).
- C. Fraser (22nd Aug. 2007). ‘Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic’. In: *PLOS ONE* 2.8, e758. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0000758 (Cited on pages 15 and 201).
- L. Fu, B. Wang, T. Yuan, X. Chen, Y. Ao, T. Fitzpatrick, P. Li, Y. Zhou, Y.-F. Lin, Q. Duan et al. (June 2020). ‘Clinical Characteristics of Coronavirus Disease 2019 (COVID-19) in China: A Systematic Review and Meta-Analysis’. In: *The Journal of Infection* 80.6, pp. 656–665. ISSN: 1532-2742. DOI: 10.1016/j.jinf.2020.03.041. pmid: 32283155 (Cited on page 10).
- T. Ganyani, C. Kremer, D. Chen, A. Torneri, C. Faes, J. Wallinga and N. Hens (30th Apr. 2020). ‘Estimating the Generation Interval for Coronavirus Disease (COVID-19) Based on Symptom Onset Data, March 2020’. In: *Eurosurveillance* 25.17, p. 2000257. ISSN: 1560-7917. DOI: 10.2807/1560-7917.ES.2020.25.17.2000257 (Cited on page 11).
- W.-J. Guan, Z.-Y. Ni, Y. Hu, W.-H. Liang, C.-Q. Ou, J.-X. He, L. Liu, H. Shan, C.-L. Lei, D. S. C. Hui et al. (Apr. 2020). ‘Clinical Characteristics of Coronavirus Disease 2019 in China’. In: *New England Journal of Medicine* 382.18, pp. 1708–1720 (Cited on page 10).
- I. Habli, T. Lawton and Z. Porter (1st Apr. 2020). ‘Artificial Intelligence in Health Care: Accountability and Safety’. In: *Bulletin of the World Health Organization* 98.4, pp. 251–256. ISSN: 0042-9686. DOI: 10.2471/BLT.19.237487. pmid: 32284648 (Cited on page 6).
- T. M. Hamill (1st Mar. 2001). ‘Interpretation of Rank Histograms for Verifying Ensemble Forecasts’. In: *Monthly Weather Review* 129.3, pp. 550–560. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2 (Cited on page 207).
- W. S. Hart, S. Abbott, A. Endo, J. Hellewell, E. Miller, N. Andrews, P. K. Maini, S. Funk and R. N. Thompson (30th May 2021a). ‘Inference of SARS-CoV-2 Gen-

- eration Times Using UK Household Data’. In: *medRxiv*, p. 2021.05.27.21257936. DOI: 10.1101/2021.05.27.21257936 (Cited on page 17).
- W. S. Hart, E. Miller, N. J. Andrews, P. Waight, P. K. Maini, S. Funk and R. N. Thompson (23rd Oct. 2021b). ‘Generation Time of the Alpha and Delta SARS-CoV-2 Variants’. In: ISSN: 2126-5216. DOI: 10.1101/2021.10.21.21265216 (Cited on page 17).
- C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu et al. (15th Feb. 2020). ‘Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China’. In: *The Lancet* 395.10223, pp. 497–506. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(20)30183-5. pmid: 31986264 (Cited on pages 8 and 10).
- N. Imai, A. Cori, I. Dorigatti, M. Baguelin, C. A. Donnelly, S. Riley and N. M. Ferguson (2020). *Report 3: Transmissibility of 2019-nCoV*. 3. MRC Centre for Global Infectious Disease Analysis: Imperial College London. URL: <https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2020-01-25-COVID19-Report-3.pdf> (Cited on page 11).
- S. Kakarmath, A. Esteva, R. Arnaout, H. Harvey, S. Kumar, E. Muse, F. Dong, L. Wedlund and J. Kvedar (16th Oct. 2020). ‘Best Practices for Authors of Healthcare-Related Artificial Intelligence Manuscripts’. In: *npj Digital Medicine* 3.1 (1), pp. 1–3. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00336-w (Cited on page 6).
- M. J. Keeling, L. Danon, M. C. Vernon and T. A. House (11th May 2010). ‘Individual Identity and Movement Networks for Disease Metapopulations’. In: *Proceedings of the National Academy of Sciences* 107.19, pp. 8866–8870. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1000416107. pmid: 20421468 (Cited on page 13).
- S. Kotz (8th May 1995). *Continuous Univariate Distributions, Vol. 2*. Wiley-Interscience. 752 pp. (Cited on page 265).
- A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday et al. (1st May 2020). ‘Early Dynamics of Transmission and Control of COVID-19: A Mathematical Modelling Study’.

- In: *The Lancet Infectious Diseases* 20.5, pp. 553–558. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(20)30144-4 (Cited on page 11).
- L. Lacasa, R. Challen, E. Brooks-Pollock and L. Danon (21st Oct. 2020). ‘A Flexible Method for Optimising Sharing of Healthcare Resources and Demand in the Context of the COVID-19 Pandemic’. In: *PLOS ONE* 15.10, e0241027. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0241027 (Cited on page 13).
- L. M. Leemis and J. T. McQueston (Feb. 2008). ‘Univariate Distribution Relationships’. In: *The American Statistician* 62.1, pp. 45–53. ISSN: 0003-1305, 1537-2731. DOI: 10.1198/000313008X270448 (Cited on page 261).
- F.-X. Lescure, L. Bouadma, D. Nguyen, M. Parisey, P.-H. Wicky, S. Behillil, A. Gaymard, M. Bouscambert-Duchamp, F. Donati, Q. L. Hingrat et al. (1st June 2020). ‘Clinical and Virological Data of the First Cases of COVID-19 in Europe: A Case Series’. In: *The Lancet Infectious Diseases* 20.6, pp. 697–706. ISSN: 1473-3099, 1474-4457. DOI: 10.1016/S1473-3099(20)30200-0. pmid: 32224310 (Cited on page 9).
- Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong et al. (26th Mar. 2020). ‘Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia’. In: *New England Journal of Medicine* 382.13, pp. 1199–1207. ISSN: 0028-4793. DOI: 10.1056/NEJMoa2001316. pmid: 31995857 (Cited on page 11).
- Y.-F. Lin, Q. Duan, Y. Zhou, T. Yuan, P. Li, T. Fitzpatrick, L. Fu, A. Feng, G. Luo, Y. Zhan et al. (2020). ‘Spread and Impact of COVID-19 in China: A Systematic Review and Synthesis of Predictions From Transmission-Dynamic Models’. In: *Frontiers in Medicine* 7, p. 321. ISSN: 2296-858X. DOI: 10.3389/fmed.2020.00321. pmid: 32626719 (Cited on page 11).
- C. Loader, J. Sun, L. Technologies and A. Liaw (25th Mar. 2020). *Locfit: Local Regression, Likelihood and Density Estimation*. Version 1.5-9.4. URL: <https://CRAN.R-project.org/package=locfit> (visited on 08/07/2020) (Cited on page 249).

- C. Loader (1999). *Local Regression and Likelihood*. Statistics and Computing. New York: Springer-Verlag. ISBN: 978-0-387-98775-0. DOI: 10.1007/b98858 (Cited on pages 33, 245 and 247).
- Local Government Association (26th July 2021). *Coping with the Rise of the Delta Variant in Bolton — Local Government Association*. URL: <https://www.local.gov.uk/coping-rise-delta-variant-bolton> (visited on 04/10/2021) (Cited on page 30).
- T. Maishman, S. Schaap, D. S. Silk, S. J. Nevitt, D. C. Woods and V. E. Bowman (12th Aug. 2021). *Statistical Methods Used to Combine the Effective Reproduction Number,  $R(t)$ , and Other Related Measures of COVID-19 in the UK*. arXiv: 2103.01742 [physics, q-bio, stat]. URL: <http://arxiv.org/abs/2103.01742> (visited on 06/01/2022) (Cited on page 16).
- R. Mann and J. Williams (July 2003). ‘Standards in Medical Record Keeping’. In: *Clinical Medicine* 3.4, pp. 329–332 (Cited on page 5).
- K. Mizumoto and G. Chowell (1st Jan. 2020). ‘Transmission Potential of the Novel Coronavirus (COVID-19) Onboard the Diamond Princess Cruises Ship, 2020’. In: *Infectious Disease Modelling* 5, pp. 264–270. ISSN: 2468-0427. DOI: 10.1016/j.idm.2020.02.003 (Cited on page 11).
- K. Mizumoto, K. Kagaya, A. Zarebski and G. Chowell (12th Mar. 2020). ‘Estimating the Asymptomatic Proportion of Coronavirus Disease 2019 (COVID-19) Cases on Board the Diamond Princess Cruise Ship, Yokohama, Japan, 2020’. In: *Eurosurveillance* 25.10, p. 2000180. ISSN: 1560-7917. DOI: 10.2807/1560-7917.ES.2020.25.10.2000180 (Cited on pages 9 and 10).
- P. Moss, G. Barlow, N. Easom, P. Lillie and A. Samson (Mar. 2020). ‘Lessons for Managing High-Consequence Infections from First COVID-19 Cases in the UK’. In: *Lancet* 395.10227, e46 (Cited on page 9).
- NHS England (2018). *Global Digital Exemplars*. URL: <https://www.england.nhs.uk/digitaltechnology/connecteddigitalsystems/exemplars/> (visited on 13/09/2021) (Cited on page 4).

- Á. O’Toole, V. Hill, O. G. Pybus, A. Watts, I. I. Bogoch, K. Khan, J. P. Messina, T. C.-. G. U. C.-U. Consortium, N. f. G. S. i. S. Africa (NGS-SA), B.-U. C. G. Network et al. (17th Sept. 2021). ‘Tracking the International Spread of SARS-CoV-2 Lineages B.1.1.7 and B.1.351/501Y-V2 with Grinch’. In: 6:121. DOI: 10.12688/wellcomeopenres.16661.2 (Cited on page 24).
- L. Pellis, F. Scarabel, H. B. Stage, C. E. Overton, L. H. K. Chappell, E. Fearon, E. Bennett, K. A. Lythgoe, T. A. House and I. Hall (19th July 2021). ‘Challenges in Control of COVID-19: Short Doubling Time and Long Delay to Effect of Interventions’. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 376.1829, p. 20200264. DOI: 10.1098/rstb.2020.0264 (Cited on pages 12 and 244).
- T. Pham-Gia and N. Turkkan (12th July 2011). ‘Distributions of Ratios: From Random Variables to Random Matrices’. In: *Open Journal of Statistics* 1.2 (2), pp. 93–104. DOI: 10.4236/ojs.2011.12011 (Cited on page 261).
- H. Pishro-Nik (24th Aug. 2014). *Introduction to Probability, Statistics, and Random Processes*. Blue Bell, PA: Kappa Research, LLC. 744 pp. ISBN: 978-0-9906372-0-2 (Cited on page 261).
- Potential Sequences That Should Be Included in B.1.617* (14th May 2021). cov-lineages/pango-designation. URL: <https://github.com/cov-lineages/pango-designation/issues/49> (visited on 11/10/2021) (Cited on page 30).
- J. Powles and H. Hodson (Mar. 2017). ‘Google DeepMind and Healthcare in an Age of Algorithms’. In: *Health Technology*, pp. 1–17 (Cited on page 5).
- Proposed New B.1 Sublineage Circulating in India* (30th Mar. 2021). cov-lineages/pango-designation. URL: <https://github.com/cov-lineages/pango-designation/issues/38> (visited on 11/10/2021) (Cited on page 30).
- Public Health England (8th July 2021a). *COVID-19 Vaccine Surveillance Report - Week 27*. Public Health England, p. 25. URL: [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/1000512/Vaccine\\_surveillance\\_report\\_-\\_week\\_27.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1000512/Vaccine_surveillance_report_-_week_27.pdf) (visited on 06/10/2021) (Cited on page 27).

- Public Health England (7th May 2021b). *SARS-CoV-2 Variants of Concern and Variants under Investigation: Technical Briefing 10*. 10. Public Health England, p. 39. URL: [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/984274/Variants\\_of\\_Concern\\_VOC\\_Technical\\_Briefing\\_10\\_England.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/984274/Variants_of_Concern_VOC_Technical_Briefing_10_England.pdf) (Cited on page 31).
- A. Rambaut, E. C. Holmes, Á. O’Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis and O. G. Pybus (Nov. 2020). ‘A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology’. In: *Nature Microbiology* 5.11 (11), pp. 1403–1407. ISSN: 2058-5276. DOI: 10.1038/s41564-020-0770-5 (Cited on page 24).
- J. M. Read, J. R. E. Bridgen, D. A. T. Cummings, A. Ho and C. P. Jewell (19th July 2021). ‘Novel Coronavirus 2019-nCoV (COVID-19): Early Estimation of Epidemiological Parameters and Epidemic Size Estimates’. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 376.1829, p. 20200265. DOI: 10.1098/rstb.2020.0265 (Cited on page 11).
- A. Remuzzi and G. Remuzzi (11th Apr. 2020). ‘COVID-19 and Italy: What Next?’ In: *The Lancet* 395.10231, pp. 1225–1228. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(20)30627-9. pmid: 32178769 (Cited on page 9).
- W. Research (2010). *Beta Prime Distribution*. Wolfram language function. URL: <https://reference.wolfram.com/language/ref/BetaPrimeDistribution.html> (visited on 05/01/2022) (Cited on page 265).
- S. R. Seaman, T. Nyberg, C. E. Overton, D. Pascall, A. M. Presanis and D. D. Angelis (18th Aug. 2021). ‘Adjusting for Time of Infection or Positive Test When Estimating the Risk of a Post-Infection Outcome in an Epidemic’. In: p. 2021.08.13.21262014. DOI: 10.1101/2021.08.13.21262014 (Cited on page 26).
- S. Siegert, J. Bhend, I. Kroener and M. D. Felice (26th Feb. 2020). *SpecsVerification: Forecast Verification Routines for Ensemble Forecasts of Weather and Climate*. Version 0.5-3. URL: <https://CRAN.R-project.org/package=SpecsVerification> (visited on 23/11/2021) (Cited on page 207).

- J. Smeaton and L. Christie (12th Oct. 2021). ‘AI and Healthcare’. In: URL: <https://post.parliament.uk/research-briefings/post-pn-0637/> (visited on 12/10/2021) (Cited on page 6).
- C. Sohrabi, Z. Alsafi, N. O’Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis and R. Agha (1st Apr. 2020). ‘World Health Organization Declares Global Emergency: A Review of the 2019 Novel Coronavirus (COVID-19)’. In: *International Journal of Surgery* 76, pp. 71–76. ISSN: 1743-9191. DOI: 10.1016/j.ijvsu.2020.02.034 (Cited on page 9).
- Somerset NHS Foundation Trust (2022). *About Us*. Somerset NHS Foundation Trust. URL: <https://www.somersetft.nhs.uk/about-us/> (visited on 11/01/2022) (Cited on page 4).
- M. Taboga (8th Dec. 2017). *Lectures on Probability Theory and Mathematical Statistics - 3rd Edition*. CreateSpace Independent Publishing Platform. 670 pp. ISBN: 978-1-981369-19-5 (Cited on page 261).
- M. Taboga (2021). *Functions of Random Variables and Their Distribution*. Lectures on probability theory and mathematical statistics, Kindle direct publishing - Online Appendix. URL: <https://www.statlect.com/fundamentals-of-probability/functions-of-random-variables-and-their-distribution> (visited on 05/01/2022) (Cited on page 261).
- R. N. Thompson, J. E. Stockwin, R. D. van Gaalen, J. A. Polonsky, Z. N. Kamvar, P. A. Demarsh, E. Dahlqwisst, S. Li, E. Miguel, T. Jombart et al. (1st Dec. 2019). ‘Improved Inference of Time-Varying Reproduction Numbers during Infectious Disease Outbreaks’. In: *Epidemics* 29, p. 100356. ISSN: 1755-4365. DOI: 10.1016/j.epidem.2019.100356 (Cited on pages 17, 31, 202, 218 and 248).
- UK Government (16th Mar. 2020a). *Coronavirus: UK Government Ventilator Challenge - GOV.UK*. URL: <https://www.gov.uk/government/topical-events/coronavirus-uk-government-ventilator-challenge> (visited on 06/01/2022) (Cited on page 13).
- UK Government (16th Mar. 2020b). *COVID-19: Guidance on Social Distancing and for Vulnerable People*. GOV.UK. URL: <https://www.gov.uk/government/>

- publications / covid - 19 - guidance - on - social - distancing - and - for - vulnerable-people (visited on 10/01/2022) (Cited on page 9).
- UK Government (23rd Mar. 2020c). *Staying at Home and Away from Others (Social Distancing)*. GOV.UK. URL: <https://www.gov.uk/government/publications/full-guidance-on-staying-at-home-and-away-from-others> (visited on 10/01/2022) (Cited on page 9).
- UK Government (19th Apr. 2021). *India Added to Red List as Cases of New Variant Rise*. GOV.UK. URL: <https://www.gov.uk/government/news/india-added-to-red-list-as-cases-of-new-variant-rise> (visited on 30/09/2021) (Cited on page 27).
- C. Vegvari (2020). *Reproduction Number (R) and Growth Rate (r) of the COVID-19 Epidemic in the UK*. Royal Society, p. 86 (Cited on page 16).
- C. Vegvari, S. Abbott, F. Ball, E. Brooks-Pollock, R. Challen, B. S. Collyer, C. Dangerfield, J. R. Gog, K. M. Gostic, J. M. Heffernan et al. (27th Sept. 2021). ‘Commentary on the Use of the Reproduction Number R during the COVID-19 Pandemic’. In: *Statistical Methods in Medical Research*, p. 9622802211037079. ISSN: 1477-0334. DOI: 10.1177/09622802211037079. pmid: 34569883 (Cited on pages 15 and 200).
- R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. T. Walker, H. Fu et al. (1st June 2020). ‘Estimates of the Severity of Coronavirus Disease 2019: A Model-Based Analysis’. In: *The Lancet Infectious Diseases* 20.6, pp. 669–677. ISSN: 1473-3099, 1474-4457. DOI: 10.1016/S1473-3099(20)30243-7. pmid: 32240634 (Cited on page 10).
- E. Volz, M. Baguelin, S. Bhatia, A. Boonyasiri, A. Cori, Z. Cucunubá, G. Cuomo-Dannenburg, C. A. Donnelly, I. Dorigatti, R. FitzJohn et al. (15th Feb. 2020). *Report 5: Phylogenetic Analysis of SARS-CoV-2*. 5. MRC Centre for Global Infectious Disease Analysis: Imperial College London. URL: <https://www.imperial.ac.uk/medicine/departments/school-public-health/infectious-disease-epidemiology/mrc-global-infectious-disease-analysis/covid-19/report-5-phylogenetics-of-sars-cov-2/> (Cited on page 12).



- J Wallinga and M Lipsitch (22nd Feb. 2007). ‘How Generation Intervals Shape the Relationship between Growth Rates and Reproductive Numbers’. In: *Proceedings of the Royal Society B: Biological Sciences* 274.1609, pp. 599–604. DOI: 10.1098/rspb.2006.3754. pmid: 17476782 (Cited on pages 11, 15, 33, 202, 204, 206, 237, 247, 248 and 254).
- E. W. Weisstein (2010). *Regularized Beta Function*. MathWorld - A Wolfram Web Resource. URL: <https://mathworld.wolfram.com/RegularizedBetaFunction.html> (visited on 05/01/2022) (Cited on page 262).
- World Health Organisation (n.d.). *Tracking SARS-CoV-2 Variants*. URL: <https://www.who.int/emergencies/what-we-do/tracking-SARS-CoV-2-variants> (visited on 06/01/2022) (Cited on page 24).
- J. Wu, J. Liu, X. Zhao, C. Liu, W. Wang, D. Wang, W. Xu, C. Zhang, J. Yu, B. Jiang et al. (Feb. 2020). ‘Clinical Characteristics of Imported Cases of COVID-19 in Jiangsu Province: A Multicenter Descriptive Study’. In: *Clinical Infectious Diseases* (Cited on page 8).
- J. T. Wu, K. Leung and G. M. Leung (Feb. 2020). ‘Nowcasting and Forecasting the Potential Domestic and International Spread of the 2019-nCoV Outbreak Originating in Wuhan, China: A Modelling Study’. In: *Lancet* 395.10225, pp. 689–697 (Cited on page 11).
- M. Zamo and P. Naveau (1st Feb. 2018). ‘Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts’. In: *Mathematical Geosciences* 50.2, pp. 209–234. ISSN: 1874-8953. DOI: 10.1007/s11004-017-9709-7 (Cited on page 208).
- S. Zhao, D. Gao, Z. Zhuang, M. K. C. Chong, Y. Cai, J. Ran, P. Cao, K. Wang, Y. Lou, W. Wang et al. (2020). ‘Estimating the Serial Interval of the Novel Coronavirus Disease (COVID-19): A Statistical Analysis Using the Public Data in Hong Kong From January 16 to February 15, 2020’. In: *Frontiers in Physics* 8, p. 347. ISSN: 2296-424X. DOI: 10.3389/fphy.2020.00347 (Cited on page 11).