Combining models to generate a consensus effective reproduction number R for the COVID-19 epidemic status in England

Josie Park[1], Luke Bevan[1,2], Alberto Sanchez-Marroquin[1], Gabriel Danelian[1], Thomas Bayley[1], Harrison Manley[1], Veronica Bowman[3], Thomas Maishman[3], Thomas Finnie[1], André Charlett[1], Nicholas A Watkins[1], Johanna Hutchinson[1], Steven Riley[1], Nowcasts Model Contributing Group[4], and Jasmina Panovska-Griffiths[1, 5, 6]

[1]UK Health Security Agency
[2]University College London
[3]Defence Science and Technology Laboratory
[4]The Nowcasts model contribution group comprises: Sebastian Funk (LSHTM), Paul Birrell and Daniela De Angelis (UK Health Security Agency and MERC Biostatistics Unit, University of Cambridge, Matt Keeling (University of Warwick), Lorenzo Pellis (University of Manchester), Marc Baguelin (Imperial College London), Graeme Ackland (University of Edinburgh), Jonathan Read and Christopher Jewell (University of Lancaster), Robert Challen (University of Exeter)
[5]The Big Data Institute and the Pandemic Sciences Institute, University of Oxford
[6]The Queen's College, University of Oxford

February 2023

Abstract

The effective reproduction number R was widely accepted as a key indicator during the early stages of the COVID-19 pandemic. In the UK, the R value published on the UK Government Dashboard has been generated as a combined value from an ensemble of fourteen epidemiological models via a collaborative initiative between academia and government. In this paper we outline this collaborative modelling approach and illustrate how, by using an established combination method, a combined R estimate can be generated from an ensemble of epidemiological models. We show that this R is robust to different model weighting methods and ensemble size and that using heterogeneous data sources for validation increases its robustness and reduces the biases and limitations associated with a single source of data. We discuss how R can be generated from different data sources and is therefore a good summary indicator of the current dynamics in an epidemic.

1 Introduction

Since the onset of the coronavirus disease in 2019 (COVID-19) as a pandemic, mathematical modelling has been widely used to generate policy-relevant evidence. Mathematical modelling provides a framework for simulating the dynamics of the pandemic. When parameterised with, and calibrated to data, this can be used to generate projections of future epidemic trajectories. Epidemiological estimates such as the reproduction number R derived from models can be useful tools for epidemic status tracking.

The reproduction number R is a measure of the infectious potential of a disease and represents the number of secondary infections that emerge from one infection [26]. At the onset of a new disease, in a naive, fully susceptible population, the basic reproduction number R_0 represents the number of secondary infections stemming from an initial case. In contrast to R_0, R is the reproduction number at any time during an epidemic - often referred to as the *effective reproduction number* R_e or *temporal reproduction number* R_t [5]. It reflects the number of secondary infections generated from a population consisting of susceptible, exposed and immune individuals, and potential changes in mixing and the presence of interventions.

The growth rate r represents the rate at which the epidemic is growing during the exponential phase of epidemic growth. In epidemiological modelling, r and R are related via the generation time (τ) of the epidemic: the longer the generation time and the higher the epidemic growth rate, the higher the value of R [5]. Mathematically, this is expressed as

$$R^{-1} = \int_0^\infty e^{-r\tau} f(\tau) d\tau, \tag{1}$$

where τ is the time since infection, and $f(\tau)$ is the probability density function for the time of infection, or the generation time distribution.

While R is reflective of the current strength of transmission, r is reflective of the transmission speed [18]. In policy, R has been used as it provides more information about the impact of control measures compared to r. For example, if an intervention is imposed and R is consequentially reduced to below the $R = 1$ threshold, this suggests that the intervention has had an impact on reducing onward transmission. Additionally, R at the onset i.e. R_0 also provides information on the likely level of herd immunity necessary, $(1 - \frac{1}{R_0})$. This suggests that the more people that become infected by each individual who has the virus, the higher the proportion of the population that needs to be immune to reach herd immunity. Further details on R and the differing methodologies for calculating the reproduction number can be found in section 2.1.

In the UK, the Scientific Advisory Group for Emergencies (SAGE), is activated in response to emergencies and is made up of several sub-groups consisting of experts relating to different scientific fields [10]. These sub-groups are often called upon in order to provide evidence to the UK government relating to key policy questions. One of these groups is the Scientific Pandemic Influenza Group on Modelling - Operational (SPI-M-O). SPI-M-O primarily consists of experts in infectious diseases modelling.

During the COVID-19 pandemic, SPI-M-O, and from June 2021 the United Kingdom Health Security Agency (UKHSA) has provided the UK government with weekly estimates of key epidemiological indicators, including the effective reproduction number R [22]. These values were generated as a combined estimate from a set of epidemiological models maintained and run by members of SPI-M-O and combined using a random effects meta analysis approach with equal weighting applied [35], with visualisation implemented using CrystalCast developed by the Defence Science and

Technology Laboratory (DSTL)[43].

Generating an across-model combined estimate, instead of single model truth, can lead to improved predictive power [48], allows an increased robustness of the outcomes and is a useful tool for policy makers [40]. Generating a combined estimate from a set of models is not a new concept; they are widely used across many disciplines; in forecasting the weather [34], hydrology [25], flood losses [20], in cancer prediction [52] and climate modelling [36]. Within infectious diseases, combined model estimates have been applied to modelling HIV [19], influenza [42] and Ebola [45, 13] transmission, and recently for outbreak analysis related to COVID-19 in the USA [47] and Europe [9].

While mathematical models have been used to offer informed advice to the scientific community and policy makers throughout the COVID-19 pandemic across a number of countries, the use of modelling has differed. For example, modellers in the United States, in conjunction with the Center for Disease Control (CDC), published ensemble forecasts using a wide variety of mathematical models [16, 41]. These models had focused on forecasting new cases, hospitalisations and deaths at a national and state-level as opposed to estimating R or r specifically. On the other hand, in New Zealand and Italy modellers advising the government have compared estimates of R obtained from different models but without producing formal combined estimates [29, 15]. In Norway, multiple data sources including confirmed cases, proportion of COVID-19 attributable hospital admissions and a national symptom survey were used to estimate r over the course of the pandemic, but only one model has been used to estimate R from these sources [38]. Similarly, the Robert Koch Institute in Germany only uses a single model to estimate R which depends on nowcasting estimates of the number of new cases [44].

In the UK, since the onset of the pandemic a set of fourteen mathematical models developed, maintained and applied by the members of SPI-M-O have been used to track epidemic status, including generating R and r. The R value published on the UK Government Dashboard [27] has been generated as a combined value from these models. Since early 2021, the United Kingdom Health Security Agency (UKHSA) Epidemiological Ensemble (Epi-Ensemble) has been working with academic modelling groups, including SPI-M-O members, and with the CrystalCast team to run the process of combining epidemic outcomes from these different models to generate a weekly or fortnightly combined estimates for R.

The value in getting a combined value from across models and datasets is not just in the averaging of those estimates with weighting, but also in the formation of a community that are constantly discussing the outcomes, the assumptions, the input data identifying the drivers behind the differences across models. This is especially important when generating R. While doubling time and r can be thought of as almost features of the data, requiring very few assumptions, the move to R requires a set of subjective assumptions. This is why there is a need to have multiple groups making different assumptions leading to heterogeneous outcomes which can be discussed, understood and combined. When R can be generated using different data sets, in addition to different models, this is particularly important.

This paper outlines how a previously established combination method, described in [35], has been applied in the UK throughout the COVID-19 pandemic. We detail our approach of generating a consensus value of R from fourteen epidemiological models applied to the English epidemic. We illustrate the process, show how a combined R estimate has been generated in April 2021 and in September 2021 and explore the robustness of the combined R value on the size and weighting of the models' combination. By comparing the change in R with the change in COVID-19 cases, hospitalisations and deaths, we also explore whether R has been a good indicator of epidemic status.

2 Methodology

2.1 Outline of epidemiological models used to produce R values

Generating an R estimate requires a model of some kind with subjective assumptions and information from other sources. Fourteen mathematical models have been developed, adapted and used throughout 2020-2022 to model the COVID-19 epidemic in England and to generate epidemic metrics including R. These models fall into three broad groups, as described in [17] and [5]: data-driven models (DDMs), population based models (PBMs) and agent-based models (ABMs). Table 1 outlines these models along with a high-level description of the method used to calculate R, with their main characteristics and differences described in table 2.

Model Name	**Description**	**R estimation**
Manchester Model (DetSEIRwithNB) [39, 11]	A deterministic compartmental ODE model that fits to hospital admissions, hospital occupancy, ICU occupancy and deaths in hospital. β, the transmission rate, varies step-wise between change points. Change points, such as policy or behavioural changes (e.g. schools returning and lockdowns) are defined by the modeller and are used to represent changes in the epidemic.	R is estimated from the most recent β
EpiEstim [14, 31, 32]	EpiEstim applies the renewal equation given a time series of incidence. We use the implementation of EpiEstim [14] described in [32] with code available at [31]. We back-calculate from an observation, e.g. cases, to time of infection, using an assumed delay distribution.	R is calculated using the renewal equation, eq. (4).
GenomicSurveillance [51]	Spatio-temporal hierarchical Bayesian model fitting to daily new cases and COVID-19 lineage counts.	R is calculated based on the derivative of cubic-spline function fitted to the incidence.
Epidemia [46]	Hierarchical semi-mechanistic Bayesian model based on the renewal equation. It can fit to multiple data streams at once, or a single data stream. UKHSA EpiEnsemble runs two separate versions of the model, one that fits to weekly admissions, and one that fits to weekly cases. The admissions version of the model was developed later in the epidemic in response to changes in the case ascertainment rate.	R is specified to vary weekly according to a random walk and is calculated using the renewal equation.
Covasim [33]	Stochastic agent-based model calibrated to COVID-19 diagnoses, hospital admissions and deaths related to COVID-19 and modelling progressive vaccine roll-outs. It also models social distancing and Test-Trace-Isolate interventions and incorporates different SARS-CoV-2 variants.	R is calculated by directly counting the number of secondary infections that are caused by each primary infection.

University of Liverpool Model [37]	Bayesian statistical model that comprises a deterministic compartmental transmission model governed by a system of ODEs and a stochastic observation model. Fits to deaths, hospital admissions, and symptomatic report data from NHS 111 online.	R is calculated from estimates of the daily number of infections, the infectious population, and the mean time for which individuals are infectious.
Lancaster Spatial Stochastic	A Bayesian spatial stochastic compartmental model, that fits to cases and uses census commuter flow data to infer mixing between Local Authorities.	R is calculated as the dominant eigenvalue of the next generation matrix.
LSHTM EpiNow2 [3, 1]	EpiNow2 uses the renewal equation to estimate R, where initial infections are estimated based on the initial number of cases or hospital admissions. The relationship between cases (or hospital admissions) and infections is obtained from a convolution of the relevant delay distributions (an uncertain incubation period and reporting delay). Similarly to Epidemia, versions fitting to cases and admissions are run.	R is derived using the renewal equation.
LSHTM ONS inc2prev [2]	A Gaussian process model that uses PCR positivity rates published by the ONS to model incidence by convolution with the curve estimating the evolution of the probability of a positive test since time of infection.	R is derived using the renewal equation.
Imperial Stochastic Compartmental (sircovid) [6]	Compartmental transmission model described by stochastic difference equations fitted to deaths, hospital admissions and prevalence, tested cases in hospital beds, ICU prevalence and serology data.	R is defined as the maximum eigenvalue of the next generation matrix.
Oxford CSML Model. Dashboard on [49]	Hierarchical semi-mechanistic Bayesian model fitted to cases, similar to Epidemia and as described in [21], but with a spatio-temporal component.	R is calculated similarly to Epidemia.
PHE/Cambridge Model [8, 7]	Deterministic age-structured compartmental model fitted to serology data and uses google mobility data. Different versions of the model have been run throughout the pandemic that fit to slightly different data streams. Two versions of the model are presented in fig. 1. Deaths/ons fits to ONS infection survey data, whereas regional/age does not. More recently during the pandemic, the regional/age model has been replaced with the admissions/ons model, which has the same model structure as deaths/ons, but fits to admissions.	R is calculated in terms of the dominant eigenvalue of the next-generation matrix.

OpenABM [28]	Stochastic agent-based model calibrated to hospital admissions, ICU bed occupancy, deaths and vaccinations. Similarly to Covasim, it also explicitly models social mixing, Test-Trace-Isolate interventions as well as different SARS-CoV-2 variants and progressive vaccination.	R is calculated similarly as in Covasim by directly counting the number of secondary infections that are caused by each primary infection.
Warwick Model [30]	A deterministic age-structured compartmental ODE model fitted to hospital and ICU admissions, and COVID-19 positivity rate data.	R is calculated from the next generation matrix.
Edinburgh WSS Model [4]	The WSS (Weight-Shift-Scale) model fits to case data to derive R but accounts for systematic reporting errors (e.g. false positives and negatives and under-reporting). Furthermore, case counts are weighted, scaled and shifted to account for the change in the size in future compartments, the delay between infection and case reporting, and to account for seasonality.	The estimated R is assumed to be the combination of the true R plus a stochastic term and is calculated from the rate of change in reported cases, scaled by the time lag between infection and the time of case report.

Table 1: Outline of the epidemiological models used to generate R outcomes for the English COVID-19 epidemic. We list the names of the models, as well as their main modelling characteristics and the data to which they are calibrated against.

While these models can be broadly stratified into these three structure-based groups, each model within the group has distinctive characteristics. For example, EpiEstim follows the methodology as described in [14], and therefore share the limitation that they assume a consistent relationship between infections and cases. The estimated R is robust when the ascertainment rate is roughly constant. While GenSur shares this same limitation, Epidemia and OxfordCSML do not make this assumption [21]. Furthermore, renewal-equation based models tend to be semi-mechanistic i.e. assuming that the effects of interventions are assumed to be absorbed into the data to which they fit. In contrast, fully mechanistic models such as the SEIR population-based and the agent-based models, explicitly model the effects of interventions such as Test-Trace-Isolate strategies and imposing and removing of social distancing measures.

In epidemiological models, the structure of the model determines the method to calculate R and depends on the assumptions and data sets used to parameterise and validate the model. For example, individual-based models count exactly the number of secondary infections at any stage of the epidemic and hence explicitly estimate R.

In the classic compartmental Susceptible-Exposed-Infected-Recovered (SEIR) model, $R_0 = \beta * c/\gamma$ where β is the transmission probability, c is the number of contacts c and $1/\gamma$ is the infectiousness period (average time that an individual is infectious for). R is typically calculated as the largest eigenvalue of the next generation matrix (NGM). The NGM can be expressed as FV^{-1}, where F represents infection rates, and V recovery rates. The exact representation of F and V vary dependant on the structure of the model [12]. Though generally,

$$\frac{dx_i}{dt} = F_i(x) - V_i(x), \quad V_i(x) = V_i^-(x) - V_i^+(x), \tag{2}$$

where x_i is the number of infected individuals in the ith infected compartment at time t, $F_i(x)$ is the rate at which individuals in compartment i are infected, where $V_i^-(x)$ and $V_i^+(x)$ represent the rate at which individuals leave and enter the ith compartment respectively through means other than infection. Expressing F and V as below, where m is the number of compartments, we can derive the NGM:

$$F(x) = (F_1(x), \cdot F_m(x)), \quad V(x) = (V_1(x), \cdot, V_m(x)). \tag{3}$$

We can express the relationship between the reproduction number, the expected incidence, and previous incidence rates with the renewal equation:

$$R(t) = \frac{E(I(t))}{\int_0^{\inf} f(\tau) I(t-\tau) d\tau} \tag{4}$$

where I_t is number of infections as time t, and $E[]$ denotes the expected value.

A selection of models in the ensemble are formulated based on eq. (1), these are typically models of the data-driven type. Where the generation time is described by a gamma distribution, with shape a and rate b, R can be expressed in terms of the growth rate r as:

$$R = \frac{(r+b)^a}{b^a}, \tag{5}$$

Another key difference between the models in the ensemble is the data to which they fit. The models are broadly split into three categories based on the data they primarily use to inform their estimates: case-based models, admissions-based models, and models that fit to both case data and hospital data. For the purposes of this study, models that fit to survey data are categorised as case-based models as they are focused on detecting incidence of the disease, though there will be differing delays associated with models that fit to cases and models that fit to survey data. There are drawbacks and advantages associated with fitting to either cases or admissions. Case data is highly sensitive to ascertainment biases. The scale of these biases have varied greatly over time. Therefore, models that fit to case counts or positivity must be interpreted in the context of testing behaviours and policies at the time. However, admissions data is not free from bias either. The likelihood of being admitted to hospital varies greatly by age. Hence, without age-stratification in the model, it is likely that community transmission is under-estimated among younger age groups. Furthermore, the delay between being infected with COVID-19 and being admitted to hospital is on average far greater than that between infection and receiving a positive case. This presents difficulties when trying to produce timely estimates of community transmission. A summary of the data to which each model fits is described in table 2.

Model Name	Model Type	Data Type	Is the model run either by the UKHSA or a Devolved Administration department?
Manchester Model	PBM	Hospital data	Yes
EpiEstim	DDM	Case data	Yes
GenomicSurveillance	DDM	Case data	Yes
Epidemia	DDM	UKHSA run two versions, one fitting to cases, one to admissions data.	Yes
Covasim	ABM	The UKHSA version fits to a mix of data.	Yes
University of Liverpool Model	PBM	Hospital data	Yes
Lancaster Spatial Stochastic	PBM	Case data	No
LSHTM EpiNow2	DDM	Two versions are run: one fitting to cases and one to admissions	Yes
LSHTM ONS inc2prev	DDM	Fits to ONS positivity, which is treated as case data	Yes
Imperial Model	PBM	A mix of data	No
Oxford CSML Model	DDM	Case data	Yes
PHE/Cam	PBM	A mix of data	Yes
OpenABM	ABM	Hospital data	Yes
Warwick Model	PBM	A mix of data	No
Edinburgh WSS Model	PBM	Cases	No

Table 2: Table detailing the UKHSA/SPI-M-O models split by model type and the data to which they fit

2.2 Combining model estimates to generate a consensus R

To generate combined R estimates from the fourteen models, we use the statistical model developed as a collaboration between DSTL, University of Southampton and University of Liverpool with the underlying methodology described in [35]. We present a high-level outline of the method below. Each of the epidemiological models described in table 1 and calibrated to the data as outlined in table 2, has generated 5th, 25th, 50th, 75th and 95th percentile estimates for R. Using these, a mean and a standard deviation for each model's R estimate was generated. The mean (y_i) is initially estimated as the median (or 50th quantile), and the standard deviation is calculated as follows:

$$s_i = \frac{\max\left(|q_i(95) - q_i(50)|, |q_i(50) - q_i(5)|\right)}{z_{95}}, \tag{6}$$

Where $q_i(x)$ represented the xth quantile of the ith model, and z_{95} is the z-score for the 90% confidence interval (CI) of the standard normal distribution. Where model estimates are highly skewed, a skewness correction calculation was applied to provide alternative estimates for the mean and standard error (see [35] for further details). Otherwise, the distribution of the model estimates for R are assumed to be symmetric.

These estimates are then combined using a random effects model allowing to account for the differences in the model's structure and not assuming that models share a common effect size. The random effects statistical model is described by:

$$y_i = \mu + \mu_i + \epsilon_i, \quad \mu_i \sim \mathcal{N}(0, \tau^2), \quad \epsilon_i \sim \mathcal{N}(0, v_i), \tag{7}$$

where the estimated mean for model i is denoted by y_i and standard error denoted by $s_i = \sqrt{v_i}$. The model was fitted to provide estimates for μ and τ which are the mean and standard deviation of the true effect size respectively. τ^2, the between model variance, was estimated using the restricted maximum likelihood method, and the confidence intervals of the mean true overall effect size are estimated using the standard Wald-type method. The models were equally weighted (see next section for more details) and a range of R rounded out to one decimal place, by using the lower and upper bounds respectively - to give a measure of uncertainty with acceptable confidence was used. Further details of other methods used for calculating the between group errors and CIs are provided in [50].

2.3 Sensitivity analysis

Two sensitivity analyses explored whether the combined R was impacted by variable weighting of the models within the ensemble and the size of the ensemble.

2.3.1 Exploring the impact of model weighting on the combined R

Firstly we explored the impact of the choice of model weighting on the consensus R. The combined estimate y is calculated from the true effect size of each model y_i. The true effect size can therefore be weighted. The simplest method, is that of equal weighting; and one that has been used to generate the published consensus R over 2020-2022. In this method, each model is assumed to have an equal contribution to the combined estimate under the assumption that all models are equally valid.

Another common alternative method of weighting, is that of inverse-variance weighting. Therefore, models with a high variance, i.e. those that are less certain, are penalised more than certain models. However, individual models have different methods of representing uncertainty, and a model that is more certain is not necessarily more likely to be accurate. Therefore this method is not applicable here.

An alternative method of model weighting is to group models by either their structure, or by the data to which they fit. For example, models that may have a different structure but use the same data form a sub-group as described in table 2. We explored the impact of this on the consensus R value by dividing the ensemble into sub-groups, so each sub-group represents a homogeneous set of models either according to structure or to the data to which they fit. Models within each sub-group

are equally weighted, and then the contributions from the sub-groups are equally weighted to give the overall combined estimate. This forms two purposes: firstly, a single data-stream or model structure will not have a larger weighting in the final combination, meaning that the combination is 'data-agnostic' or 'model-agnostic' and models such as EpiEstim, with a larger representation in the ensemble, do not bias the final estimates; secondly, it allows us to compare the difference in trends between admissions and case data and therefore learn about the epidemic dynamics by inspection. Similarly as for the equal weighting models method, a consensus R value was derived with this alternative, variable - weighting method as a range for April and September 2021. We present the results as rounded to two decimal places. However, the range is published to only one decimal place to avoid presenting a false sense of precision. The range is rounded out, as opposed to rounding to the nearest decimal place, as to increase the uncertainty instead of possibly reducing it.

2.3.2 Exploring the impact of ensemble size on the combined R

The models included within the ensemble varied throughout the pandemic; as new models were developed and introduced, some were phased out and others were updated in response to the changing epidemic. This could hypothetically result in inconsistent estimates through time. Furthermore, as the UKHSA moves from a 'response' to a 'business-as-usual' phase there is a need to reduce the resource dedicated to modelling COVID-19.

We investigate the implications of reducing the size of the ensemble on Nowcasts during the period January 2022 - May 2022. UKHSA models are labelled in table 2. We re-calculate the combined estimate from a reduced ensemble of models run internally in UKHSA and using equal weighting, and compare this to the published consensus R number in England.

2.4 R as an epidemic indicator

The R timeseries is a transform of the epidemic metrics such as cases incidence or hospitalisations. Hence we expect it would be statistically correlated to the epidemic metrics, but quantifying the degree of correlation with different metrics is interesting.

Here we explore the correlation between the consensus R as published on the UK government COVID-19 dashboard and the key public data sources relating to the COVID-19 pandemic; namely, cases, admissions and deaths. We calculate the Pearson's correlation coefficient of the published R number against the weekly rate of change of cases, admissions and deaths over the August 2021 - June 2022 period.

We note that it is important to differentiate between the published R number and the actual R number. Modellers provide R estimates for two weeks prior to combination. COVID-19 data has clear weekly seasonality due to reporting patterns. In order to adjust for weekly seasonality, each source of data is transformed to a centered weekly moving average. For each date that an R number is published, the slope of the data is calculated over a centered weekly window. We use the same length and position of windows over which to perform the analysis in order to ensure consistency, otherwise additional artificial lag would be introduced into the analysis. The R number is published as a range.

In order to simplify the correlation coefficient calculation, the correlation is calculated against the mid-point of the range. The correlation between the weekly rate of change in cases, admissions or deaths and the published R might have an inherent lag given the delays in the publication of R, or due to delays between cases, deaths and admissions. In order to investigate that, we examined

the correlation between R and the rate of change assuming various delays. This was done by shifting backwards the data window we used to calculate the weekly rate of change by 0 - 29 days and observing how the Pearson correlation coefficient and the associated p-values change with an increasing lag.

The R number represents a rate in change, and therefore we expect it to be linearly correlated with the rate of change of cases, admissions and deaths. Therefore, the Pearson's correlation coefficient is an appropriate choice of correlation.

3 Results

3.1 Generating a consensus R range in April and September 2021 using different weighting methods

Whisker plots of the 90th confidence intervals of R for each model are plotted alongside the resulting combinations from the different methods, and shown in fig. 1. The numerical values for the 90% confidence intervals for each weighting method are given in table 3.

Using the equal weighting method, and combining the R outcomes from the fourteen epidemiological models (7 SEIR -type, 2 agent-based models and 5 data-driven models) we generated combined R estimates of $[0.81, 0.93]$ in April 2021 and $[0.91, 1.07]$ in September 2021. These represent the 90% confidence interval that was published on the UK Government dashboard at the time.

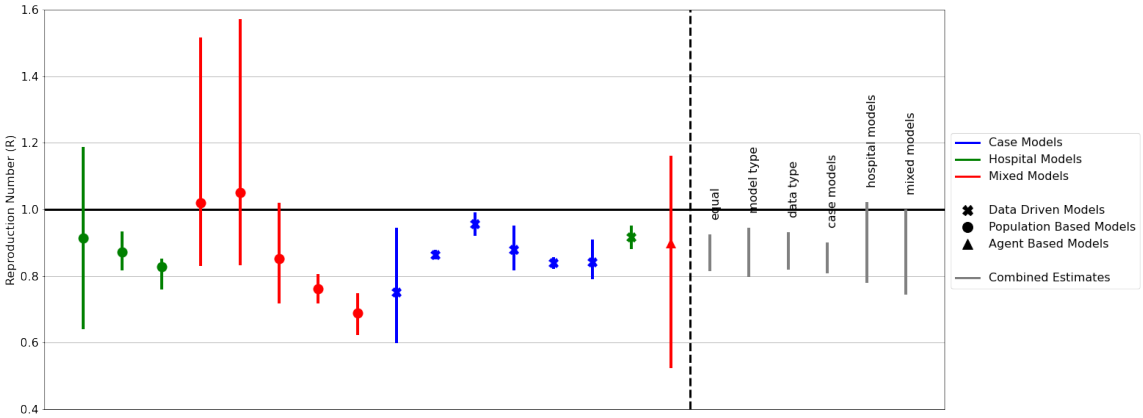
Using different weighting for the combination of models produces very similar combined R values at the two snapshots in time we studied: in April 2021 and in September 2021. Weighting by data resulted in R combination of $[0.82, 0.93]$ and $[0.86, 1.04]$ for the April 2021 and September 2021 estimates respectively. Weighting by model structure resulted in a combination of $[0.8, 0.94]$ and $[0.9, 1.07]$ for the April 2021 and the September 2021 estimates respectively.

Weighting methodology	April 6, 2021	September 14, 2021
Equal weighting	[0.81, 0.93]	[0.91, 1.07]
Weighting by data	[0.82, 0.93]	[0.86, 1.04]
Weighting by model	[0.80, 0.94]	[0.90, 1.07]
Equal weighting of case models	[0.81, 0.90]	[0.98, 1.15]
Equal weighting of hospital models	[0.78, 1.02]	[0.80, 0.93]
Equal weighting of mixed models	[0.74, 1.00]	[0.72, 1.12]

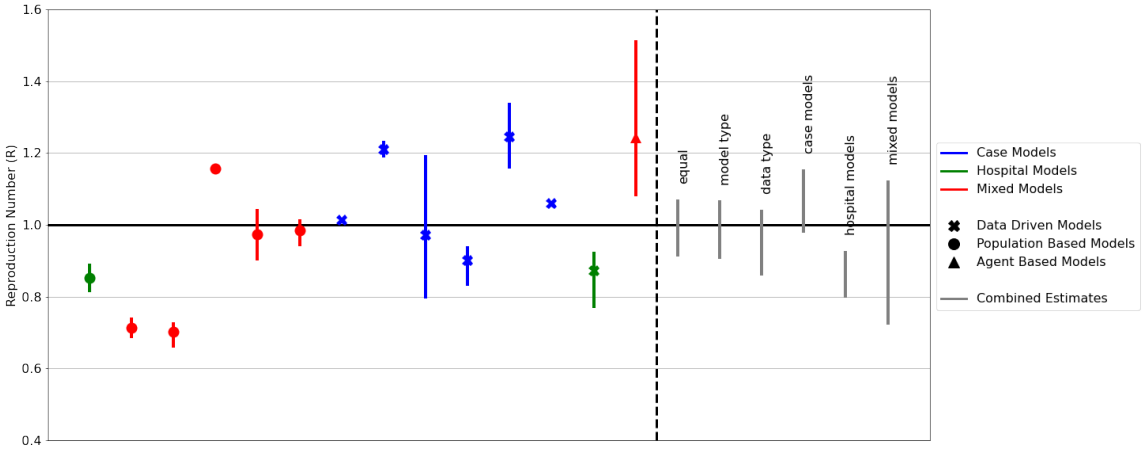
Table 3: 90% confidence intervals for combined R estimates using different weighting methods.

3.2 The effect of ensemble size on the combined estimate

Figure 2 compares what the un-rounded combined R number would be if the size of the ensemble of models was reduced only to include models run internally within UKHSA and mentioned in section 2.3.2. Our results show that the two combined R value time series are almost identical, particularly when rounding out to the nearest decimal place.



(a) April 21, 2021



(b) September 29, 2021

Figure 1: Error bars of the reproduction number R for the models included in the model ensemble on April 21, 2021 and September 29, 2021. The width of the error bar for each model represents the 90% confidence interval. To account for reporting delays and the time it takes to run the models, the R values that are reported are estimates for April 6, 2021 and September 14, 2021.

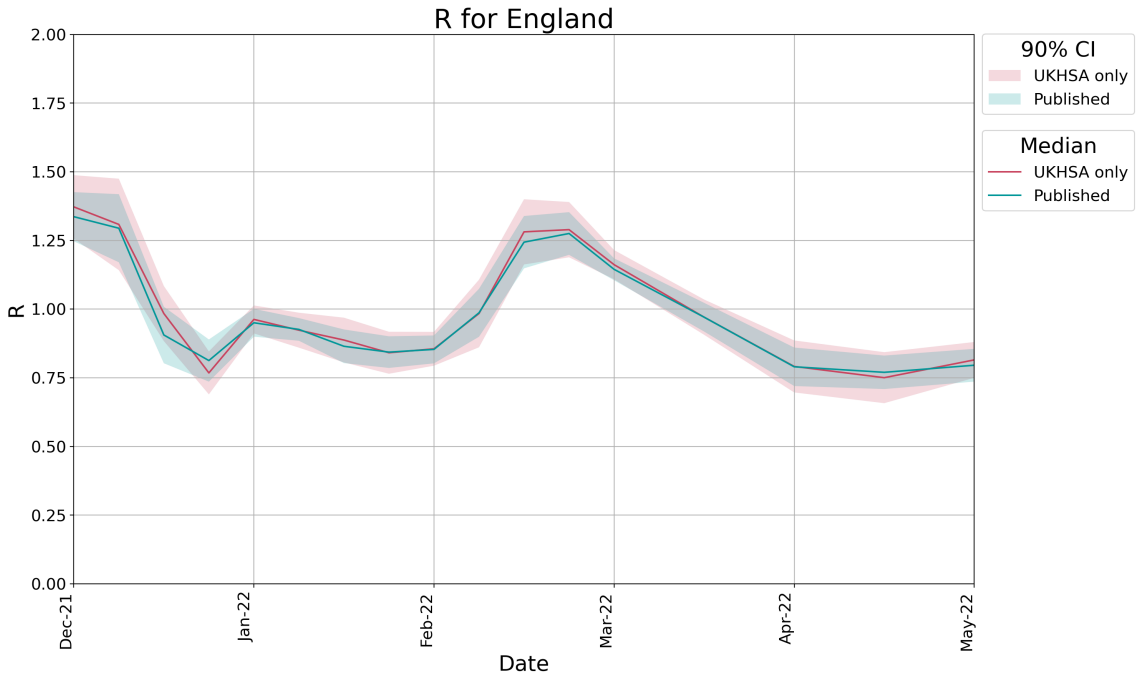


Figure 2: The combined R number in the period December 2021 - May 2022 for the published and reduced ensemble of models for England.

3.3 The combined R is a good epidemic indicator

We can visually inspect the relationship between cases, admissions and deaths with R in the top panels of fig. 3. Note that the R number is plotted against the day it was produced, not the day for which estimate was taken. This analysis shows that positive or negative R values (shown in red and blue respectively) do not necessarily occur when the number of cases and admissions is increasing or decreasing, suggesting there is a delay between the published R and these metrics.

The bottom panels of fig. 3 show the correlation between the timeseries of R with weekly change rate of cases, hospital admissions and deaths. The correlations have been calculated using different lags (0 to 29 days) between the published R number and the open source COVID-19 data. In each case, the left panels corresponds to the Pearson's correlation coefficient and the right panel corresponds to the corresponding p-value.

Our results suggest that there is a maximum and positive correlation between the published R number and the rate of change of data at varying lags. The maximum correlation is approximately 0.9 for cases and admissions and 0.8 for deaths and occurs at lags of 17, 14 and 2 days respectively. The correlation at each of these maxima is also found to be statistical significant. This means that we would expect that once the published R number is high, it is likely that cases were increasing roughly 17 days prior, hospitalisations roughly 14 days prior and deaths roughly 2 days prior. This lag is expected since the published R represents a two weeks prior estimated R. The actual R would instead lead case data once the delay for infections to present as cases and the generation time are taken into account.

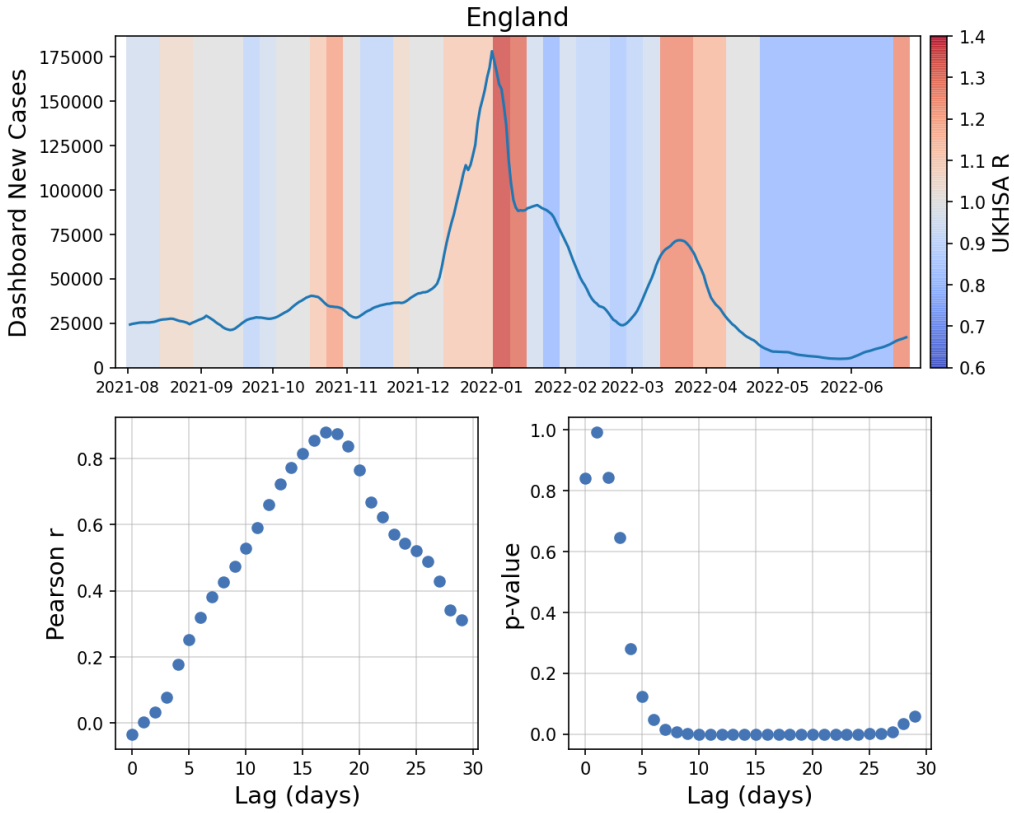
4 Summary and Conclusion

This study outlined a collaborative approach to generate the combined R value for England using a previously established combination method [35] and R estimates from fourteen separate epidemiological models. The combined R value was used to track the epidemic status over the COVID-19 epidemic in England, and was produced by SPI-M-O in 2020, and by the epidemiological ensemble group within UKHSA since early 2021.

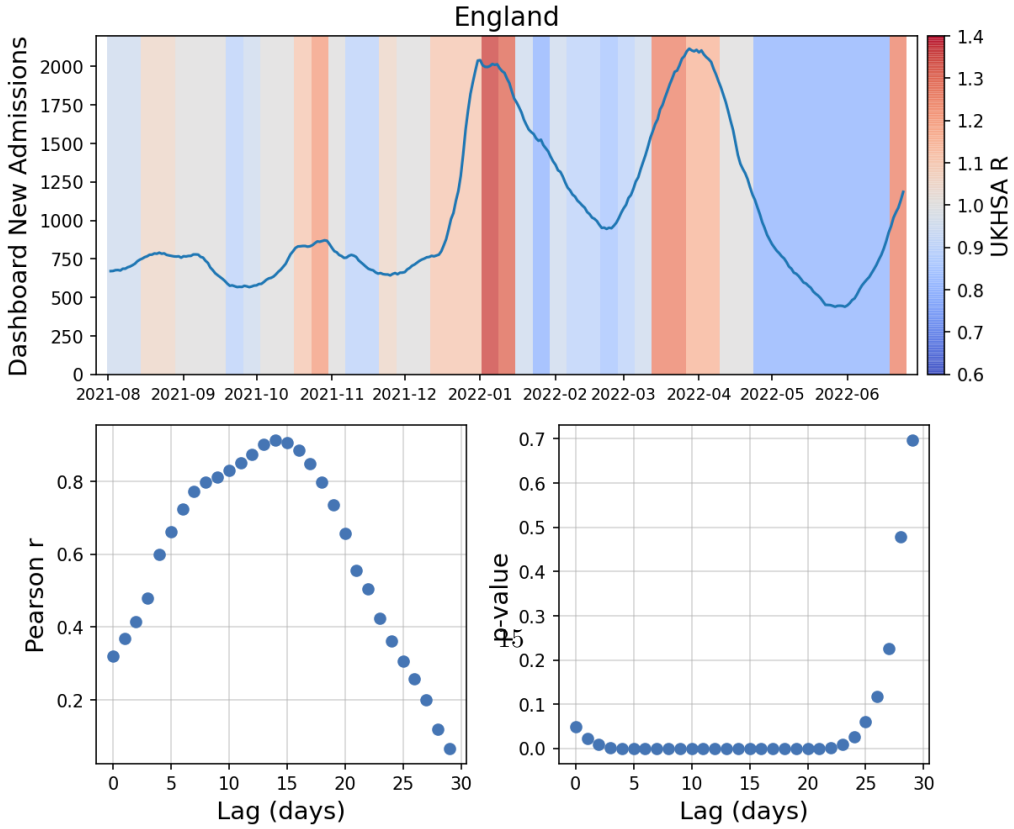
In this paper we describe the fourteen epidemiological models (7 SEIR-type, 2 agent-based models and 5 data-driven models) used to generate individual R values in England, highlighting their key structural characteristics and the data they use, as well as the method to individually derive an R value. We also outline the methodology developed in [35] of combining the individual R values to generate a combined consensus R value and illustrate this by generating the published R values of $[0.81, 0.93]$ on April 21, 2021 and $[0.91, 1.07]$ on September 29, 2021.

We showed that the combined R is robust to different weighting of the combining models' values. Weighting by data and weighting by model structure resulted in very similar combined R estimates. While for the generated value on April 21, 2021 the ensemble had homogeneous R values across the models, the individual values from September 29, 2021 had notable heterogeneity.

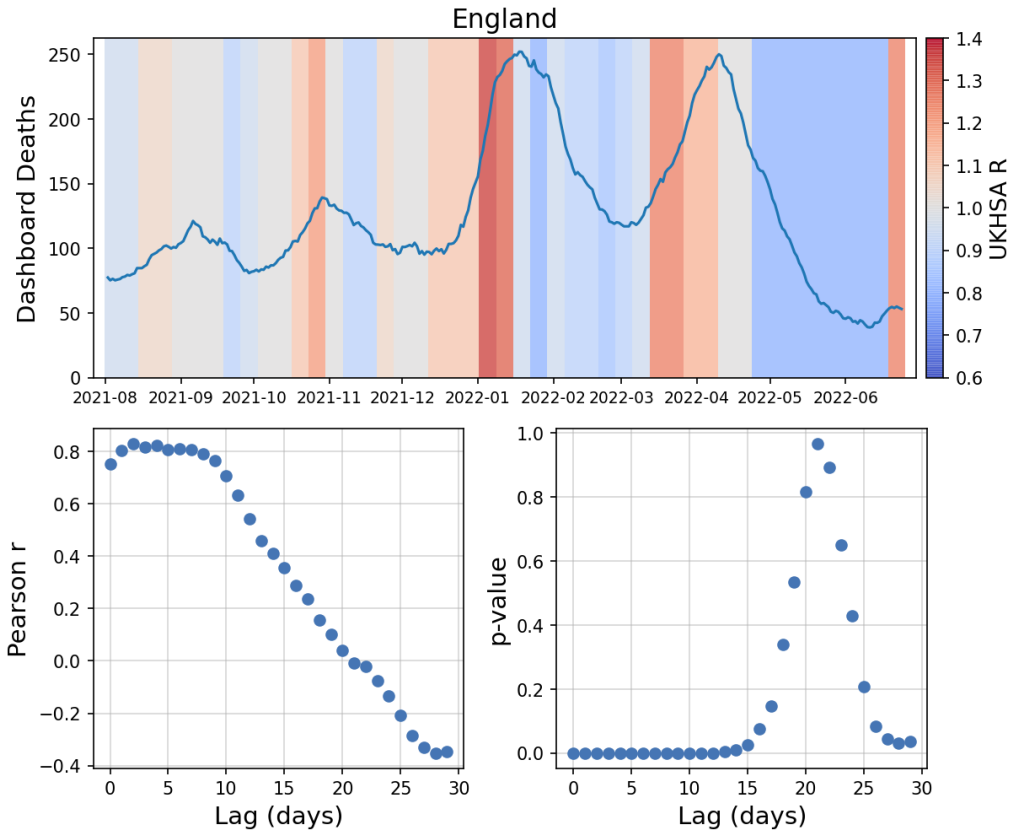
April 21, 2021 and September 29, 2021 are very different epidemic points in time. April 21, 2021 followed the third national lockdown in England imposed to control the transmission of the alpha variant [24]. Incidence and prevalence within the population were low and large scale vaccination against COVID-19 had only started to be rolled out, with roughly half the population having received a first dose, and only 8% having received a second dose. Against this mostly homogeneous immunity, susceptibility and vaccine backdrop, the assumptions within the models would have been similar, producing similar R values across models.



(a) Lag between new cases and the published R



(b) Lag between new admissions and the published R



(c) Lag between new deaths and the published R

Figure 3: Figures comparing the published R number to data published on the public government COVID-19 dashboard. The top panel of each sub-figures superimposes the time series of the weekly moving average of the dashboard data on top of the published R number for England. Where the shading is red, the most recent published R number was greater than 1. Where it is blue, the most recent published R number was less than 1. The bottom two panels plot the Pearson correlation coefficient and the associated p-value against the lag in days for the published R number against the dashboard data.

In September 2021, the immunity, susceptibility and vaccination levels were very different. There was a backdrop of population immunity from either vaccination or previous infection, with a large proportion of the population aged 12 and over either having received two doses of the vaccines or having been infected by the large Delta epidemic wave over the summer 2021. The COVID-19 case rate remained high with schools just returning, and this period preceded the arrival of the Omicron variant.

Different models would have made different assumptions on the impact of the large Delta wave on population immunity, and would have incorporated different assumptions around vaccination and social mixing associated with returning to school. All of these assumptions would impact individual R values, illustrated by the varying R values across models at this time.

Furthermore, different models fit to different data and this can generate different estimates. For example, the two LSHTM EpiNow2 models, one that fits to cases, and the second that fits to admissions, have vastly different R estimates. This difference is also reflected in the combinations from models that fit to cases (reporting a range of [0.98, 1.15]) and from models that fit to hospital data (reporting a range of [0.8, 0.93]). If we were only to use models that fit to cases, this would imply that the epidemic was increasing. However, models that fit to hospital data imply that the epidemic was decreasing. Models that fit to both, report a central estimate in between the two, albeit with larger uncertainty. Models that try to fit to multiple data streams tend to be more complex, mechanistic, or hierarchical, and hence more difficult to calibrate. Therefore, it is important that the ensemble features models that fit to a range of different data sources.

Reducing the size of the model ensemble, to include only models run internally within UKHSA, made negligible difference. We note that the ensemble of models on April 21, 2021 and September 29, 2021 are not identical, and the model ensemble has been changing over time. New models were introduced to the ensemble throughout the epidemic, models were omitted from or not submitted to the ensemble due to technical issues, such as calibration error or computer outage. Furthermore, in periods of change, such as the introduction of a new variant, some models had required extensive development work before re-inclusion into the ensemble. Exploring these aspects and whether having the same combination of models - or deriving an optimal combination - is important, but this is beyond the scope of this work, and will be explored further in future.

Finally, we have shown that the published R is an indicator of epidemic trends and a proxy for epidemic status as it is correlated with recorded epidemic outcomes, albeit with a delay of 17, 14 and 2 days with the rate of change in COVID-19 cases, hospital admissions and deaths. This delay is expected as it includes the two weeks between the date of the actual R estimate, and the published R.

Throughout the pandemic, R has been used to inform policy advisors, politicians and the general public of the current epidemic status. Our results suggest that this is justifiable as, once taking into account the lag, R is correlated with the key COVID-19 data streams (cases, hospitalisations and deaths). From a public health and health service capacity planning perspective, admissions and deaths are often the data streams of most interest. However, hospital admissions typically effect the elderly or more vulnerable, cases throughout the pandemic have more often been recorded among younger populations. Therefore, throughout the pandemic, the number of cases have also been used as an indicator of extra strain on the health service and the wider economy.

Bearing in mind these delays between R and the other epidemic outcomes and to mitigate uncertainty associated with nowcasting, since March 2021, the R value from each model was taken on a single day in time 2 weeks before the day on which models were combined. Incorporating these delays in R is important as not all models are always able to report estimates up to the day that

they are run as they do not possess the ability to forecast. For example, the simplest model, an application of *EpiEstim*, uses a delay distribution between infection and the observation to which it is fit, to back-calculate and infer the incidence time series. The R number is then estimated directly from the back-calculated time series for incidence. Therefore, the model is only able to provide estimates lagged to the order of the length of the delay distribution. Even where models are able to estimate current R numbers, due to the delay between infection and observation, the infections occurring today correspond to data that will be observed in the future, and hence, are in essence, projections. Prior to March 2021, SPI-M-O combined the most recent R numbers for each model, which may be estimates for different dates. SPI-M-O found that there was little difference in the combined estimates produced by using estimates from two weeks prior [23].

While combining multiple models, particularly in epidemic modelling, has proven to be very useful during the COVID-19 epidemic, there are lessons from this that should be considered in future.

It should be ensured that confidence intervals calculated by each of the models, represent the same sources of uncertainty. Do they capture the underlying uncertainty present in the data, the parametric uncertainty or the structural uncertainty? The forecast hub at the CDC treats models primarily as black boxes, though model details are published and models are assessed for accuracy, and there is not explicit treatment of the resulting uncertainty. For future pandemics, there should be a clear definition of uncertainty and what it should represent.

The combination method used to generate a consensus R is insensitive to the performance of individual models. Whereas for forecasts, model performance can be calculated by comparing model estimates with observed data, the R number is a latent variable and therefore is not observed. We rely on the expertise of modellers to ensure that models fit well to the data and make sound assumptions. In the future, developing an unbiased scoring method for individual models would help in ensuring the robustness and reliability of the individual models.

Finally, running an ensemble of models is resource intensive and relies on a significant amount of external expertise. If models are not to be treated as black boxes specialist expertise of academic groups continues to be required.

4.1 Summary

In this paper we have described the collaborative approach between academia and government to generate a combined R value in England from an ensemble of epidemiological models. An ensemble made up of models of different structures, that make different assumptions and fit to different data has clear benefits. It differs from a single model approach, and can avoid some of the complexities associated with trying to form a consensus view of the epidemic with a single model. No single model can perfectly encapsulate all the dynamics in an epidemic. The more data streams that a model fits to, and the more interventions it models, the more difficult it can be to fit the model. By using multiple models, uncertainty stemming from different sources can be captured. We show that the combined R is robust to different model weighting methods and ensemble size. We also show that the combined published R is correlated with the rate of change in COVID-19 cases, hospital admissions and deaths, albeit with a delay that is reflecting using R value from two weeks earlier. Our findings suggest R can be good indicator of future epidemic trends and justify its use as a proxy for epidemic status throughout the COVID-19 epidemic in England.

Acknowledgements

Estimates for R have been provided and combined as part of the UK wide respond to COVID-19. We would like to thank members of the SPI-M-O modelling group named as co-authors for continuing to provide estimates. In addition, we also acknowledge Yee-Whye Teh, Robert Hinch and Christophe Fraser (University of Oxford), Michelle Kendall and Louise Dyson (University of Warwick), Axel Gandy and Neil Ferguson (Imperial College London), Robert Moore, Conor Rosato and Simon Maskell (University of Liverpool), Leon Danon and Ellen Brooks-Pollock (University of Bristol), Sam Abbott and John Edmunds (London School of Hygiene and Tropical Medicine), Robert Shaw, Ewan Wakeman, Nicholas Groves-Kirkby and Seema Patel (NHS England) and Ross Burton (Cardiff University) for insightful ongoing discussions and input to this modelling work. We would in particular like to thank the SPI-M-O secretariat, and Graham Medley (LSHTM and SPI-M-O chair) for their support and continued discussions while conducting this analysis and while producing consensus R estimates.

References

[1] S Abbott, J Hellewell, RN Thompson, K Sherratt, HP Gibbs, NI Bosse, JD Munday, S Meakin, EL Doughty, JY Chun, YWD Chan, F Finger, P Campbell, A Endo, CAB Pearson, A Gimma, T Russell, null null, S Flasche, AJ Kucharski, RM Eggo, and S Funk. Estimating the time-varying reproduction number of sars-cov-2 using national and subnational case counts [version 1; peer review: awaiting peer review]. *Wellcome Open Research*, 5(112), 2020.

[2] Sam Abbott and Sebastian Funk. Estimating epidemiological quantities from repeated cross-sectional prevalence measurements. *medRxiv*, 2022.

[3] Sam Abbott, Joel Hellewell, Katharine Sherratt, Katelyn Gostic, Joe Hickson, Hamada S. Badr, Michael DeWitt, Robin Thompson, EpiForecasts, and Sebastian Funk. *EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters*, 2020.

[4] Graeme J. Ackland, James A. Ackland, Mario Antonioletti, and David J. Wallace. Fitting the reproduction number from uk coronavirus case data and why it is close to 1. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 380(2233):20210301, 2022.

[5] Roy Anderson, Christl Donnelly, Deirdre Hollingsworth, Matt Keeling, Carolin Vegvari, Rebecca Baggaley, and Rosie Maddren. Reproduction number (r) and growth rate (r) of the covid-19 epidemic in the uk: methods of estimation, data sources, causes of heterogeneity, and use as a guide in policy formulation, 2020.

[6] Marc Baguelin, Sangeeta Bhatia, Edward Knock, Lilith Whittles, Rich FitzJohn, John Lees an OJ Watson, Anne Cori, and Pablo Perez-Guzman. sircovid. Available at `https://mrc-ide.github.io/sircovid/`.

[7] Paul Birrell, Joshua Blake, Edwin van Leeuwen, Nick Gent, and Daneila De Angelis. Real-time nowcasting and forecasting of covid-19 dynamics in england: the first wave? *medRxiv*, 2020.

[8] Paul Birrell, Joshua Blake, Edwin van Leeuwen, MRC Biostatistics Unit COVID-19 Working Group, and Daniela De Angelis. Covid-19: nowcast and forecast. Available at `https://www.mrc-bsu.cam.ac.uk/now-casting/`.

[9] Johannes Bracher, Daniel Wolffram, Jannik Deuschel, Konstantin Görgen, Jakob L. Ketterer, Alexander Ullrich, Sam Abbott, Maria V. Barbarossa, Dimitris Bertsimas, Sangeeta Bhatia, Marcin Bodych, Nikos I. Bosse, Jan Pablo Burgard, Lauren Castro, Geoffrey Fairchild, Jochen Fiedler, Jan Fuhrmann, Sebastian Funk, Anna Gambin, Krzysztof Gogolewski, Stefan Heyder, Thomas Hotz, Yuri Kheifetz, Holger Kirsten, Tyll Krueger, Ekaterina Krymova, Neele Leithäuser, Michael L. Li, Jan H. Meinke, Błażej Miasojedow, Isaac J. Michaud, Jan Mohring, Pierre Nouvellet, Jedrzej M. Nowosielski, Tomasz Ozanski, Maciej Radwan, Franciszek Rakowski, Markus Scholz, Saksham Soni, Ajitesh Srivastava, Tilmann Gneiting, and Melanie Schienle. National and subnational short-term forecasting of covid-19 in germany and poland during early 2021. *Communications Medicine*, 2(1):136, Oct 2022.

[10] Ellen Brooks-Pollock, Leon Danon, Thibaut Jombart, and Lorenzo Pellis. Modelling that shaped the early covid-19 pandemic response in the uk. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1829):20210001, 2021.

[11] Joshua Burton. Detseirwithnb_mcmc. Available at `https://github.com/burtonjosh/DetSEIRwithNB_MCMC`.

[12] Carlos W Castillo-Garsow and Carlos Castillo-Chavez. A tour of the basic reproductive number and the next generation of researchers. In *Foundations for Undergraduate Research in Mathematics*, pages 87–124. Springer International Publishing, Cham, 2020.

[13] G. Chowell, R. Luo, K. Sun, K. Roosa, A. Tariq, and C. Viboud. Real-time forecasting of epidemic trajectories using computational dynamic ensembles. *Epidemics*, 30:100379, Mar 2020.

[14] Anne Cori, Neil M. Ferguson, Christophe Fraser, and Simon Cauchemez. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of Epidemiology*, 178(9):1505–1512, 09 2013.

[15] CovidStat. Contagion rate rt, 2022. Available at `https://covid19.infn.it/sommario/rt.html`.

[16] Estee Y Cramer, Evan L Ray, Velma K Lopez, Johannes Bracher, Andrea Brennen, Alvaro J Castro Rivadeneira, et al. Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the us. *medXriv*, 2021.

[17] Saskia den Boon, Mark Jit, Marc Brisson, Graham Medley, Philippe Beutels, Richard White, Stefan Flasche, T. Déirdre Hollingsworth, Tini Garske, Virginia E. Pitzer, and et al. Guidelines for multi-model comparisons of the impact of infectious disease interventions. *BMC Medicine*, 17(1):163, Aug 2019.

[18] Jonathan Dushoff and Sang Woo Park. Speed and strength of an epidemic intervention. *Proceedings of the Royal Society B: Biological Sciences*, 288(1947):20201556, 2021.

[19] Jeffrey W Eaton, Nicolas A Menzies, John Stover, Valentina Cambiano, Leonid Chindelevitch, Anne Cori, Jan A C Hontelez, Salal Humair, Cliff C Kerr, Daniel J Klein, and et al. Health benefits, costs, and cost-effectiveness of earlier eligibility for adult antiretroviral therapy and expanded treatment coverage: a combined analysis of 12 mathematical models. *The Lancet Global Health*, 2(1):e23–e34, Jan 2014.

[20] Rui Figueiredo, Kai Schröter, Alexander Weiss-Motz, Mario L. V. Martina, and Heidi Kreibich. Multi-model ensembles for assessment of flood losses and associated uncertainty. *Natural Hazards and Earth System Sciences*, 18(5):1297–1314, May 2018.

[21] Seth Flaxman, Swapnil Mishra, Axel Gandy, H. Juliette T. Unwin, Thomas A. Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W. Eaton, Mélodie Monod, Pablo N. Perez-Guzman, Nora Schmit, Lucia Cilloni, Kylie E. C. Ainslie, Marc Baguelin, Adhiratha Boonyasiri, Olivia Boyd, Lorenzo Cattarino, Laura V. Cooper, Zulma Cucunubá, Gina Cuomo-Dannenburg, Amy Dighe, Bimandra Djaafara, Ilaria Dorigatti, Sabine L. van Elsland, Richard G. FitzJohn, Katy A. M. Gaythorpe, Lily Geidelberg, Nicholas C. Grassly, William D. Green, Timothy Hallett, Arran Hamlet, Wes Hinsley, Ben Jeffrey, Edward Knock, Daniel J. Laydon, Gemma Nedjati-Gilani, Pierre Nouvellet, Kris V. Parag, Igor Siveroni, Hayley A. Thompson, Robert Verity, Erik Volz, Caroline E. Walters, Haowei Wang, Yuanrong Wang, Oliver J. Watson, Peter Winskill, Xiaoyue Xi, Patrick G. T. Walker, Azra C. Ghani, Christl A. Donnelly, Steven Riley, Michaela A. C. Vollmer, Neil M. Ferguson, Lucy C. Okell, Samir Bhatt, and Imperial College COVID-19 Response Team. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584(7820):257–261, Aug 2020.

[22] Scientific Advisory Group for Emergencies. Spi-m-o consensus statement on covid-19. `https://www.gov.uk/government/publications/spi-m-o-consensus-statement-on-covid-19-27-may-2020`.

[23] Scientific Advisory Group for Emergencies. Spi-m-o consensus statement on covid-19, 24th march 2021. `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/976328/S1164_SPI-M-O_Consensus_Statement.pdf`.

[24] Institute for Government. Timeline of uk government coronavirus lockdowns and restrictions. `https://www.instituteforgovernment.org.uk/charts/uk-government-coronavirus-lockdowns`. Accessed: 2022-11-11.

[25] Konstantine P Georgakakos, Dong-Jun Seo, Hoshin Gupta, John Schaake, and Michael B Butts. Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *Journal of Hydrology*, 298(1):222–241, Oct 2004.

[26] Katelyn M. Gostic, Lauren McGough, Edward B. Baskerville, Sam Abbott, Keya Joshi, Christine Tedijanto, Rebecca Kahn, Rene Niehus, James A. Hay, Pablo M. De Salazar, Joel Hellewell, Sophie Meakin, James D. Munday, Nikos I. Bosse, Katharine Sherrat, Robin N. Thompson, Laura F. White, Jana S. Huisman, Jérémie Scire, Sebastian Bonhoeffer, Tanja Stadler, Jacco Wallinga, Sebastian Funk, Marc Lipsitch, and Sarah Cobey. Practical considerations for measuring the effective reproductive number, rt. *PLOS Computational Biology*, 16(12):e1008409, Dec 2020.

[27] UK Government. Gov.uk coronavirus dashboard. `https://coronavirus.data.gov.uk/`. Accessed: 2022-11-11.

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

[29] Alex James, Rachelle N. Plank, Michael J.and Binny, Shaun C. Hendy, Audrey Lustig, and Nicholas Steyn. Auckland's august 2020 covid-19 outbreak – cabinet advice, 2020. Available at `https://www.covid19modelling.ac.nz/aucklands-august-outbreak/`.

[30] Matt J. Keeling, Louise Dyson, Glen Guyver-Fletcher, Alex Holmes, Malcolm G. Semple, Michael J. Tildesley, and Edward M. Hill. Fitting to the uk covid-19 outbreak, short-term forecasts and estimating the reproductive number. *medRxiv*, page 2020.08.04.20163782, Jan 2020.

[31] Michelle Kendall. Uk recent r estimate. Available at `https://github.com/MichelleKendall/UK_recent_R_estimate`.

[32] Michelle Kendall, Luke Milsom, Lucie Abeler-Dörner, Chris Wymant, Luca Ferretti, Mark Briers, Chris Holmes, David Bonsall, Johannes Abeler, and Christophe Fraser. Epidemiological changes on the isle of wight after the launch of the NHS test and trace programme: a preliminary analysis. *Lancet Digit Health*, 2(12):e658–e666, October 2020.

[33] Cliff C. Kerr, Robyn M. Stuart, Dina Mistry, Romesh G. Abeysuriya, Katherine Rosenfeld, Gregory R. Hart, Rafael C. Núñez, Jamie A. Cohen, Prashanth Selvaraj, Brittany Hagedorn, Lauren George, Michał Jastrzębski, Amanda Izzo, Greer Fowler, Anna Palmer, Dominic Delport, Nick Scott, Sherrie Kelly, Caroline S. Bennette, Bradley Wagner, Stewart Chang, Assaf P. Oron, Edward Wenger, Jasmina Panovska-Griffiths, Michael Famulare, and Daniel J. Klein. Covasim: an agent-based model of covid-19 dynamics and interventions. *medRxiv*, 2021.

[34] T. N. Krishnamurti, C. M. Kishtawal, Zhan Zhang, Timothy LaRow, David Bachiochi, Eric Williford, Sulochana Gadgil, and Sajani Surendran. Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate*, 13(23):4196–4216, Dec 2000.

[35] Thomas Maishman, Stephanie Schaap, Daniel S Silk, Sarah J Nevitt, David C Woods, and Veronica E Bowman. Statistical methods used to combine the effective reproduction number, [formula: see text], and other related measures of COVID-19 in the UK. *Stat Methods Med Res*, 31(9):1757–1777, July 2022.

[36] Gerald A. Meehl, Curt Covey, Bryant McAvaney, Mojib Latif, and Ronald J. Stouffer. Meeting summaries: Overview of the coupled model intercomparison project. *Bulletin of the American Meteorological Society*, 86(1):89–96, 2005.

[37] Robert E. Moore, Conor Rosato, and Simon Maskell. Refining epidemiological forecasts with simple scoring rules. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 380(2233):20210305, 2022.

[38] Norwegian Institute of Public Health. Weekly reports for coronavirus and covid-19, 2022. Available at specifically,translatedinfoinhere:https://www.fhi.no/contentassets/8a971e7b0a3c4a06bdbf381ab52e6157/vedlegg/2022/ukerapport-uke-12-21.03---27.03.22.pdf.

[39] Christopher E. Overton, Lorenzo Pellis, Helena B. Stage, Francesca Scarabel, Joshua Burton, Christophe Fraser, Ian Hall, Thomas A. House, Chris Jewell, Anel Nurtay, Filippo Pagani, and Katrina A. Lythgoe. Epibeds: Data informed modelling of the covid-19 hospital burden in england. 2021.

[40] Evan L Ray, Nutcha Wattanachit, Jarad Niemi, Abdul Hannan Kanji, Katie House, Estee Y Cramer, Johannes Bracher, Andrew Zheng, Teresa K Yamana, Xinyue Xiong, Spencer Woody, Yuanjia Wang, Lily Wang, Robert L Walraven, Vishal Tomar, Katharine Sherratt, Daniel Sheldon, Robert C Reiner, B. Aditya Prakash, Dave Osthus, Michael Lingzhi Li, Elizabeth C Lee, Ugur Koyluoglu, Pinar Keskinocak, Youyang Gu, Quanquan Gu, Glover E. George, Guido España, Sabrina Corsetti, Jagpreet Chhatwal, Sean Cavany, Hannah Biegel, Michal Ben-Nun, Jo Walker, Rachel Slayton, Velma Lopez, Matthew Biggerstaff, Michael A Johansson, and Nicholas G Reich. Ensemble forecasts of coronavirus disease 2019 (covid-19) in the u.s. *medRxiv*, 2020.

[41] Evan L Ray, Nutcha Wattanachit, Jarad Niemi, Abdul Hannan Kanji, Katie House, Estee Y Cramer, Johannes Bracher, Andrew Zheng, Teresa K Yamana, Xinyue Xiong, Spencer Woody, Yuanjia Wang, Lily Wang, Robert L Walraven, Vishal Tomar, Katharine Sherratt, Daniel Sheldon, Robert C Reiner, B. Aditya Prakash, Dave Osthus, Michael Lingzhi Li, Elizabeth C Lee, Ugur Koyluoglu, Pinar Keskinocak, Youyang Gu, Quanquan Gu, Glover E. George, Guido España, Sabrina Corsetti, Jagpreet Chhatwal, Sean Cavany, Hannah Biegel, Michal Ben-Nun, Jo Walker, Rachel Slayton, Velma Lopez, Matthew Biggerstaff, Michael A Johansson, and Nicholas G Reich. Ensemble forecasts of coronavirus disease 2019 (covid-19) in the u.s. *medRxiv*, 2020.

[42] Nicholas G. Reich, Craig J. McGowan, Teresa K. Yamana, Abhinav Tushar, Evan L. Ray, Dave Osthus, Sasikiran Kandula, Logan C. Brooks, Willow Crawford-Crudell, Graham Casey Gibson, and et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the u.s. *PLOS Computational Biology*, 15(11):e1007486, Nov 2019.

[43] RISKAWARE. Crystalcast : Advanced disease forecasting to predict outbreaks and anticipate population impact. https://www.riskaware.co.uk/wp-content/uploads/BioAware-CrystalCast-Product-Sheet.pdf.

[44] Robert Koch Institut. Epidemiologisches bulletin, 2020. Available at https://www.rki.de/DE/Content/Infekt/EpidBull/Archiv/2020/Ausgaben/17_20.pdf?__blob=publicationFile.

[45] Kimberlyn Roosa, Amna Tariq, Ping Yan, James M. Hyman, and Gerardo Chowell. Multi-model forecasts of the ongoing ebola epidemic in the democratic republic of congo, march–october 2019. *Journal of The Royal Society Interface*, 17(169):20200447, Aug 2020.

[46] James A. Scott, Axel Gandy, Swapnil Mishra, Juliette Unwin, Seth Flaxman, and Samir Bhatt. epidemia: Modeling of epidemics using hierarchical bayesian models, 2020. R package version 1.0.0.

[47] Katriona Shea, Rebecca K. Borchering, William J. M. Probert, Emily Howerton, Tiffany L. Bogich, Shouli Li, Willem G. van Panhuis, Cecile Viboud, Ricardo Aguás, Artur Belov, and et al. Covid-19 reopening strategies at the county level in the face of uncertainty: Multiple models for outbreak decision support. *medRxiv*, page 2020.11.03.20225409, Nov 2020.

[48] K. Sherratt, H. Gruson, R. Grah, H. Johnson, R. Niehus, B. Prasse, F. Sandman, J. Deuschel, D. Wolffram, S. Abbott, A. Ullrich, G. Gibson, EL. Ray, NG. Reich, D. Sheldon, Y. Wang, N. Wattanachit, L. Wang, J. Trnka, G. Obozinski, T. Sun, D. Thanou, L. Pottier, E. Krymova, MV. Barbarossa, N. Leithäuser, J. Mohring, J. Schneider, J. Wlazlo, J. Fuhrmann, B. Lange, I. Rodiah, P. Baccam, H. Gurung, S. Stage, B. Suchoski, J. Budzinski, R. Walraven, I. Villanueva, V. Tucek, M. Šmíd, M. Zajícek, C. Pérez Álvarez, B. Reina, NI. Bosse, S. Meakin, P. Alaimo Di Loro, A. Maruotti, V. Eclerová, A. Kraus, D. Kraus, L. Pribylova, B. Dimitris, ML. Li, S. Saksham, J. Dehning, S. Mohr, V. Priesemann, G. Redlarski, B. Bejar, G. Ardenghi, N. Parolini, G. Ziarelli, W. Bock, S. Heyder, T. Hotz, D. E. Singh, M. Guzman-Merino, JL. Aznarte, D. Moriña, S. Alonso, E. Álvarez, D. López, C. Prats, JP. Burgard, A. Rodloff, T. Zimmermann, A. Kuhlmann, J. Zibert, F. Pennoni, F. Divino, M. Català, G. Lovison, P. Giudici, B. Tarantino, F. Bartolucci, G. Jona Lasinio, M. Mingione, A. Farcomeni, A. Srivastava, P. Montero-Manso, A. Adiga, B. Hurt, B. Lewis, M. Marathe, P. Porebski, S. Venkatramanan, R. Bartczuk, F. Dreger, A. Gambin, K. Gogolewski, M. Gruziel-Slomka, B. Krupa, A. Moszynski, K. Niedzielewski, J. Nowosielski, M. Radwan, F. Rakowski, M. Semeniuk, E. Szczurek, J. Zielinski, J. Kisielewski, B. Pabjan, K. Holger, Y. Kheifetz, M. Scholz, M. Bodych, M. Filinski, R. Idzikowski, T. Krueger, T. Ozanski, J. Bracher, and S. Funk. Predictive performance of multi-model ensemble forecasts of covid-19 across european nations. *medRxiv*, 2022.

[49] Yee Whye Teh, Avishkar Bhoopchand, Peter Diggle, Bryn Elesedy, Bobby He, Michael Hutchinson, Ulrich Paquet, Sang Woo Park, Jonathan Read, Nenad Tomasev, and Sheheryar Zaidi. Uk local covid map.

[50] Areti Angeliki Veroniki, Dan Jackson, Wolfgang Viechtbauer, Ralf Bender, Jack Bowden, Guido Knapp, Oliver Kuss, Julian P. T. Higgins, Dean Langan, and Georgia Salanti. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods*, 7(1):55–79, Mar 2016. 26332144[pmid].

[51] Harald Vöhringer, Matthew Sinnott, Theo Sanderson, Roberto Amato, Inigo Martincorena, Dominic Kwiatkowski, Jeffrey C. Barrett, and Moritz Gerstung. Genomic surveillance: collection of statistical algorithms to monitor local prevalence of viral lineages.

[52] Yawen Xiao, Jun Wu, Zongli Lin, and Xiaodong Zhao. A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine*, 153:1–9, Jan 2018.