Collaborative outbreak modelling for decision support: evaluating trade-offs from multi-model combination

Katharine Sherratt

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy of the University of London

June 2024

Centre for Mathematical Modelling of Infectious Diseases

Department of Infectious Disease Epidemiology and Dynamics

Faculty of Epidemiology and Population Health

London School of Hygiene and Tropical Medicine

Declaration

I, Katharine Sherratt, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Name: Katharine Sherratt

Student ID Number: 170369

Date: 1 June 2024

Abstract

Infectious disease modelling is a useful tool for supporting outbreak control, offering to interpret the complex uncertainty of epidemiological dynamics. This uncertainty allows for many approaches, choices, and interpretations during modelling work, producing a wide variety of modelling results. Multi-model collaborations create comparability across this diversity, and the opportunity for evidence synthesis. This often includes a quantitative combination of numerical model results.

This thesis evaluates collaborative modelling work during the COVID-19 response in the UK and Europe. Four papers draw from the UK's Scientific Pandemic Infections group on Modelling (SPI-M), and the European COVID-19 Forecast and Scenario Hubs. First, I found that outbreak detection may be confounded by combining estimates of the reproduction number, aggregating over relevant heterogeneity. Next, I evaluated ensemble projections of both short- and long-term COVID-19 incidence, characterising predictive performance, representation of uncertainty, and policy relevance. I then identified tensions in the structural sustainability of modelling work for outbreak response.

A thematic analysis draws out shared challenges in collaborative outbreak modelling. Modelling collaborations are vulnerable to sampling biases that may limit the validity of multi-model combinations, while also facing varied and competing stakeholder needs. Meanwhile, collaborative decision support may face a fundamental trade-off between offering consensus versus context: creating a single evidence synthesis risks losing insight into heterogeneous epidemic dynamics. A further trade-off challenges collaboration with capacity: multi-model combinations depend on consistent model components, despite collaborators' constrained capacity during emergency response.

Selecting an appropriate strategy to resolve these tensions likely depends on the purpose, timing, and scale of outbreak decision-making. Future work should explore the validity of epidemiological inference from multi-model combinations, and clarify both capacity constraints and stakeholder needs at the science-policy interface.

Acknowledgements

Contents

List of tables

Analytic commentary

Background

Collaborative outbreak modelling

Computational modelling is a useful tool for interpreting the complex uncertainty of infectious disease dynamics. It becomes particularly relevant during disease outbreaks, with limited data available and when modelling may be the only such tool at hand [1]. Modelling can be used to support a variety of outbreak response needs. These might include characterising epidemiological processes, predicting future trends, or exploring different scenarios for intervention [2]. At the same time, modellers face many layers of uncertainty: both in complex epidemic dynamics, and in the multiplicity of methods for modelling them [3]. This results in a diversity of models which are worthy of comparison. This comparison may in turn offer the opportunity for multi-model combination.

Modelling work draws across multiple forms of knowledge and combines this with an understanding of underlying mechanisms into a single modelling framework. This process requires many uncertain choices among multiple plausible and interrelated options. Firstly, the dynamics of infectious disease transmission are fundamentally uncertain because they are typically stochastic, non-linear, and remain largely unobserved in data [4]. Furthermore, there are now many available computational modelling methods to estimate these uncertainties [5].

For example, modellers may vary in the extent to which they rely on inevitably lagged and biased data, or incorporate current understandings of natural history of infection or immunity [6]; in how this understanding is parameterised and calibrated to data [7]; or in which tools are used to implement a particular model strategy [8]. As a result, modelling approaches can vary widely, and different choices made by modellers to address uncertainty may lead to different results, even if they are equally valid. Rather than competing to provide a single best option, each model may be thought of as adding another perspective [9].

Faced with this diversity, working in collaboration allows modellers to compare among these approaches, choices, and results. Collaborations bring modellers together around one or more epidemiological targets, and collate multiple model results in a standardised format. This offers comparability across the diversity of modelling work. Modelling collaborations may aim to enable expert elicitation among modellers [10], clarify the extent and policy relevance of uncertainty [11], and provide a synthesis of modelling evidence [12]. This could include producing a

qualitative narrative synthesis, or directly processing model outputs into a quantitative multi-model combination, such as a summary range or an ensemble projection.

Opportunities from collaborative modelling

In the context of a rapidly evolving outbreak, working in collaboration may be essential for supporting modelling responders. Collaborations can allow modellers to better access and interpret context-dependent data and decision needs, allowing for model development in response to rapidly evolving epidemic dynamics and insights [13–15], or identify data issues and biases [1]. As well as improving quality, collaborative work can support broader modelling capacity building [16,17], while open collaborations specifically promote a transparent approach to epidemic modelling [11]. This is particularly important where modelling work has traditionally only been reported through the academic publication process, which is too slow for policy use [18]. At the same time, both modelling methods and code remain underreported, making it difficult to assess either quality or comparability [19,20].

At the same time, modelling collaborations typically define themselves as increasing the relevance of modelling to outbreak decision making [17,21,22]. Model stakeholders typically include public health decision makers amid a wider landscape of outbreak response work [22]. Model stakeholders may draw on different types of modelling evidence over time and can themselves inform model development. Best practice guidance for individual modellers suggests decision support should include a clear understanding of stakeholder needs and a timely and ongoing cycle of feedback between model development, interpretation, and use [22–24]. A collaborative structure can centralise this activity. This may support modellers otherwise unable to build relationships with model stakeholders, reduce research waste, and increase both trust and diversity in the modelling used for decision support [22].

However, communicating the multi-layered uncertainties between and within models is highly challenging [23,25], and potential users of modelling evidence can be overwhelmed by its diversity [9]. For example, using different methods to confront the same policy question can create contrasting results, with potentially conflicting policy implications [26]. Stakeholders in modelling evidence may deal with this by relying on a single model, underestimating the full extent of uncertainty and potentially missing decision-relevant information [10]. This creates an important role for well structured comparison and communication of modelling evidence [27,28].

In approaching this, collaborative projects often centre around model combinations as offering a synthesis of modelling evidence. Specifically, quantitative combinations are justified by arguing that the combined result from multiple models produces a more robust characterisation of the modelling target than any single model alone [11,27,29]. This argument may work by analogy to the greater accuracy of forecast prediction using ensemble models [30]. Repeated studies across fields have shown the increased accuracy of predicting observed data when multiple model predictions are combined [31,32], including in outbreak forecasting across multiple pathogens, outbreaks, and years (e.g. [13,33,34]). An alternative analogy is to meta-analysis [35]. In this approach, each model is viewed as giving some information about an effect and statistical combination across comparable model outputs gives a more precise, or less biased, effect estimate.

Modelling collaborations played a prominent role in the research and policy environment of the COVID-19 response, while building on a history of collaborative responses to infectious disease outbreaks. These have included endemic pathogens, such as dengue [36] or influenza [37], and emerging epidemics, such as Foot & Mouth Disease in 2001 [38], H1N1 pandemic influenza in 2009 [1], or in post-response evaluation of Ebola after 2014 [13,39]. Collaborative modelling became highly salient during COVID-19. These included scientific advisory groups for national policy, for example in the UK [40] and Europe [41]; international comparisons based on region or resource context [16,28,42]; or open source collaborations, including for real-time outbreak monitoring [43], short-term prediction [34,44], or comparing scenarios for intervention [45].

Similar collaborations have been highlighted as a way forward for the field of policy-relevant infectious disease modelling [11], and collaborations are now targets for substantial investment. At an international level, the World Health Organisation (WHO) has invested in a Hub for Pandemic and Epidemic Intelligence [46]. At a regional scale, the European Centre for Disease Prevention and Control (ECDC) is extending cross-European collaborations built during COVID-19 to seasonal respiratory infections [47]; while the US CDC has launched a long-term Infectious Disease Modeling and Analytics Initiative [17]. As a global open-source initiative, the "hubverse" [24] is developing a suite of publicly reusable infrastructure to support the creation of new forecast and scenario modelling hub collaborations.

Challenges to collaborative modelling

While opportunities are well recognised, there has been relatively little analysis of the challenges of model comparison and combination work. It is likely that collaborative work involves both challenges and trade-offs. Evaluating evidence across multiple models requires considerable care [27]: for example, quantitatively similar modelling outcomes could arise from very different underlying assumptions about epidemiological mechanisms, invalidating their comparison [6]. At the same time, collaborative structures may raise new operational and ethical concerns compared to individual modelling for policy. These should be recognised and justified.

This work and commentary attempts to identify some of these challenges, focussing on the credibility of model combinations while expanding to the relevance and legitimacy of collaborative structures. The work presented here draws from three contrasting outbreak modelling collaborations responding to COVID-19 in the UK and Europe.

The Scientific Pandemic Influenza Group on Modelling-Operational (SPI-M)

In January 2020, the UK's Scientific Advisory Group for Emergencies (SAGE) was activated to support the government response to COVID-19. SAGE drew on multiple forms of scientific advice, including modelling work from its SPI-M sub-group, formally established in response to 2009 H1N1 influenza. From January 2020 through 2022, SPI-M met at least weekly. It is chaired by policy and academic co-leads with modellers invited to contribute. Membership varied, but expanded to around 70 contributors over time. Much early work included estimating key epidemiological parameters, with weekly consensus estimates of the growth rate and the reproduction number. SPI-M also produced short and medium term projections and responded to specific policy questions [40].

The European COVID-19 Forecast Hub

From late 2020, the European Centre for Disease Control and Prevention (ECDC) commissioned our team to lead scientific and technical development of a European COVID-19 Forecast Hub. The Hub collated and combined 1-4 week forecasts of COVID-19 for cases, deaths, and hospital admissions, for 32 countries across Europe. Up to 48 independent modelling teams contributed weekly forecasts, although this varied over time and by forecast target, and we held a weekly open exchange via videoconference among all Hub participants, including the ECDC. Forecasts were combined into a single ensemble projection, and all

forecasts were published in an publicly accessible online platform for forecast comparison, visualisation, and evaluation. This structure drew on similar Forecast Hubs established for the US [34], and Germany and Poland [48].

The European COVID-19 Scenario Hub

From March 2022, the ECDC similarly commissioned the development of the European COVID-19 Scenario Hub. This supported the ECDC to explore European policy interventions, such as targeted vaccine distribution. The Scenario Hub attempted to adopt a more intensively collaborative style of expert elicitation, modelled on similar work in the US [49]. In setting scenarios, we aimed to explore the relative impact of varying policy options, while expressing the wide range of plausible epidemiological uncertainty driving COVID-19. Twelve teams collaborated to co-create six sets of scenarios. Projections for each scenario were visualised and compared alongside a narrative interpretation of policy implications. This focussed on highlighting differences between models, such as between underlying assumptions, and interpreting across sources of uncertainty.

Research objectives

This PhD by publication addresses the overarching aim of evaluating collaborative outbreak modelling, focussing on tensions and trade-offs in collaborative model combination. In this thesis, these are associated with four papers representing four phases of outbreak response: from outbreak investigation, short-term forecasting, long-term scenario planning, to post-outbreak capacity building.

The remainder of this work presents a critical commentary discussing four selected papers in terms of their development and original contribution, placing this in the context of the wider field and further work. A concluding evaluation draws out common themes across the work.

Overview of the thesis

Outbreak investigation depends on accurate understanding of current epidemic transmission, but this is difficult given lagged and biased real-time data. This is often addressed by estimating the time-varying reproduction number (R_t), a key contribution of SPI-M. In a first piece of work, I identified the potential for comparing estimates of R_t to provide insight into policy-relevant transmission dynamics, while suggesting the potential for bias in comparisons between multiple

models. A further discussion interprets this result in light of methods and context for real-time R_t estimation, and SPI-M's use of comparison and combination of R_t estimates for policy.

Short-term forecasting is useful for operational planning and situational awareness and may be sought after by a diverse range of users and decision makers. The second research work evaluated the ensemble combination of forecasting models contributed to the European COVID-19 Forecast Hub, finding increased predictive accuracy of this combination compared to the set of individual component models. This is discussed in light of challenges interpreting an ensemble made against multiple epidemiological targets with a changing and unclear sample of models.

In contrast, decision makers focussed on long term planning require an understanding of options for policy intervention, for which scenario modelling can be used. A third paper described identifying policy needs and producing appropriate model combinations to support them, given high within- and between-model uncertainty. The commentary sets this in context of how collaborations may respond to stakeholder needs to control, expand, and contract representations of uncertainty.

Finally, post-outbreak evaluations often suggest the need for building modelling capacity. A fourth piece of work explored the longer term sustainability of collaborative outbreak response work, bringing together a diverse range of modellers' experiences contributing to SPI-M. This work is discussed in the context of continued calls for improvements across the field of outbreak modelling, and the potential for bias in the process of post-outbreak evaluation.

A concluding evaluation draws out common themes throughout this work. This focuses on challenges and trade-offs in collaborative outbreak modelling and multi-model combination, before suggesting opportunities to both explore and resolve tensions in collaborative outbreak modelling.

Critical commentary

Contextualising comparisons of the real-time reproduction number

This commentary addresses the paper: "Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of COVID-19 in England".

Development of the work

The time-varying reproduction number (R_t) is a summary measure of infectious disease transmission over time, representing the average number of expected secondary cases from a single infectious individual under current conditions [50]. R_t estimates are useful for characterising and tracking the underlying epidemic transmission process, in contrast to lagged and biased trends in epidemiological data [51]. R_t became a central focus of the UK's response to COVID-19 [52], and a key task of the modelling advisory group SPI-M, published weekly as a "consensus range" [53]. Meanwhile, R_t estimates are highly variable due to both heterogeneous transmission, and a diversity of methods for estimation [50].

This work originated in real-time response providing R_t estimates to SPI-M from March 2020. We noted that estimates showed differing results based on data source, both within our own model and across models contributing to SPI-M [54]. We used three sources of data to estimate R_t: COVID-19 positive test results; hospitalisations; and deaths. We estimated separate delay distributions for these data sources, and used a renewal equation based approach to estimate R_t [55]. Assuming homogeneous transmission, the resulting R_t estimates should be essentially the same as they track back to the same data generating process.

We hypothesised that the observed differences in R_t between data sources were a meaningful indication of variable transmission patterns between different source populations for surveillance data. While data from test-positive cases represented all those presenting at a test centre, hospitalisation and death data drew from populations at risk of severe disease. Crucially, this risk was not homogeneous distributed across the population, for example with age as a key determinant of severity. We suggested that when each data source was biased in this way, tracking differences in R_t could be a real-time indicator of differences in subpopulation transmission, for example identifying outbreaks among residential care settings. This also challenged SPI-M's contemporary use of a model combination, suggesting that this was unrepresentative of any underlying transmission process.

Individual contribution

Development of the Rt estimation and forecasting procedure using EpiNow and EpiNow2 was led by Sam Abbott and Sebastian Funk. I initially contributed data cleaning, processing, and analysis for UK and global surveillance data. I also supported running the end to end Rt estimation pipeline for regular SPI-M submission, and contributed to model development with testing, evaluation, and documentation. In May 2020, I worked together with Sam Abbott and Sebastian Funk on the initial hypothesis of exploring differences in Rt estimation by data source. I led work to develop this idea into this analysis, including developing code to run the Rt estimation pipeline using each data source and visualising the comparative results. I also led the write up of this into a short briefing note, presented to SPI-M first in early June and in three further iterations.

I then led work to develop this into a paper. I developed the original work by using public data for estimating delay distributions, in order for all work to be presented in the public domain. I also designed and conducted further analysis to quantify the comparisons of Rt estimates, for example assessing peaks and wave durations. The specific further hypotheses for retrospectively understanding differences between Rt estimates over time were discussed jointly with Sam Abbott and Sebastian Funk, in turn drawing on wider discussions among SPI-M. I led work to source relevant data and formally add comparisons with these additional data sources, for example data on deaths in residential facilities. I led the initial and subsequent drafts of the paper. Sam Abbott and Sebastian Funk provided supervision and review before submission for peer review. All code contributions are documented on Github: https://github.com/epiforecasts/rt-comparison-uk-public.

Themes and context

Comparing real-time Rt estimates

By building on carefully standardised real-time R_t estimates, this paper suggested the role of comparison across multiple estimates to provide additional epidemiological insights. Exploring comparisons between models was a crucial element of the SPI-M collaboration [12]. However, comparisons of R_t estimates can be difficult to interpret due to the number and diversity of analytical choices involved [56]. Methodological choices include, among many others, the definition of R_t itself, in whether it estimates forward-looking or backward (instantaneous) transmission from an infection [50]; the timescale over which R_t is estimated; data

pre-processing; or the parameterisation of the generation time, incubation period, and reporting delays [56,57]. Many of these choices varied among SPI-M models of R_t [53].

Making valid comparisons among R_t estimates is a particular challenge during real-time response because estimation must account for multiple biases and lags in real-time data. For example, in this work we specified separate delay distributions to account for the lag between transmission and reporting a test-positive case, versus hospitalisation or death. Misspecifying these delays could have explained some of the differences we observed in R_t between data sources.

The extent of this issue is such that the uncertainty around a real-time R_t estimate may not contain the revised, corrected estimate even after complete data are available [56]. In addition to epidemiological delays, R_t estimation must also account for reporting effects in the observation process, including interval censoring, right truncation, and dynamic bias depending on epidemic phase [51]. In continuing to develop the model used in this paper, recent work has developed new methods to better account for these problems, for example allowing for a changing distribution of reporting delays over time [58]. This is an active area of further research, with an emerging consensus on how best to account for these issues [59,60].

Contextualising real-time R_t estimates

After comparison, the ability to draw insight relies on interpreting the contextual relevance and meaning of any differences. In this work, we relied on highly contextualised knowledge about the spatiotemporal dynamics of vulnerability in the UK epidemic. This led us to triangulate our hypotheses with a range of alternative data sources, including test positivity, patient demographics, and data from care homes, as well as knowledge of targeted testing sites and restrictions. This was supported by SPI-M discussions of the difference in R_t estimates between cases and hospital admissions, which suggested different explanations: first considering localised transmission clusters in hospitals before those in long term residential facilities.

Explaining differences between R_t estimates is likely to be challenging in real-time when alternative data sources may be unavailable and contextual knowledge may be limited. This issue complements calls for disaggregated, timely, and accessible data sources during novel outbreaks [61] [62]. For example, in this work data used on deaths from care homes was confidential until long after the first epidemic among this population had peaked [63]. The availability of such disaggregated data may not be replicable in future outbreaks or in settings

with fewer resources for data collection [15,64], while processing such data can require substantial infrastructure [65].

The challenge of triangulating real-time R_t estimates with contextual knowledge suggests a useful role for collaboration. Contextualising model outputs was a central function of SPI-M, where weekly meetings brought modellers together in moderated discussion [12,66]. However, this process relies on the diversity, quality, and management of participants in the collaboration, while avoiding "groupthink" [27]. Concerns with fairly balancing across collaborators may have contributed to the production of combined R_t estimates across data sources when this was known to be epidemiologically incoherent [67].

Combining Rt estimates

Our approach to comparison of R_t estimates contrasted with SPI-M's approach to reporting a single combined estimate across multiple models. Over 2020 to 2021, ten SPI-M modelling groups estimated R_t, using a variety of model structures and assumptions for key parameters. Estimates across all models were combined with equal weighting in a random effects meta analysis [35]. Individual and combined estimates were discussed at weekly SPI-M meetings, and the consensus range of R_t estimates was agreed and published alongside a narrative summary of discussion [53].

This approach reflected a view of uncertainty where each individual model captured some aspect of a single underlying R_t value, with uncertainty about the most appropriate model framework, parameterisation, or data source to use for estimation [53]. Combining uncertainty across multiple models was therefore seen as increasing the robustness of evidence [66,67]. A justification for this view might be that different methods for R_t estimation are not easy to evaluate. Since it is an unobserved quantity, there is no objective method for assessing the accuracy of different R_t estimates in real-time. Methods for evaluation such as simulation studies or comparisons to R_t estimates from gold standard data (e.g. seroprevalence studies) are challenged by real-world validity and availability at the pace of real-time response [62].

On the other hand, the combined estimate represented an average across all modelling assumptions and methodologies. This loses the ability to link estimates to key epidemiological assumptions, and may mischaracterise sources of uncertainty [57]. This work demonstrated one aspect of this, with misspecification to heterogeneous transmission patterns between data sources. Further differences are also important. For example, combination using meta analysis

could not account for dependency between models in fitting to the same inputs [68]. In particular, the generation time [69] is a key quantity in many models underpinning how the reproduction number relates to the speed of transmission, represented by the growth rate. With many different options for parameterising the generation time, this can produce R_t estimates that are inconsistent [70,71] and overconfident [57].

Following the work presented here, better understanding sources of variation between estimates may make it possible to interpret apparently conflicting estimates of R_t. Further work has characterised methodological differences in R_t estimation [50], and explored its impact. For example, one study systematically explored the effect of methodological differences among multiple models of R_t in Germany [56]. This progressively created increasingly aligned estimates from different models by adapting each model's estimation system to use a standardised choice among a set of common methodological differences. The most substantial source of differences was in the choice of generation time distribution, and estimation window size. Further influential choices were standardising the temporal shifts of incubation period, and accounting for reporting delays. Reconciling across these methodological differences was able to produce more closely similar estimates.

This process of methodological reconciliation suggests the possibility for more meaningful combinations of R_t estimates. Pre-specifying and standardising data sources and model paramaterisations across models before combination would ensure the consistency of key assumptions and treatment of uncertainties underlying combined estimates. One step towards this might include creating shared references for best practices, or databases of relevant parameters, to enable standardisation across multiple models. For example, work is underway identifying best practices in delay estimation [59,72], or key parameters for the nine WHO blue-print priority pathogens [73–75]. Work could also explore more epidemiologically informed methods for model combination. For example, one study separately pooled information on the generation time distribution and exponential growth rate, before combining models hierarchically [57]. From a wider perspective, this requires an approach to model building that follows principles of interoperability: developing model estimation systems that are modular, reproducible, and robust to being used and repurposed [68].

Summary

In this paper, we focussed on comparing and contextualising real-time R_t estimation across data sources, showing the potential for comparison as a source of early insight into heterogeneous

epidemiological dynamics. This faced the difficulties of interpreting limited and biased real-time outbreak data, in both estimating R_t and in interpreting resulting transmission estimates. Further work has continued to improve methods specific to real-time estimation, and considered the availability of outbreak data to interpret differences as they are observed. Second, this work suggested that the combination of R_t estimates based on multiple data sources could produce a misspecified or epidemiologically incoherent estimate. Further work has evaluated across multiple sources of differences in R_t estimates, and suggested directions for the appropriate use of comparative or combined R_t estimates, with a promising approach from modularised model development.

Interpreting accuracy in evaluating multi-model ensemble forecasts

This commentary addresses the paper: "Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations".

Development of the work

Epidemic forecasting is a useful tool for supporting current situational awareness and short-term resource planning for outbreak control. Many models may produce forecasts, and these can be combined into a single ensemble projection. This practice is supported by the finding that ensemble forecasts are on average more accurate in predicting future observations than individual models. This result has been noted since the 1960s [31], both across fields [32] and repeatedly observed in outbreak forecasting [30].

From 2021, the primary output of the European Forecast Hub was the creation of an ensemble model producing prospective real-time forecasts each week. The ensemble took the median average across multiple models' probabilistic quantile intervals to produce a combined probabilistic projection. The central contribution of this work was to characterise the performance of the ensemble forecast against observed data across multiple forecast targets. Complementing existing work, we found the Hub ensemble more robust in providing consistently strong forecast accuracy than any individual component model. A secondary contribution was to compare weighted and unweighted mean and median combinations. We found the strongest performance from an unweighted median. This agreed with a history of mixed successes in finding a better performing alternative to simple ensemble methods [76–78].

Individual contribution

My contribution to this work was twofold. Firstly, from January 2021 I jointly developed, built, and maintained the infrastructure for the European COVID-19 Forecast Hub, together with Sebastian Funk, Johannes Bracher, Hugo Gruson, and many collaborators. I led the adaptation of code for forecast processing and validation; data sourcing, processing, and validation; all documentation; and the initial ensemble. I also contributed code for the forecast evaluation procedures. This drew on existing software from both the US and Germany/Poland COVID-19 Forecast Hubs, and was planned and reviewed jointly with Sebastian Funk. From March 2021, I led work supporting forecasters to contribute to the Hub, and continuously maintained the weekly updating of data sources, forecasts, ensemble, and evaluation.

Second, I led work developing this paper focussing on the performance of the resulting Hub ensemble. I led this work from conceptualisation to analysis and drafting, with supervision by Sebastian Funk. I designed the plan for this work and developed all code to conduct analysis, including accessing forecasts, observed data, and evaluation scores; descriptive summaries of individual model and ensemble performance; and visualisations. Sebastian Funk contributed code for weighted ensembles and their evaluation. I wrote the first and subsequent drafts of the paper. We invited initial feedback from collaborators at the ECDC and in Germany, and a further round of review from all individuals who were named in metadata of the forecasts contributed to the Hub, who had therefore contributed component forecasts to the ensemble. I led work on all subsequent drafts, submission for peer review, and revisions. Code contributions are documented on Github: https://github.com/epiforecasts/euro-hub-ensemble.

Themes and context

Evaluating ensemble performance

A key challenge in developing this work was evaluating an ensemble across widely varying epidemiological targets from an uneven sample of models. To present a single evaluation, we needed to aggregate across differing magnitudes of epidemiological targets between countries, and the number of forecasters and forecasts contributed for each target. This contrasted with evaluations of Hubs such as the US or that for Germany and Poland, which were able to focus on reporting performance against a national level target for which most models contributed [34,45,79].

In this work, we approached this by using a pairwise comparison of the weighted interval score (relative WIS) [80]. The WIS represents both accuracy against an observed data point, with probabilistic precision (sharpness). The relative WIS mitigates the challenge of unequal contributions between forecasting models, by allowing a pairwise comparison between individual models. We compared against a baseline forecast that assumes no change with expanding uncertainty. At the time, this relative use was a newer development in evaluating probabilistic predictions, and this paper contributed to applying this method in practice. The relative WIS has been adopted in subsequent evaluations of collaborative modelling projects, including US and European nowcasting, forecasting, and scenario projections [45,81,82].

The wider context for this work is the fundamental challenge of understanding the generalisability of ensemble performance.

Relating ensemble performance to real-time epidemiological dynamics

By summarising performance relatively, our evaluation obscured the epidemiological conditions of forecast performance. In this work, we found that performance was more stable over a longer time horizon for forecasts of deaths compared to cases, and suggested this was due to longer lags and fewer fluctuations in the observation process. However, beyond this finding we did not quantify accuracy in epidemiological terms and were limited to qualitative observations of poorer performance around local peaks and new variant introductions.

This meant we were unable to specify when over the course of an outbreak an ensemble is the most appropriate choice of forecast [83]. On the other hand, assessing prospective forecasts against retrospectively identified characteristics of the forecast target risks biassing evaluation (the "forecaster's dilemma") [84]. For example, assessing models by performance at the start of an uncontrolled outbreak that later causes a large epidemic would overly reward those that always predict exponential growth.

A more principled strategy to relate forecast performance to epidemiological dynamics might be to use an evaluation metric that better captures the data generating process. Since this paper was written, it has been suggested that the underlying comparison of a forecast to data should be assessed on the logarithmic scale [85]. Using a log transformation is closer to assessing forecast accuracy in terms of the time-varying epidemic growth rate, where forecasts are targeting an underlying exponential process. Meanwhile, further analysis found that using the log transformation did not change the direction of our result demonstrating outperformance of the European Hub ensemble [85]. This may be a more useful way to rank among forecasting models, depending on the needs and expectations of forecast users.

Recent work has also explored adapting methods for forecast evaluation based on relevance to forecast user needs. For example, alternative evaluation methods may summarise forecast performance relative to a specific policy context, including against resource allocation needs [86], or a user-defined threshold for the forecasted target [87]. However, defining the utility of a forecast requires clear decision points, and opportunity for interaction between modellers and stakeholders. Furthermore, defining a utility threshold runs into the issue of propriety (the ability to "game" the forecast score), so unlike using a transformation, these methods may be more suitable for retrospective evaluations.

We have also explored a reverse-engineering approach to understanding the epidemiological context of ensemble outperformance. We identified six aspects of ensemble performance, and incorporated these into a model with known assumptions about epidemiological dynamics [88]. For example, we observed that the Hub ensemble reacted slowly but accurately when a trend turned towards stability; we replicated this by including a multiplicative decay parameter. We evaluated this "surrogate" model's forecasts, finding qualitatively similar behaviour to the Hub ensemble, although with increased uncertainty and poorer performance around periods of peak incidence. This work suggested the possibility that the performance of the ensemble could be at least partly replicated using a simpler, epidemiologically interpretable method.

Relating ensemble performance to component models

While we found that in aggregate across forecast targets an ensemble was the most frequent top performer, we did not investigate ensemble composition to identify the reasons behind its performance. The ensemble was "opportunistic" [89], drawing from an uneven sample of contributing models' forecasts and with little information about model characteristics. Better understanding the composition of the ensemble would support the interpretation of its performance. This is a focus of current in-progress work.

First, evaluating ensemble performance with varying numbers of model components could indicate the impact of sample size. Work analysing US influenza and COVID-19 forecasts has characterised this [82]. This found that among a random selection of models, including more models improved ensemble performance, particularly by decreasing the variability of ensemble performance. We are conducting similar analyses of the European Hub, with compatible early findings that increasing the sample size of models improves performance.

Secondly, an ensemble may be composed of many modelling approaches and methodological strategies for forecasting. Forecasters vary widely in modelling approaches, with more or less ability to tune models to specific targets. Past work has suggested little difference between forecast performance from differing methodological types [13,36]. In further in-progress work, we classify European Hub models by methodological structure and the target specificity of the forecasting model, and model the impact on the interval score. This could suggest differences among forecast methodologies, which future collaborative efforts may consider in recruiting for and interpreting resulting ensembles.

Enabling forecast evaluation

More generally, the standardisation and scale of Modelling Hubs creates a valuable opportunity for systematic evaluations among multiple modelling methods. For example, data from previous collaborative Hubs in the US continues to be re-analysed in recent work assessing individual models and ensemble techniques [82,90] or using new methods of evaluation [86,87]. An important aim of the Forecast Hub was for its data to be re-used in independent analyses of epidemiological forecast skill.

However, such evaluation relies on transparent, ideally reproducible modelling methods being documented alongside model outputs. In the European Forecast Hub, model metadata was limited to authorship and a brief voluntary description of methods, leaving forecasting models as essentially "black boxes". This meant that in further work we could only link forecast performance to two simplified aspects of component models. More detailed evaluations currently rely on published papers as the most extensive form of model documentation, adopting a systematic review process to classify reported methods [18,91]. However, this is subject to publication bias and poor reporting practices [92].

One best practice for reporting modelling work is provided by the EPIFORGE 2020 guidelines for reporting model projections. While designed for forecasting models, this is generally applicable to modelling. The guidelines cover documentation of: data sources, availability, and processing; methods, assumptions, and code; model validation and forecast evaluation; and interpretation of uncertainty, limitations, and generalisability. These guidelines could be adopted by collaborative modelling efforts in their process of collecting metadata. A stronger approach would be for modelling Hubs to adopt the Findable, Accessible, Interoperable, and Reusable (FAIR) principles across contributed models themselves, including model metadata. This follows the example of systems biology, where models are encoded, published, exchanged, and annotated with highly detailed and structured metadata [93].

Collecting better metadata would enable higher quality retrospective evaluations to draw deeper learnings across the field, and could support model users to interpret across multiple models in real-time. It is important to collect this prospectively, given each team's methods can change over time and to avoid biases from retrospective data collection. Supporting modellers to use the more detailed guidelines or share code would also make use of the high profile of modelling collaborations to promote a best practice in the field more widely. While this entails time investment to document models, this could be mitigated by ensuring model documentation is

easy to contribute, for example designing metadata templates based on the EPIFORGE guidelines with standardised options wherever possible.

Summary

This commentary discussed the evaluation of short-term ensemble forecasts, emphasising the challenges of interpreting underlying epidemiological dynamics and component models. Future directions included assessing more epidemiologically relevant methods for evaluation and model selection, and supporting best practices in model documentation to enable more detailed forecast evaluation.

Structuring information and uncertainty among multi-model scenarios

This commentary addresses the paper: "Characterising information gains and losses when collecting multiple epidemic model outputs".

Development of the work

Scenario models are an essential method for accounting for uncertainty while demonstrating the effects of manipulating different variables in a system. COVID-19 saw a proliferation of such modelling exploring both epidemiological and policy interventions influencing health outcomes [94]. Some collaborative work specifically focussed on structuring models' representation of uncertainty for decision support [21].

In March 2022, the European COVID-19 Scenario Hub was launched to support the ECDC to explore uncertainty in scenarios for European epidemiology and intervention. The Hub supported 12 teams to collaborate with the ECDC to co-create several rounds of scenarios that reflected both scientific uncertainty and policy options for COVID-19 control. In this paper, we aimed to understand the potential for information gains and losses when collecting results from multiple scenario models.

This paper was developed after noticing greater differences in results between models than between different scenarios. In the pilot round of the Hub, we wished to explore each model's probabilistic distribution in more detail. However, we were frustrated in this aim by our method of collecting scenario projections summarised by quantile. From Round 1 of the European Scenario Hub, we instead asked modellers to submit up to 100 simulations of equal probability. This study was motivated by exploring both the policy impact and scientific potential of this choice, while it pointed towards wider challenges of handling and representing uncertainty in scenario modelling work.

Individual contribution

I first contributed building and maintaining the European Scenario Hub infrastructure. My work included adapting and developing software for collecting, validating, and visualising scenario projections, and producing all documentation. This was supported by Hugo Gruson with supervision from Sebastian Funk and collaboration with the US Scenario Hub. During the project, I led four rounds of collaborative scenario modelling. I facilitated stakeholder workshops

to codesign scenarios, and designed procedures for collecting metadata, visualising, and narratively synthesising results from the scenario projections.

For this paper, I conceived and conducted the research questions and conducted all analysis, visualisations, and drafting, supervised by Sebastian Funk. First, I led work analysing trajectories for policy relevant characteristics, including writing code and creating visualisations, as part of the development of the Scenario Hub. This aspect of the work was also reviewed by the team at the ECDC. Second, I planned and conducted all analysis and visualisation for work comparing an ensemble from trajectories to the Vincentised ensemble from quantile summaries. In later drafts I added the Linear Opinion Pool ensemble, drawing on code developed by Emily Howerton and Evan Ray. The third aspect of this work was to ensemble scenario trajectories using a weighting procedure based on ongoing evaluation against observed data. I developed all code and visualisation. I led all draft writing, revisions, and paper management. We sought review from all those who had contributed to the Scenario Hub before submission for peer review. Code contributions are at: https://github.com/epiforecasts/multi-model-information.

Themes and context

Identifying policy relevant information

Our work first demonstrated that understanding decision objectives changed the fundamental method of collaborative data collection. This produced information that was both more credible, by creating a more robust visualisation, combination, and assessment of individual trajectories; and more relevant, by using trajectories to show a different set of epidemic characteristics [25,95]. For example, assessing thresholds for outbreak size was a particular concern of the ECDC, which could not be robustly analysed from collecting model quantiles. We noted that while this difference was well recognised informally, there was relatively little note of it in the scientific record. We therefore aimed to add our experience in adapting model collection to maximise policy relevance. This work also supported the development of the US Scenario Hub to similarly collect and analyse modelled trajectories.

This echoes methods for participatory modelling, in which the modelling process is integrated with stakeholder feedback and understanding the decision points for which model outputs are relevant [16,96,97]. Similarly, a growing body of work is now formalising methods for decision support throughout the process of real-time epidemiological modelling. This has focussed on

comparative analysis of multi-model projections in scenario design [98]. A consistent emphasis across this literature is on prospectively identifying decision-maker goals and priorities [99].

In the wider context of the European Scenario Hub, this element of the work could have been substantially improved. European policy making is distributed between transnational bodies, such as the European Commission, and national public health agencies. Our process for identifying stakeholder needs was based on unstructured input from a small technical group within the ECDC, who were limited in their ability to identify, access, or engage with these potential policy users. As a result, there was no direct exchange between the Scenario Hub collaborators and the intended policy users. This compromised the Scenario Hub where information was not tailored to the needs of any specific user or decision.

This mattered specifically for the Hub's focus on evaluating interventions. Both the design of the interventions themselves, and the credibility of model outputs, is dependent on correctly capturing context-specific factors and priorities, such as intervention acceptability and feasibility [98]. The challenge of identifying stakeholder needs could be mitigated by having a clearer process for stakeholder mapping and engagement throughout collaborative work [22,45]. This should both identify potential users of scenario modelling outputs and engage with their specific information needs and priorities, such as which epidemic indicators are already in use for planning health policies. This would ensure both scenario design and model output create relevant, usable information.

Controlling for confounding

A central element of the Scenario Hub collaboration was assessing differences between multiple scenarios. However, correctly understanding the uncertainty in these differences depends on maintaining epidemiological consistency between models. This requires harmonising among potentially conflicting assumptions within each model. For this we adopted a structure of expert elicitation among participants to cover shared sources of uncertainty [10]. For example, we aimed to identify where scenarios were likely to be confounded by additional epidemiological dynamics, such as vaccine effectiveness, and find a shared consensus on how these should be parameterised. This meant that resulting model comparisons should reflect "true" differences between scenarios rather than between differences in model assumptions.

We used a facilitation process that emphasised equal contributions among participants, with the explicit aim stated in each meeting to "share approaches, support other modellers, and discuss

interpretations". On the other hand, a risk of the group deliberation process was that of creating excessive conformity, at the expense of individual modellers' valid knowledge and beliefs (groupthink) [100]. One strategy to handle this would be to include an additional stage of independent work, including model development and testing, before further comparison [10,27,101].

At the same time, it is unclear how effective our process was in balancing between standardisation versus diversifying representations of uncertainty. To assess this, we also collected detailed qualitative model metadata in each scenario round, asking for descriptions of key assumptions with additional comments and interpretation of their impact on model outputs. However, this was not further analysed due to limited time and resources between producing multiple scenario rounds. Further work could consider formalising such retrospective evaluations of confounding.

Expanding and reducing uncertainty: contrasting ensemble methods

This paper addressed a broader concern in the appropriate representation of uncertainty in combining multiple model projections. Selecting among methods for quantitative combinations involves a balance between accurately representing heterogeneity, versus precision. For example, central estimates may be more relevant to the task of short-term prediction rather than representing possible extremes relevant to long-term scenario planning [102].

With this study, we added the perspective of varying data collection before combination, showing that collecting model trajectories retained the greatest degree of heterogeneity at the point at which these were collected. We concluded that while central estimates of an ensemble remained largely similar between ensembles from quantiles or trajectories, either ensembling from trajectories or using a linear opinion pool (LOP) method of combination was important in representing extreme values. This corresponded with parallel work comparing ensemble methods from quantile model outputs in the US [103].

In a further contribution of this work, we started to explore the idea of reducing uncertainty by evaluating individual trajectories as forecasts in a combined ensemble. Progressively comparing these trajectories to observed data created an increasingly sharp consensus prediction, with little loss to the accuracy of representing epidemiological context. To our knowledge this conversion of scenario trajectories to forecasts among multiple models had not been previously explored in a formal evaluation framework.

This aspect of the work also contributed to a wider debate over the circumstances in which future projections should be evaluated against observation. Any model that projects an observable quantity can be held accountable to comparisons against observed data. Projections identified as forecasts are explicitly offered to this account. However, this could be considered inappropriate for scenario models exploring a range of possible futures rather than a single outcome. Nonetheless, scenario projections are frequently evaluated against data retrospectively, often qualitatively [104] and sometimes quantitatively [45,105]. For example, work from the US Scenario Hub evaluated ensemble projections against a retrospective assessment of the plausibility of a variety of scenarios over time, finding ensemble accuracy was slightly increased under more plausible scenario assumptions [45]. Our work took this approach further in suggesting a dynamic method for evaluating and combining model trajectories prospectively, independently of their scenarios in real-time.

In contrast to this work, an alternative approach may reduce the representation of uncertainty even further by presenting ordinal rankings among categorical scenario options. A set of related work has shown that multiple models tend to agree on the ranking of different interventions, even while disagreeing on the magnitude of projected outcomes [26,106]. This may have further implications, for example that relatively few models would be needed in order to arrive at the same recommendation for policy action. At the same time, the above argument for model diversity suggests that an apparent agreement in rankings could be a function of the similarity of model assumptions and approaches.

Further work could develop the method of comparing scenario rankings to ensure it is robust for use. In addition to model similarity, agreement between ranks might also depend on the method used for either measuring the level of agreement [26] or combining across rankings into a single order [107]. It is also not clear at what level or on what scenario rankings should agree. The level of possible agreement between models will differ based on the specific set of scenarios, with more complex scenarios seeing less agreement. Model agreement may differ by the outcome of interest, for example, minimising case incidence at the peak compared to final epidemic size [26]. Promising areas for further work could include defining the size and structure of component models to achieve a stable agreed ranking, as well as the appropriate decision context and methods for ordering among scenarios.

Summary

This work identified the importance of collecting and combining models in a way that is both relevant to policy needs and appropriately represents heterogeneity within and between models. This suggested a need for further work in stakeholder identification, as well as the challenge of controlling for confounding among multiple models together with a consideration for methods of combining and evaluating uncertainty in multi-model work.

Representing experiences of crisis response

This commentary addresses the paper: "Improving modelling for epidemic responses: reflections from members of the UK infectious disease modelling community on their experiences during the COVID-19 pandemic".

Development of the work

The experiences of the modelling community throughout the COVID-19 pandemic highlighted substantial challenges in the field, including psychological pressures, gaps in institutional support, and systemic issues across the academic research landscape. This work was motivated by experiencing these challenges and a shared commitment to improving future response efforts. In this work, we focussed on the modelling community contributing to the UK emergency response, largely channelled through SPI-M-O [40]. We aimed to explore the diversity of responders' experiences, having noted that contemporary evaluations may not reflect the full range of contributions to modelling work.

This work originally developed as an informal survey among the modelling community, before developing into a more structured workshop that could support a fuller and more collaborative exploration of the issues. We then synthesised the experiences and recommendations documented throughout the workshop, and invited all participants to collaborate on this paper. This represented our consensus view of challenges in infrastructure for outbreak modelling and a set of clear recommendations to make future outbreak response modelling more efficient, ethical, and sustainable.

The main contribution of this work was its evaluation of epidemic response work in terms of modelling capacity, together with the development of a set of actionable recommendations for improving sustainability. A specifically novel aspect of this contribution was its inclusive approach to representing and interpreting the experiences of a diverse set of participants across the field of response work. This method contrasted with existing and contemporary evaluations, while producing similar conclusions for necessary action to improve capacity.

Individual contribution

This work was developed in collaboration with Anna Carnegie (co-first author), Sam Abbott, and Yang Liu. The original concept of surveying modellers involved in the UK COVID-19 response was developed jointly and then led by Sam Abbott and Anna Carnegie. I provided feedback on

the survey design, together with Yang Liu, and supported its promotion. We then jointly designed the workshop, with implementation led by Sam Abbott and Anna Carnegie. I reviewed this together with Yang Liu, participated in the workshop, and gathered materials after the workshop completed. I then led on synthesising the experiences and recommendations documented throughout. This was initially intended as a workshop report before developing further into the academic output presented here.

In this phase of the work, I first led work gathering and digitising all workshop outputs, for example notes, mind maps, and dot plots from the day, converting these into usable data. I then led work ordering and synthesising these qualitative data into themes and recommendations, with review and feedback from Anna Carnegia, Sam Abbott and Yang Liu. Anna Carnegie then wrote an initial outline for a workshop report. I first reviewed this and then with joint agreement I led work to draft an academic paper reporting our findings. I led on the development of this write up, including the background, formalising the structure of the methods, writing the results and discussion. I then managed several rounds of review from the core authorship group and those who participated in the workshop.

Themes and context

Participatory post-crisis evaluation

Infectious disease outbreaks can occur with little warning and require rapid action to prevent an exponential crisis. At the earliest stage, modelling may be the best, if not the only, decision support tool, demanding modellers rapidly adapt methods to each outbreak's novel characteristics [1,12]. This "trial by fire" can identify a range of weaknesses and limitations with available capacity for outbreak response, prompting retrospective evaluations aimed at improving outbreak modelling [108]. In the UK, qualitative evaluations of COVID-19 modelling have focussed on its role relative to crisis policymaking: for example, in the UK public COVID-19 Inquiry [109], or academic research [110–112]. In developing this work, we questioned whether these evaluations could improve the field of modelling as a whole.

We particularly noted that evaluation exercises focussed on sampling from senior career stages and traditional modelling work, poorly reflecting the diversity of experience among those who contributed to the response. Scientific contributions to the response were both formal and informal [113] and the composition of contributing modelling teams was highly variable between

research groups and over time. Meanwhile, a diverse range of career levels and skill sets were needed to create and support modelling work [40,65].

However, gathering these views is particularly challenged by accessibility and survivor bias. Not all who contributed were acknowledged or publicly identified, and evaluations have relied on recruiting more easily accessible individuals, typically also more senior: for example, by surveying those on the public list of SAGE and SPI-M attendees [111]. While this sampling may be useful for representing policy interaction, this does not represent the wider field of outbreak modelling. Further, those with negative experiences are more likely to change fields, becoming less easy to contact, or have less motivation to participate in an evaluation. At the same time, we observed that negative experiences were differentiated by seniority, as early career researchers and support staff (already more likely to be on short term contracts) left the field. Both accessibility and survivor biases in sampling would likely influence any evaluation to represent the status quo, missing opportunities to address issues experienced by the wider field of contributors to modelling work.

In this work, we attempted to ensure that participants represented a range of experiences contributing to the COVID-19 response. We targeted individuals across institutions, including a range of UK universities and the UK public health agency (UKHSA), as well as across career stages: from doctoral students relatively new to the field, to senior academics with long-standing policy relationships. We attempted to address survivor bias by deliberately inviting participation from those who had moved away from infectious disease modelling. Our efforts were only partially successful, with several individuals unresponsive or unable to attend the workshop.

Future evaluations could develop the use of real-time methods for evaluating capacity during outbreak response. Previous prospective evaluations of UK modelling were policy-oriented and limited by the capacity and willingness of modellers to contribute to such evaluative research in an ongoing crisis [111,112]. However, prospectively evaluating and reporting challenges in modelling capacity across the field could provide immediate feedback to modelling teams and collaborations. This could be used to rebalance capacity and preventatively address burnout, and thereby improve both the quality of short-term response and long-term sustainability. This is likely to be particularly relevant when the length of the crisis response is uncertain [12].

Evaluating capacity for epidemic response

Our findings from this work complemented the body of both pre- and post-COVID-19 modelling evaluations [108]. These suggest the intense resource requirements of contributing to collaborative outbreak response. For example, the rapid pace of SPI-M policy interaction created extreme pressure and substantial time away from ongoing research [12,113] while modelling teams often reallocated all possible resources to voluntary response work [40].

This trade-off between time contributing to collaborative work versus long-term academic commitments is a common feature of working at the science-policy interface [114,115]. This is intensified in work which carries political and moral urgency, often resulting in individual burnout [116]. In this work, these impacts were exacerbated by the unusually long period of crisis response (2020-2022), together with highly varied and often unclear policy needs [112]. Our work also demonstrated the differential impact of this trade-off by career stage and gender. While appearing to be highly collaborative, science during COVID-19 may have been an intensification of existing hierarchical structures [117], with similar differentials noted across scientific fields [118] [119].

To address these issues, in this work we called for a stable "critical mass" of modelling expertise, maintained at an institutional level independent of individual grant funding. In an emergency, adapting existing structures is more efficient than creating new ones [113]; at the same time, outbreaks are both unpredictable, and becoming more likely [120]. This makes it crucial to create standing capacity for modelling work that can be rapidly and sustainably deployed when needed. Similar calls to improve capacity have been made repeatedly across evaluations of outbreak responses, for example after 2001 Foot and Mouth Disease [121], 2009 H1N1 [1], or 2014 Ebola [122,123]. Likewise, and as in our work, similar recommendations have been repeated since COVID-19, with capacity constraints noted in at least western Europe [124], Switzerland and Germany [125], Australia [126].

Despite the urgency and necessity of our and others' recommendations, there is a notable gap in translating them into actionable steps. This is likely due to the rapid onset of outbreaks requiring immediate action, the subsequent focus on immediate crisis management, and the lack of attention and resources when outbreaks are not in progress. Meanwhile, these recommendations raise the question of exactly where and for what for what such capacity should be created.

While modelling has traditionally been an academic pursuit, outbreak response involves a constellation of actors, who all may draw on modelling work and therefore be implicated in calls to increase capacity. For example, modelling pipelines involve many different tasks, skills, and specialisms, of which only some intersect with traditional research [123]. This can make for an unclear division of labour between the technical development and operational implementation of modelling work [65,102]. For example, in the UK's COVID-19 response various elements of modelling pipelines used in policy shifted over time between the academic contributors to SPI-M, the UK Health Security Agency, and the Joint Biosecurity Centre [66].

The need for such standing capacity has become more widely recognised amid a wave of post-COVID-19 investment in modelling. There is some evidence that modelling roles have been expanded into national public health rapid support teams in Canada [127], the UK [128]. In the US, this has taken the form of a new unit within the US Centre for Disease Control, the Centre for Forecasting Analytics; this was explicitly based on previous recommendations to improve outbreak modelling [123,129]. At an international level, the WHO has invested in a Hub for Pandemic and Epidemic Intelligence [46]. Meanwhile the "Hubverse" initiative aims to facilitate standardised infrastructure to collect multiple model outputs [130].

However, it is unclear whether new capacity is appropriately or effectively allocated or across needs for modelling for outbreak response. Firstly, it is not evident how new initiatives map to existing recommendations, or whether proposed targets for improving modelling are clear, quantifiable, and set independently from the institutions responsible for implementation. Further, identifying and expanding capacity may be challenging in settings where modelling work is not seen as a traditional part of outbreak response, for example in highly localised outbreaks [131].

We have started some further work to address this in a follow up to the work presented here. This uses a repeated survey of the same set of workshop participants to explore how the current landscape of outbreak modelling reflects the five priority recommendations and actions suggested in the original workshop. While limited to our existing sample, from this work we hope to identify potential current and future changes in outbreak modelling response work.

Summary

In this work, we explored issues in capacity for outbreak response using a participatory method. Our inclusive approach to evaluation led to a broader reflection of the heterogeneous landscape of modelling work, showing where this created unequal and inefficient impacts from the intensity

of emergency response. While issues with modelling capacity during outbreak response have been widely noted, efforts to address this remain potentially biased. Future work should support ongoing evaluation of capacity using a prospective and participatory perspective.

Evaluating collaborative modelling and multi-model combination

The works discussed in this thesis explored collaborative outbreak modelling across four phases of outbreak response: from outbreak detection and investigation, through evaluating short term forecasts, assessing outbreak control options with scenario modelling, and capacity building from post-outbreak evaluation. This built on three modelling collaborations: the UK's SPI-M, and the European COVID-19 Forecast and Scenario Modelling Hubs. This concluding discussion draws across this work to suggest common challenges, trade-offs, and future directions for the use of model combinations arising from collaborative outbreak modelling.

Challenges

Outbreak response relies on understanding rapidly changing epidemic dynamics. Model combinations offer a focal point for this understanding, by synthesising multiple expert insights into a single estimate to be used for planning and action. However, this work repeatedly challenged the practice of aggregating across models with the need for their epidemiological interpretation.

The key challenge for modelling collaborations is their vulnerability to selection effects in the sample of component models. The ability to interpret results from a model combination depends on the validity of epidemiological mechanisms driving its component models. However, the complexity of epidemic dynamics creates both structural and parametric uncertainty in these mechanisms. This allows for a wide range of plausible assumptions and methods in the modelling process. To retain validity, a model combination should sample across this heterogeneity systematically.

However, modelling collaborations typically recruit modellers opportunistically [89], without a pre-specified sampling strategy to account for either the diversity or quality of modelling representations of the target quantity. For example, recruitment may come from existing networks of expertise, with self-selected voluntary participation that fluctuates over time. This opportunistic sampling of model components leaves model combinations open to both bias and confounding. This can affect the validity as well as reliability of resulting estimates. This tension underpinned the collaborations discussed here, where such threats to epidemiological validity played out in different ways across different modelling tasks (table 1).

Table 1. Factors affecting the epidemiological interpretation of model combinations across three outbreak modelling collaborations represented in this work.

	Sample	Model harmonisation	Published combination	Epidemiological interpretation
UK SPI-M: R_t estimation *Aim*: Characterise current state of UK transmission	Selective, by invitation from chair	Informal: unstructured discussion including comparison of model design	Consensus range of numerical estimates from meta-analysis, plus narrative synthesis	Likely confounded by incompatible use of data sources, creating estimates not representative of any single transmission process.
European COVID-19 Forecast Hub *Aim*: Predict 1-4 weeks' observed cases and deaths	Open call	None. Model documentation included voluntary method description.	Probabilistic quantile projection based on median ensemble	Typically greater predictive accuracy compared to individual models, but no ability to interpret mechanisms driving this outperformance.
European COVID-19 Scenario Hub *Aim*: Demonstrate differences between policy scenarios	Open call	Semi-structured discussion of shared parameters, resulting in shared quantitative parameter values, then documented in metadata.	Narrative synthesis with comparative visualisations	Variability within and between models obscured differences between scenarios, suggesting uncontrolled confounding and/or structural uncertainty.

Firstly, model combinations may be biased at the level of component models. At a theoretical level, model combinations can provide at most a lower bound on the uncertainty of the target quantity. This is because it is inherently unknown whether such an opportunistic sample captures the full range of structural uncertainty [132]. Similarly, opportunistic recruitment creates potential for bias in resulting model combinations. For example, where recruitment comes from an existing network sharing similar expertise, a model combination can only reflect this shared view of epidemiological dynamics. A particular example is where collaborations recruit from modelling expertise alone, even while modelling might draw on evidence from multiple disciplines: from within-host immunological dynamics to the social acceptability of interventions [133–135].

Secondly, model combinations may be biased at the level of model parameterisation. Model combinations aim to capture across the range of uncertainty expressed by equally valid parameterisations between models. At the same time, this may be confounded by heterogeneity in underlying epidemic dynamics. For example, aggregating across models that draw from multiple source populations with differing transmission patterns may confound a resulting combined estimate. This risks biassing a combined estimate, making it less accurate than any individual model representing homogenous epidemiological dynamics.

One strategy to control for such confounding involves harmonising across potentially conflicting model parameterisations. This process attempts to identify the wider epidemiological context of the target quantity, and control model parameters that may confound a resulting combined estimate: for example, vaccine efficacy in comparing interventions for vaccine coverage [136], or the generation time in estimates of the reproduction number [57]. Harmonisation might be achieved by a process of expert elicitation among modelling collaborators [27], or by adapting interoperable and modularised model code [56]. At the same time, harmonisation is challenged by the risk of "groupthink", against the competing need to maintain modellers' independence where there are multiple plausible assumptions [21].

Assessing the epidemiological validity of any model combination is critical to evaluating its role in decision support. At the same time, these issues are exacerbated when model combinations are repeated over time with fluctuating component models, for example in a longitudinal ensemble forecast. In this case, a superficially stable model combination may be fundamentally unreliable, and become more or less biased or confounded over time. Meanwhile, it is often difficult to assess these issues at all: for example, in large open call collaborations, model

outputs may be often collected with minimal associated metadata and little interaction among modellers. These issues with internal validity and reliability create additional uncertainty in how supportive a model combination can be for outbreak decision making.

This challenge to decision support can be widened further to the targets and process of model combination itself. Any model combination must be both appropriate and timely to be relevant. In order to standardise across multiple models, collaborations must collect and present results from a small set of pre-specified quantitative targets. However, both the scientific and policy realities of an emerging outbreak evolve quickly, creating constant potential for misalignment between modelling work and stakeholder needs [22]. For example, possible decision targets might include assessing risk factors and response needs, involving both continuous and categorical outcomes, such as thresholds with some degree of tolerance [102,131,137]. Collaborations may struggle to keep pace with this demand, given the need to adapt both contributing models and supporting infrastructure to create model combinations for different targets.

A final challenge to the use of model combination is by comparison to the decision support provided by an individual model. Stakeholders may gain the most benefit when modelling is simple and offers a narrative bridge between theory and practice, such as offering heuristic demonstrations of counterintuitive relationships or exploring a model's parameters and uncertainties [131,138,139]. Individual modellers can offer this opportunity, for example via discussion [96] or interactive tools [140]. By contrast, modelling collaborations add a mediating layer between modellers and stakeholders, potentially slowing this process and making it more difficult to offer this exploratory form of decision support.

Trade-offs

This discussion has raised the possibility of several trade-offs in collaborative outbreak modelling and the production of model combinations. A focus among the work presented here is balancing between creating the clarity of consensus, versus communicating the complexity of epidemic dynamics to model stakeholders; and the need for consistency among components of a model combination, versus drawing on the limited capacity of modelling collaborators.

Context/consensus

A central offer of collaborative work is to create a consensus, offering clarity and simplicity for decision support. However, in synthesising across multiple models, collaborative projects may face the competing aim of providing epidemiological context. Firstly, the uncertainty and heterogeneity of epidemic dynamics create inevitable variability among modelling evidence. This can itself offer useful context for decision support. On the other hand, this diversity may be overwhelming or irrelevant. A model combination offers to simplify across this diversity. However, model combinations with unclear model components risk becoming epidemiologically uninterpretable, biased, or confounded, while the process of model combination itself is a choice in representing greater or lesser uncertainty.

This theme was represented throughout this work. For example, we demonstrated the importance of retaining context in outbreak investigation. Given high methodological uncertainty in estimating the reproduction number, model combination was thought to create a more robust consensus. In this work, we showed that this was confounded by heterogeneous epidemic dynamics among populations represented in different data sources. In this case, model combination missed the opportunity for policy-relevant epidemiological context. On the other hand, we found the benefits of consensus in short-term predictions of observable epidemic dynamics. Model combination created relatively more accurate predictions, with some consistency across targets. However, this was challenged by the difficulty of evaluating reliability in the face of changing epidemic dynamics, particularly with unidentifiable ensemble components. In work evaluating long-term scenario projections, we contrasted these approaches directly: collecting model trajectories enabled us to retain the greatest degree of heterogeneity at the point at which these were collected. Meanwhile, progressively comparing these trajectories to observed data created an increasingly sharp consensus prediction, with little loss to the accuracy of representing epidemiological context.

The trade-off between offering context or consensus in evidence synthesis is at its clearest in the choice of model combination method. At one extreme, using qualitative narrative synthesis can provide the greatest context to communicate uncertainties. For example, SPI-M only published quantitative estimates within a longer consensus statement that summarised group discussion of differences among models [12]. Using quantitative combinations can also choose to aggregate uncertainty within a greater or lesser probabilistic range, such as Vincentised averaging or the linear opinion pool ensemble [103]. At another extreme, model results might be

reduced to combining ordinal rankings of categorical outcomes [106,107], for example summarising to a majority vote using posterior estimates [27]. This choice of method requires a subjective judgement on the importance of representing heterogeneity.

The subjectivity of this trade-off might also suggest an ethical responsibility in collaborative evidence synthesis. Against a constant context of underlying uncertainty, collaborations can select a method for model combination that indicates more or less strength of consensus. This suggests an ethical, as well as scientific, challenge in navigating between evidence and advice [141,142], which is intensified during emergency response [143,144]. For example, offering a single quantitative summary may be afforded greater trust than qualitative interpretation [145,146]; while combining categorical rather than continuous model outcomes can create the strongest and simplest perception of quantitative consensus. Meanwhile, without an understanding of context a model combination may directly breach ethical criteria by aggregating across subpopulations with known inequalities in the risks of intervention [147].

Collaboration/capacity

Interpreting results from a model combination at least partly depends on understanding its component models. To create a consistent interpretation, a model combination therefore requires a consistent sample of component models. This demand is both for temporally stable model components, and epidemiologically consistent model parameterisations, for example by harmonisation. At the same time, modellers are typically capacity constrained, facing limited resources with competing demands for both model development and application. Modelling collaborations may therefore face a trade-off between the need for consistent contributions while not depleting the capacity of collaborators.

This trade-off affected each of the collaborations represented here in different ways. For example, the European Forecast Hub used an open call strategy and had the largest sample of modelling teams among these collaborations, with over fifty models contributing over time. However, this was highly inconsistent as modelling teams joined and left. By contrast, the SPI-M collaboration included ten selectively invited models of the reproduction number, and all contributed outputs over two years [66]. However, in this work we also saw that maintaining such consistent contributions created burnout from the intensity and duration of this collaborative work. In the middle of these extremes, the European Scenario Hub gained internal consistency among models, with adaptive rounds of model harmonisation. However, the

resource demands on modellers from this process led to a smaller sample of modellers, and, alongside lack of clear policy relevance, resulted in modellers leaving the collaboration.

The key concern of balancing consistent collaboration with capacity is that the costs and benefits from collaboration are asymmetric. For example, collaboration may create a particular issue for modellers working in traditional academic roles, who must face the opportunity cost of voluntarily contributing to short term collaborative work instead of traditionally rewarded longer term research [114,115]. This both affects the sustainability of individual research careers, and hampers methodological advances in model development [14]. This can devalue the legitimacy of modelling collaborations, particularly where, as we observed in this work, this asymmetry of cost/benefit is unequal even within the field of modelling work.

Criteria

Forming and presenting a collaborative model combination requires some criteria for choosing among these trade-offs. Such criteria might include the complexity of epidemiological dynamics underlying the target quantity, combined with an understanding of stakeholder needs for reliable epidemiological interpretation. For example, one approach might favour retaining the maximum extent of epidemiological context, such as adopting a process for between-model harmonisation or using a combination that represents the widest possible view of heterogeneity. This might be preferred when models cannot be easily objectively calibrated, such as in scenario modelling or estimates of the reproduction number, or when planning requirements are highly sensitive to anomalies and extreme events, such as in early outbreak detection.

On the other hand, collaborations might approach model combination by exchanging epidemiological interpretation with a stronger degree of consensus, such as aggregating across quantiles rather than trajectories, or presenting a single ensemble regardless of fluctuating model components. This approach might be used if the epidemiological context is relatively well defined, for example when future projections are limited to one or two generations of transmission; when there is a shared understanding of data generating processes; or when operational decisions are not likely to be impacted by underlying heterogeneity in epidemic dynamics.

Meanwhile, the ability to implement either of these strategies for model combination depends on the capacity of both modellers and a central coordinating team; adequate infrastructure and processes to support different approaches to model combination; a clear understanding of

stakeholder needs from a specific model combination; and the wider ethical context of decision making during outbreak response.

Conclusions

Summary

This thesis drew together approaches to collaborative outbreak modelling and model combination during COVID-19 in the UK and Europe. This built on three modelling collaborations: the UK's Scientific Pandemic Infections group on Modelling, and the European COVID-19 Forecast and Scenario Modelling Hubs. Each paper here reflected a different use of modelling across the lifecycle of outbreak decision support, while a discussion drew out common themes and challenges in the use of multi-model combination.

Starting with outbreak investigation, I showed that model combinations of the reproduction number in the UK could be confounded by misspecification to heterogeneous epidemic dynamics. This created over confident uncertainty, while missing the opportunity for policy-relevant insights. Next, I demonstrated that short-term predictions of COVID-19 across Europe became more accurate with model combination, while challenged by the difficulty of evaluating accuracy in terms of epidemiological dynamics and mechanisms. Third, I explored long term European scenario projections. This identified the importance of collecting and combining models in a way that was both relevant to policy needs and appropriately identified and represented heterogeneity both within and between models. Finally, I evaluated issues in the long term sustainability of outbreak response modelling, identifying both the challenges of response work and the consensus of recommendations for improvements to be found in retrospective evaluations of outbreak modelling.

This suggests substantial challenges in building and evaluating modelling collaborations and resulting evidence syntheses. The validity of methods for modelling synthesis may be challenged by creating inappropriate precision and loss of mechanistic interpretability, arising from the sample of contributing models and the extent of harmonisation across them. At the same time, collaborations for outbreak modelling operate at the boundary of science and policy and must manage stakeholders from both these groups. This is challenged by conflicting needs, interests, and resource constraints, and raises concerns in selecting and communicating methods of evidence synthesis. These challenges in both the structure and process of collaborative modelling interact; they are also intensified under the pressures of both high uncertainty and time sensitivity during real-time outbreak analysis and control.

The validity of multi-model combination could be addressed by restricting the target of analysis, harmonising model assumptions, and selecting a method of combination appropriate to the decision context. Infrastructural concerns could be addressed by better aligning incentives towards collaborative work, including expanding modelling capacity. In light of these trade-offs, appropriate strategies for outbreak modelling collaborations likely depend on the timing, scale, and purpose of outbreak decision making as well as the existing capacity and resources available to modelling collaborators.

Limitations

A limitation throughout this work is that it offers little comparison to collaborative modelling across settings. This work focused on three collaborations tailored to the COVID-19 response. In particular, these focussed on high-resource settings that attracted a large amount of research capital and labour. This may not be repeated in future outbreaks or alternate settings where modelling is used for public health use [117]. In these cases, collaborative trade-offs are likely to change based on differences in both epidemiological dynamics and political pressures. Collaborations for less salient infectious disease outbreaks may be unlikely to see the scale and diversity of data sources, models, and modellers, and it may be more or less difficult to bring these together in collaboration. To address this limitation, future work could specifically focus on contrasting both the contribution and challenges of collaborative work in different outbreak contexts. This would help identify which aspects of collaboration should be prioritised for capacity building investments.

Relatedly, this work often lacked direct insight into the decision-making context of modelling collaborations. However, this context is key to justifying the relevance of such projects. For example, collaborative work specifically targeted to lower and middle income countries has reported a much higher degree of stakeholder involvement and influence from direct decision makers in comparison to the work presented here [16,22]. This limitation meant the challenges identified here focussed on the internal validity of model combinations rather than clarifying the impact of this on applied use.

Future directions

This discussion has moved in parallel between assessing the infrastructure of collaborative work, and the information from model combination that such collaborations produce. This suggests two directions for future work.

Epidemiological inference and model transparency

This body of work converges on the importance of understanding model selection, both within and between models participating in modelling collaborations. This suggests potential for future work to consider sampling, selection effects, and confounding affecting the validity of model combination. For example, this could include exploring the accuracy and/or precision of combined estimates with varying degrees of diversity among component models; or identifying the impact of differential attrition of models over time, for example by model structure or performance. Further work could also develop and evaluate more epidemiologically principled model combination methods: such as using stratification on the basis of epidemiological assumptions [57]; more systematically evaluating methods for prospective model harmonisation [27]; or continue developing model combinations for decision making targets that are not continuous projections, such as for crossing thresholds or among scenario choices [107].

A general conclusion from this aspect of the work suggests a focus on developing a systematic approach to model documentation and reproducibility. This is the crucial link between modelling contributions and the ability to compare and interpret across them. Real-time collaborative work can enable discussion of this context. However, model documentation enables transparency and longer-term reproducibility, as well as further retrospective analyses. Further work could support adopting the EPIFORGE reporting checklist [19] into collaborative work, support shared model code, and link collaborative outputs to such model documentation, for example by visualised summaries of key parameters. Future collaborative work presenting combined model estimates should increase their transparency and interpretability by clearly reporting the sample of models, criteria, and method of combination, following guidelines for best practice [27].

Capacity constraints and stakeholder needs

Further work could better characterise and evaluate the demand and supply dynamic of collaborative work for outbreak decision support. To address the challenge of collaborating under capacity constraints, future work could more systematically understand modellers' incentives and disincentives for collaborating. This work specifically suggested the need for both establishing and evaluating standing modelling capacity outside of an outbreak setting, and continuously evaluating this in a way that appropriately reflects the diversity of modelling work during an outbreak. To mirror this, future work could more thoroughly assess stakeholder needs and expectations from model combinations, for example with prospective stakeholder mapping

and co-design of collaborative infrastructure. This could better align collaborative contributions with their use.

This work points towards some of the broader tensions of collaboration at the science-policy interface during emergencies. Further evaluative work could draw from established frameworks for such work, for example explicitly accounting for trade-offs between the credibility, relevance, and legitimacy of such work [114,148]. In this work, we specifically observed issues with legitimacy, including the inequalities and moral dilemmas of creating consensus in emergency outbreak response. Such issues could be better recognised. For example, one call for improving evidence synthesis suggests analyses should be inclusive, rigorous, transparent, and accessible [149]; while multi-model combinations might consider concepts of ownership, justification, and robustness [150]. Meanwhile, some efforts have been made to define ethical frameworks for individual modelling in public health policy, based on principles from biomedical ethics such as independence and beneficence [147,151]. Future work should translate and operationalise such principles into the specific setting of modelling synthesis during real-time emergency response.

Bibliography

1. Van Kerkhove MD, Ferguson NM. Epidemic and intervention modelling: a scientific rationale for policy decisions? Lessons from the 2009 influenza pandemic. Bull World Health Organ. 2012;90: 306–310. Available: https://www.scielosp.org/pdf/bwho/v90n4/v90n4a15.pdf

2. Metcalf CJE, Morris DH, Park SW. Mathematical models to guide pandemic response. Science. 2020. pp. 368–369. doi:10.1126/science.abd1668

3. Diekmann O, Heesterbeek H, Britton T. Mathematical Tools for Understanding Infectious Disease Dynamics. Princeton University Press; 2013.

4. Heesterbeek H, Anderson RM, Andreasen V, Bansal S, De Angelis D, Dye C, et al. Modeling infectious disease dynamics in the complex landscape of global health. Science. 2015;347: aaa4339. doi:10.1126/science.aaa4339

5. Baguelin M, Medley GF, Nightingale ES, O'Reilly KM, Rees EM, Waterlow NR, et al. Tooling-up for infectious disease transmission modelling. Epidemics. 2020;32: 100395. doi:10.1016/j.epidem.2020.100395

6. Huppert A, Katriel G. Mathematical modelling and prediction in infectious disease epidemiology. Clin Microbiol Infect. 2013;19: 999–1005. doi:10.1111/1469-0691.12308

7. Swallow B, Birrell P, Blake J, Burgman M, Challenor P, Coffeng LE, et al. Challenges in estimation, uncertainty quantification and elicitation for pandemic modelling. Epidemics. 2022;38: 100547. doi:10.1016/j.epidem.2022.100547

8. Funk S, King AA. Choices and trade-offs in inference with infectious disease models. Epidemics. 2019;30: 100383. doi:10.1016/j.epidem.2019.100383

9. Berger L, Berger N, Bosetti V, Gilboa I, Hansen LP, Jarvis C, et al. Rational policymaking during a pandemic. Proceedings of the National Academy of Sciences. 2021;118: e2012704118. doi:10.1073/pnas.2012704118

10. Shea K, Runge MC, Pannell D, Probert WJM, Li S-L, Tildesley M, et al. Harnessing multiple models for outbreak management. Science. 2020;368: 577–579. doi:10.1126/science.abb9934

11. Reich NG, Lessler J, Funk S, Viboud C, Vespignani A, Tibshirani RJ, et al. Collaborative Hubs: Making the Most of Predictive Epidemic Modeling. Am J Public Health. 2022;112: 839–842. doi:10.2105/AJPH.2022.306831

12. Medley GF. A consensus of evidence: The role of SPI-M-O in the UK COVID-19 response. Adv Biol Regul. 2022;86: 100918. doi:10.1016/j.jbior.2022.100918

13. Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, et al. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. Epidemics. 2018;22: 13–21. doi:10.1016/j.epidem.2017.08.002

14. Kucharski AJ, Funk S, Eggo RM. The COVID-19 response illustrates that traditional

academic reward structures and metrics do not reflect crucial contributions to modern science. PLoS Biol. 2020;18: e3000913. doi:10.1371/journal.pbio.3000913

15. Cori A, Lassmann B, Nouvellet P. Data needs for better surveillance and response to infectious disease threats. Epidemics. 2023;43: 100685. doi:10.1016/j.epidem.2023.100685

16. Aguas R, White L, Hupert N, Shretta R, Pan-Ngum W, Celhay O, et al. Modelling the COVID-19 pandemic in context: an international participatory approach. BMJ Global Health. 2020;5: e003126. doi:10.1136/bmjgh-2020-003126

17. Biggerstaff M, Slayton RB, Johansson MA, Butler JC. Improving Pandemic Response: Employing Mathematical Modeling to Confront Coronavirus Disease 2019. Clin Infect Dis. 2022;74: 913–917. doi:10.1093/cid/ciab673

18. Kobres P-Y, Chretien J-P, Johansson MA, Morgan JJ, Whung P-Y, Mukundan H, et al. A systematic review and evaluation of Zika virus forecasting and prediction research during a public health emergency of international concern. PLoS Negl Trop Dis. 2019;13: e0007451. doi:10.1371/journal.pntd.0007451

19. Pollett S, Johansson MA, Reich NG, Brett-Major D, Valle SYD, Venkatramanan S, et al. Recommended reporting items for epidemic forecasting and prediction research: The EPIFORGE 2020 guidelines. PLoS Med. 2021;18: e1003793. doi:10.1371/journal.pmed.1003793

20. Zavalis EA, Ioannidis JPA. A meta-epidemiological assessment of transparency indicators of infectious disease models. PLoS One. 2022;17: e0275380. doi:10.1371/journal.pone.0275380

21. Shea K, Borchering RK, Probert WJM, Howerton E, Bogich TL, Li S-L, et al. Multiple models for outbreak decision support in the face of uncertainty. Proc Natl Acad Sci U S A. 120: e2207537120. doi:10.1073/pnas.2207537120

22. Teerawattananon Y, Kc S, Chi Y-L, Dabak S, Kazibwe J, Clapham H, et al. Recalibrating the notion of modelling for policymaking during pandemics. Epidemics. 2022;38: 100552. doi:10.1016/j.epidem.2022.100552

23. Metcalf CJE, Edmunds WJ, Lessler J. Six challenges in modelling for public health policy. Epidemics. 2015;10: 93–96. doi:10.1016/j.epidem.2014.08.008

24. Hadley L, Challenor P, Dent C, Isham V, Mollison D, Robertson DA, et al. Challenges on the interaction of models and policy for pandemic control. Epidemics. 2021;37: 100499. doi:10.1016/j.epidem.2021.100499

25. McCabe R, Kont MD, Schmit N, Whittaker C, Løchen A, Walker PGT, et al. Communicating uncertainty in epidemic models. Epidemics. 2021;37: 100520. doi:10.1016/j.epidem.2021.100520

26. Wade-Malone LK, Howerton E, Probert WJM, Runge MC, Viboud C, Shea K. When do we need multiple infectious disease models? Agreement between projection rank and magnitude in a multi-model setting. Epidemics. 2024; 100767. doi:10.1016/j.epidem.2024.100767

27. den Boon S, Jit M, Brisson M, Medley G, Beutels P, White R, et al. Guidelines for multi-model comparisons of the impact of infectious disease interventions. BMC Med. 2019;17: 163. doi:10.1186/s12916-019-1403-9

28. Clapham H, Gad M, Gheorghe A, Hutubessy R, Megiddo I, Painter C, et al. Assessing fitness-for-purpose and comparing the suitability of COVID-19 multi-country models for local contexts and users. Gates Open Research; 2021 May. doi:10.12688/gatesopenres.13224.1

29. Hollingsworth TD, Medley GF. Learning from multi-model comparisons: Collaboration leads to insights, but limitations remain. Epidemics. 2017;18: 1–3. doi:10.1016/j.epidem.2017.02.014

30. Buckee CO, Johansson MA. Individual model forecasts can be misleading, but together they are useful. Eur J Epidemiol. 2020;35: 731–732. doi:10.1007/s10654-020-00667-8

31. Bates JM, Granger CWJ. The Combination of Forecasts. OR. 1969;20: 451–468. doi:10.2307/3008764

32. Chen L. A review of the applications of ensemble forecasting in fields other than meteorology. Weather. 2024. doi:10.1002/wea.4584

33. Del Valle SY, McMahon BH, Asher J, Hatchett R, Lega JC, Brown HE, et al. Summary results of the 2014-2015 DARPA Chikungunya challenge. BMC Infect Dis. 2018;18: 245. doi:10.1186/s12879-018-3124-7

34. Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proceedings of the National Academy of Sciences. 2022;119: e2113561119. doi:10.1073/pnas.2113561119

35. Maishman T, Schaap S, Silk DS, Nevitt SJ, Woods DC, Bowman VE. Statistical methods used to combine the effective reproduction number, R(t), and other related measures of COVID-19 in the UK. Stat Methods Med Res. 2022;31: 1757–1777. doi:10.1177/09622802221109506

36. Johansson MA, Apfeldorf KM, Dobson S, Devita J, Buczak AL, Baugher B, et al. An open challenge to advance probabilistic forecasting for dengue epidemics. Proceedings of the National Academy of Sciences. 2019;116: 24268–24274. doi:10.1073/pnas.1909865116

37. McGowan CJ, Biggerstaff M, Johansson M, Apfeldorf KM, Ben-Nun M, Brooks L, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. Sci Rep. 2019;9: 683. doi:10.1038/s41598-018-36361-9

38. Green LE, Medley GF. Mathematical modelling of the foot and mouth disease epidemic of 2001: strengths and weaknesses. Res Vet Sci. 2002;73: 201–205. doi:10.1016/s0034-5288(02)00106-6

39. Rivers C, Pollett S, Viboud C. The opportunities and challenges of an Ebola modeling research coordination group. PLoS Negl Trop Dis. 2020;14: e0008158. doi:10.1371/journal.pntd.0008158

40. Brooks-Pollock E, Danon L, Jombart T, Pellis L. Modelling that shaped the early COVID-19

pandemic response in the UK. Philos Trans R Soc Lond B Biol Sci. 2021;376: 20210001. doi:10.1098/rstb.2021.0001

41. Delfraissy J-F, Horgan M, Mølbak K, Simón FS, Stadler T, van Dissel J, et al. Scientific advisory councils in the COVID-19 response. Lancet. 2024;403: 510–512. doi:10.1016/S0140-6736(23)01846-9

42. Adib K, Hancock PA, Rahimli A, Mugisa B, Abdulrazeq F, Aguas R, et al. A participatory modelling approach for investigating the spread of COVID-19 in countries of the Eastern Mediterranean Region to support public health decision-making. BMJ global health. 2021;6: e005207. doi:10.1136/bmjgh-2021-005207

43. Daniel Wolffram, Sam Abbott, Matthias an der Heiden, Sebastian Funk, Felix Günther, Davide Hailer, et al. Collaborative nowcasting of COVID-19 hospitalization incidences in Germany. medRxiv. 2023; 2023.04.17.23288668. doi:10.1101/2023.04.17.23288668

44. Adiga A, Hurt B, Kaur G, Lewis B, Marathe M, Porebski P, et al. A Multi-Team Multi-Model Collaborative Covid-19 Forecasting Hub for India. 2023 Winter Simulation Conference (WSC). 2023. pp. 994–1005. doi:10.1109/WSC60868.2023.10407748

45. Howerton E, Contamin L, Mullany LC, Qin M, Reich NG, Bents S, et al. Evaluation of the US COVID-19 Scenario Modeling Hub for informing pandemic response under uncertainty. Nat Commun. 2023;14: 7260. doi:10.1038/s41467-023-42680-x

46. World Health Organization. Workshop Report: Advanced Analytics to Inform Decision Making During Public Health Emergencies. Berlin, Germany; 2024 Jan. doi:10.25561/108600

47. RespiCast, European Respiratory Diseases Forecasting Hub. RespiCast. [cited 14 Jun 2024]. Available: https://respicast.ecdc.europa.eu/

48. Bracher J, Wolffram D, Deuschel J, Görgen K, Ketterer JL, Ullrich A, et al. A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. Nat Commun. 2021;12: 5173. doi:10.1038/s41467-021-25207-0

49. Loo SL, Howerton E, Contamin L, Smith CP, Borchering RK, Mullany LC, et al. The US COVID-19 and Influenza Scenario Modeling Hubs: Delivering long-term projections to guide policy. Epidemics. 2024;46: 100738. doi:10.1016/j.epidem.2023.100738

50. Gostic KM, McGough L, Baskerville EB, Abbott S, Joshi K, Tedijanto C, et al. Practical considerations for measuring the effective reproductive number, Rt. PLoS Comput Biol. 2020;16: e1008409. doi:10.1371/journal.pcbi.1008409

51. Park SW, Akhmetzhanov AR, Charniga K, Cori A, Davies NG, Dushoff J, et al. Estimating epidemiological delay distributions for infectious diseases. medRxiv; 2024. doi:10.1101/2024.01.12.24301247

52. Adam D. A guide to R - the pandemic's misunderstood metric. Nature. 2020;583: 346–348. doi:10.1038/d41586-020-02009-w

53. Anderson R, Donnelly C, Hollingsworth D, Keeling M, Vegvari C, Baggaley R, et al. Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK: methods

of estimation, data sources, causes of heterogeneity, and use as a guide in policy formulation. The Royal Society; 2020. Available: https://royalsociety.org/news/2020/09/set-c-covid-r-rate/

54. SPI-M-O: Consensus statement on COVID-19, 1 July 2020. In: GOV.UK [Internet]. 17 Jul 2020 [cited 20 May 2024]. Available: https://www.gov.uk/government/publications/spi-m-o-consensus-statement-on-covid-19-1-july-2020

55. Abbott S, Hellewell J, Thompson RN, Sherratt K, Gibbs HP, Bosse NI, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Research. 2020;5: 112. doi:10.12688/wellcomeopenres.16006.1

56. Brockhaus EK, Wolffram D, Stadler T, Osthege M, Mitra T, Littek JM, et al. Why are different estimates of the effective reproductive number so different? A case study on COVID-19 in Germany. PLoS Comput Biol. 2023;19: e1011653. doi:10.1371/journal.pcbi.1011653

57. Park SW, Bolker BM, Champredon D, Earn DJD, Li M, Weitz JS, et al. Reconciling early-outbreak estimates of the basic reproductive number and its uncertainty: framework and applications to the novel coronavirus (SARS-CoV-2) outbreak. J R Soc Interface. 2020;17: 20200144. doi:10.1098/rsif.2020.0144

58. Abbott S, Lison A, Funk S, Pearson C, Gruson H. Epinowcast: Flexible hierarchical nowcasting. Zenodo. 2021. Available: https://samabbott.co.uk/presentations/2023/royal-society-epinowcast.pdf

59. Charniga K, Park SW, Akhmetzhanov AR, Cori A, Dushoff J, Funk S, et al. Best practices for estimating and reporting epidemiological delay distributions of infectious diseases using public health surveillance and healthcare data. 2024. Available: https://hal.science/hal-04572940

60. Lison A, Abbott S, Huisman J, Stadler T. Generative Bayesian modeling to nowcast the effective reproduction number from line list data with missing symptom onset dates. PLoS Comput Biol. 2024;20: e1012021. doi:10.1371/journal.pcbi.1012021

61. Cori A, Donnelly CA, Dorigatti I, Ferguson NM, Fraser C, Garske T, et al. Key data for outbreak evaluation: building on the Ebola experience. Philos Trans R Soc Lond B Biol Sci. 2017;372: 20160371. doi:10.1098/rstb.2016.0371

62. Donnici C, Ilincic N, Cao C, Zhang C, Deveaux G, Clifton D, et al. Timeliness of reporting of SARS-CoV-2 seroprevalence results and their utility for infectious disease surveillance. Epidemics. 2022;41: 100645. doi:10.1016/j.epidem.2022.100645

63. Daly M. COVID-19 and care homes in England: What happened and why? Social Policy & Administration. 2020;54: 985–998. doi:10.1111/spol.12645

64. Polonsky JA, Baidjoe A, Kamvar ZN, Cori A, Durski K, Edmunds WJ, et al. Outbreak analytics: a developing data science for informing the response to emerging pathogens. Philos Trans R Soc Lond B Biol Sci. 2019;374: 20180276. doi:10.1098/rstb.2018.0276

65. Gaythorpe KAM, Fitzjohn RG, Hinsley W, Imai N, Knock ES, Perez Guzman PN, et al. Data pipelines in a public health emergency: The human in the machine. Epidemics. 2023;43:

100676. doi:10.1016/j.epidem.2023.100676

66. Manley H, Park J, Bevan L, Sanchez-Marroquin A, Danelian G, Bayley T, et al. Combining models to generate a consensus effective reproduction number for the COVID-19 epidemic status in England. Epidemiology & Infection. 2024; 1–35. doi:10.1017/S0950268824000347

67. UK Covid-19 Inquiry Archives. INQ000260643 - Witness Statement of Professor Graham Medley (Co-Chair of SPI-M-O), dated 04/09/2023. In: UK Covid-19 Inquiry [Internet]. 12 Oct 2023 [cited 19 May 2024]. Available: https://covid19.public-inquiry.uk/documents/inq000260643-witness-statement-of-professor-graham-medley-co-chair-of-spi-m-o-dated-04-09-2023/

68. Nicholson G, Blangiardo M, Briers M, Diggle PJ, Fjelde TE, Ge H, et al. Interoperability of Statistical Models in Pandemic Preparedness: Principles and Reality. Stat Sci. 2022;37: 183–206. doi:10.1214/22-STS854

69. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proceedings of the Royal Society B: Biological Sciences. 2007;274: 599–604. doi:10.1098/rspb.2006.3754

70. Wearing HJ, Rohani P, Keeling MJ. Appropriate Models for the Management of Infectious Diseases. PLoS Med. 2005;2: e174. doi:10.1371/journal.pmed.0020174

71. Park SW, Champredon D, Weitz JS, Dushoff J. A practical generation-interval-based approach to inferring the strength of epidemics from their speed. Epidemics. 2019;27: 12–18. doi:10.1016/j.epidem.2018.12.002

72. Park SW, Abbott S, Howes A. epidist: Estimate epidemiological delay distributions for infectious diseases. Github; doi:10.5281/zenodo.5637165

73. Cuomo-Dannenburg G, McCain K, McCabe R, Unwin HJT, Doohan P, Nash RK, et al. Marburg virus disease outbreaks, mathematical models, and disease parameters: a systematic review. Lancet Infect Dis. 2024;24: e307–e317. doi:10.1016/S1473-3099(23)00515-7

74. epireview. [cited 21 May 2024]. Available: https://mrc-ide.github.io/epireview/

75. Lambert J, Kucharski A. epiparameter: Library of Epidemiological Parameters. doi:10.5281/zenodo.11110882

76. Claeskens G, Magnus JR, Vasnev AL, Wang W. The forecast combination puzzle: A simple theoretical explanation. Int J Forecast. 2016;32: 754–762. doi:10.1016/j.ijforecast.2015.12.005

77. Ray EL, Wattanachit N, Niemi J, Kanji AH, House K, Cramer EY, et al. Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. medRxiv. 2020; 2020.08.19.20177493. doi:10.1101/2020.08.19.20177493

78. Funk S, Abbott S, Atkins BD, Baguelin M, Baillie JK, Birrell P, et al. Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. medRxiv. 2020; 2020.11.11.20220962. doi:10.1101/2020.11.11.20220962

79. Bracher J, Wolffram D, Deuschel J, Görgen K, Ketterer JL, Ullrich A, et al. National and subnational short-term forecasting of COVID-19 in Germany and Poland during early 2021. Communications Medicine. 2022;2: 1–17. doi:10.1038/s43856-022-00191-8

80. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. PLoS Comput Biol. 2021;17: e1008618. doi:10.1371/journal.pcbi.1008618

81. Wolffram D, Abbott S, Heiden M an der, Funk S, Günther F, Hailer D, et al. Collaborative nowcasting of COVID-19 hospitalization incidences in Germany. PLoS Comput Biol. 2023;19: e1011394. doi:10.1371/journal.pcbi.1011394

82. Fox SJ, Kim M, Meyers LA, Reich NG, Ray EL. Optimizing the number of models included in outbreak forecasting ensembles. medRxiv; 2024. doi:10.1101/2024.01.05.24300909

83. Oidtman RJ, Omodei E, Kraemer MUG, Castañeda-Orjuela CA, Cruz-Rivera E, Misnaza-Castrillón S, et al. Trade-offs between individual and ensemble forecasts of an emerging infectious disease. Nat Commun. 2021;12: 5379. doi:10.1038/s41467-021-25695-0

84. Lerch S, Thorarinsdottir TL, Ravazzolo F, Gneiting T. Forecaster's Dilemma: Extreme Events and Forecast Evaluation. Stat Sci. 2017;32: 106–127. Available: https://www.jstor.org/stable/26408123

85. Bosse NI, Abbott S, Cori A, van Leeuwen E, Bracher J, Funk S. Scoring epidemiological forecasts on transformed scales. PLoS Comput Biol. 2023;19: e1011393. doi:10.1371/journal.pcbi.1011393

86. Gerding A, Reich NG, Rogers B, Ray EL. Evaluating infectious disease forecasts with allocation scoring rules. arXiv; 2023. doi:10.48550/arXiv.2312.16201

87. Marshall M, Parker F, Gardner LM. When are predictions useful? a new method for evaluating epidemic forecasts. medRxiv; 2023. doi:10.1101/2023.06.29.23292042

88. Abbott S, Sherratt K, Bosse N, Gruson H, Bracher J, Funk S. Evaluating an epidemiologically motivated surrogate model of a multi-model ensemble. medRxiv; 2022. doi:10.1101/2022.10.12.22280917

89. Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA. Challenges in Combining Projections from Multiple Climate Models. J Clim. 2010;23: 2739–2758. doi:10.1175/2009JCLI3361.1

90. Taylor KS, Taylor JW. Interval forecasts of weekly incident and cumulative COVID-19 mortality in the United States: A comparison of combining methods. PLoS One. 2022;17: e0266096. doi:10.1371/journal.pone.0266096

91. Nixon K, Jindal S, Parker F, Reich NG, Ghobadi K, Lee EC, et al. An evaluation of prospective COVID-19 modelling studies in the USA: from data to science translation. The Lancet Digital Health. 2022;4: e738–e747. doi:10.1016/S2589-7500(22)00148-0

92. Pollett S, Johansson M, Biggerstaff M, Morton LC, Bazaco SL, Brett Major DM, et al. Identification and evaluation of epidemic prediction and forecasting reporting guidelines: A systematic review and a call for action. Epidemics. 2020;33: 100400. doi:10.1016/j.epidem.2020.100400

93. Ramachandran K, König M, Scharm M, Nguyen TVN, Hermjakob H, Waltemath D, et al. FAIR sharing of reproducible models of epidemic and pandemic forecast. Preprints. 2022. doi:10.20944/preprints202206.0137.v1

94. Crawford MM, Wright G. The value of mass-produced COVID-19 scenarios: A quality evaluation of development processes and scenario content. Technol Forecast Soc Change. 2022;183: 121937. doi:10.1016/j.techfore.2022.121937

95. Juul JL, Græsbøll K, Christiansen LE, Lehmann S. Fixed-time descriptive statistics underestimate extremes of epidemic curve ensembles. Nat Phys. 2021;17: 5–8. doi:10.1038/s41567-020-01121-y

96. Moss R, Fielding JE, Franklin LJ, Stephens N, McVernon J, Dawson P, et al. Epidemic forecasts as a tool for public health: interpretation and (re)calibration. Aust N Z J Public Health. 2018;42: 69–76. doi:10.1111/1753-6405.12750

97. Gaydos DA, Petrasova A, Cobb RC, Meentemeyer RK. Forecasting and control of emerging infectious forest disease through participatory modelling. Philos Trans R Soc Lond B Biol Sci. 2019;374: 20180283. doi:10.1098/rstb.2018.0283

98. Runge MC, Shea K, Howerton E, Yan K, Hochheiser H, Rosenstrom E, et al. Scenario Design for Infectious Disease Projections: Integrating Concepts from Decision Analysis and Experimental Design. medRxiv; 2023. doi:10.1101/2023.10.11.23296887

99. Probert WJM, Shea K, Fonnesbeck CJ, Runge MC, Carpenter TE, Dürr S, et al. Decision-making for foot-and-mouth disease control: Objectives matter. Epidemics. 2016;15: 10–19. doi:10.1016/j.epidem.2015.11.002

100. Ellis DG, Fisher BA. Small group decision making: communication and the group process. 4ème édition. New York: McGraw-Hill; 1994.

101. Humphrey-Murto S, de Wit M. The Delphi method—more research please. J Clin Epidemiol. 2019;106: 136–139. doi:10.1016/j.jclinepi.2018.10.011

102. Morgan O. How decision makers can use quantitative approaches to guide outbreak responses. Philos Trans R Soc Lond B Biol Sci. 2019;374: 20180365. doi:10.1098/rstb.2018.0365

103. Howerton E, Runge MC, Bogich TL, Borchering RK, Inamine H, Lessler J, et al. Context-dependent representation of within- and between-model uncertainty: aggregating probabilistic predictions in infectious disease epidemiology. J R Soc Interface. 2023;20: 20220659. doi:10.1098/rsif.2022.0659

104. Keeling MJ, Dyson L, Tildesley MJ, Hill EM, Moore S. Comparison of the 2021 COVID-19 roadmap projections against public health data in England. Nat Commun. 2022;13: 4924. doi:10.1038/s41467-022-31991-0

105. Bay C, St-Onge G, Davis JT, Chinazzi M, Howerton E, Lessler J, et al. Ensemble2: Scenarios ensembling for communication and performance analysis. Epidemics. 2024;46: 100748. doi:10.1016/j.epidem.2024.100748

106. Li S-L, Bjørnstad ON, Ferrari MJ, Mummah R, Runge MC, Fonnesbeck CJ, et al.

Essential information: Uncertainty and optimal control of Ebola outbreaks. Proceedings of the National Academy of Sciences. 2017;114: 5659–5664. doi:10.1073/pnas.1617482114

107.   Probert WJM, Nicol S, Ferrari MJ, Li S-L, Shea K, Tildesley MJ, et al. Vote-processing rules for combining control recommendations from multiple models. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2022;380: 20210314. doi:10.1098/rsta.2021.0314

108.   Becker AD, Grantz KH, Hegde ST, Bérubé S, Cummings DAT, Wesolowski A. Development and dissemination of infectious disease dynamic transmission models during the COVID-19 pandemic: what can we learn from other pathogens and how can we move forward? The Lancet Digital Health. 2021;3: e41–e50. doi:10.1016/S2589-7500(20)30268-5

109.   Terms of Reference. In: UK Covid-19 Inquiry [Internet]. 20 Jul 2022 [cited 23 May 2024]. Available: https://covid19.public-inquiry.uk/documents/terms-of-reference/

110.   Cairney P. The UK Government's COVID-19 Policy: What Does "Guided by the Science" Mean in Practice? Frontiers in Political Science. 2021;3. doi:10.3389/fpos.2021.624068

111.   McCabe R, Donnelly CA. Disease transmission and control modelling at the science–policy interface. Interface Focus. 2021;11: 20210013. doi:10.1098/rsfs.2021.0013

112.   Atkinson P, Mableson H, Sheard S, Martindale A-M, Solomon T, Borek A, et al. How did UK policymaking in the COVID-19 response use science? Evidence from scientific advisers. Evid Policy. 2022;18: 633–650. doi:10.1332/174426421x16388976414615

113.   Whitty CJM, Collet-Fenson LB. Formal and informal science advice in emergencies: COVID-19 in the UK. Interface Focus. 2021;11: 20210059. doi:10.1098/rsfs.2021.0059

114.   Sarkki S, Niemelä J, Tinch R, van den Hove S, Watt A, Young J. Balancing credibility, relevance and legitimacy: A critical assessment of trade-offs in science–policy interfaces. Sci Public Policy. 2014;41: 194–206. doi:10.1093/scipol/sct046

115.   Oliver K, Kothari A, Mays N. The dark side of coproduction: do the costs outweigh the benefits for health research? Health Res Policy Syst. 2019;17: 33. doi:10.1186/s12961-019-0432-3

116.   Graffy E. Enhancing Policy-Relevance without Burning Up or Burning Out: A Strategy for Scientists. 1999.

117.   Hook DW, Wilsdon JR. The pandemic veneer: COVID-19 research as a mobilisation of collective intelligence by the global research community. Collective Intelligence. 2023;2: 26339137221146482. doi:10.1177/26339137221146482

118.   Herman E, Nicholas D, Watkinson A, Rodríguez-Bravo B, Abrizah A, Boukacem-Zeghmouri C, et al. The impact of the pandemic on early career researchers: What we already know from the internationally published literature. Profesional de la Informacion. 2021;30: 1–16. doi:10.3145/epi.2021.mar.08

119.   Lee KGL, Mennerat A, Lukas D, Dugdale HL, Culina A. The effect of the COVID-19 pandemic on the gender gap in research productivity within academia. Rodgers P, editor. Elife. 2023;12: e85427. doi:10.7554/eLife.85427

120. Marani M, Katul GG, Pan WK, Parolari AJ. Intensity and frequency of extreme novel epidemics. Proc Natl Acad Sci U S A. 2021;118. doi:10.1073/pnas.2105482118

121. Nick Taylor. Review of the use of models in informing disease control policy development and adjustment: a report for DEFRA. Veterinary Epidemiology and Economics Research Unit; 2003 May.

122. Moon S, Sridhar D, Pate MA, Jha AK, Clinton C, Delaunay S, et al. Will Ebola change the game? Ten essential reforms before the next pandemic. The report of the Harvard-LSHTM Independent Panel on the Global Response to Ebola. Lancet. 2015;386: 2204–2221. doi:10.1016/S0140-6736(15)00946-0

123. Rivers C, Chretien J-P, Riley S, Pavlin JA, Woodward A, Brett-Major D, et al. Using "outbreak science" to strengthen the use of models during epidemics. Nat Commun. 2019;10: 3102. doi:10.1038/s41467-019-11067-2

124. Jit M, Ainslie K, Althaus C, Caetano C, Colizza V, Paolotti D, et al. Reflections On Epidemiological Modeling To Inform Policy During The COVID-19 Pandemic In Western Europe, 2020–23. Health Aff. 2023;42: 1630–1636. doi:10.1377/hlthaff.2023.00688

125. Le Rutte EA, Shattock AJ, Zhao C, Jagadesh S, Balać M, Müller SA, et al. A case for ongoing structural support to maximise infectious disease modelling efficiency for future public health emergencies: A modelling perspective. Epidemics. 2024;46: 100734. doi:10.1016/j.epidem.2023.100734

126. Mccaw JM, Plank MJ. THE ROLE OF THE MATHEMATICAL SCIENCES IN SUPPORTING THE COVID-19 RESPONSE IN AUSTRALIA AND NEW ZEALAND. ANZIAM J. 2022;64: 315–337. doi:10.1017/S1446181123000123

127. Tariq M, Haworth-Brockman M, Moghadas SM. Ten years of Pan-InfORM: modelling research for public health in Canada. AIMS public health. 2021;8: 265–274. doi:10.3934/publichealth.2021020

128. Raftery P, Hossain M, Palmer J. An innovative and integrated model for global outbreak response and research - a case study of the UK Public Health Rapid Support Team (UK-PHRST). BMC Public Health. 2021;21: 1378. doi:10.1186/s12889-021-11433-0

129. Rivers C, Martin E, Meyer D, Inglesby TV, Cicero AJ, Cizek J. Modernizing and expanding outbreak science to support better decision making during public health crises: Lessons for COVID-19 and beyond. The Johns Hopkins Center for Health Security; 2020.

130. Consortium of Infectious Disease Modeling Hubs. The hubverse: open tools for collaborative modeling — Hubverse. 2023. Available: https://hubverse.io/en/latest/

131. Muscatello DJ, Chughtai AA, Heywood A, Gardner LM, Heslop DJ, MacIntyre CR. Translation of Real-Time Infectious Disease Modeling into Routine Public Health Practice. Emerg Infect Dis. 2017;23: e161720. doi:10.3201/eid2305.161720

132. Parker WS. Ensemble modeling, uncertainty and robust predictions. WIREs Climate Change. 2013;4: 213–223. doi:10.1002/wcc.220

133. Whitty CJM. The contribution of biological, mathematical, clinical, engineering and social

sciences to combatting the West African Ebola epidemic. Philos Trans R Soc Lond B Biol Sci. 2017;372. doi:10.1098/rstb.2016.0293

134.   Eggo RM, Dawa J, Kucharski AJ, Cucunuba ZM. The importance of local context in COVID-19 models. Nature Computational Science. 2021;1: 6–8. doi:10.1038/s43588-020-00014-7

135.   Rhodes T, Lancaster K, Lees S, Parker M. Modelling the pandemic: attuning models to their contexts. BMJ Global Health. 2020;5: e002914. doi:10.1136/bmjgh-2020-002914

136.   Borchering RK. Modeling of Future COVID-19 Cases, Hospitalizations, and Deaths, by Vaccination Rates and Nonpharmaceutical Intervention Scenarios — United States, April–September 2021. MMWR Morb Mortal Wkly Rep. 2021;70. doi:10.15585/mmwr.mm7019e3

137.   Probert WJM, Jewell CP, Werkman M, Fonnesbeck CJ, Goto Y, Runge MC, et al. Real-time decision-making during emergency disease outbreaks. PLoS Comput Biol. 2018;14: e1006202. doi:10.1371/journal.pcbi.1006202

138.   Kao RR. The role of mathematical modelling in the control of the 2001 FMD epidemic in the UK. Trends Microbiol. 2002;10: 279–286. doi:10.1016/S0966-842X(02)02371-5

139.   Whitty CJM. What makes an academic paper useful for health policy? BMC Med. 2015;13: 301. doi:10.1186/s12916-015-0544-8

140.   Noll NB, Aksamentov I, Druelle V, Badenhorst A, Ronzani B, Jefferies G, et al. COVID-19 Scenarios: an interactive tool to explore the spread and associated morbidity and mortality of SARS-CoV-2. bioRxiv. medRxiv; 2020. doi:10.1101/2020.05.05.20091363

141.   Thorén H, Gerlee P. Model uncertainty, the COVID-19 pandemic, and the science-policy interface. Royal Society Open Science. 2024;11: 230803. doi:10.1098/rsos.230803

142.   Parkhurst J. The Politics of Evidence: From evidence -based policy to the good governance of evidence. Taylor & Francis; 2017. Available: https://library.oapen.org/handle/20.500.12657/31002

143.   Birch J. Science and policy in extremis: the UK's initial response to COVID-19. Eur J Philos Sci. 2021;11: 90. doi:10.1007/s13194-021-00407-z

144.   Benessia A, De Marchi B. When the earth shakes … and science with it. The management and communication of uncertainty in the L'Aquila earthquake. Futures. 2017;91: 35–45. doi:10.1016/j.futures.2016.11.011

145.   Leach M, Scoones I. The social and political lives of zoonotic disease models: Narratives, science and policy. Soc Sci Med. 2013;88: 10–17. doi:10.1016/j.socscimed.2013.03.017

146.   Porter TM. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press; 2020. Available: https://play.google.com/store/books/details?id=53nTDwAAQBAJ

147.   Boden LA, McKendrick IJ. Model-Based Policymaking: A Framework to Promote Ethical

"Good Practice" in Mathematical Modeling for Public Health Policymaking. Frontiers in Public Health. 2017;5. doi:10.3389/fpubh.2017.00068

148. Cash DW, Clark WC, Alcock F, Dickson NM, Eckley N, Guston DH, et al. Knowledge systems for sustainable development. Proc Natl Acad Sci U S A. 2003;100: 8086–8091. doi:10.1073/pnas.1231332100

149. Donnelly CA, Boyd I, Campbell P, Craig C, Vallance P, Walport M, et al. Four principles to make evidence synthesis more useful for policy. In: Nature Publishing Group UK [Internet]. 20 Jun 2018 [cited 8 Jun 2024]. doi:10.1038/d41586-018-05414-4

150. Parker WS. Whose Probabilities? Predicting Climate Change with Ensembles of Models. Philos Sci. 2010;77: 985–997. doi:10.1086/656815

151. Zachreson C, Savulescu J, Shearer FM, Plank MJ, Coghlan S, Miller JC, et al. Ethical frameworks should be applied to computational modelling of infectious disease interventions. PLoS Comput Biol. 2024;20: e1011933. doi:10.1371/journal.pcbi.1011933

Portfolio of selected publications

Contents

Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of COVID-19 in England

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1701639	**Title**	Ms
First Name(s)	Katharine		
Surname/Family Name	Sherratt		
Thesis Title	Collaborative outbreak modelling for decision support: evaluating trade-offs from multi-model combination		
Primary Supervisor	Sebastian Funk		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Phil. Trans. R. Soc. B		
When was the work published?	July 2021		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	PhD by Publication		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	This work originated in real-time response work contributing Rt estimates to SPI-M from March 2020. Development of the Rt estimation and forecasting procedure using EpiNow and EpiNow2 was led by Sam Abbott and Sebastian Funk. I initially contributed data cleaning, processing, and analysis for UK and global surveillance data. I also supported running the end to end Rt estimation pipeline for regular SPI-M submission, and contributed to model development with testing, evaluation, and documentation. In May 2020, I worked together with Sam Abbott and Sebastian Funk on the initial hypothesis of exploring differences in Rt estimation by data source. I led work to develop this idea into this analysis, including developing code to run the Rt estimation pipeline using each data source and visualising the comparative results. I also led the write up of this into a short briefing note, presented to SPI-M first in early June and in three further iterations. I then led work to develop this into a paper. I developed the original work by using public data for estimating delay distributions, in order for all work to be presented in the public domain. I also designed and conducted further analysis to quantify the comparisons of Rt estimates, for example assessing peaks and wave durations. The specific further hypotheses for retrospectively understanding differences between Rt estimates over time were discussed jointly with Sam Abbott and Sebastian Funk, in turn drawing on wider discussions among SPI-M. I led work to source relevant data and formally add comparisons with these additional data sources, for example data on deaths in residential facilities. I led the initial and subsequent drafts of the paper. Sam Abbott and Sebastian Funk provided supervision and review before submission for peer review. All contributions are documented on Github: https://github.com/epiforecasts/rt-comparison-uk-public.

SECTION E

Student Signature	Katharine Sherratt
Date	14 June 2024

Supervisor Signature	Sebastian Funk
Date	14 June 2024

Author for correspondence:
Katharine Sherratt
e-mail: katharine.sherratt@lshtm.ac.uk

Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of COVID-19 in England

Katharine Sherratt[†], Sam Abbott[†], Sophie R. Meakin, Joel Hellewell, James D. Munday, Nikos Bosse, CMMID COVID-19 Working Group, Mark Jit and Sebastian Funk

Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine, London, UK

KS, 0000-0003-2049-3423; SRM, 0000-0002-6385-2652; JH, 0000-0003-2683-0849; NB, 0000-0002-7750-5280; SF, 0000-0002-2842-3406

The time-varying reproduction number (R_t: the average number of secondary infections caused by each infected person) may be used to assess changes in transmission potential during an epidemic. While new infections are not usually observed directly, they can be estimated from data. However, data may be delayed and potentially biased. We investigated the sensitivity of R_t estimates to different data sources representing COVID-19 in England, and we explored how this sensitivity could track epidemic dynamics in population sub-groups. We sourced public data on test-positive cases, hospital admissions and deaths with confirmed COVID-19 in seven regions of England over March through August 2020. We estimated R_t using a model that mapped unobserved infections to each data source. We then compared differences in R_t with the demographic and social context of surveillance data over time. Our estimates of transmission potential varied for each data source, with the relative inconsistency of estimates varying across regions and over time. R_t estimates based on hospital admissions and deaths were more spatio-temporally synchronous than when compared to estimates from all test positives. We found these differences may be linked to biased representations of subpopulations in each data source. These included spatially clustered testing, and where outbreaks in hospitals, care homes, and young age groups reflected the link between age and severity of the disease. We highlight that policy makers could better target interventions by considering the source populations of R_t estimates. Further work should clarify the best way to combine and interpret R_t estimates from different data sources based on the desired use.

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

1. Background

Within six months of its emergence in late 2019, the novel coronavirus SARS-CoV-2 had caused over six million cases of disease (COVID-19) worldwide [1]. Its rapid initial spread and high death rate prompted global policy interventions to prevent continued transmission, with widespread temporary bans on social interaction outside the household [2]. Introducing and adjusting such policy measures depend on a judgement in balancing continued transmission potential with the multidimensional consequences of interventions. It is,

therefore, critical to inform the implementation of policy measures with a clear and timely understanding of ongoing epidemic dynamics [3,4].

In principle, transmission could be tracked by directly recording all new infections. In practice, real-time monitoring of the COVID-19 epidemic relies on surveillance of indicators that are subject to different levels of bias and delay. In England, widely available surveillance data across the population include: (i) the number of positive tests, biased by changing test availability and practice, and delayed by the time from infection to symptom onset (if testing is symptom-based), from symptom onset to a decision to be tested and from test to test result; (ii) the number of new hospital admissions, biased by differential severity that triggers care seeking and hospitalization, and additionally delayed by the time to develop severe diseases; and (iii) the number of new deaths due to COVID-19, biased by the differential risk of death and the exact definition of a COVID-19 death, and further delayed by the time to death.

Each of these indicators provides a different view on the epidemic and therefore contains potentially useful information. However, any interpretation of their behaviour needs to reflect these biases and lags and is best done in combination with the other indicators. One approach that allows this in a principled manner is to use the different datasets to separately track the time-varying reproduction number, R_t, the average number of secondary infections generated by each new infected person [5]. Because R_t quantifies changes in infection levels, it is independent of the level of overall ascertainment as long as this does not change over time or is explicitly accounted for [6]. At the same time, the underlying observations in each data source may result from different lags from infection to observation. However, if these delays are correctly specified then transmission behaviour over time can be consistently compared via estimates of R_t.

Different methods exist to estimate the time-varying reproduction number, and in the UK a number of mathematical and statistical methods have been used to produce estimates used to inform policy [7–9]. Empirical estimates of R_t can be achieved by estimating time-varying patterns in transmission events from mapping to a directly observed time-series indicator of infection such as reported symptomatic cases. This can be based on the probabilistic assignment of transmission pairs [10], the exponential growth rate [11] or the renewal equation [12,13]. Alternatively, R_t can be estimated via mechanistic models that explicitly compartmentalize the disease transmission cycle into stages from susceptible through exposed, infectious and recovered [14,15]. This can include accounting for varying population structures and context-specific biases in observation processes, before fitting to a source of observed cases. Across all methods, key parameters include the time after an infection to the onset of symptoms in the infecting and infected, and the source of data used as a reference point for earlier transmission events [16,17].

In this study, we used a modelling framework based on the renewal equation, adjusting for delays in observation to estimate regional and national reproduction numbers of SARS-Cov-2 across England. The same method was repeated for each of three sources of data that are available in real time. After assessing differences in R_t estimates by data source, we explored why this variation may exist. We compared the

divergence between R_t estimates with spatio-temporal variation in case detection, and the proportion at risk of severe disease, represented by the age distribution of test-positive cases and hospital admissions and the proportion of deaths in care homes.

2. Methods

(a) Data management

Three sources of data provided the basis for our R_t estimates. Time-series case data were available by specimen date of test. This was a de-duplicated dataset of COVID-19 positive tests notified from all National Health Service (NHS) settings (Pillar One of the UK Government's testing strategy) [18] and by commercial partners in community settings outside of healthcare (Pillar Two). Hospital admissions were also available by date of admission if a patient had tested positive prior to admission, or by the day preceding diagnosis if they were tested after admission. Death data were available by date of death and included only those that occurred within 28 days of a positive COVID-19 test in any setting. All data were publicly available and taken from the UK government source [19,20], and were aggregated to the seven English regions used by the NHS.

To provide context for R_t estimates, we sourced weekly data on regional and national test positivity (percentage positive tests of all tests conducted) from Public Health England [21], available as weekly average percentages from 10 May. From the same source, we also identified the age distributions of cases admitted to the hospital and all test-positive cases. Hospital admissions by age were available as age bands with rates per 100 000, so we used regional population data from 2019 [22] to approximate the raw count. We separately sourced daily data on the number of deaths in care homes by region from March 2020, available from 12 April [23]. Care homes are defined as supported living facilities (residential homes, nursing homes, rehabilitation units and assisted living units). Data were available by date of notification, which included an average 2–3 days' lag after the date of death. We also drew on a database that tracked COVID-19 UK policy updates by date and area [24].

(b) R_t estimation

We estimated R_t using EpiNow2 v. 1.2.0, an open-source package in R [13,25,26]. This package implements a Bayesian latent variable approach using the probabilistic programming language Stan [27]. To initialize the model, infections were imputed prior to the first observed case using a log-linear model with priors based on the first week of observed cases. This means that the initial observations both inform the initial parameters and are then also fit, which makes the initial R_t estimates less reliable than later estimates. This was a pragmatic choice to allow the model to be identifiable when only estimating part of the observed epidemic. We explored other parameterizations, but these suffered from poor model identification. For each subsequent time step with observed cases, new infections were imputed using the sum of previous modelled infections weighted by the generation time probability mass function, and combined with an estimate of R_t, to give the prevalence at time t [12]. The generation time was assumed to follow a gamma distribution that was fixed over time but varied between samples, with priors drawn from the literature for the mean and standard deviation [28].

These infection trajectories were mapped to reported case counts (D_t) by convolving over an incubation period distribution and report delay distribution (ξ). We assumed a negative binomial observation model for observed reported case counts (C_t),

with overdispersion ϕ using an exponential prior with mean 1 and mean D_t. We combined this with a multiplicative day of the week effect (ω(tmod7)) with an independent effect for each day of the week. We controlled temporal variation using an approximate Gaussian process [29] with a squared exponential kernel (GP).

In mathematical notation:

$$R_t \sim R_{t-1} \times \mathrm{GP},$$

$$I_t = R_t \sum_\tau w_\tau I_{t-\tau},$$

$$D_t = \sum_\tau \xi_\tau I_{t-\tau},$$

$$C_t \sim \mathrm{NB}(D_{t^{\omega(\mathrm{tmod7})}}, \phi).$$

The length scale and magnitude of the kernel were estimated during model fitting. We used an inverse gamma prior for the length scale, optimizing shape and scale values to give a distribution with 98% of the density between 2 and 21 days, and the prior on the magnitude was standard normal. Each region was fitted independently using Markov-chain Monte Carlo (MCMC). Eight chains were used with a warmup of 1000 samples and 2000 samples post warmup. Convergence was assessed using the R hat diagnostic.

We used a gamma-distributed generation time with mean 3.6 days (standard deviation (s.d.) 0.7), and s.d. of 3.1 days (s.d. 0.8), sourced from [28]. Instead of the incubation period used in the original study (which was based on fewer data points), we refitted using a lognormal incubation period with a mean of 5.2 days (s.d. 1.1) and s.d. of 1.52 days (s.d. 1.1) [30]. This incubation period was also used to convolve from unobserved infections to unobserved symptom onsets (or a corresponding viral load in asymptomatic cases) in the model. When fitting the model, the time interval distributions had independent priors placed on the mean and standard deviation of their respective lognormal distributions.

We estimated both the delay from symptom onset to positive test (either in the community or in hospital) and the delay from symptom onset to death as lognormal distributions using a sub-sampled Bayesian bootstrapping approach (with 100 subsamples each using 250 samples) from given data on these delays. Our delay from the date of onset to date of positive test (either in the community or in hospital) was taken from a publicly available linelist of international cases [31]. We removed countries with outlying delays (Mexico and the Philippines). The resulting delay data had a mean of 4.4 days and s.d. 5.6. Delays for hospital admissions and test positives were treated as having the same delay from infection to onset and observation. For the delay from onset to death we used data taken from a large observational UK study [32]. We re-extracted the delay from confidential raw data, with a mean delay of 14.3 days (s.d. 9.5). There were insufficient data available on the various reporting delays to estimate spatially or temporally varying delays, so they were considered to be static over the course of the epidemic, although we discuss the effects of this assumption. We have also discussed this approach more extensively in [25].

(c) Comparison of R_t estimates

We compared R_t estimates by data source, plotting each by region over time. To avoid the first epidemic wave obscuring visual differences, all plots were limited to the earliest date that any R_t estimate for England crossed below 1 after the peak. We also identified the time at which each R_t estimate fell below 1, the local minima and maxima of median R_t estimates and the number of times in the time-series that each R_t estimate crossed its own median, before comparing these across regions and against the total count of the raw data.

We investigated correlations between R_t estimates and the demographic and social context of transmission. We used linear regression to assess whether the level of raw data count influenced oscillations in R_t. We assessed the influence of local outbreaks using test positivity. We used a 5% threshold for positivity as the level at which testing is either insufficient to keep pace with widespread community transmission [33], or where outbreaks have already been detected and tests targeted to those more likely to be positive. We plotted this against raw data and R_t, and also used linear regression to test the association. We interpreted results in light of known outbreaks and policy changes. We plotted and qualitatively assessed variation in R_t estimates against the age distribution of cases over time, and similarly explored patterns in R_t estimates against the qualitative proportion of cases to all deaths. The latter was not assessed quantitatively due to differences in reference dates [23]. With the exception of fitting the delay from onset to death (held confidentially), code and data to reproduce this analysis are available [34].

3. Results

Across England, the COVID-19 epidemic peaked at 4798 reported test-positive cases (on 22 April 2020), 3099 admissions (1 April 2020) and 975 deaths (8 April 2020) per day (figure 1a). Following the peak, a declining trend continued for daily counts of admissions and deaths, while daily case counts from all reported test-positive cases increased from July and had more than tripled by August (from 571 on 30 June to 1929 on 1 September). Regions followed similar patterns over time to national trends. However, in the North East and Yorkshire, Midlands and North West, the incidence of test-positive cases did not decline to near the count of admissions as in other regions, and also saw a small temporary increase during the overall rise in case of counts in early August.

Following the initial epidemic peak in mid-March 2020, the date at which R_t estimates crossed below 1 varied by both data source and geography (figures 1b and 2). The first region to cross into a declining epidemic was London, on 26 March according to an R_t estimated from deaths (where the lower 90% credible interval (CrI) crossed below 1 on 24 March and the upper CrI on 28 March). However, the data source used to estimate R_t was as important as any regional variation in estimating the earliest date of epidemic decline. R_t estimated from hospital admissions gave the earliest estimate of a declining epidemic, while using all test-positive cases to estimate R_t took the longest time to reach a declining epidemic, in all but one region (East of England). This difference by data source varied by up to 21 days in the North East and Yorkshire, where hospital admissions gave a median R_t estimate under 1 on 1 April (90%CrIs 31 March, 2 April), but the median R_t estimate from test-positive cases crossed 1 on only the 22 April (90%CrIs 1 April, 25 April).

When not undergoing a clear state change, R_t estimates from all data sources oscillate, with oscillations damped when R_t estimates were transitioning to new levels. In England and all NHS regions, test-positive cases showed evidence of larger damped oscillations from July when a state change occurred to R_t over 1. In England, R_t estimates from test-positive cases increased from 0.99 (90%CrI 0.94–1.04) on 30 June to 1.37 (90%CrI 1.31–1.1.44) on 27 August. Meanwhile, the timing and duration of oscillations did not align
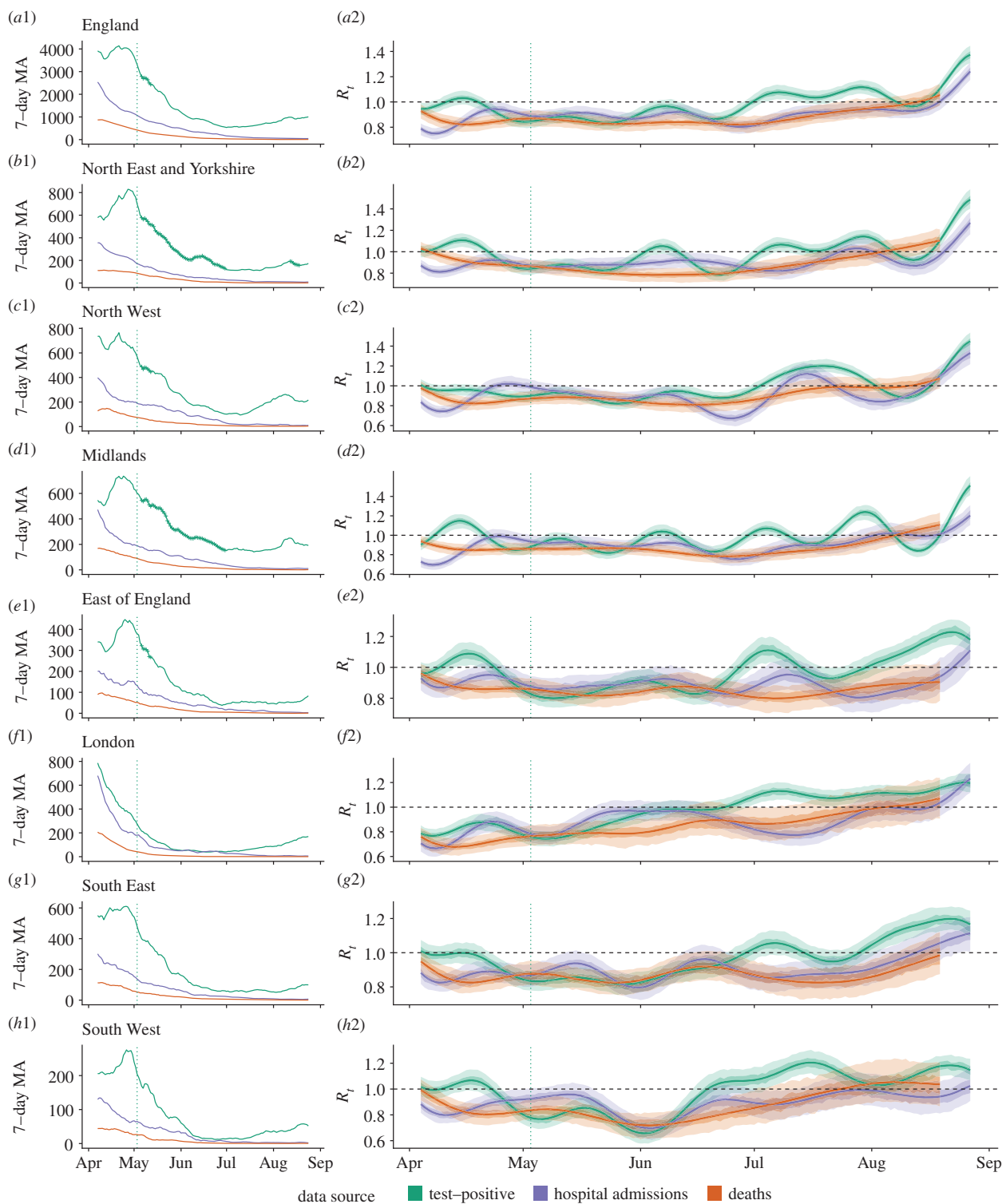
Figure 1. Epidemic dynamics across (*a*) England and (*b*–*h*) seven English National Health Service regions, 5 April through 27 August 2020. (*a*1–*h*1): Daily counts of confirmed cases by data source, as centred 7 days moving average. Counts marked with vertical dashes (on the green lines—see figure parts (*a*1,*b*1,*c*1,*d*1,*e*1)) indicate dates within weeks that averaged greater than 5% test-positivity (positive/all tests per week). Vertical dotted line indicates the start of national mass community testing on 3 May. (*a*2–*h*2): Estimates of R_t (median, with 50% (darker shade) and 90% (lightest shade) credible interval), derived from each data source. Data sources include all test-positive cases, hospital admissions and deaths with a positive test in the previous 28 days.

between R_t estimates (figure 1*b*). In some regions, the difference between R_t estimates was consistent over time, such as between R_t from admissions and deaths in the South East. In other regions such as the Midlands, this was not the case, with the divergence between the R_t estimates from test-positive cases, admissions, and deaths each varying over time. R_t estimates from test-positive cases were the

most likely to differ from estimates derived from other data sources across all regions. Across all regions, R_t estimates from deaths had slower damped oscillations compared to estimates from test-positive cases or hospital admissions. However, oscillations in R_t estimates did not appear to be linked to the level of raw data counts in each source (electronic supplementary material, figure S2).
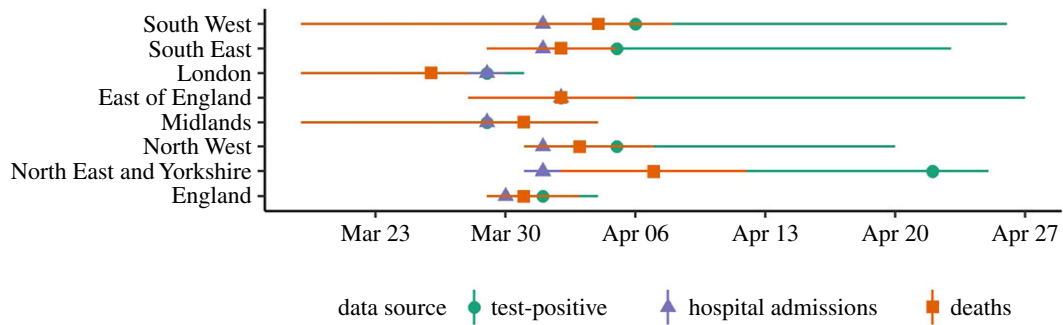
Figure 2. Dates in 2020 on which R_t estimate crossed 1 after first epidemic peak, median and 90% credible interval, by the data source for England and seven NHS regions.

More rapid oscillations in R_t estimates from test-positive cases appeared to be linked to targeted testing of case clusters, seen in high test positivity (electronic supplementary material, table SI2). Both the North East and Yorkshire and the Midlands saw more frequent oscillations in R_t estimates from test-positive cases than other regions. The R_t estimates from cases crossed its own median 10 times over the time-series in both regions, while in all other NHS regions this averaged 6 times, and oscillations in R_t estimates from cases also had a shorter duration in the North East and Yorkshire and the Midlands compared to other regions (electronic supplementary material, table SI1). Across all regions, 84% of weeks with over 5% positivity ($N = 19$) were in the North East and Yorkshire and the Midlands. In these regions, positivity peaked on the week of 9 May 2020 at 14% and 12%, respectively, and overall averaged 6% (95%CI 4.4–7.6%) and 5.9% (95%CI 4.6–7.2%, weeks of 10 May to 22 August), respectively. High test positivity is likely to have resulted from targeted testing among known local outbreaks in these regions. In the Midlands, these included local restrictions and increased testing across Leicester and in a Luton factory (restrictions between 4 and 25 July [35]). In Yorkshire case clusters were detected with local restrictions in Bradford, Calderdale and Kirklees (with restrictions from 5 August [36]).

In England, a divergence between R_t estimated from cases versus R_t estimated from deaths and admissions coincided with a decline in the age distribution among all test-positive cases in England to a younger population (electronic supplementary material, figure SI2A). From mid-April to June 2020, national estimates of R_t from test-positive cases remained around the same level as those from admissions or deaths, while after this, cases diverged to a higher steady state (figure 1a). On 23 May, the median R_t estimated from cases matched that of deaths at 0.83 (both with 90%CrIs 0.78–0.89), but this was followed by a 78 day period before the two estimates were again comparable, on 8 August. Over this period the median R_t estimate from cases was on average 14% higher (95%CI 12–15%). Meanwhile, the share of test-positive cases under age 50 increased from under one-quarter of cases in the week of 28 March (24%, $N = 16$ 185), to accounting for nearly three-quarters of cases by 22 August (77%, $N = 6733$). While the percentage of test-positive cases aged 20–49 increased consistently from April to August, the 0–19 age group experienced a rapid increase over mid-May through July, increasing by a mean 1% each week over 9 May through 1 August (from 4% of 18 774 cases to 14.8% of 5017 cases).

Similarly, R_t estimates from admissions in England oscillated over June through July 2020, potentially linked to the age distribution of hospital admissions. From 0.92 (90%CrI 0.87–0.98) on 11 June, R_t estimated from admissions fell to 0.8 (90%CI 0.75–0.85)) on 27 June. By contrast, this transition was not observed in the R_t estimate based on test-positive cases (figure 1a). Older age groups dominated COVID-19 hospital admissions, where 0–44 years never accounted for more than 12.8% of hospital-based cases (a maximum in the week of 22 August, $N = 690$; electronic supplementary material, figure SI2B). While the proportion of hospital admissions aged 75+ remained steady over May through mid-June, this proportion appeared to oscillate over July through August (standard deviation of weekly percentage at 6.1 over June–August, compared to 5.4 in months March–May). These variations were not seen in the proportion aged 70+ in the test-positive case data, which saw a continuous decline from 30% at the start of June to 7% by August.

R_t estimated from either admissions or deaths experienced near-synchronous local peaks across regions over April and May 2020. We compared this R_t estimated from deaths with its source data and a separate regional dataset of deaths in care homes. In the South East and South West, the R_t estimates from deaths rose over April, with a peak in early May. In the South West, the median R_t estimate from deaths increased by 0.04 from 22 April to 7 May (from 0.8 (90%CrI 0.72–0.88) to 0.84 (90%CrI 0.76–0.95)); and by 0.06 from 17 April to 4 May in the South East (from 0.82 (90% CrI 0.77–0.9) to 0.88 (90%CrI 0.72–0.88)). In both these regions, this early May peak in R_t estimates from deaths coincided with similarly rising R_t estimates from hospital admissions, while the reverse trend was seen in R_t estimates from cases. In all regions, care home deaths peaked over 22–29 April (by date of notification; electronic supplementary material, figure SI3). This was later than regional peaks in the raw count of all deaths in any setting (which peaked between 8 and 16 April, by date of death), even accounting for a 2–3 day reporting lag. This meant that the proportion of deaths from care homes varied over time, where in the South East and South West, deaths in care homes appeared to account for nearly all deaths for at least the period mid-May to July.

4. Discussion

We estimated the time-varying reproduction number for COVID-19 over March through August 2020 across England and English NHS regions, using test-positive cases, hospital admissions and deaths with confirmed COVID-19. Our estimates of transmission potential varied for each of these sources of infections, and the divergence between estimates from each data source was not consistent within or across

regions over time, although estimates based on hospital admissions and deaths were more spatio-temporally synchronous than compared to estimates from cases. We compared differences in R_t estimates to the extent and context of transmission and found that the difference between R_t estimated from cases, admissions and deaths may be linked to uneven rates of testing, the changing age distribution of cases and outbreaks in care home populations.

R_t estimates varied by data source, and the extent of variation itself differed by region and over time. Following the initial epidemic peak in mid-March, the date at which R_t estimates crossed below 1 varied by both data source and geography, following which R_t estimates from all data sources varied when not undergoing a clear state change. The differences in these oscillations by data source may indicate different underlying causes. This implies that each data source was influenced differently by changes in subpopulations over time.

Increasingly rapid oscillations in R_t estimates from test-positive cases were associated with higher test-positivity rates. Increasing test-positivity rates could be an indication of inconsistent community testing, with the observation of an initial rise in transmission amplified by expanded testing and local interventions where a cluster of new, mild cases had been identified [18]. This targeted testing may have driven regionally localized instability in case detection and resulting R_t estimates but may not reflect changes in underlying transmission. This is a limitation of monitoring epidemic dynamics using test-positive surveillance data in areas where testing rates vary across the population and over time. This also suggests that R_t estimates from admissions may be more reliable than that from all test-positive cases for indicating the relative intensity of an epidemic over time [37].

We hypothesized that variations in R_t estimates were also related to changes in the age distribution of cases over time, because age is associated with severity [38,39]. If each data source represented a different sample of this age-severity gradient, and transmission also varied by age or severity, R_t estimates from each source would diverge. Early in the epidemic, tests were largely limited to hospital settings, and disproportionately represented healthcare workers compared to the general population. This sampling bias would be reflected in the R_t from test-positive cases. The early peak in R_t could then represent a substantial separate route of transmission in healthcare settings, in a wave of nosocomial infections [40]. If healthcare workers were less susceptible to severe disease than those older than working age, an early peak in R_t estimated from test-positive cases would not have been represented in R_t estimated from hospital admissions or deaths. Meanwhile, either hospital admissions or deaths data would be more representative of sampling a separate route of transmission among the general population. If infections spread through the general population later than nosocomial infections, then the timing of peaks in R_t estimates from each data source would not have matched.

From late spring, outbreaks in care homes may have contributed to a divergence between R_t estimates from test-positive cases and other data sources. All regions saw a near-synchronous local peak in R_t estimated from hospital admissions over spring, which was not seen in R_t estimated from test-positive cases. This may have reflected the known widespread regional outbreaks in care homes. The care home population is on average older and more clinically vulnerable than the general population, while also being less likely to appear for community testing [41,42]. Increased transmission in care homes would then be seen in an increased R_t from hospital admissions, but not observed in an R_t from test-positive cases.

Similarly, the age-severity gradient may have impacted transmission estimates later in the epidemic when community testing became more widely available. We found that from June 2020 onwards, R_t estimates from all test-positive cases appeared to increasingly diverge away from R_t estimates from admissions and deaths, transitioning into a separate, higher, steady state. This was followed by the observed age distribution of all test-positive cases becoming increasingly younger, while the age distribution of admissions remained approximately level. Because of the severity gradient, this suggested that the R_t estimates from all test-positive cases and admissions were more biased by the relative proportion of younger cases and older cases, respectively, than the R_t estimates from admissions or deaths.

Our analysis was limited where data or modelling assumptions did not reflect underlying differences in transmission. R_t estimates can become increasingly uncertain and unstable with lower case counts. Further, estimated unobserved infections were mapped to reported cases or deaths using two delay distributions: the time from infection to test in the community or hospital, and a longer delay from infection to death. Mis-specification of the priors would have created bias in the temporal distribution of all resulting R_t estimates, with estimated dates of infection and R_t incorrectly shifted too much or too little in time compared to the true infection curve, and decreased accuracy of R_t estimates [43].

We used the same distribution priors for both delays after symptom onset to positive test, and to hospital admission. This may be inaccurate where cases with mild symptoms take longer to present for testing than severe cases presenting for hospital admission, or vice versa. The difference between the two delays over time may also have varied, with a possible decrease in delay to reported tests when mass community testing became available over the summer of 2020. This would have had a differential impact on the accuracy of R_t estimates over time in either direction, which could explain some of the oscillations in R_t estimates from test-positive case data compared to hospital admissions. We had no data over time on delays from symptom onset to reporting in each data source with which to test this hypothesis. However, we have mitigated some of the impact of this by using a sub-sampled bootstrap of the available delay data when estimating the delay distribution priors. This inflated the uncertainty of these priors in line with the hypothesis that they varied over time. This adjustment may be conservative if the delay distributions are stable over time.

Spatial dependence in delay distributions may also have contributed to their mis-specification and increased uncertainty in R_t estimates. We observed that the variation in R_t estimates from admissions and deaths often showed comparable levels and patterns in oscillations over time but were out of phase with each other. This may have been due to using data sources from different populations for each delay estimate. To estimate the delay between symptom onset to either a positive test or hospitalization, we used a linelist of all patients publicly reported globally, which had a mean delay of 5.4 days (s.d. 5.6). This varied only slightly from an early estimate in the UK epidemic, where the delay from

onset to hospitalization had a mean of 5.14 days (s.d. 4.2) in confidential Public Health England (FF100) data [44]. Meanwhile, the same global public linelist contained few records with delay from onset to death, with mean 11.4 (s.d. 16.5). We compared this to confidential UK data from an observational study that had mean delay 14.3 days (s.d. 9.5) [32].

Comparing each type and source of delay, we judged the benefits of using open data to outweigh the minor observed spatial variation of the delay from onset to test or admission, although at the expense of increased uncertainty. However, we judged that the difference in delay from onset to death in the UK compared to public (international) data was sufficiently meaningful to justify using confidential UK data in order to maintain the accuracy of the R_t estimate from deaths. The difference in the geographical source of delay distributions should not have substantially altered our conclusions about discrepancies between central estimates of R_t from either test-positives or admissions, compared to R_t estimated from deaths. However, using the international public linelist for the delay to test or admission may have introduced additional uncertainty around the respective R_t estimates, compared to greater accuracy (reduced uncertainty) in estimates of R_t from deaths based on a UK-specific delay distribution.

The data sources themselves may also have been inaccurate or biased, which would change the representation of the population we have assumed here. For example, we excluded data from other nations of the UK (Wales, Scotland and Northern Ireland) in our analysis, as these differed in both availability over time and in data collection and reporting practices [19,45]. English regional data may also contain bias where new parts of the population might be under focus for testing efforts, or the population characteristics of hospital admissions from COVID-19 may have changed over time with changes in clinical criteria or hospital capacity for admission. This would mean that an R_t estimate from these data sources would represent different source populations over time, limiting our ability to reliably compare against R_t estimates from other data sources. Where possible we highlighted this by comparing R_t estimates to known biases and changes in case detection and reporting.

Our approach is unable to make strong causal conclusions about varying transmission, and assumptions about sampling and the representation of subpopulations remain implicit. Alternatively, varying epidemics in subpopulations could have been addressed with mechanistic models that explicitly consider transmission in different settings and are fitted to multiple data sources. However, these require additional assumptions, detailed data to parameterize and may be time-consuming to develop. In the absence of data, the number of assumptions required for these models can introduce inherent structural biases. Our approach contains few structural assumptions and therefore may be more robust when data are sparse, or information is required in real-time.

We conclude that when estimating R_t, the choice of data source should be guided by the policy context in which the estimates will be used and interpreted. This work highlights that there is no clear superior choice of data source, while R_t estimates are sensitive to assumptions about the underlying population of each data source. This means that both producers and users of R_t estimates should understand relevant biases in the data source's population sampling strategy, such as by community case detection or patient severity,

before drawing conclusions about transmission in the population as a whole.

We also recommend presenting concurrent R_t estimates jointly, rather than pooling estimates of R_t from different data sources. Pooling estimates would both suffer from unclear weighting and lose useful information about variation in subpopulation transmission. Although the reconstruction of the underlying transmission process from the reporting processes is robust, it is unclear how weights would be assigned based on likelihood to estimates from different data sources. Further, the variation in concurrent R_t estimates provides more information about population transmission than any single estimate, when considered in light of the sampling biases of each data source. This additional information can be useful to identify transmission intensity by subpopulation where access to high quality disaggregated data may not be available in real time. While this can be difficult to interpret without specific knowledge of population structure and dynamics, this information would be lost altogether in a single or pooled estimate of R_t. By contrast, if the policy were to be based on either a single or an averaged R_t estimate, it would be unclear what any recommendation should be and for whom.

Future work could explore systematic differences in the influence of data sources on R_t estimates by extending the comparison of R_t by data source to other countries or infectious diseases. Additionally, work should also clarify the potential for comparing R_t estimates in real-time tracking of outbreaks and explore the inconsistencies in case detection over time and space, where a cluster of cases leads to a highly localized expansion of community testing, creating an uneven spatial bias in transmission estimates. These findings may be used to improve R_t estimation and identify findings of use for epidemic control. Based on the work presented here we now provide R_t estimates, updated each day, for test positive cases, admissions, and deaths in each NHS region and in England. Our estimates are visualized on our website, are available for download, and are produced using publicly accessible code [46,47].

Tracking differences by data source can improve understanding of variation in testing bias in data collection, highlight outbreaks in new subpopulations, indicate differential rates of transmission among vulnerable populations and clarify the strengths and limitations of each data source. Our approach can quickly identify such patterns in developing epidemics that might require further investigation and early policy intervention. Our method is simple to deploy and scale over time and space using existing open-source tools, and all code and estimates used in this work are available to be used or re-purposed by others.

5. In context

In the UK, public policy and the media have prominently used the effective reproduction number (R_t) of COVID-19 to summarise ongoing pandemic transmission. Several teams in the UK have been contributing estimates of R_t that are aggregated into a consensus range, but the methods, approaches, and data sources for estimating transmission have varied among teams and over time. For example, data sources could, amongst others, include counts of test-positive cases, hospital admissions, or deaths due to COVID-19. In

our team's submissions to the Scientific Pandemic Influenza Group on Modelling (SPI-M) from March onwards, we saw that even when using a consistent method, R_t estimates were not a single, clear-cut number, but varied depending on the source of data.

In late May, we started to explore whether these differences in transmission estimates from each data source could be a policy-relevant indicator of biased data sampling and subpopulation epidemics. We first presented a summary of the differences in our team's R_t estimates by data source to SPI-M as a short note in early June. From June onwards we used all three data sources to estimate R_t and contributed them separately to the weekly reproduction number estimates published by SPI-M and considered by the Scientific Advisory Group for Emergencies (SAGE). Over this time, we have adapted our work to support the changing UK policy context. This has meant there are several differences in available data, methods, and implications of this work between the time we first generated the SPI-M report and the time of this publication.

As COVID-19 data became more openly accessible, we started to publish a daily comparison of UK R_t estimates by data source (epiforecasts.io/covid/posts/national/united-kingdom). This had initially been impossible as there were very few sources of public subnational data. Thanks to the Public Health England dashboard (coronavirus.data.gov.uk), public data sources for England increased in both quantity and quality and from October we were able to produce subnational R_t estimates using a variety of public data sources. We felt that presenting these estimates publicly would be useful given the high level of interest in the government's claimed use of R_t as a policy decision tool.

Between generating the original SPI-M submission and this publication, we significantly developed and improved the software we have built to estimate R_t ("EpiNow2"). We continue to refine our methods for estimating R_t, although the improved methods did not substantially change the trend or direction of differences between estimates and our resulting conclusions.

Our interpretation of the differences in R_t estimates has changed over time as we saw new evidence for concentrated transmission in subpopulations. In the earliest paper presented to SPI-M, discussion centred on the likely effects of hospital-acquired infection and testing availability on differences between R_t from test-positives compared to admissions or deaths over March and May. However, increasing evidence for a widespread and severe epidemic in care homes provided an alternative explanation for such differences. We realised that, even without disaggregated data by age or residence, simply identifying the differences in R_t estimates could have been an early indicator of the epidemic in this vulnerable subpopulation. We therefore continued to track these differences, which once again became wider over the summer as transmission moved between age groups after restrictions were lifted and mass testing became available.

Most importantly, we continue to find new insights into the state of the UK pandemic from comparing R_t estimates. One of the clearest trends we have seen in varying R_t estimates by data source has followed from the National Health Service vaccination campaign. R_t estimates from deaths are now consistently below those from hospitalisations and cases. This is a strong indicator of the positive impact of vaccination, and an encouraging further use for this work.

References

1. European Centre for Disease Prevention and Control. 2020 COVID-19 situation update worldwide, as of 6 June 2020. See https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases.

2. World Health Organisation. 2020 Strengthening and adjusting public health measures throughout the COVID-19 transition phases. Policy considerations for the WHO European Region. WHO Regional Office for Europe; 2020 May. See http://www.euro.who.int/en/countries/hungary/publications/strengthening-and-adjusting-public-health-measures-throughout-the-covid-19-transition-phases.-policy-considerations-for-the-who-european-region,-24-april-2020.

3. HM Government. 2020 Our Plan to Rebuild: The UK Government's COVID-19 recovery strategy. 2020 May. (CP:239). See https://www.gov.uk/government/publications/our-plan-to-rebuild-the-uk-governments-covid-19-recovery-strategy.

4. Michael Parker. 2020 Ethics and value judgements involved in developing policy for lifting physical distancing measures. 2020 Apr. (SAGE 30). See https://www.gov.uk/government/publications/ethics-and-value-judgements-involved-in-developing-policy-for-lifting-physical-distancing-measures-29-april-2020.

5. Thompson RN. 2020 Epidemiological models are important tools for guiding COVID-19 interventions. *BMC Med.* **18**, 152. (doi:10.1186/s12916-020-01628-4)

6. Pitzer VE, Chitwood M, Havumaki J, Menzies NA, Perniciaro S, Warren JL, Weinberger DM, Cohen T. 2020 The impact of changes in diagnostic testing practices on estimates of COVID-19 transmission in the United States. *medRxiv.* 2020.04.20.20073338. See (doi:10.1101/2020.04.20.20073338)

7. The Royal Society. 2020 Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK. See https://royalsociety.org/-/media/policy/projects/set-c/set-covid-19-R-estimates.pdf.

8. Scientific Advisory Group for Emergencies. 2020 Scientific evidence supporting the government response to coronavirus (COVID-19). See https://www.gov.uk/government/collections/scientific-evidence-supporting-the-government-response-to-coronavirus-covid-19.

9. Funk S *et al.* 2020 Short-term forecasts to inform the response to the COVID-19 epidemic in the UK. *medRxiv.* 2020 Jan 1;2020.11.11.20220962. (doi:10.1101/2020.11.11.20220962)

10. Wallinga J, Teunis P. 2004 Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* **160**, 509–516. (doi:10.1093/aje/kwh255)

11. Wallinga J, Lipsitch M. 2007 How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B* **274**, 599–604. (doi:10.1098/rspb.2006.3754)

12. 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)

13. Abbott S Hellewell J, Sherratt K, Gostic K, Hickson J, Badr HS, DeWitt M, Thompson R, EpiForecasts, Funk S. 2020 EpiNow2: estimate real-time case counts and time-varying epidemiological parameters. Available at https://epiforecasts.io/EpiNow2/dev. (doi:10.5281/zenodo.3957489)

14. Keeling MJ *et al.* 2020 Fitting to the UK COVID-19 outbreak, short-term forecasts and estimating the reproductive number. *medRxiv.* 2020.08.04.20163782. (doi:10.1101/2020.08.04.20163782)

15. Kucharski AJ *et al.* 2020 Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet Infect. Dis.* **20**, 553–558. (doi:10.1016/S1473-3099(20)30144-4)

16. Cori A *et al.* 2017 Key data for outbreak evaluation: building on the Ebola experience. *Phil. Trans. R. Soc. B* **372**, 20160371. (doi:10.1098/rstb.2016.0371)

17. Gostic KM *et al.* 2020 Practical considerations for measuring the effective reproductive number, Rt. *medRxiv.* 2020.06.18.20134858. (doi:10.1101/2020.06.18.20134858)

18. Department of Health and Social Care. 2020 Coronavirus (COVID-19) - Scaling up our testing programmes. See https://www.gov.uk/government/publications/coronavirus-covid-19-scaling-up-testing-programmes.

19. Public Health England, NHSX. 2020 Coronavirus (COVID-19) in the UK. See https://coronavirus.data.gov.uk/about-data.

20. Abbott S, Sherratt K, Bevan J, Gibbs H, Hellewell J, Munday J, Campbell P, Funk S. 2020 COVIDregionaldata: subnational data for the COVID-19 outbreak. Version 0.7.0. Available at: https://github.com/epiforecasts/covidregionaldata/releases. See (doi:10.5281/zenodo.3957539).

21. Public Health England. 2020 National COVID-19 surveillance reports. GOV.UK. See https://www.gov.uk/government/publications/national-covid-19-surveillance-reports.

22. Office for National Statistics. 2020 Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland. See https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland.

23. Office for National Statistics. 2020 Number of deaths in care homes notified to the Care Quality Commission, England. See https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/numberofdeathsincarehomesnotifiedtothecarequalitycommissionengland.

24. The Health Foundation. 2020 COVID-19 policy tracker. A timeline of national policy and health system responses to COVID-19 in England. See https://www.health.org.uk/news-and-comment/charts-and-infographics/covid-19-policy-tracker.

25. Abbott S, Hickson J, Ellis P, Badr HS, Munday JD, Allen J, Funk S. 2020 EpiNow2 v1.2.0: estimate realtime case counts and time-varying epidemiological parameters. Available at: https://github.com/epiforecasts/EpiNow2/releases. See (doi:10.5281/zenodo.4088545).

26. R Core Team. 2020 *R: a language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. See http://www.R-project.org/.

27. Stan Development Team. 2020 RStan: the R interface to Stan. See http://mc-stan.org/.

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. Riutort-Mayol G, Bürkner P-C, Andersen MR, Solin A, Vehtari A. 2020 Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. ArXiv Prepr. arXiv:2004.11408. See https://arxiv.org/abs/2004.11408v1.

30. 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)

31. Xu B *et al.* 2020 Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci. Data.* **7**, 106. (doi:10.1038/s41597-020-0448-0)

32. Docherty AB *et al.* 2020 Features of 16,749 hospitalised UK patients with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol.

medRxiv. See http://medrxiv.org/lookup/doi/10.1101/2020.04.23.20076042.

33. World Health Organisation. 2020 Considerations in adjusting public health and social measures in the context of COVID-19. World Health Organisation; 2020 May. Report No.: WHO/2019-nCoV/Adjusting_PH_measures/2020. See https://www.who.int/publications-detail-redirect/public-health-criteria-to-adjust-public-health-and-social-measures-in-the-context-of-covid-19.

34. Sherratt K, Abbott S. 2020 Rt comparison by data source in the UK. GitHub; 2020. (GitHub). See https://github.com/epiforecasts/rt-comparison-uk-public.

35. Acts of Parliament. 2020 The Health Protection (Coronavirus, Restrictions) (Leicester) Regulations 2020. 2020 No. 685 Jul 4, 2020. See https://www.legislation.gov.uk/uksi/2020/685/.

36. Acts of Parliament. 2020 The Health Protection (Coronavirus, Restrictions on Gatherings) (North of England) Regulations 2020. 2020 No. 828 Aug 5, 2020. See https://www.legislation.gov.uk/uksi/2020/828.

37. Smith DR, Duval A, Pouwels KB, Guillemot D, Fernandes J, Huynh B-T, Temime L, Opatowski L. 2020 How best to use limited tests? Improving COVID-19 surveillance in long-term care. (Scientific Advisory Group for Emergencies (SAGE)). See http://medrxiv.org/lookup/doi/10.1101/2020.04.19.20071639.

38. Verity R *et al*. 2020 Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect. Dis.* **20**, 669–677. (doi:10.1016/S1473-3099(20)30243-7)

39. Levin AT, Meyerowitz-Katz G, Owusu-Boaitey N, Cochran KB, Walsh SP. 2020 Assessing the age specificity of infection fatality rates for COVID-19: Systematic review, meta-analysis, and public policy implications. *Eur. J. Epidemiol.* **35**, 1123–1138. (doi:10.1007/s10654-020-00698-1).

40. Evans S, Agnew E, Vynnycky E, Robotham J. 2020 The impact of testing and infection prevention and control strategies on within-hospital transmission dynamics of COVID-19 in English hospitals. *medRxiv*. See https://www.medrxiv.org/content/10.1101/2020.05.12.20095562v2.

41. Gordon AL *et al.* 2020 Commentary: COVID in care homes—challenges and dilemmas in healthcare delivery. *Age Ageing* **49**, 701–705. (doi:10.1093/ageing/afaa113)

42. Scientific Advisory Group for Emergencies. 2020 SAGE 33 minutes: Coronavirus (COVID-19) response, 5 May 2020. See https://www.gov.uk/government/publications/sage-minutes-coronavirus-covid-19-response-5-may-2020.

43. Guenther F, Bender A, Katz K, Kuechenhoff H, Hoehle M. 2021 Nowcasting the COVID-19 Pandemic in Bavaria. *Biometrical J.* **63**, 490–502. (doi:10.1002/bimj.202000112)

44. Pellis L *et al.* 2020 Challenges in control of COVID-19: short doubling time and long delay to effect of interventions. Eprint ArXiv200400117 Q-Bio. See http://arxiv.org/abs/2004.00117.

45. Department of Health and Social Care. 2020 COVID-19 testing data: methodology note. See https://www.gov.uk/government/publications/coronavirus-covid-19-testing-data-methodology/covid-19-testing-data-methodology-note.

46. Abbott S, Hickson J, Ellis P, Badr HS, Munday J, Allen J. *et al.* 2020 National and subnational estimates of the time-varying reproduction number for COVID-19. See https://github.com/epiforecasts/covid-rt-estimates.

47. Abbott S, Hickson J, Ellis P, Badr HS, Allen J, Munday JD *et al.* 2020 COVID-19: National and Subnational estimates for the United Kingdom. Epiforecasts. See https://epiforecasts.io/covid/posts/national/united-kingdom/.

Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of Covid-19 in England

Katharine Sherratt*, Sam Abbott*, Sophie R Meakin, Joel Hellewell, James D Munday, Nikos Bosse, CMMID Covid-19 working group, Mark Jit, Sebastian Funk

Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine
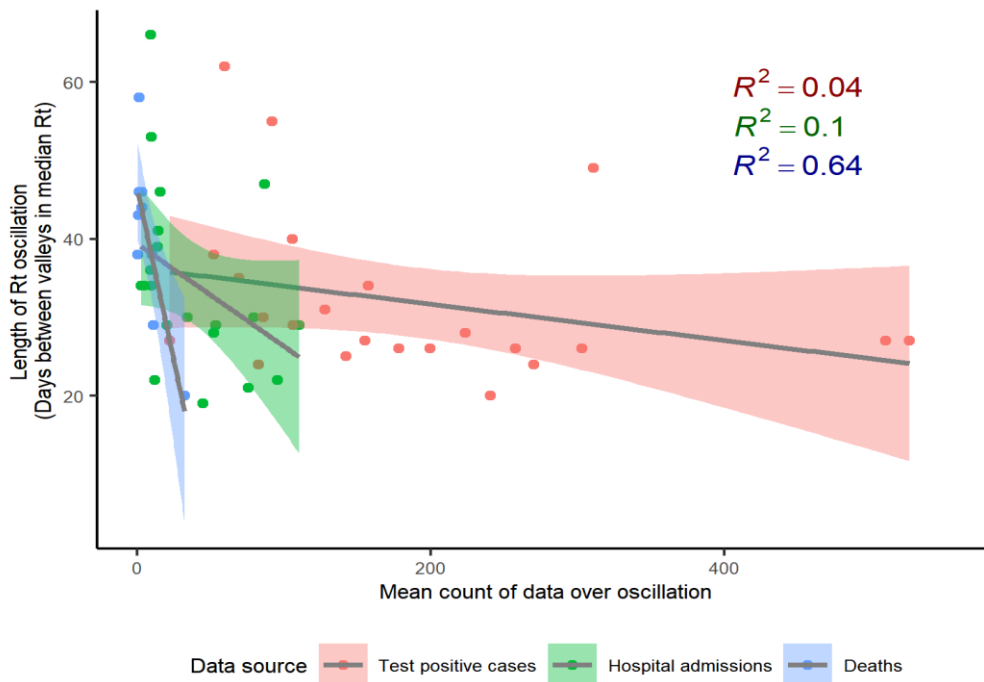
* Equal contributors

Figure SI1. Duration of each R_t oscillation against the mean of data over the oscillation. Shown with fitted linear models by data source. Scatter points represent each oscillation across 7 English NHS regions. For R_t derived from deaths, longer durations of oscillations appeared to be related to lower raw counts (coefficient: -0.88, R2 0.64), but this had a very small sample size (N oscillations = 9) and this relationship was not meaningful for cases or hospital admissions.
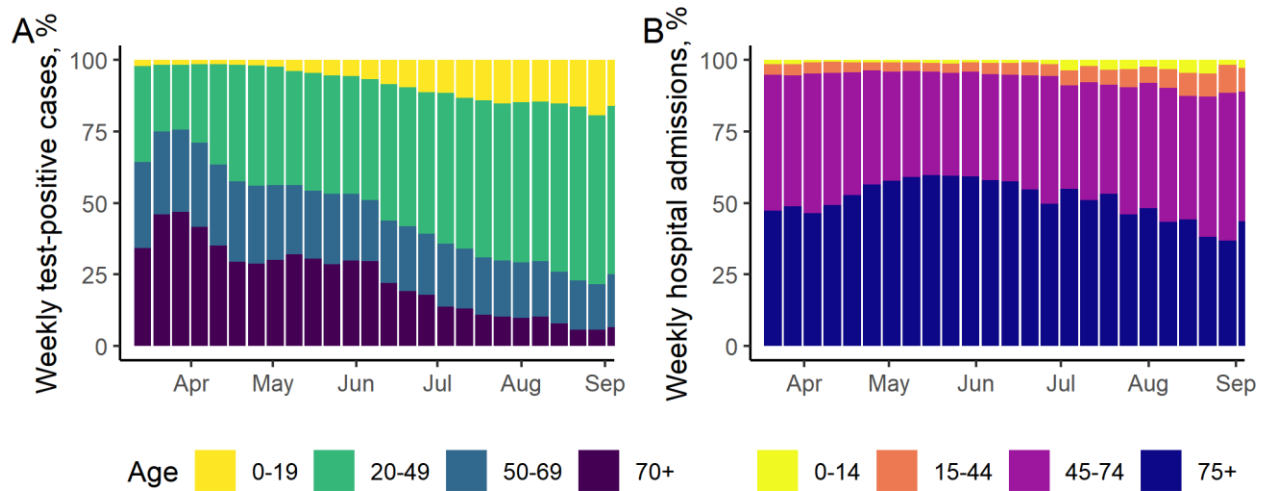
Figure SI2. Weekly percentage of cases in England by age, among all test-positive cases (A) and newly diagnosed hospital admissions (B).
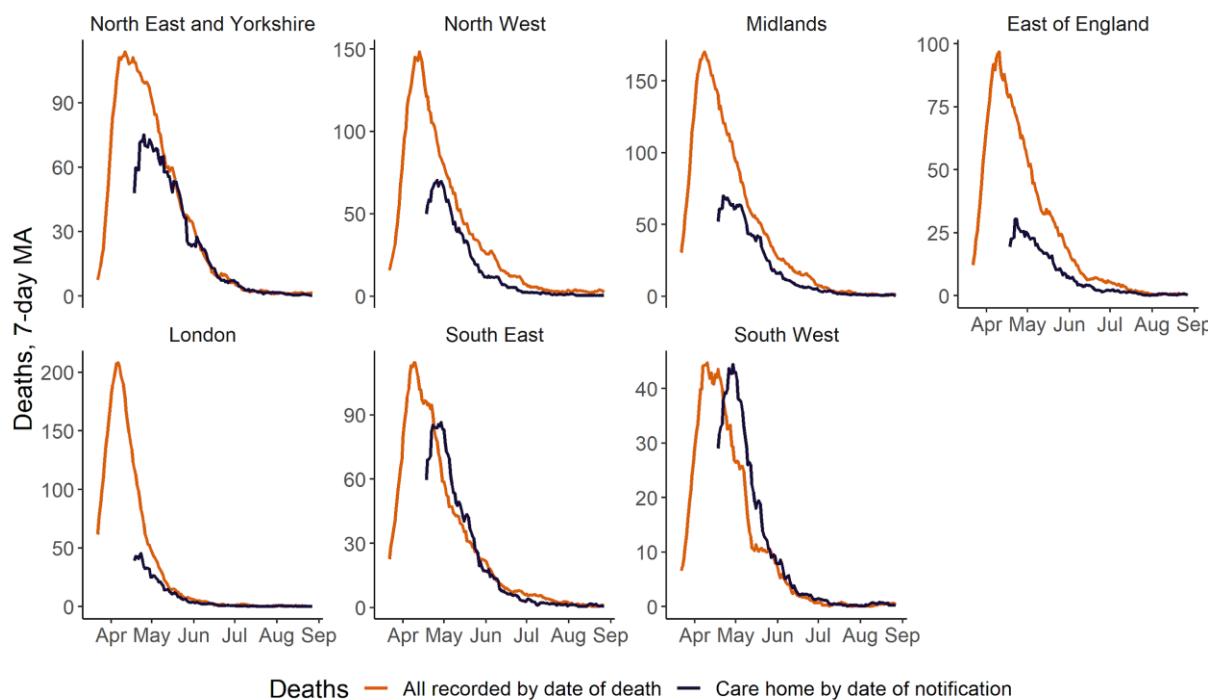
Figure SI3. Deaths in all settings by date of death, and deaths in care homes by date of notification to Clinical Quality Commission, shown by English NHS region, April to August 2020. Date of notification is typically 2-3 days after the date of death [22]. Counts are shown as a seven-day moving average. Care homes data reported from April 2020.

Region	Data source for Rt estimation	Earliest date median Rt <1	Median Rt after first wave* (90% CrI)	Minimum median Rt after first wave* (date, median [90%CrI])	Maximum median Rt after first wave* (date, median [90%CrI])	Number of days where median Rt crossed series median	Mean number of days between troughs** (mean, 95% CI)
England	*Test-positive cases*	01 Apr	0.97 (0.85-1.11)	23 May (0.83 [0.78-0.89])	27 Aug (1.37 [1.31-1.44])	6	32 (21-44, n=4)
	Hospital admissions	30 Mar	0.9 (0.81-0.98)	08 Apr (0.75 [0.7-0.79])	27 Aug (1.24 [1.17-1.31])	8	27 (24-29, n=3)
	Deaths	31 Mar	0.86 (0.8-0.98)	27 Jun (0.82 [0.76-0.89])	19 Aug (1.06 [0.97-1.14])	4	40 (n=1)
North East and Yorkshire	*Test-positive cases*	22 Apr	0.96 (0.82-1.13)	22 Jun (0.79 [0.73-0.85])	27 Aug (1.49 [1.4-1.58])	10	26 (23-29, n=5)
	Hospital admissions	01 Apr	0.89 (0.82-1.01)	09 Apr (0.81 [0.76-0.87])	27 Aug (1.27 [1.16-1.38])	8	31 (23-39, n=4)
	Deaths	07 Apr	0.86 (0.77-1.02)	04 Jun (0.78 [0.72-0.85])	19 Aug (1.11 [1-1.21])	2	NA
North West	*Test-positive cases*	05 Apr	0.94 (0.85-1.19)	26 May (0.82 [0.76-0.87])	27 Aug (1.45 [1.37-1.54])	4	34 (22-46, n=3)
	Hospital admissions	01 Apr	0.91 (0.74-1.09)	25 Jun (0.67 [0.6-0.75])	27 Aug (1.33 [1.22-1.44])	8	38 (30-47, n=3)
	Deaths	03 Apr	0.88 (0.8-1.02)	15 Jun (0.81 [0.73-0.88])	19 Aug (1.08 [0.97-1.17])	4	NA
Midlands	*Test-positive cases*	29 Mar	0.96 (0.84-1.17)	23 May (0.82 [0.75-0.88])	27 Aug (1.51 [1.41-1.61])	12	27 (25-28, n=5)
	Hospital admissions	29 Mar	0.91 (0.76-1.03)	07 Apr (0.7 [0.64-0.75])	27 Aug (1.2 [1.11-1.31])	6	26 (22-30, n=4)
	Deaths	31 Mar	0.86 (0.78-0.99)	24 Jun (0.78 [0.71-0.85])	19 Aug (1.11 [1.01-1.23])	8	33 (12-54, n=2)
East of England	*Test-positive cases*	02 Apr	0.96 (0.82-1.15)	09 May (0.8 [0.74-0.86])	22 Aug (1.23 [1.15-1.31])	6	36 (34-38, n=3)
	Hospital admissions	02 Apr	0.89 (0.81-0.99)	28 Jul (0.8 [0.71-0.9])	27 Aug (1.11 [0.99-1.24])	8	36 (28-45, n=3)
	Deaths	02 Apr	0.85 (0.78-0.94)	06 Jul (0.8 [0.71-0.88])	02 Apr (0.99 [0.93-1.05])	6	38 (24-51, n=2)
London	*Test-positive cases*	29 Mar	0.98 (0.77-1.15)	07 May (0.74 [0.68-0.8])	25 Aug (1.2 [1.13-1.27])	4	35 (27-43, n=2)
	Hospital admissions	29 Mar	0.89 (0.73-1.04)	08 Apr (0.67 [0.6-0.73])	27 Aug (1.23 [1.12-1.36])	4	47 (16-78, n=2)
	Deaths	26 Mar	0.84 (0.7-1.02)	13 Apr (0.68 [0.61-0.75])	19 Aug (1.07 [0.91-1.24])	4	41 (36-46, n=2)
South East	*Test-positive cases*	05 Apr	0.96 (0.82-1.15)	28 May (0.82 [0.75-0.87])	22 Aug (1.2 [1.14-1.27])	4	35 (19-51, n=3)
	Hospital admissions	01 Apr	0.89 (0.81-1.02)	31 May (0.8 [0.73-0.87])	27 Aug (1.11 [1-1.21])	8	29 (23-34, n=3)
	Deaths	02 Apr	0.86 (0.79-0.95)	25 May (0.82 [0.76-0.9])	19 Aug (0.99 [0.87-1.12])	6	48 (32-64, n=2)
South West	*Test-positive cases*	06 Apr	1.02 (0.73-1.18)	02 Jun (0.66 [0.58-0.74])	16 Jul (1.21 [1.12-1.3])	6	39 (21-58, n=3)
	Hospital admissions	01 Apr	0.9 (0.76-1.02)	03 Jun (0.7 [0.61-0.78])	27 Aug (1.02 [0.91-1.18])	6	44 (28-59, n=2)
	Deaths	04 Apr	0.85 (0.72-1.06)	03 Jun (0.72 [0.62-0.81])	07 Aug (1.05 [0.9-1.24])	2	43 (n=1)

Table SI1. Key summary statistics for R_t from when R_t crossed 1, indicating the end of the first wave of the epidemic, to the most recent estimate. 90%CrI = credible interval (quantiles around median); 95%CI = confidence interval (standard error around mean)

*End of first wave = earliest date Rt <1. Last estimate of Rt from cases and admissions: 28th August, from deaths: 19th August 2020

**Troughs defined as local minima within a sequential fall and rise in the median Rt estimate

Effect on regional average duration between troughs in Rt	Influence of % test positive		
Source of Rt	Intercept (95% CI)	Coefficient (95% CI)	Adjusted R^2
Test positive cases	39 (36.7 to 42)	-2.2 (-2.8 to -1.4)	0.87 (p <0.001)
Hospital admissions	41 (29.5 to 52)	-2.1 (-5.5 to 1.2)	0.18 (p 0.2)
Deaths	45.4 (38 to 52.5)	-3.7 (-4.7 to 0.37)	0.48 (p 0.07)

Table SI2. Results of individual linear regressions of test positive on each Rt estimate from test-positive cases, hospital admissions, and deaths separately. Test positivity is the average % positive of all tests conducted. Data for seven English NHS regions and England (N=8).

Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations

Sherratt K, Gruson H, Grah R, Johnson H, Niehus R, Prasse B, Sandmann F, Deuschel J, Wolffram D, Abbott S, Ullrich A, Gibson G, Ray EL, Reich NG, Sheldon D, Wang Y, Wattanachit N, Wang L, Trnka J, Obozinski G, Sun T, Thanou D, Pottier L, Krymova E, Meinke JH, Barbarossa MV, Leithauser N, Mohring J, Schneider J, Wlazlo J, Fuhrmann J, Lange B, Rodiah I, Baccam P, Gurung H, Stage S, Suchoski B, Budzinski J, Walraven R, Villanueva I, Tucek V, Smid M, Zajicek M, Perez Alvarez C, Reina B, Bosse NI, Meakin SR, Castro L, Fairchild G, Michaud I, Osthus D, Alaimo Di Loro P, Maruotti A, Eclerova V, Kraus A, Kraus D, Pribylova L, Dimitris B, Li ML, Saksham S, Dehning J, Mohr S, Priesemann V, Redlarski G, Bejar B, Ardenghi G, Parolini N, Ziarelli G, Bock W, Heyder S, Hotz T, Singh DE, Guzman-Merino M, Aznarte JL, Morina D, Alonso S, Alvarez E, Lopez D, Prats C, Burgard JP, Rodloff A, Zimmermann T, Kuhlmann A, Zibert J, Pennoni F, Divino F, Catala M, Lovison G, Giudici P, Tarantino B, Bartolucci F, Jona Lasinio G, Mingione M, Farcomeni A, Srivastava A, Montero-Manso P, Adiga A, Hurt B, Lewis B, Marathe M, Porebski P, Venkatramanan S, Bartczuk RP, Dreger F, Gambin A, Gogolewski K, Gruziel-Slomka M, Krupa B, Moszyński A, Niedzielewski K, Nowosielski J, Radwan M, Rakowski F, Semeniuk M, Szczurek E, Zielinski J, Kisielewski J, Pabjan B, Holger K, Kheifetz Y, Scholz M, Przemyslaw B, Bodych M, Filinski M, Idzikowski R, Krueger T, Ozanski T, Bracher J, Funk S. Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. Elife. 2023 Apr 21;12:e81916. doi: 10.7554/eLife.81916.

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1701639	**Title**	Ms
First Name(s)	Katharine		
Surname/Family Name	Sherratt		
Thesis Title	Collaborative outbreak modelling for decision support: evaluating trade-offs from multi-model combination		
Primary Supervisor	Sebastian Funk		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	eLife		
When was the work published?	June 2023		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	PhD by Publication		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	My contribution to this work is twofold. Firstly, from January 2021 I jointly developed, built, and maintained the infrastructure for the European COVID-19 Forecast Hub, together with Sebastian Funk, Johannes Bracher, Hugo Gruson, and many collaborators. I led the adaptation of code for forecast processing and validation; data sourcing, processing, and validation; all documentation; and the initial ensemble. I also contributed code for the forecast evaluation procedures, and website design. This drew on existing software from both the US and Germany/Poland COVID-19 Forecast Hubs, and was planned and reviewed jointly with Sebastian Funk. From March 2021, I led work supporting forecasters to contribute to the Hub, and continuously maintained the weekly updating of data sources, forecasts, ensemble, and evaluation. Second, I led work developing this paper focussing on the performance of the resulting Hub ensemble. I led this work from conceptualisation to analysis and drafting, with supervision by Sebastian Funk. I designed the plan for this work and developed all code to conduct analysis, including accessing forecasts, observed data, and evaluation scores; descriptive summaries of individual model and ensemble performance; and visualisations. Sebastian Funk contributed code for weighted ensembles and their evaluation. I wrote the first and subsequent drafts of the paper. We invited initial feedback from close collaborators at the ECDC and in Germany (Rene Niehus, Rok Grah, Frank Sandmann, Bastian Prasse, and Johannes Bracher). We invited a further round of review from all individuals who were named in metadata of the forecasts contributed to the Hub, who had therefore contributed component forecasts to the ensemble. I led work on all subsequent drafts, submission for peer review, and revisions. All contributions are documented on Github: https://github.com/epiforecasts/euro-hub-ensemble.

SECTION E

Student Signature	Katharine Sherratt
Date	14 June 2024

Supervisor Signature	Sebastian Funk
Date	14 June 2024

Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations

Katharine Sherratt[1]*, Hugo Gruson[1], Rok Grah[2], Helen Johnson[2], Rene Niehus[2], Bastian Prasse[2], Frank Sandmann[2], Jannik Deuschel[3], Daniel Wolffram[3], Sam Abbott[1], Alexander Ullrich[4], Graham Gibson[5], Evan L Ray[5], Nicholas G Reich[5], Daniel Sheldon[5], Yijin Wang[5], Nutcha Wattanachit[5], Lijing Wang[6], Jan Trnka[7], Guillaume Obozinski[8], Tao Sun[8], Dorina Thanou[8], Loic Pottier[9], Ekaterina Krymova[10], Jan H Meinke[11], Maria Vittoria Barbarossa[12], Neele Leithauser[13], Jan Mohring[13], Johanna Schneider[13], Jaroslaw Wlazlo[13], Jan Fuhrmann[14], Berit Lange[15], Isti Rodiah[15], Prasith Baccam[16], Heidi Gurung[16], Steven Stage[17], Bradley Suchoski[16], Jozef Budzinski[18], Robert Walraven[19], Inmaculada Villanueva[20], Vit Tucek[21], Martin Smid[22], Milan Zajicek[22], Cesar Perez Alvarez[23], Borja Reina[23], Nikos I Bosse[1], Sophie R Meakin[1], Lauren Castro[24], Geoffrey Fairchild[24], Isaac Michaud[24], Dave Osthus[24], Pierfrancesco Alaimo Di Loro[25], Antonello Maruotti[25], Veronika Eclerova[26], Andrea Kraus[26], David Kraus[26], Lenka Pribylova[26], Bertsimas Dimitris[27], Michael Lingzhi Li[27], Soni Saksham[27], Jonas Dehning[28], Sebastian Mohr[28], Viola Priesemann[28], Grzegorz Redlarski[29], Benjamin Bejar[30], Giovanni Ardenghi[31], Nicola Parolini[31], Giovanni Ziarelli[31], Wolfgang Bock[32], Stefan Heyder[33], Thomas Hotz[33], David E Singh[34], Miguel Guzman-Merino[34], Jose L Aznarte[35], David Morina[36], Sergio Alonso[37], Enric Alvarez[37], Daniel Lopez[37], Clara Prats[37], Jan Pablo Burgard[38], Arne Rodloff[39], Tom Zimmermann[39], Alexander Kuhlmann[40], Janez Zibert[41], Fulvia Pennoni[42], Fabio Divino[43], Marti Catala[44], Gianfranco Lovison[45], Paolo Giudici[46], Barbara Tarantino[46], Francesco Bartolucci[47], Giovanna Jona Lasinio[48], Marco Mingione[48], Alessio Farcomeni[49], Ajitesh Srivastava[50], Pablo Montero-Manso[51], Aniruddha Adiga[52], Benjamin Hurt[52], Bryan Lewis[52], Madhav Marathe[52], Przemyslaw Porebski[52], Srinivasan Venkatramanan[52], Rafal P Bartczuk[53], Filip Dreger[53], Anna Gambin[53], Krzysztof Gogolewski[53], Magdalena Gruziel-Slomka[53], Bartosz Krupa[53], Antoni Moszyński[53], Karol Niedzielewski[53], Jedrzej Nowosielski[53], Maciej Radwan[53], Franciszek Rakowski[53], Marcin Semeniuk[53], Ewa Szczurek[53], Jakub Zielinski[53], Jan Kisielewski[53,54], Barbara Pabjan[55], Kirsten Holger[56], Yuri Kheifetz[56], Markus Scholz[56], Biecek Przemyslaw[57]†, Marcin Bodych[58], Maciej Filinski[58], Radoslaw Idzikowski[58], Tyll Krueger[58], Tomasz Ozanski[58], Johannes Bracher[3], Sebastian Funk[1]

*For correspondence:
katharine.sherratt@lshtm.ac.uk

Present address: †University of Warsaw, Warsaw, Poland

[1]London School of Hygiene & Tropical Medicine, London, United Kingdom; [2]European Centre for Disease Prevention and Control (ECDC), Stockholm, Sweden; [3]Karlsruhe Institute of Technology, Karlsruhe, Germany; [4]Robert Koch Institute, Berlin, Germany; [5]University of Massachusetts Amherst, Amherst, United States; [6]Boston Children's Hospital and Harvard Medical School, Boston, United States; [7]Third Faculty of Medicine, Charles University, Prague, Czech Republic; [8]Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland; [9]Éducation nationale, Valbonne, France; [10]Eidgenossische Technische Hochschule, Zurich, Switzerland; [11]Forschungszentrum

Jülich GmbH, Jülich, Germany; [12]Frankfurt Institute for Advanced Studies, Frankfurt, Germany; [13]Fraunhofer Institute for Industrial Mathematics, Kaiserslautern, Germany; [14]Heidelberg University, Heidelberg, Germany; [15]Helmholtz Centre for Infection Research, Braunschweig, Germany; [16]IEM, Inc, Bel Air, United States; [17]IEM, Inc, Baton Rouge, United States; [18]Independent researcher, Vienna, Austria; [19]Independent researcher, Davis, United States; [20]Institut d'Investigacions Biomèdiques August Pi i Sunyer, Universitat Pompeu Fabra, Barcelona, Spain; [21]Institute of Computer Science of the CAS, Prague, Czech Republic; [22]Institute of Information Theory and Automation of the CAS, Prague, Czech Republic; [23]Inverence, Madrid, Spain; [24]Los Alamos National Laboratory, Los Alamos, United States; [25]LUMSA University, Rome, Italy; [26]Masaryk University, Brno, Czech Republic; [27]Massachusetts Institute of Technology, Cambridge, United States; [28]Max-Planck-Institut für Dynamik und Selbstorganisation, Göttingen, Germany; [29]Medical University of Gdansk, Gdańsk, Poland; [30]Paul Scherrer Institute, Villigen, Switzerland; [31]Politecnico di Milano, Milan, Italy; [32]Technical University of Kaiserlautern, Kaiserslautern, Germany; [33]Technische Universität Ilmenau, Ilmenau, Germany; [34]Universidad Carlos III de Madrid, Leganes, Spain; [35]Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain; [36]Universitat de Barcelona, Barcelona, Spain; [37]Universitat Politècnica de Catalunya, Barcelona, Spain; [38]Universitat Trier, Trier, Germany; [39]University of Cologne, Cologne, Germany; [40]University of Halle, Halle, Germany; [41]University of Ljubljana, Ljubljana, Slovenia; [42]University of Milano-Bicocca, Milano, Italy; [43]University of Molise, Pesche, Italy; [44]University of Oxford, Oxford, United Kingdom; [45]University of Palermo, Palermo, Italy; [46]University of Pavia, Pavia, Italy; [47]University of Perugia, Perugia, Italy; [48]University of Rome "La Sapienza", Rome, Italy; [49]University of Rome "Tor Vergata", Rome, Italy; [50]University of Southern California, Los Angeles, United States; [51]University of Sydney, Sydney, Australia; [52]University of Virginia, Charlottesville, United States; [53]University of Warsaw, Warsaw, Poland; [54]University of Bialystok, Warsaw, Poland; [55]University of Wroclaw, Wroclaw, Poland; [56]Universtät Leipzig, Leipzig, Germany; [57]Warsaw University of Technology, Warsaw, Poland; [58]Wroclaw University of Science and Technology, Wroclaw, Poland

Abstract

Background: Short-term forecasts of infectious disease burden can contribute to situational awareness and aid capacity planning. Based on best practice in other fields and recent insights in infectious disease epidemiology, one can maximise the predictive performance of such forecasts if multiple models are combined into an ensemble. Here, we report on the performance of ensembles in predicting COVID-19 cases and deaths across Europe between 08 March 2021 and 07 March 2022.

Methods: We used open-source tools to develop a public European COVID-19 Forecast Hub. We invited groups globally to contribute weekly forecasts for COVID-19 cases and deaths reported by a standardised source for 32 countries over the next 1–4 weeks. Teams submitted forecasts from March 2021 using standardised quantiles of the predictive distribution. Each week we created an ensemble forecast, where each predictive quantile was calculated as the equally-weighted average (initially the mean and then from 26th July the median) of all individual models' predictive quantiles. We measured the performance of each model using the relative Weighted Interval Score (WIS), comparing models' forecast accuracy relative to all other models. We retrospectively explored alternative methods for ensemble forecasts, including weighted averages based on models' past predictive performance.

Results: Over 52 weeks, we collected forecasts from 48 unique models. We evaluated 29 models' forecast scores in comparison to the ensemble model. We found a weekly ensemble had a consistently strong performance across countries over time. Across all horizons and locations, the

ensemble performed better on relative WIS than 83% of participating models' forecasts of incident cases (with a total N=886 predictions from 23 unique models), and 91% of participating models' forecasts of deaths (N=763 predictions from 20 models). Across a 1–4 week time horizon, ensemble performance declined with longer forecast periods when forecasting cases, but remained stable over 4 weeks for incident death forecasts. In every forecast across 32 countries, the ensemble outperformed most contributing models when forecasting either cases or deaths, frequently outperforming all of its individual component models. Among several choices of ensemble methods we found that the most influential and best choice was to use a median average of models instead of using the mean, regardless of methods of weighting component forecast models.

Conclusions: Our results support the use of combining forecasts from individual models into an ensemble in order to improve predictive performance across epidemiological targets and populations during infectious disease epidemics. Our findings further suggest that median ensemble methods yield better predictive performance more than ones based on means. Our findings also highlight that forecast consumers should place more weight on incident death forecasts than incident case forecasts at forecast horizons greater than 2 weeks.

Editor's evaluation

This large-scale collaborative study is a timely contribution that will be of interest to researchers working in the fields of infectious disease forecasting and epidemic control. This paper provides a comprehensive evaluation of the predictive skills of real-time COVID-19 forecasting models in

Europe. The conclusions of the paper are well supported by the data and are consistent with findings from studies in other countries.

Introduction

Epidemiological forecasts make quantitative statements about a disease outcome in the near future. Forecasting targets can include measures of prevalent or incident disease and its severity, for some population over a specified time horizon. Researchers, policy makers, and the general public have used such forecasts to understand and respond to the global outbreaks of COVID-19 (*Van Basshuysen et al., 2021*; *CDC, 2020*; *European Centre for Disease Prevention and Control, 2021c*). At the same time, forecasters use a variety of methods and models for creating and publishing forecasts, varying in both defining the forecast outcome and in reporting the probability distribution of outcomes (*Zelner et al., 2021*; *James et al., 2021*).

Within Europe, comparing forecasts across both models and countries can support a range of national policy needs simultaneously. European public health professionals operate across national, regional, and continental scales, with strong existing policy networks in addition to rich patterns of cross-border migration influencing epidemic dynamics. A majority of European countries also cooperate in setting policy with inter-governmental European bodies such as the European Centre for Disease Prevention and Control (ECDC). In this case, a consistent approach to forecasting across the continent as a whole can support accurately informing cross-European monitoring, analysis, and guidance (*European Centre for Disease Prevention and Control, 2021c*). At a regional level, multi-country forecasts can support a better understanding of the impact of regional migration networks. Meanwhile, where there is limited capacity for infectious disease forecasting at a national level, forecasters generating multi-country results can provide an otherwise-unavailable opportunity for forecasts to inform national situational awareness. Some independent forecasting models have sought to address this by producing multi-country results (*Aguas et al., 2020*; *Adib et al., 2021*; *Agosto and Giudici, 2020*; *Agosto et al., 2021*).

Variation in forecast methods and presentation makes it difficult to compare predictive performance between forecast models, and from there to derive objective arguments for using one forecast over another. This confounds the selection of a single representative forecast and reduces the reliability of the evidence base for decisions based on forecasts. A 'forecast hub' is a centralised effort to improve the transparency and usefulness of forecasts, by standardising and collating the work of many independent teams producing forecasts (*Reich et al., 2019a*). A hub sets a commonly agreed-upon structure for forecast targets, such as type of disease event, spatio-temporal units, or the set of quantiles of the probability distribution to include from probabilistic forecasts. For instance, a hub may collect predictions of the total number of cases reported in a given country for each day in the next 2 weeks. Forecasters can adopt this format and contribute forecasts for centralised storage in the public domain.

This shared infrastructure allows forecasts produced from diverse teams and methods to be visualised and quantitatively compared on a like-for-like basis, which can strengthen public and policy use of disease forecasts. The underlying approach to creating a forecast hub was pioneered in climate modelling and adapted for collaborative epidemiological forecasts of dengue (*Johansson et al., 2019*) and influenza in the USA (*Reich et al., 2019a*; *Reich et al., 2019b*). This infrastructure was adapted for forecasts of short-term COVID-19 cases and deaths in the US (*Cramer et al., 2021a*; *Ray et al., 2020*), prompting similar efforts in some European countries (*Bracher et al., 2021c*; *Funk et al., 2020*; *Bicher et al., 2020*).

Standardising forecasts allows for combining multiple forecasts into a single ensemble with the potential for an improved predictive performance. Evidence from previous efforts in multi-model infectious disease forecasting suggests that forecasts from an ensemble of models can be consistently high performing compared to any one of the component models (*Johansson et al., 2019*; *Reich et al., 2019b*; *Viboud et al., 2018*). Elsewhere, weather forecasting has a long-standing use of building ensembles of models using diverse methods with standardised data and formatting in order to improve performance (*Buizza, 2019*; *Moran et al., 2016*).

The European COVID-19 Forecast Hub (*European Covid-19 Forecast Hub, 2023d*) is a project to collate short-term forecasts of COVID-19 across 32 countries in the European region. The Hub is

funded and supported by the ECDC, with the primary aim to provide reliable information about the near-term epidemiology of the COVID-19 pandemic to the research and policy communities and the general public (*European Centre for Disease Prevention and Control, 2021c*). Second, the Hub aims to create infrastructure for storing and analysing epidemiological forecasts made in real time by diverse research teams and methods across Europe. Third, the Hub aims to maintain a community of infectious disease modellers underpinned by open science principles.

We started formally collating and combining contributions to the European Forecast Hub in March 2021. Here, we investigate the predictive performance of an ensemble of all forecasts contributed to the Hub in real time each week, as well as the performance of variations of ensemble methods created retrospectively.

Materials and methods

We developed infrastructure to host and analyse prospective forecasts of COVID-19 cases and deaths. The infrastructure is compatible with equivalent research software from the US (*Cramer et al., 2021c*; *Wang et al., 2021*) and German and Polish COVID-19 (*Bracher et al., 2020*) Forecast Hubs, and easy to replicate for new forecasting collaborations.

Forecast targets and models

We sought forecasts for the incidence of COVID-19 as the total reported number of cases and deaths per week. We considered forecasts for 32 countries in Europe, including all countries of the European Union, European Free Trade Area, and the United Kingdom. We compared forecasts against observed data reported for each country by Johns Hopkins University (JHU, *Dong et al., 2020*). JHU data sources included a mix of national and aggregated subnational data. We aggregated incidence over the Morbidity and Mortality Weekly Report (MMWR) epidemiological week definition of Sunday through Saturday.

Teams could express their uncertainty around any single forecast target by submitting predictions for up to 23 quantiles (from 0.01 to 0.99) of the predictive probability distribution. Teams could also submit a single point forecast. At the first submission, we asked teams to add a pre-specified set of metadata briefly describing the forecasting team and methods (provided online and in supplementary information). No restrictions were placed on who could submit forecasts. To increase participation, we actively contacted known forecasting teams across Europe and the US and advertised among the ECDC network. Teams submitted a broad spectrum of model types, ranging from mechanistic to empirical models, agent-based and statistical models, and ensembles of multiple quantitative or qualitative models (described at *European Covid-19 Forecast Hub, 2023a*). We maintain a full project specification with a detailed submissions protocol (*European Covid-19 Forecast Hub, 2023c*).

We collected forecasts submitted weekly in real time over the 52-week period from 08 March 2021 to 07 March 2022. Teams submitted at latest 2 days after the complete dataset for the latest forecasting week became available each Sunday. We implemented an automated validation programme to check that each new forecast conformed to standardised formatting. Forecast validation ensured a monotonic increase of predictions with each increasing quantile, integer-valued non-negative counts of predicted cases, as well as consistent date and location definitions.

Each week we used all available valid forecasts to create a weekly real-time ensemble model (referred to as 'the ensemble' from here on), for each of the 256 possible forecast targets: incident cases and deaths in 32 locations over the following one through 4 weeks. The ensemble method was an unweighted average of all models' forecast values, at each predictive quantile for a given location, target, and horizon. From 08 March 2021, we used the arithmetic mean. However we noticed that including highly anomalous forecasts in a mean ensemble produced extremely wide uncertainty. To mitigate this, from 26th July 2021 onwards the ensemble instead used a median of all predictive quantiles.

We created an open and publicly accessible interface to the forecasts and ensemble, including an online visualisation tool allowing viewers to see past data and interact with one or multiple forecasts for each country and target for up to 4 weeks' horizon (*European Covid-19 Forecast Hub, 2023b*). All forecasts, metadata, and evaluations are freely available and held on Github (*European Covid-19 Forecast Hub, 2023d*) (archived in real-time at *Sherratt, 2022*), and Zoltar, a platform for hosting

epidemiological forecasts (*EpiForecasts, 2021*; *Reich et al., 2021*). In the codebase for this study (*covid19-forecast-hub-europe, 2022*) we provide a simple method and instructions for downloading and preparing these data for analysis using R. We encourage other researchers to freely use and adapt this to support their own analyses.

Forecast evaluation

In this study, we focused only on the comparative performance of forecasting models relative to each other. Performance in absolute terms is available on the Hub website (*European Covid-19 Forecast Hub, 2023b*). For each model, we assessed calibration and overall predictive performance. We evaluated all previous forecasts against actual observed values for each model, stratified by the forecast horizon, location, and target. We calculated scores using the *scoringutils* R package (*Bosse et al., 2023*). We removed any forecast surrounding (both the week of, and the first week after) a strongly anomalous data point. We defined anomalous as where any subsequent data release revised that data point by over 5%.

To investigate calibration, we assessed coverage as the correspondence between the forecast probability of an event and the observed frequency of that event. This usage follows previous work in epidemic forecasting (*Bracher et al., 2021a*), and is related to the concept of reliability for binary forecasts. We established the accuracy of each model's prediction boundaries as the coverage of the predictive intervals. We calculated coverage at a given interval level k, where $k \in [0, 1]$, as the proportion p of observations that fell within the corresponding central predictive intervals across locations and forecast dates. A perfectly calibrated model would have $p = k$ at all 11 levels (corresponding to 22 quantiles excluding the median). An underconfident model at level k would have $p > k$, i.e. more observations fall within a given interval than expected. In contrast, an overconfident model at level k would have $p < k$, i.e. fewer observations fall within a given interval than expected. We here focus on coverage at the $k = 0.5$ and $k = 0.95$ levels.

We also assessed the overall predictive performance of weekly forecasts using the Weighted Interval Score~(WIS) across all available quantiles. The WIS represents a parsimonious approach to scoring forecasts based on uncertainty represented as forecast values across a set of quantiles (*Bracher et al., 2021a*), and is a strictly proper scoring rule, that is, it is optimal for predictions that come from the data-generating model. As a consequence, the WIS encourages forecasters to report predictions representing their true belief about the future (*Gneiting and Raftery, 2007*). Each forecast for a given location and date is scored based on an observed count of weekly incidence, the median of the predictive distribution and the predictive upper and lower quantiles corresponding to the central predictive interval level.

Not all models provided forecasts for all locations and dates, and we needed to compare predictive performance in the face of various levels of missingness across each forecast target. Therefore we calculated a relative WIS. This is a measure of forecast performance which takes into account that different teams may not cover the same set of forecast targets (i.e. weeks and locations). The relative WIS is computed using a *pairwise comparison tournament* where for each pair of models a mean score ratio is computed based on the set of shared targets. The relative WIS of a model with respect to another model is then the ratio of their respective geometric mean of the mean score ratios, such that smaller values indicate better performance.

We scaled the relative WIS of each model with the relative WIS of a baseline model, for each forecast target, location, date, and horizon. The baseline model assumes case or death counts stay the same as the latest data point over all future horizons, with expanding uncertainty, described previously in *Cramer et al., 2021b*. In this study, we report the relative WIS of each model with respect to the baseline model.

Retrospective ensemble methods

We retrospectively explored alternative methods for combining forecasts for each target at each week. A natural way to combine probability distributions available in the quantile format *Genest, 1992* used here is

$$F^{-1}(\alpha) = \sum_{i=1}^{n} w_i F_i^{-1}(\alpha),$$

Where $F_1 \ldots F_n$ are the cumulative distribution functions of the individual probability distributions (in our case, the predictive distributions of each forecast model i contributed to the hub), w_i are a set of weights in $[0, 1]$; and α are the quantile levels, such that following notation introduced in *Genest, 1992*,

$$F^{-1}(\alpha) = \inf\{t : F_i(t) \geq \alpha\}.$$

Different ensemble choices then mainly translate to the choice of weights w_i. An arithmetic mean ensemble uses weights at $w_i = 1/n$, where all weights are equal and sum up to 1.

Alternatively, we can choose a set of weights to apply to forecasts before they are combined. Numerous options exist for choosing these weights with the aim to maximise predictive performance, including choosing weights to reflect each forecast's past performance (thereby moving from an untrained to a trained ensemble). A straightforward choice is so-called inverse score weighting. In this case, the weights are calculated as

$$w_i = \frac{1}{S_i},$$

where S_i reflects the forecasting skill calculated as the relative WIS of forecaster i, calculated over all available model data, and normalised so that weights sum to 1. This method of weighting was found in the US to outperform unweighted scores during some time periods (*Taylor and Taylor, 2023*) but this was not confirmed in a similar study in Germany and Poland (*Bracher et al., 2021c*).

When constructing ensembles from quantile means, a single outlier can have an oversized effect on the ensemble forecast. Previous research has found that a median ensemble, replacing the arithmetic mean of each quantile with a median of the same values, yields competitive performance while
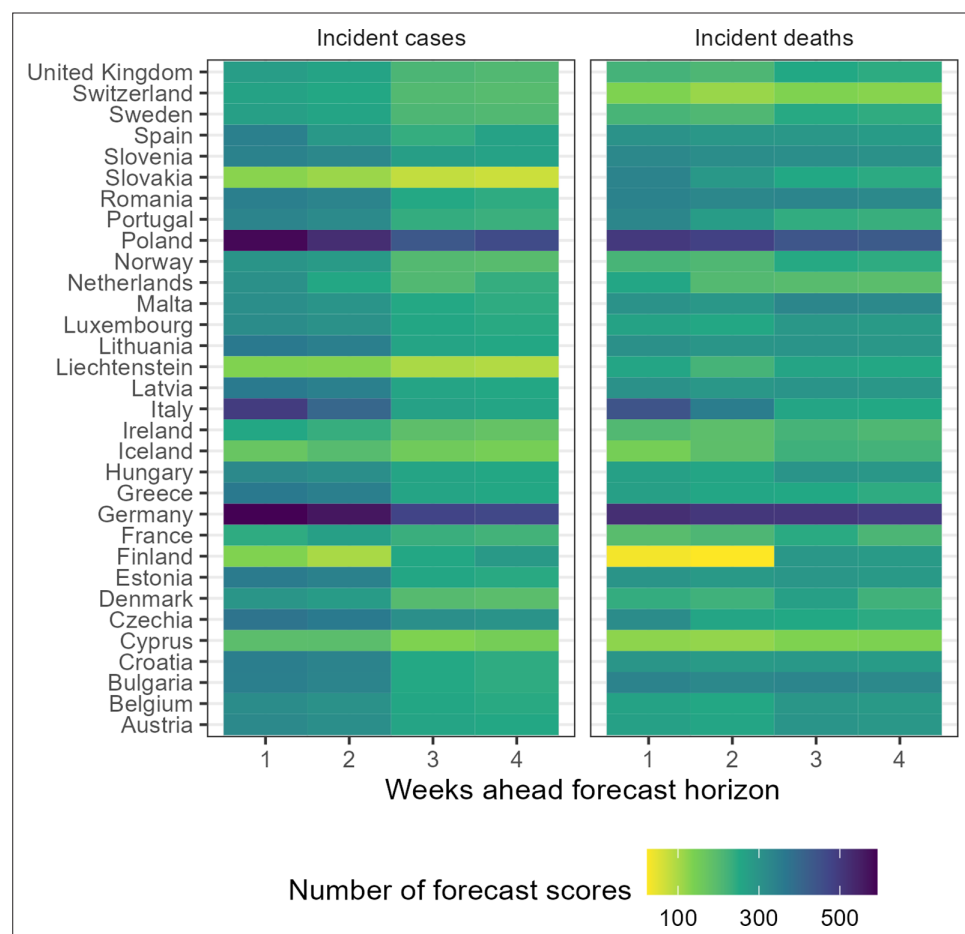

Figure 1. Total number of forecasts included in evaluation, by target location, week ahead horizon, and variable.

maintaining robustness to outlying forecasts (*Ray et al., 2022*). Building on this, we also created weighted median ensembles using the weights described above and a Harrel-Davis quantile estimator with a beta function to approximate the weighted percentiles (*Harrell and Davis, 1982*). We then compared the performance of unweighted and inverse relative WIS weighted mean and median ensembles, comparing the ratio of interval scores between each ensemble model relative to the baseline model.

Results

For 32 European countries, we collected, visualised, and made available online weekly COVID-19 forecasts and observed data (*Sherratt, 2022*). Over the whole study period, we collected forecasts from 48 unique models. Modellers created forecasts choosing from a set of 32 possible locations, four time horizons, and two variables, and modellers variously joined and left the Hub over time. This meant the number of models contributing to the Hub varied over time and by forecasting target. Using all models and the ensemble, we created 2139 forecasting scores, where each score summarises a unique combination of forecasting model, variable, country, and week ahead horizon (*Figure 1*).

Of the total 48 models, we received the most forecasts for Germany, with 29 unique models submitting 1-week case forecasts, while only 12 models ever submitted 4-week case or death forecasts for Liechtenstein. Modelling teams also differed in how they expressed uncertainty. Only three models provided point forecasts with no estimate of uncertainty around their predictions, while 41 models provided the full set of 23 probabilistic quantiles across the predictive distribution for each target.

In this evaluation we included 29 models in comparison to the ensemble forecast (*Figure 1*). We have included metadata provided by modellers in the supplement and online (*Sherratt, 2022*). In this evaluation, at most 15 models contributed forecasts for cases in Germany at the 1 week horizon, with an accumulated 592 forecast scores for that single target over the study period. In contrast, deaths in Finland at the 2 week horizon saw the smallest number of forecasts, with only 6 independent models contributing 24 forecast scores at any time over the 52-week period. Of the 29 models included in this evaluation, 5 models provided less than the full set of 23 quantiles, and were excluded when creating the ensemble. No ensemble forecast was composed of less than 3 independent models.

We visually compared the absolute performance of forecasts in predicting numbers of incident cases and deaths. We observed that forecasts predicted well in times of stable epidemic behaviour,
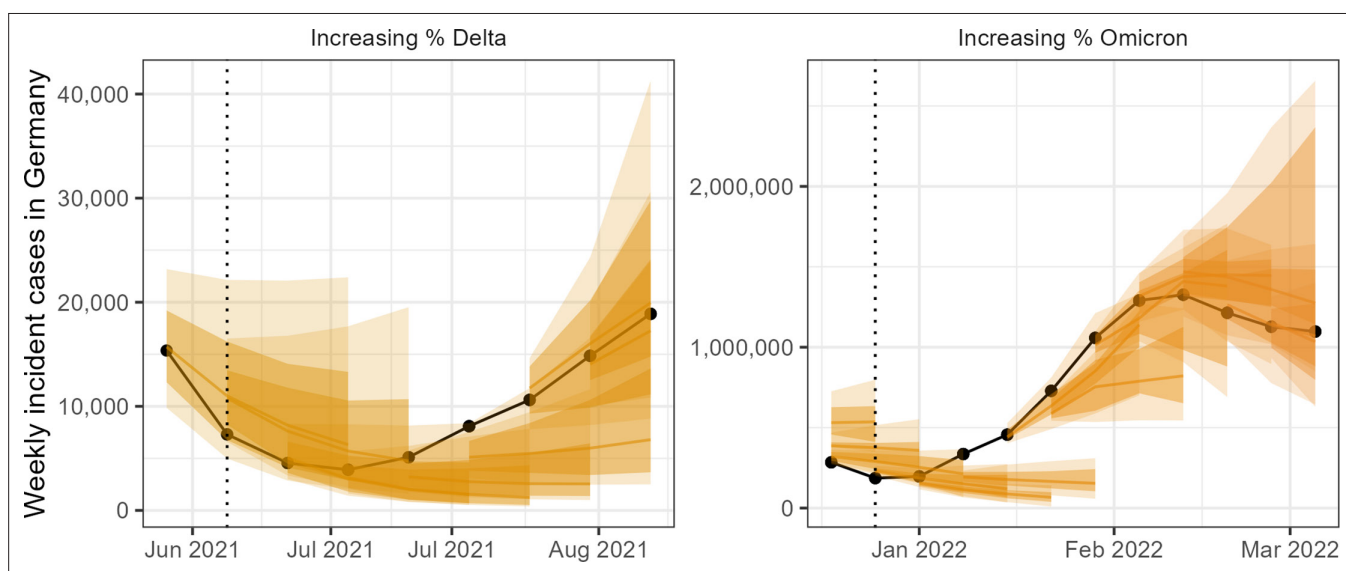

Figure 2. Ensemble forecasts of weekly incident cases in Germany over periods of increasing SARS-CoV-2 variants Delta (B.1.617.2, left) and Omicron (B.1.1.529, right). Black indicates observed data. Coloured ribbons represent each weekly forecast of 1–4 weeks ahead (showing median, 50%, and 90% probability). For each variant, forecasts are shown over an x-axis bounded by the earliest dates at which 5% and 99% of sequenced cases were identified as the respective variant of concern, while vertical dotted lines indicate the approximate date that the variant reached dominance (>50% sequenced cases).

while struggling to accurately predict at longer horizons around inflection points, for example during rapid changes in population-level behaviour or surveillance. Forecast models varied widely in their ability to predict and account for the introduction of new variants, giving the ensemble forecast over these periods a high level of uncertainty. An example of weekly forecasts from the ensemble model is shown in *Figure 2*.

In relative terms, the ensemble of all models performed well compared to both its component models and the baseline. By relative WIS scaled against a baseline of 1 (where a score <1 indicates outperforming the baseline), the median score of forecasts from the Hub ensemble model was 0.71, within an interquartile range of 0.61 at 25% probability to 0.88 at 75% probability. Meanwhile the median score of forecasts across all participating models (excluding the Hub ensemble) was 1.04 (IQR 0.82–1.36).

Across all horizons and locations, the ensemble performed better on scaled relative WIS than 83% of forecast scores when forecasting cases (with a total N=886 from 23 unique models), and 91% of scores for forecasts of incident deaths (N=763 scores from 20 models). We also saw high performance from the ensemble when evaluating against all models including those who did not submit the full set of probabilistic quantile predictions (80% for cases with N=1006 scores from 28 models, and 88% for deaths, N=877 scores from 24 models).

The performance of individual and ensemble forecasts varied by length of the forecast horizon (*Figure 3*). At each horizon, the typical performance of the ensemble outperformed both the baseline model and the aggregated scores of all its component models, although we saw wide variation between individual models in performance across horizons. Both individual models and the ensemble saw a trend of worsening performance at longer horizons when forecasting cases with the median scaled relative WIS of the ensemble across locations worsened from 0.62 for 1-week ahead forecasts to 0.9 when forecasting 4 weeks ahead. Performance for forecasts of deaths was more stable over one through 4 weeks, with median ensemble performance moving from 0.69 to 0.76 across the 4-week horizons.

We observed similar trends in performance across horizon when considering how well the ensemble was calibrated with respect to the observed data. At 1 week ahead the case ensemble was well calibrated (ca. 50% and 95% nominal coverage at the 50% and 95% levels, respectively). This did not hold at longer forecast horizons as the case forecasts became increasingly over-confident. Meanwhile, the ensemble of death forecasts was well calibrated at the 95% level across all horizons, and the calibration of death forecasts at the 50% level improved with lengthening horizons compared to being underconfident at shorter horizons.

The ensemble also performed consistently well in comparison to individual models when forecasting across countries (*Figure 4*). In total, across 32 countries forecasting for 1 through 4 weeks, when forecasting cases the ensemble outperformed 75% of component models in 22 countries, and outperformed all available models in 3 countries. When forecasting deaths, the ensemble outperformed 75% and 100% of models in 30 and 8 countries, respectively. Considering only the the 2-week horizon shown in *Figure 4*, the ensemble of case forecasts outperformed 75% models in 25 countries and all models in only 12 countries. At the 2-week horizon for forecasts of deaths, the ensemble outperformed 75% and 100% of its component models in 30 and 26 countries, respectively.

We considered alternative methods for creating ensembles from the participating forecasts, using either a mean or median to combine either weighted or unweighted forecasts. We evaluated each alternative ensemble model against the baseline model, taking the mean score ratio across all targets (*Table 1*). Across locations we observed that the median outperformed the mean across all one through 4 week horizons and both cases and death targets, for all but cases at the 1 week horizon. This held regardless of whether the component forecasts were weighted or unweighted by their individual past performance. Between methods of combination, weighting made little difference to the performance of the median ensemble, but appeared to improve performance of a mean ensemble in forecasting deaths.

Discussion

We collated 12 months of forecasts of COVID-19 cases and deaths across 32 countries in Europe, collecting from multiple independent teams and using a principled approach to standardising both forecast targets and the predictive distribution of forecasts. We combined these into an ensemble
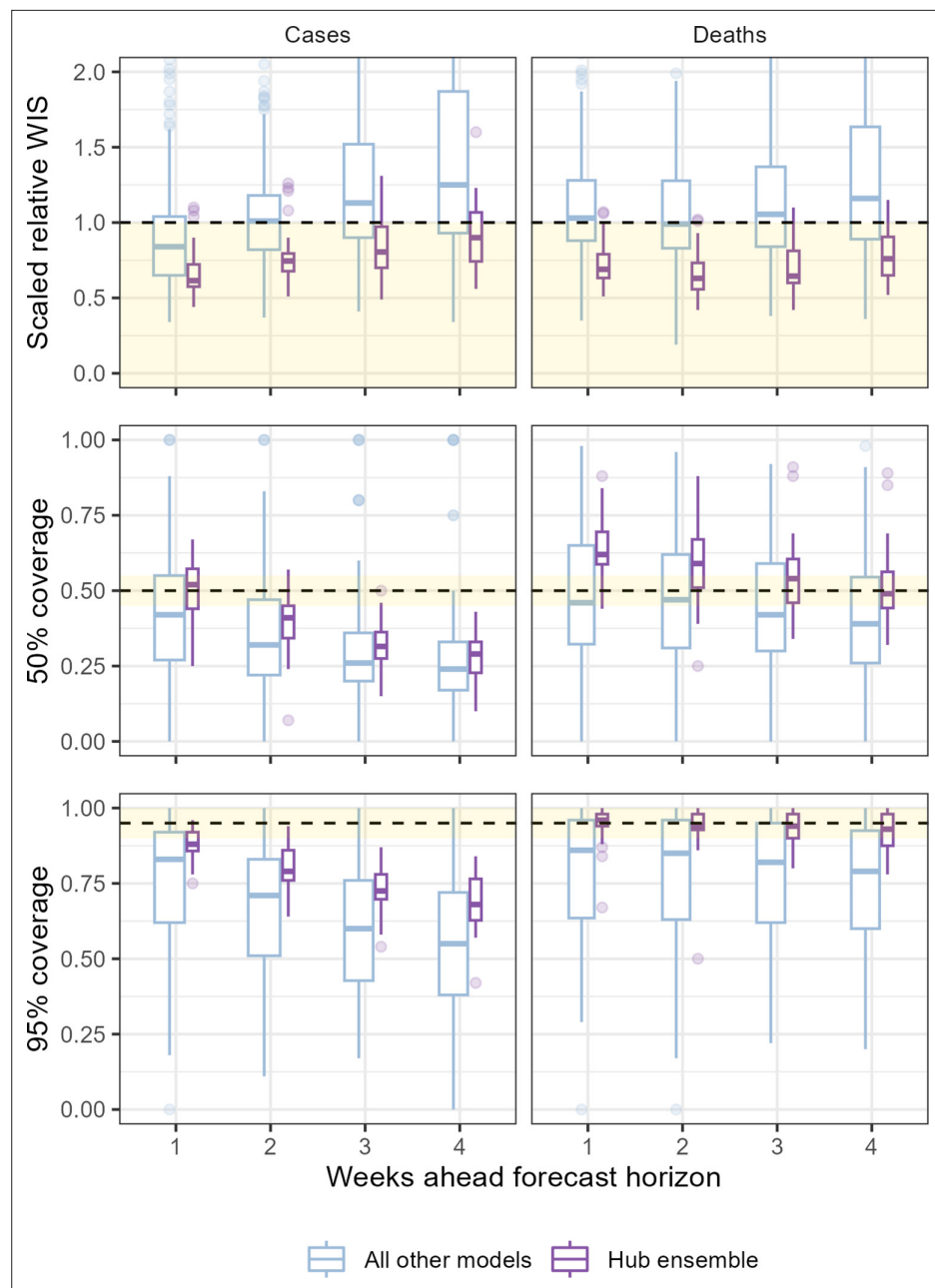
Figure 3. Performance of short-term forecasts aggregated across all individually submitted models and the Hub ensemble, by horizon, forecasting cases (left) and deaths (right). Performance measured by relative weighted interval score scaled against a baseline (dotted line, 1), and coverage of uncertainty at the 50% and 95% levels. Boxplot, with width proportional to number of observations, show interquartile ranges with outlying scores as faded points. The target range for each set of scores is shaded in yellow.

forecast and compared the relative performance of forecasts between models, finding that the ensemble forecasts outperformed most individual models across all countries and horizons over time.

Across all models we observed that forecasting changes in trend in real time was particularly challenging. Our study period included multiple fundamental changes in viral-, individual-, and population-level factors driving the transmission of COVID-19 across Europe. In early 2021, the introduction of vaccination started to change population-level associations between infections, cases, and deaths (*European Centre for Disease Prevention and Control, 2021b*), while the Delta variant emerged and became dominant (*European Centre for Disease Prevention and Control, 2021a*). Similarly from
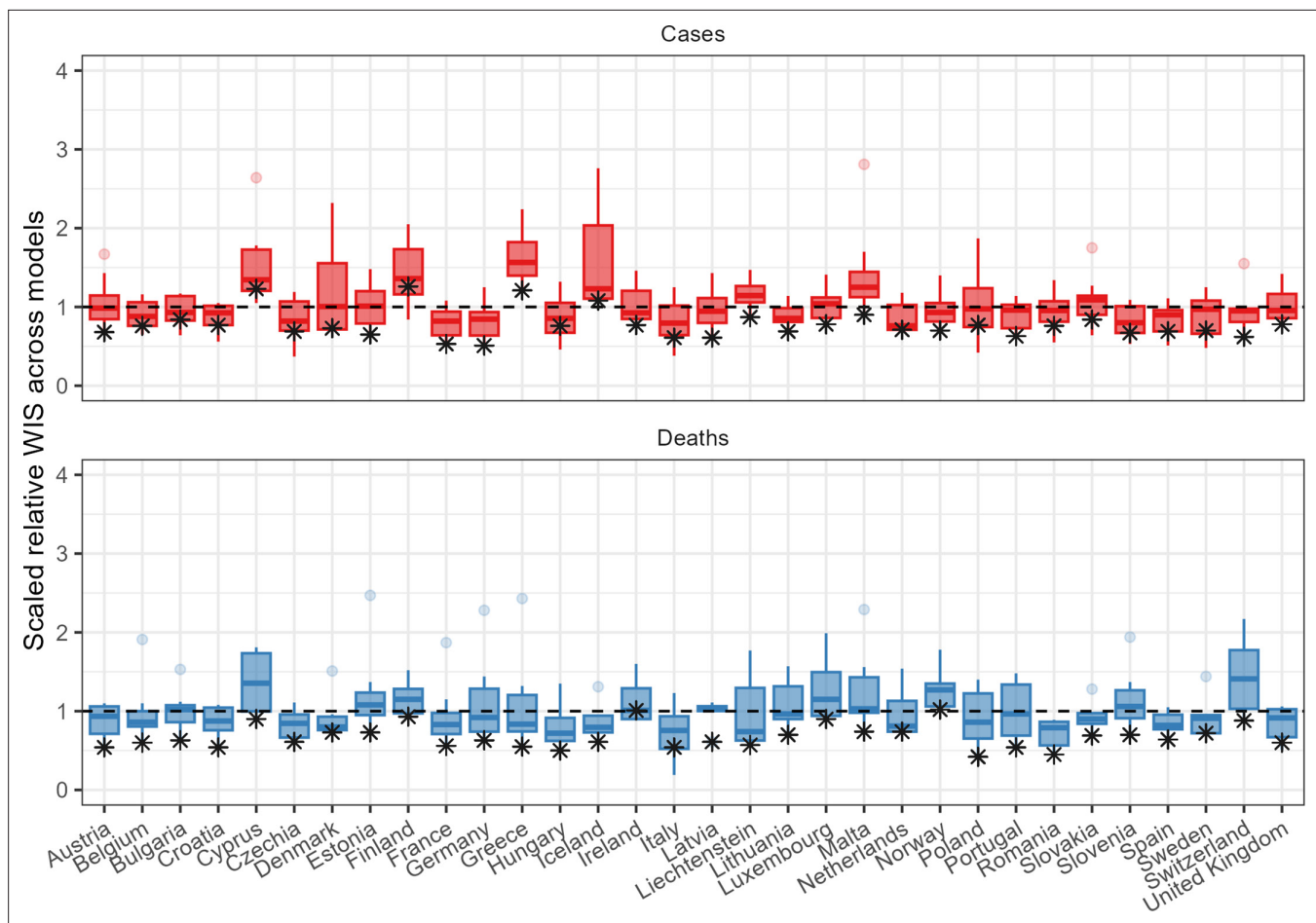
Figure 4. Performance of short-term forecasts across models and median ensemble (asterisk), by country, forecasting cases (top) and deaths (bottom) for 2-week ahead forecasts, according to the relative weighted interval score. Boxplots show interquartile ranges, with outliers as faded points, and the ensemble model performance is marked by an asterisk. y-axis is cut-off to an upper bound of 4 for readability.

Table 1. Predictive performance of main ensembles, as measured by the mean ratio of interval scores against the baseline ensemble.

Horizon	Weighted mean	Weighted median	Unweighted mean	Unweighted median
Cases				
1 week	0.63	0.64	0.61	0.64
2 weeks	0.72	0.71	0.69	0.69
3 weeks	0.82	0.76	0.82	0.72
4 weeks	1.07	0.86	1.12	0.78
Deaths				
1 week	0.65	0.61	1.81	0.61
2 weeks	0.58	0.54	1.29	0.54
3 weeks	0.64	0.57	1.17	0.53
4 weeks	0.82	0.67	0.84	0.62

late 2021 we saw the interaction of individually waning immunity during the emergence and global spread of the Omicron variant (*European Centre for Disease Prevention and Control, 2022b*). Neither the extent nor timing of these factors were uniform across European countries covered by the Forecast Hub (*European Centre for Disease Prevention and Control, 2023*). This meant that the performance of any single forecasting model depended partly on the ability, speed, and precision with which it could adapt to new conditions for each forecast target.

We observed a contrast between a more stable performance of forecasting deaths further into the future compared to forecasts of cases. Previous work has found rapidly declining performance for case forecasts with increasing horizon (*Cramer et al., 2021b*; *Castro et al., 2020*), while death forecasts can perform well with up to 6 weeks lead time (*Friedman et al., 2021*). We can link this to the specific epidemic dynamics in this study.

First, COVID-19 has a typical serial interval of less than a week (*Alene et al., 2021*). This implies that case forecasts of more than 2 weeks only remain valid if rates of both transmission and detection remain stable over the entire forecast horizon. In contrast, we saw rapid changes in epidemic dynamics across many countries in Europe over our study period, impacting the longer term case forecasts.

Second, we can interpret the higher reliability of death forecasts as due to the different lengths and distributions of time lags from infection to case and death reporting (*Jin, 2021*). For example, a spike in infections may be matched by a consistently sharp increase in case reporting, but a longer tailed distribution of the subsequent increase in death reports. This creates a lower magnitude of fluctuation in the time-series of deaths compared to that of cases. Similarly, surveillance data for death reporting is substantially more consistent, with fewer errors and retrospective corrections, than case reporting (*Català et al., 2021*).

Third, we also note that the performance of trend-based forecasts may have benefited from the slower changes to trends in incident deaths caused by gradually increasing vaccination rates. These features allow forecasters to incorporate the effect of changes in transmission more easily when forecasting deaths, compared to cases.

We found the ensemble in this study continued to outperform both other models and the baseline at up to 4 weeks ahead. Our results support previous findings that ensemble forecasts are the best or nearly the best performing models with respect to absolute predictive performance and appropriate coverage of uncertainty (*Funk et al., 2020*; *Viboud et al., 2018*; *Cramer et al., 2021b*). While the ensemble was consistently high performing, it was not strictly dominant across all forecast targets, reflecting findings from previous comparable studies of COVID-19 forecasts (*Bracher et al., 2021c*; *Brooks, 2020*). Our finding suggests the usefulness of an ensemble as a robust summary when forecasting across many spatio-temporal targets, without replacing the importance of communicating the full range of model predictions.

When exploring variations in ensemble methods, we found that the choice of median over means yielded the most consistent improvement in predictive performance, regardless of the method of weighting. Other work has supported the importance of the median in providing a stable forecast that better accounts for outlier forecasts than the mean (*Brooks, 2020*), although this finding may be dependent on the quality of the individual forecast submissions. In contrast, weighing models by past performance did not result in any consistent improvement in performance. This is in line with existing mixed evidence for any optimal ensemble method for combining short term probabilistic infectious disease forecasts. Many methods of combination have performed competitively in analyses of forecasts for COVID-19 in the US, including the simple mean and weighted approaches outperforming unweighted or median methods (*Taylor and Taylor, 2023*). This contrasts with later analyses finding weighted methods to give similar performance to a median average (*Ray et al., 2020*; *Brooks, 2020*). We can partly explain this inconsistency if performance of each method depends on the outcome being predicted (cases, deaths), its count (incident, cumulative) and absolute level, the changing disease dynamics, and the varying quality and quantity of forecasting teams over time.

We note several limitations in our approach to assessing the relative performance of an ensemble among forecast models. While we have described differences in model scores, we have not used any formal statistical test for comparing forecast scores, such as the Diebold-Mariano test (*Diebold and Mariano, 1995*), recognising that it is unclear how this is best achieved across many models. Our results are the outcome of evaluating forecasts against a specific performance metric and baseline, where multiple options for evaluation exist and the choice reflects the aim of the evaluation process.

Further, our choice of baseline model affects the given performance scores in absolute terms, and more generally the choice of appropriate baseline for epidemic forecast models is not obvious when assessing infectious disease forecasts. The model used here is supported by previous work (*Cramer et al., 2021b*), yet previous evaluation in a similar context has suggested that choice of baseline affects relative performance in general (*Bracher et al., 2021b*), and future research should be done on the best choices of baseline models in the context of infectious disease epidemics.

Our assessment of forecast performance may further have been inaccurate due to limitations in the observed data against which we evaluated forecasts. We sourced data from a globally aggregated database to maintain compatibility across 32 countries (*Dong et al., 2020*). However, this made it difficult to identify the origin of lags and inconsistencies between national data streams, and to what extent these could bias forecasts for different targets. In particular, we saw some real time data revised retrospectively, introducing bias in either direction where the data used to create forecasts was not the same as that used to evaluate it. We attempted to mitigate this by using an automated process for determining data revisions, and excluding forecasts made at a time of missing, unreliable, or heavily revised data. We also recognise that evaluating forecasts against updated data is a valid alternative approach used elsewhere (*Cramer et al., 2021b*). More generally it is unclear if the expectation of observation revisions should be a feature built into forecasts. Further research is needed to understand the perspective of end-users of forecasts in order to assess this.

The focus of this study was describing and summarising an ensemble of many models. We note that we have little insight into the individual methods and wide variety of assumptions that modellers used. While we asked modellers to provide a short description of their methods, we did not create a rigorous framework for this, and we did not document whether modellers changed the methods for a particular submitted model over time. Both the content of and variation in modelling methods and assumptions are likely to be critical to explaining performance, rather than describing or summarising it. Exploring modellers' methods and relating this to forecast performance will be an important area of future work.

In an emergency setting, access to visualised forecasts and underlying data is useful for researchers, policymakers, and the public (*CDC, 2020*). Previous European multi-country efforts to forecast COVID-19 have included only single models adapted to country-specific parameters (*Aguas et al., 2020*; *Adib et al., 2021*; *Agosto et al., 2021*).

The European Forecasting Hub acted as a unique tool for creating an open-access, cross-country modelling network, and connecting this to public health policy across Europe. By opening participation to many modelling teams and with international high participation, we were able to create robust ensemble forecasts across Europe. This also allows comparison across forecasts built with different interpretations of current data, on a like for like scale in real time. The European Hub has supported policy outputs at an international, regional, and national level, including Hub forecasts cited weekly in ECDC Communicable Disease Threats Reports (*European Centre for Disease Prevention and Control, 2022a*).

For forecast producers, an easily accessible comparison between results from different methods can highlight individual strengths and weaknesses and help prioritise new areas of work. Collating time-stamped predictions ensures that we can test true out-of-sample performance of models and avoid retrospective claims of performance. Testing the limits of forecasting ability with these comparisons forms an important part of communicating any model-based prediction to decision makers. For example, the weekly ECDC Communicable Disease Threats reports include the specific results of this work by qualitatively highlighting the greater uncertainty around case forecasts compared to death forecasts.

This study raises many further questions which could inform epidemic forecast modellers and users. The dataset created by the European Forecast Hub is an openly accessible, standardised, and extensively documented catalogue of real time forecasting work from a range of teams and models across Europe (*European Covid-19 Forecast Hub, 2023b*), and we recommend its use for further research on forecast performance. In the code developed for this study, we provide a worked example of downloading and using both the forecasts and their evaluation scores (*covid19-forecast-hub-europe, 2022*).

Future work could explore the impact on forecast models of changing epidemiology at a broad spatial scale by combining analyses of trends and turning points in cases and deaths with forecast

94

performance, or extending to include data on vaccination, variant, or policy changes over time. There is also much scope for future research into methods for combining forecasts to improve performance of an ensemble. This includes altering the inclusion criteria of forecast models based on different thresholds of past performance, excluding or including only forecasts that predict the lowest and highest values (trimming) (*Taylor and Taylor, 2023*), or using alternative weighting methods such as quantile regression averaging (*Funk et al., 2020*). Exploring these questions would add to our understanding of real time performance, supporting and improving future forecasting efforts.

We see additional scope to adapt the Hub format to the changing COVID-19 situation across Europe. We have extended the Forecast Hub infrastructure to include short term forecasts for hospitalisations with COVID-19, which is a challenging task due to limited data across the locations covered by the hub. As the policy focus shifts from immediate response to anticipating changes brought by vaccinations or the geographic spread of new variants (*European Centre for Disease Prevention and Control, 2023*), we are also separately investigating models for longer term scenarios in addition to the short term forecasts in a similar framework to existing scenario modelling work in the US (*Borchering et al., 2021*).

In conclusion, we have shown that during a rapidly evolving epidemic spreading through multiple populations, an ensemble forecast performed highly consistently across a large matrix of forecast targets, typically outperforming the majority of its separate component models and a naive baseline model. In addition, we have linked issues with the predictability of short-term case forecasts to underlying COVID-19 epidemiology, and shown that ensemble methods based on past model performance were unable to reliably improve forecast performance. Our work constitutes a step towards both unifying COVID-19 forecasts and improving our understanding of them.

Additional information

Author contributions

Katharine Sherratt, Conceptualization, Data curation, Software, Formal analysis, Investigation, Methodology, Writing - original draft, Writing – review and editing; Hugo Gruson, Software, Writing – review

and editing; Rok Grah, Helen Johnson, Rene Niehus, Bastian Prasse, Frank Sandmann, Funding acquisition, Project administration, Writing – review and editing; Jannik Deuschel, Daniel Wolffram, Graham Gibson, Evan L Ray, Nicholas G Reich, Daniel Sheldon, Yijin Wang, Nutcha Wattanachit, Nikos I Bosse, Johannes Bracher, Software, Methodology, Writing – review and editing; Sam Abbott, Validation, Methodology, Writing – review and editing; Alexander Ullrich, Software, Visualization; Lijing Wang, Jan Trnka, Guillaume Obozinski, Tao Sun, Dorina Thanou, Loic Pottier, Ekaterina Krymova, Jan H Meinke, Maria Vittoria Barbarossa, Neele Leithauser, Jan Mohring, Johanna Schneider, Jaroslaw Wlazlo, Jan Fuhrmann, Berit Lange, Isti Rodiah, Prasith Baccam, Heidi Gurung, Steven Stage, Bradley Suchoski, Jozef Budzinski, Robert Walraven, Inmaculada Villanueva, Vit Tucek, Martin Smid, Milan Zajicek, Cesar Perez Alvarez, Sophie R Meakin, Lauren Castro, Geoffrey Fairchild, Isaac Michaud, Dave Osthus, Pierfrancesco Alaimo Di Loro, Antonello Maruotti, Veronika Eclerova, Andrea Kraus, David Kraus, Lenka Pribylova, Bertsimas Dimitris, Michael Lingzhi Li, Soni Saksham, Jonas Dehning, Sebastian Mohr, Viola Priesemann, Grzegorz Redlarski, Benjamin Bejar, Giovanni Ardenghi, Nicola Parolini, Giovanni Ziarelli, Wolfgang Bock, Stefan Heyder, Thomas Hotz, David E Singh, Miguel Guzman-Merino, Jose L Aznarte, David Morina, Sergio Alonso, Enric Alvarez, Daniel Lopez, Clara Prats, Jan Pablo Burgard, Arne Rodloff, Tom Zimmermann, Alexander Kuhlmann, Janez Zibert, Fulvia Pennoni, Fabio Divino, Marti Catala, Gianfranco Lovison, Paolo Giudici, Barbara Tarantino, Francesco Bartolucci, Giovanna Jona Lasinio, Marco Mingione, Alessio Farcomeni, Ajitesh Srivastava, Pablo Montero-Manso, Aniruddha Adiga, Benjamin Hurt, Bryan Lewis, Madhav Marathe, Przemyslaw Porebski, Srinivasan Venkatramanan, Rafal P Bartczuk, Filip Dreger, Anna Gambin, Krzysztof Gogolewski, Magdalena Gruziel-Slomka, Bartosz Krupa, Antoni Moszyński, Karol Niedzielewski, Jedrzej Nowosielski, Maciej Radwan, Franciszek Rakowski, Marcin Semeniuk, Ewa Szczurek, Jakub Zielinski, Jan Kisielewski, Barbara Pabjan, Kirsten Holger, Yuri Kheifetz, Markus Scholz, Biecek Przemyslaw, Marcin Bodych, Maciej Filinski, Radoslaw Idzikowski, Tyll Krueger, Tomasz Ozanski, Methodology, Writing – review and editing; Borja Reina, Methodology, Writing – review and editing, Conceptualization; Sebastian Funk, Conceptualization, Software, Supervision, Writing – review and editing

Author ORCIDs
Katharine Sherratt ⬤ http://orcid.org/0000-0003-2049-3423
Daniel Wolffram ⬤ http://orcid.org/0000-0003-0318-3669
Yijin Wang ⬤ http://orcid.org/0000-0003-4438-6366
Jan Trnka ⬤ http://orcid.org/0000-0002-1786-7562
Tao Sun ⬤ http://orcid.org/0000-0001-6357-6726
Johanna Schneider ⬤ http://orcid.org/0000-0002-9330-2838
Jan Fuhrmann ⬤ http://orcid.org/0000-0002-7091-3740
Inmaculada Villanueva ⬤ http://orcid.org/0000-0003-4940-085X
Milan Zajicek ⬤ http://orcid.org/0000-0002-3226-7266
Antonello Maruotti ⬤ http://orcid.org/0000-0001-8377-9950
Veronika Eclerova ⬤ http://orcid.org/0000-0001-8476-7740
Viola Priesemann ⬤ http://orcid.org/0000-0001-8905-5873
Sergio Alonso ⬤ http://orcid.org/0000-0002-3989-8757
Clara Prats ⬤ http://orcid.org/0000-0002-1398-7559
Jan Pablo Burgard ⬤ http://orcid.org/0000-0002-5771-6179
Alessio Farcomeni ⬤ http://orcid.org/0000-0002-7104-5826
Bryan Lewis ⬤ http://orcid.org/0000-0003-0793-6082
Przemyslaw Porebski ⬤ http://orcid.org/0000-0001-8012-5791
Rafal P Bartczuk ⬤ http://orcid.org/0000-0002-0433-7327
Krzysztof Gogolewski ⬤ http://orcid.org/0000-0001-5523-5198
Jakub Zielinski ⬤ http://orcid.org/0000-0001-8935-8137
Sebastian Funk ⬤ http://orcid.org/0000-0002-2842-3406

Decision letter and Author response
Decision letter https://doi.org/10.7554/eLife.81916.sa1
Author response https://doi.org/10.7554/eLife.81916.sa2

Additional files

Supplementary files
• Supplementary file 1. EPIFORGE reporting guidelines Completed checklist following reporting guidelines on epidemic forecasting research.
• Supplementary file 2. Participating team metadata Team metadata for teams participating in the European Forecast Hub and evaluated in this study.
• MDAR checklist

Data availability
All source data were openly available before the study, originally available at: https://github.com/covid19-forecast-hub-europe/covid19-forecast-hub-europe (copy archived at swh:1:rev:b4d66c495e-07c12d88384506154cf58f08592365). All data and code for this study are openly available on Github: covid19-forecast-hub-europe/euro-hub-ensemble.

References

Adib K, Hancock PA, Rahimli A, Mugisa B, Abdulrazeq F, Aguas R, White LJ, Hajjeh R, Al Ariqi L, Nabeth P. 2021. A participatory modelling approach for investigating the spread of covid-19 in countries of the eastern Mediterranean region to support public health decision-making. *BMJ Global Health* **6**:e005207. DOI: https://doi.org/10.1136/bmjgh-2021-005207, PMID: 33762253

Agosto A, Giudici P. 2020. A poisson autoregressive model to understand COVID-19 contagion dynamics. *Risks* **8**:77. DOI: https://doi.org/10.3390/risks8030077

Agosto A, Campmas A, Giudici P, Renda A. 2021. Monitoring COVID-19 contagion growth. *Statistics in Medicine* **40**:4150–4160. DOI: https://doi.org/10.1002/sim.9020, PMID: 33973656

Aguas R, White L, Hupert N, Shretta R, Pan-Ngum W, Celhay O, Moldokmatova A, Arifi F, Mirzazadeh A, Sharifi H, Adib K, Sahak MN, Franco C, Coutinho R, CoMo Consortium. 2020. Modelling the COVID-19 pandemic in context: an international participatory approach. *BMJ Global Health* **5**:e003126. DOI: https://doi.org/10.1136/bmjgh-2020-003126, PMID: 33361188

Alene M, Yismaw L, Assemie MA, Ketema DB, Gietaneh W, Birhan TY. 2021. Serial interval and incubation period of COVID-19: a systematic review and meta-analysis. *BMC Infectious Diseases* **21**:257. DOI: https://doi.org/10.1186/s12879-021-05950-x, PMID: 33706702

Bicher M, Zuba M, Rainer L, Bachner F, Rippinger C, Ostermann H, Popper N, Thurner S, Klimek P. 2020. Supporting COVID-19 Policy-Making with a Predictive Epidemiological Multi-Model Warning System. [medRxiv]. DOI: https://doi.org/10.1101/2020.10.18.20214767

Borchering RK, Viboud C, Howerton E, Smith CP, Truelove S, Runge MC, Reich NG, Contamin L, Levander J, Salerno J, van Panhuis W, Kinsey M, Tallaksen K, Obrecht RF, Asher L, Costello C, Kelbaugh M, Wilson S, Shin L, Gallagher ME, et al. 2021. Modeling of future COVID-19 cases, hospitalizations, and deaths, by vaccination rates and nonpharmaceutical intervention scenarios-United States, april-september 2021. *MMWR. Morbidity and Mortality Weekly Report* **70**:719–724. DOI: https://doi.org/10.15585/mmwr.mm7019e3, PMID: 33988185

Bosse NI, Gruson H, Funk S, Abbott S. 2023. Scoringutils: utilities for scoring and assessing predictions. CRAN. https://github.com/epiforecasts/scoringutils

Bracher J, Wolffram D, Deuschel J, Görgen K, Ketterer J, SchienleM M. 2020. The German and Polish COVID-19 forecast hub. Github. https://github.com/KITmetricslab/covid19-forecast-hub-de

Bracher J, Ray EL, Gneiting T, Reich NG. 2021a. Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology* **17**:e1008618. DOI: https://doi.org/10.1371/journal.pcbi.1008618, PMID: 33577550

Bracher J, Wolffram D, Deuschel J, Görgen K, Ketterer JL, Ullrich A, Abbott S, Barbarossa MV, Bertsimas D, Bhatia S, Bodych M, Bosse NI, Burgard JP, Castro L, Fairchild G, Fiedler J, Fuhrmann J, Funk S, Gambin A, Gogolewski K, et al. 2021b. National and Subnational Short-Term Forecasting of COVID-19 in Germany and Poland during Early 2021. [medRxiv]. DOI: https://doi.org/10.1101/2021.11.05.21265810

Bracher J, Wolffram D, Deuschel J, Görgen K, Ketterer JL, Ullrich A, Abbott S, Barbarossa MV, Bertsimas D, Bhatia S, Bodych M, Bosse NI, Burgard JP, Castro L, Fairchild G, Fuhrmann J, Funk S, Gogolewski K, Gu Q, Heyder S, et al. 2021c. A pre-registered short-term forecasting study of covid-19 in Germany and Poland during the second wave. *Nature Communications* **12**:5173. DOI: https://doi.org/10.1038/s41467-021-25207-0, PMID: 34453047

Brooks L. 2020. Comparing ensemble approaches for short-term probabilistic COVID-19 forecasts in the U.S. https://forecasters.org/blog/2020/10/28/comparing-ensemble-approaches-for-short-term-probabilistic-covid-19-forecasts-in-the-u-s/ [Accessed April 1, 2022].

Buizza R. 2019. Introduction to the special issue on `` 25 years of ensemble forecasting.'' *Quarterly Journal of the Royal Meteorological Society* **145**:1–11. DOI: https://doi.org/10.1002/qj.3370

Castro M, Ares S, Cuesta JA, Manrubia S. 2020. The turning point and end of an expanding epidemic can not be precisely forecast. *PNAS* **117**:26190–26196. DOI: https://doi.org/10.1073/pnas.2007868117, PMID: 33004629

Català M, Pino D, Marchena M, Palacios P, Urdiales T, Cardona P-J, Alonso S, López-Codina D, Prats C, Alvarez-Lacalle E. 2021. Robust estimation of diagnostic rate and real incidence of COVID-19 for European policymakers. *PLOS ONE* **16**:e0243701. DOI: https://doi.org/10.1371/journal.pone.0243701, PMID: 33411737

CDC. 2020. Coronavirus disease 2019. COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting.html

covid19-forecast-hub-europe. 2022. Predictive performance of multi-model ensemble forecasts of covid-19 across European nations. Github. https://github.com/covid19-forecast-hub-europe/euro-hub-ensemble

Cramer EY, Huang Y, Wang Y, Ray EL, Cornell M, Bracher J, Brennen A, Castero Rivadeneira AJ, Gerding A, House K, Jayawardena D, Kanji AH, Khandelwal A, Le K, Niemi J, Stark A, Shah A, Wattanchit N, Zorn MW, Reich NG. 2021a. The United States COVID-19 Forecast Hub Dataset. [medRxiv]. DOI: https://doi.org/10.1101/2021.11.04.21265886

Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, Gerding A, Gneiting T, House KH, Huang Y, Jayawardena D, Kanji AH, Khandelwal A, Le K, Mühlemann A, Niemi J, Shah A, Stark A, Wang Y, Wattanachit N, et al. 2021b. Evaluation of Individual and Ensemble Probabilistic Forecasts of COVID-19 Mortality in the US. [medRxiv]. DOI: https://doi.org/10.1101/2021.02.03.21250974

Cramer E, Wang SY, Reich NG, Hanna A, Niem J, House K, Huang YD. 2021c. Reichlab/covid19-forecast-hub: release for zenodo, 20210816. Zenodo. https://doi.org/10.5281/zenodo.5208210DOI: https://doi.org/10.5281/zenodo.5208210

Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* **13**:253–263. DOI: https://doi.org/10.1080/07350015.1995.10524599

Dong E, Du H, Gardner L. 2020. An interactive web-based dashboard to track covid-19 in real time. *The Lancet. Infectious Diseases* **20**:533–534. DOI: https://doi.org/10.1016/S1473-3099(20)30120-1, PMID: 32087114

EpiForecasts. 2021. Project: ECDC European COVID-19 forecast hub. 0.1. Zoltar. https://www.zoltardata.com/project/238

European Centre for Disease Prevention and Control. 2021a. Threat assessment brief: implications for the EU/EEA on the spread of the SARS-cov-2 delta (B.1.617.2) variant of concern. https://www.ecdc.europa.eu/en/publications-data/threat-assessment-emergence-and-impact-sars-cov-2-delta-variant [Accessed April 1, 2022].

European Centre for Disease Prevention and Control. 2021b. Interim guidance on the benefits of full vaccination against COVID-19 for transmission and implications for non-pharmaceutical interventions. https://www.ecdc.europa.eu/en/publications-data/interim-guidance-benefits-full-vaccination-against-covid-19-transmission [Accessed April 1, 2022].

European Centre for Disease Prevention and Control. 2021c. Forecasting COVID-19 cases and deaths in Europe-new hub will support European pandemic planning. https://www.ecdc.europa.eu/en/news-events/forecasting-covid-19-cases-and-deaths-europe-new-hub [Accessed April 1, 2022].

European Centre for Disease Prevention and Control. 2022a. Weekly threats reports (CDTR). https://www.ecdc.europa.eu/en/publications-and-data/monitoring/weekly-threats-reports [Accessed April 1, 2023].

European Centre for Disease Prevention and Control. 2022b. Assessment of the further spread and potential impact of the SARS-CoV-2 Omicron variant of concern in the EU/EEA, 19th update. https://www.ecdc.europa.eu/en/publications-data/covid-19-omicron-risk-assessment-further-emergence-and-potential-impact [Accessed April 1, 2022].

European Centre for Disease Prevention and Control. 2023. Overview of the implementation of COVID-19 vaccination strategies and deployment plans in the EU/EEA. https://www.ecdc.europa.eu/en/publications-data/overview-implementation-covid-19-vaccination-strategies-and-deployment-plans [Accessed April 1, 2023].

European Covid-19 Forecast Hub. 2023a. Community. https://covid19forecasthub.eu/community.html [Accessed April 1, 2023].

European Covid-19 Forecast Hub. 2023b. European Covid-19 Forecast Hub. https://covid19forecasthub.eu/index.html [Accessed April 1, 2023].

European Covid-19 Forecast Hub. 2023c. Covid19-forecast-hub-europe. 9d13832. Github. https://github.com/covid19-forecast-hub-europe/covid19-forecast-hub-europe

European Covid-19 Forecast Hub. 2023d. Covid19-forecast-hub-europe, 2021. 9d13832. Github. https://github.com/covid19-forecast-hub-europe/covid19-forecast-hub-europe

Friedman J, Liu P, Troeger CE, Carter A, Reiner RC, Barber RM, Collins J, Lim SS, Pigott DM, Vos T, Hay SI, Murray CJL, Gakidou E. 2021. Predictive performance of international COVID-19 mortality forecasting models. *Nature Communications* **12**:2609. DOI: https://doi.org/10.1038/s41467-021-22457-w, PMID: 33972512

Funk S, Abbott S, Atkins BD, Baguelin M, Baillie JK, Birrell P, Blake J, Bosse NI, Burton J, Carruthers J, Davies NG, De Angelis D, Dyson L, Edmunds WJ, Eggo RM, Ferguson NM, Gaythorpe K, Gorsich E, Guyver-Fletcher G, Hellewell J, et al. 2020 Short-Term Forecasts to Inform the Response to the Covid-19 Epidemic in the UK. medRxiv. DOI: https://doi.org/10.1101/2020.11.11.20220962

Genest C. 1992. Vincentization revisited. *The Annals of Statistics* **20**:1137–1142. DOI: https://doi.org/10.1214/aos/1176348676

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**:359–378. DOI: https://doi.org/10.1198/016214506000001437

Harrell FE, Davis CE. 1982. A new distribution-free quantile estimator. *Biometrika* **69**:635–640. DOI: https://doi.org/10.1093/biomet/69.3.635

James LP, Salomon JA, Buckee CO, Menzies NA. 2021. The use and misuse of mathematical modeling for infectious disease policymaking: lessons for the COVID-19 pandemic. *Medical Decision Making* **41**:379–385. DOI: https://doi.org/10.1177/0272989X21990391, PMID: 33535889

Jin R. 2021. The lag between daily reported covid-19 cases and deaths and its relationship to age. *Journal of Public Health Research* **10**:2049. DOI: https://doi.org/10.4081/jphr.2021.2049, PMID: 33709641

Johansson MA, Apfeldorf KM, Dobson S, Devita J, Buczak AL, Baugher B, Moniz LJ, Bagley T, Babin SM, Guven E, Yamana TK, Shaman J, Moschou T, Lothian N, Lane A, Osborne G, Jiang G, Brooks LC, Farrow DC, Hyun S, et al. 2019. An open challenge to advance probabilistic forecasting for dengue epidemics. *PNAS* **116**:24268–24274. DOI: https://doi.org/10.1073/pnas.1909865116, PMID: 31712420

Moran KR, Fairchild G, Generous N, Hickmann K, Osthus D, Priedhorsky R, Hyman J, Del Valle SY. 2016. Epidemic forecasting is messier than weather forecasting: the role of human behavior and Internet data streams in epidemic forecast. *The Journal of Infectious Diseases* **214**:S404–S408. DOI: https://doi.org/10.1093/infdis/jiw375, PMID: 28830111

Ray EL, Wattanachit N, Niemi J, Kanji AH, House K, Cramer EY, Bracher J, Zheng A, Yamana TK, Xiong X, Woody S, Wang Y, Wang L, Walraven RL, Tomar V, Sherratt K, Sheldon D, Reiner RC, Prakash BA, Osthus D, et al. 2020. Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. [medRxiv]. DOI: https://doi.org/10.1101/2020.08.19.20177493

Ray EL, Brooks LC, Bien J, Biggerstaff M, Bosse NI, Bracher J, Cramer EY, Funk S, Gerding A, Johansson MA, Rumack A, Wang Y, Zorn M, Tibshirani RJ, Reich NG. 2022. Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. *International Journal of Forecasting* **1**:005. DOI: https://doi.org/10.1016/j.ijforecast.2022.06.005, PMID: 35791416

Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, Osthus D, Ray EL, Tushar A, Yamana TK, Biggerstaff M, Johansson MA, Rosenfeld R, Shaman J. 2019a. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *PNAS* **116**:3146–3154. DOI: https://doi.org/10.1073/pnas.1812594116, PMID: 30647115

Reich NG, McGowan CJ, Yamana TK, Tushar A, Ray EL, Osthus D, Kandula S, Brooks LC, Crawford-Crudell W, Gibson GC, Moore E, Silva R, Biggerstaff M, Johansson MA, Rosenfeld R, Shaman J. 2019b. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLOS Computational Biology* **15**:e1007486. DOI: https://doi.org/10.1371/journal.pcbi.1007486, PMID: 31756193

Reich NG, Cornell M, Ray EL, House K, Le K. 2021. The zoltar forecast archive, a tool to standardize and store interdisciplinary prediction research. *Scientific Data* **8**:59. DOI: https://doi.org/10.1038/s41597-021-00839-5, PMID: 33574342

Sherratt K. 2022. European covid-19 forecast hub. Zenodo. DOI: https://doi.org/10.5281/zenodo.7356267

Taylor JW, Taylor KS. 2023. Combining probabilistic forecasts of covid-19 mortality in the United States. *European Journal of Operational Research* **304**:25–41. DOI: https://doi.org/10.1016/j.ejor.2021.06.044, PMID: 34219901

Van Basshuysen P, White L, Khosrowi D, Frisch M. 2021. Three ways in which pandemic models may perform a pandemic. *Erasmus Journal for Philosophy and Economics* **14**:10–127. DOI: https://doi.org/10.23941/ejpe.v14i1.582

Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, Zhang Q, Chowell G, Simonsen L, Vespignani A, RAPIDD Ebola Forecasting Challenge group. 2018. The RAPIDD Ebola forecasting challenge: synthesis and lessons learnt. *Epidemics* **22**:13–21. DOI: https://doi.org/10.1016/j.epidem.2017.08.002, PMID: 28958414

Wang SY, Stark A, Ray E, Bosse N, Reich NG, Sherratt K, Shah A. 2021. Reichlab/covidhubutils: Repository release for zenodo. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.5207940

Zelner J, Riou J, Etzioni R, Gelman A. 2021. Accounting for uncertainty during a pandemic. *Patterns* **2**:100310. DOI: https://doi.org/10.1016/j.patter.2021.100310, PMID: 34405155

Section of manuscript	Item	Checklist item	Reported on page
Title/Abstract	1	Describe the study as forecast or prediction research in at least the title or abstract	1
Introduction	2	Define the purpose of study and forecasting targets	4
Methods	3	Fully document the methods	4,5,6,7,8
Methods	4	Identify whether the forecast was performed prospectively, in real time, and/or retrospectively	5
Methods	5	Explicitly describe the origin of input source data, with references	5
Methods	6	Provide source data with publication, or document reasons as to why this was not possible	see Github epiforecasts/euro-hub-ensemble
Methods	7	Describe input data processing procedures in detail	5,6
Methods	8	State and describe the model type, and document model assumptions, including references	5,6, Supplement Table 1
Methods	9	Make the model code available, or document the reasons why this was not possible	see Github epiforecasts/euro-hub-ensemble
Methods	10	Describe the model validation, and justify the approach	5,6
Methods	11	Describe the forecast accuracy evaluation method used, with justification	6,7
Methods	12	Where possible, compare model results to a benchmark or other comparator model, with justification of comparator choice	6,7
Methods	13	Describe the forecast horizon, with justification of its length	5
Results	14	Present and explain uncertainty of forecasting results	8,9,10,11,12
Results	15	Briefly summarize the results in nontechnical terms, including a nontechnical interpretation of forecast uncertainty	12,13,14
Results	16	If results are published as a data object, encourage a time-stamped version number	see Github epiforecasts/euro-hub-ensemble
Discussion	17	Describe the weaknesses of the forecast, including weaknesses specific to data quality and methods	12,13,14
Discussion	18	If the research is applicable to a specific epidemic, comment on its potential implications and impact for public health action and decision-making	14,15
Discussion	19	If the research is applicable to a specific epidemic, comment on how generalizable it may be across populations	15

Supplement: Team metadata

Team	Model	Authors	Methods	Website	Metadata
BIOCOMSC	BIOCOMSC-Gompertz	Martí Català, Enric Álvarez, Sergio Alonso, Daniel López, Clara Prats	Empirical model based on cases and deaths dynamics.	https://biocomsc.upc.edu/en/covid-19	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/BIOCOMSC-Gompertz.yml
University of Cologne Covid Metrics	CovidMetrics-epiBATS	Tom Zimmermann, Arne Rodloff	Forecasts are based on TBATS - models (DeLivera, Hyndman and Snyder (2011)) and are updated daily for each German state.	https://tomz.shinyapps.io/coronaLandkreise/	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/CovidMetrics-epiBATS.yml
Epiforecasts / London School of Hygiene and Tropical Medicine	epiforecasts-EpiNow2	Nikos Bosse, Sam Abbott, Sebastian Funk	Semi-mechanistic estimation of the time-varying reproduction number for latent infections mapped to reported cases/deaths.	https://epiforecasts.io/EpiNow2	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/epiforecasts-EpiNow2.yml
epiforecasts	epiforecasts-weeklygrowth	Sam Abbott	A Bayesian autoregressive model using weekly incidence data, application of the forecast.vocs R package.	https://samabbott.co.uk	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/epiforecasts-weeklygrowth.yml
epiMOX	epiMOX-SUIHTER	Giovanni Ardenghi, Giovanni Ziarelli, Luca Dede', Nicola Parolini, Alfio Quarteroni	Compartmental model SUIHTER	https://www.epimox.polimi.it	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/epiMOX-SUIHTER.yml
European COVID-19 Forecast Hub	EuroCOVIDhub-ensemble	Katharine Sherratt, Nikos Bosse, Sebastian Funk	An ensemble, or model average, of submitted forecasts to the European COVID-19 Forecast Hub.	https://covid19forecasthub.eu/	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/EuroCOVIDhub-ensemble.yml
Frankfurt Institute for Advanced Studies & Forschungszentrum Jülich	FIAS_FZJ-Epi1Ger	Maria V. Barbarossa, Jan Fuhrmann, Stefan Krieg, Jan H. Meinke	An extended SEIR model with additional compartments for undetected cases	https://www.fz-juelich.de/SharedDocs/Meldungen/IAS/JSC/DE/2021/2021-01-covid-19.html;jsessionid=F4D5FB4027E871A6F4C2FCAF0F08FC35	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/FIAS_FZJ-Epi1Ger.yml

Helmholtz Zentrum fuer Infektionsforschung	HZI-AgeExtendedSEIR	Isti Rodiah, Berit Lange, Pratizio Vanella, Alexander Kuhlmann, Wolfgang Bock	Deterministic SEIR type model	https://www.helmholtz-hzi.de/en/nc/research/research-topics/bacterial-and-viral-pathogens/epidemiology/team/	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/HZI-AgeExtendedSEIR.yml
ICM / University of Warsaw	ICM-agentModel	Rafał Bartczuk, Łukasz Górski, Magdalena Gruziel-Słomka, Artur Kaczorek, Jan Kisielewski, Antoni Moszyński, Karol Niedzielewski, Jędrzej Nowosielski, Maciej Radwan, Franciszek Rakowski, Marcin Semeniuk, Jakub Zieliński	Agent-based model	https://covid-19.icm.edu.pl/en/model-description/	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/ICM-agentModel.yml
IEM Health	IEM_Health-CovidProject	Brad Suchoski, Steve Stage, Heidi Gurung, Sid Baccam	SEIR model projections for daily incident confirmed COVID cases and deaths by using AI to fit actual cases observed.	https://iem.com/	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/IEM_Health-CovidProject.yml
ILM	ILM-EKF	Stefan Heyder, Thomas Hotz	Extended Kalman filter based on reproduction equation	https://github.com/Stochastik-TU-Ilmenau	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/ILM-EKF.yml
Fraunhofer Institute for Industrial Mathematics ITWM	itwm-dSEIR	Jan Mohring, Neele Leithäuser, Michael Helmling	Integral equation model based on age cohorts taking into account vaccination and testing. The parameters are adjusted to the counted cases and deaths.	https://www.itwm.fraunhofer.de/	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/itwm-dSEIR.yml
ITWW	ITWW-county_repro	Przemyslaw Biecek, Viktor Bezborodov, Marcin Bodych, Jan Pablo Burgard, Stefan Heyder, Thomas Hotz, Tyll Krüger	Forecasts of county level incidence based on regional reproduction numbers.	https://github.com/Stochastik-TU-Ilmenau	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/ITWW-county_repro.yml
JBUD	JBUD-HMXK	Jozef Budzinski	Heavily modified infection-age SIR-X model with waning immunity, vaccinations, seasonality and undetected cases.	https://joebud.github.io/covid-19-analyses/	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/JBUD-HMXK.yml

MOCOS group	MOCOS-agent1	Marek Bawiec, Marcin Bodych, Tyll Krueger, Tomasz Ozanski, Barbara Pabjan, Agata Migalska, Przemyslaw Biecek, Viktor Bezborodov, Ewa Szczurek, Ewaryst Rafajłowicz, Ewa Rafajłowicz, Wojciech Rafajłowicz	Agent-based microsimulation model	https://mocos.pl/	https://raw.githubusercontent.com/epif orecasts/covid19-forecast-hub-europe/main/model-metadata/MOCOS-agent1.yml
Masaryk University	MUNI-ARIMA	Andrea Kraus, David Kraus	ARIMA model with outlier detection fitted to transformed weekly aggregated series.	https://krausstat.shinyapps.io/covid19global/	https://raw.githubusercontent.com/epif orecasts/covid19-forecast-hub-europe/main/model-metadata/MUNI-ARIMA.yml
Department of Mathematics and Statistics Masaryk University Team	MUNI_DMS-SEIAR	Veronika Eclerova, Lenka Pribylova	SEIAR model with A compartment of absent unobserved infected estimated from hospital data with incorporated mobility data dependence; optimized to the compartment of all exposed (unobserved included)	https://webstudio.shinyapps.io/MAMES/	https://raw.githubusercontent.com/epif orecasts/covid19-forecast-hub-europe/main/model-metadata/MUNI_DMS-SEIAR.yml
Grzegorz Redlarski	PL_GRedlarski-DistrictsSum	Grzegorz Redlarski	Modified SIR method, applied to all districts. Forecasts for districts are summed up.	https://docs.google.com/spre adsheets/d/e/2PACX-1vRpH4yhKRts7Co5tydhZhojl PTcTTybms1PqJ9j1tmSBzzPLo U2U9XjUWDwiKYxnE6gMLayl 71rpGC8/pubhtml?gid=49325 1550&single=true	https://raw.githubusercontent.com/epif orecasts/covid19-forecast-hub-europe/main/model-metadata/PL_GRedlarski-DistrictsSum.yml
prolix	prolix-euclidean	Loïc Pottier	Offsets obtained by correlations, best linear approximation of reproduction rates (using vaccination approximation) by least euclidean distance, and linear prediction.	https://cp.lpmib.fr/medias/co vid19/_synthese.html	https://raw.githubusercontent.com/epif orecasts/covid19-forecast-hub-europe/main/model-metadata/prolix-euclidean.yml
Robert Walraven	RobertWalrave n-ESG	Robert Walraven	Multiple skewed gaussian distribution peaks fit to raw data	http://rwalraven.com/COVID 19	https://raw.githubusercontent.com/epif orecasts/covid19-forecast-hub-europe/main/model-metadata/RobertWalraven-ESG.yml

Supplement: Team metadata

Swiss Data Science Center / University of Geneva	SDSC_ISG-TrendModel	Ekaterina Krymova, Dorina Thanou, Benjamin Bejar Haro, Tao Sun, Gavin Lee, Elisa Manetti, Christine Choirat, Antoine Flahault, Guillaume Obozinski	The Trend Model predicts daily cases and deaths using linear extrapolation on the linear or log scale of the underlying trend estimated by a robust LOESS seasonal-trend decomposition model.	https://renkulab.shinyapps.io/COVID-19-Epidemic-Forecasting/	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/SDSC_ISG-TrendModel.yml
Statgroup19	Statgroup19-richards	Pierfrancesco Alaimo Di Loro, Fabio Divino, Alessio Farcomeni, Giovanna Jona Lasinio, Antonello Maruotti, Marco Mingione, Gianfranco Lovison	Richards' curve based generalized growth model	https://statgroup19.shinyapps.io/Covid19App/	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/Statgroup19-richards.yml
Statgroup19	Statgroup19-spatialrichards	Pierfrancesco Alaimo Di Loro, Fabio Divino, Alessio Farcomeni, Giovanna Jona Lasinio, Antonello Maruotti, Marco Mingione, Gianfranco Lovison	Richards' curve based generalized growth model taking into account spatial dependence	https://statgroup19.shinyapps.io/Covid19App/	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/Statgroup19-spatialrichards.yml
Universidad Carlos III de Madrid	UC3M-EpiGraph	David E. Singh, Miguel Guzman Merino, Maria Cristina Marinescu, Jesus Carretero, Alberto Cascajo Garcia	Agent-based parallel simulator that models individual interactions extracted from social networks and demographical data.	https://www.arcos.inf.uc3m.es/epigraph/	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/UC3M-EpiGraph.yml
University of Ljubljana, Faculty of Health Sciences Team	ULZF-SEIRC19SI	Janez Zibert	SEIHR model extended with compartments for hospitals, intensive care units, asymptomatic cases, separate submodels for vaccinated and unvaccinated, divided to 5 age subgroups of population	https://apps.lusy.fri.uni-lj.si	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/ULZF-SEIRC19SI.yml
UMass-Amherst	UMass-MechBayes	Dan Sheldon, Graham Gibson, Nick Reich	Bayesian compartmental model with observations on cumulative case counts and cumulative deaths. Model is fit independently to each state. Model includes observation noise and a case detection rate.	https://github.com/dsheldon/covid	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/UMass-MechBayes.yml
UNED	UNED-PreCoV2	José L. Aznarte, César Pérez, José Almagro, Pedro Álvarez, Álvaro Ortiz, Fernando Blat	Bayesian time series models with ARIMA noise and fixed transfer functions for each input.	https://precov2.org	https://raw.githubusercontent.com/epiforecasts/covid19-forecast-hub-europe/main/model-metadata/UNED-PreCoV2.yml

Supplement: Team metadata

University of Perugia / University of Milano-Bicocca / Università della Svizzera Italiana	UpgUmibUsi-MultiBayes	Francesco Bartolucci, Fulvia Pennoni, Antonietta Mira	Bayesian Dirichlet-Multinomial models for counts of patients in mutually exclusive and exhaustive categories such as hospitalized in regular wards and in intensive care units, deceased and recovered	https://github.com/francesco bartolucci/ARMultinomial	https://raw.githubusercontent.com/epif orecasts/covid19-forecast-hub-europe/main/model-metadata/UpgUmibUsi-MultiBayes.yml
University of Southern California	USC-SIkJalpha	Ajitesh Srivastava, Frost Tianjian Xu	A heterogeneous infection rate model with human mobility for epidemic modeling. Our model adapts to changing trends and provide predictions of confirmed cases and deaths.	https://scc-usc.github.io/ReCOVER-COVID-19	https://raw.githubusercontent.com/epif orecasts/covid19-forecast-hub-europe/main/model-metadata/USC-SIkJalpha.yml
University of Virginia, Biocomplexity COVID-19 Response Team	UVA-Ensemble	Aniruddha Adiga, Lijing Wang, Srinivasan Venkatramanan, Akhil Sai Peddireddy, Benjamin Hurt, Przemyslaw Porebski, Bryan Lewis, Madhav Marathe, Jiangzhou Chen, Anil Vullikanti	An ensemble of multiple methods such as auto-regressive (AR)models with exogenous variables, Long short-term memory (lSTM) models,Kalman filter and PatchSim (an SEIR model).	https://biocomplexity.virginia .edu/	https://raw.githubusercontent.com/epif orecasts/covid19-forecast-hub-europe/main/model-metadata/UVA-Ensemble.yml

Materials Design Analysis Reporting (MDAR)
Checklist for Authors

The [MDAR framework](#) establishes a minimum set of requirements in transparent reporting mainly applicable to studies in the life sciences.

eLife asks authors to **provide detailed information within their article** to facilitate the interpretation and replication of their work. Authors can also upload supporting materials to comply with relevant reporting guidelines for health-related research (see [EQUATOR Network](#)), life science research (see the [BioSharing Information Resource](#)), or animal research (see the [ARRIVE Guidelines](#) and the [STRANGE Framework](#); for details, see *eLife*'s [Journal Policies](#)). Where applicable, authors should refer to any relevant reporting standards materials in this form.

For all that apply, please note **where in the article** the information is provided. Please note that we also collect information about data availability and ethics in the submission form.

Materials:

Newly created materials	Indicate where provided: section/figure legend	N/A
The manuscript includes a dedicated "materials availability statement" providing transparent disclosure about availability of newly created materials including details on how materials can be accessed and describing any restrictions on access.		N/A

Antibodies	Indicate where provided: section/figure legend	N/A
For commercial reagents, provide supplier name, catalogue number and [RRID](#), if available.		N/A

DNA and RNA sequences	Indicate where provided: section/figure legend	N/A
Short novel DNA or RNA including primers, probes: Sequences should be included or deposited in a public repository.		N/A

Cell materials	Indicate where provided: section/figure legend	N/A
Cell lines: Provide species information, strain. Provide accession number in repository OR supplier name, catalog number, clone number, OR RRID.		N/A
Primary cultures: Provide species, strain, sex of origin, genetic modification status.		N/A

107

Experimental animals	Indicate where provided: section/figure legend	N/A
Laboratory animals or Model organisms: Provide species, strain, sex, age, genetic modification status. Provide accession number in repository OR supplier name, catalog number, clone number, OR RRID.		N/A
Animal observed in or captured from the field: Provide species, sex, and age where possible.		N/A

Plants and microbes	Indicate where provided: section/figure legend	N/A
Plants: provide species and strain, ecotype and cultivar where relevant, unique accession number if available, and source (including location for collected wild specimens).		N/A
Microbes: provide species and strain, unique accession number if available, and source.		N/A

Human research participants	Indicate where provided: section/figure legend) or state if these demographics were not collected	N/A
If collected and within the bounds of privacy constraints report on age, sex, gender and ethnicity for all study participants.		N/A

Design:

Study protocol	Indicate where provided: section/figure legend	N/A
If the study protocol has been pre-registered, provide DOI. For clinical trials, provide the trial registration number OR cite DOI.		N/A

Laboratory protocol	Indicate where provided: section/figure legend	N/A
Provide DOI OR other citation details if detailed step-by-step protocols are available.		N/A

Experimental study design (statistics details) *		
For in vivo studies: State whether and how the following have been done	**Indicate where provided: section/figure legend. If it could have been done, but was not, write "not done"**	**N/A**
Sample size determination		N/A
Randomisation		N/A
Blinding		N/A
Inclusion/exclusion criteria		N/A

Sample definition and in-laboratory replication	**Indicate where provided: section/figure legend**	**N/A**
State number of times the experiment was replicated in the laboratory.		N/A
Define whether data describe technical or biological replicates.		N/A

Ethics	**Indicate where provided: section/submission form**	**N/A**
Studies involving human participants: State details of authority granting ethics approval (IRB or equivalent committee(s), provide reference number for approval.		N/A
Studies involving experimental animals: State details of authority granting ethics approval (IRB or equivalent committee(s), provide reference number for approval.		N/A
Studies involving specimen and field samples: State if relevant permits obtained, provide details of authority approving study; if none were required, explain why.		N/A

Dual Use Research of Concern (DURC)	**Indicate where provided: section/submission form**	**N/A**
If study is subject to dual use research of concern regulations, state the authority granting approval and reference number for the regulatory approval.		N/A

Analysis:

Attrition	Indicate where provided: section/figure legend	N/A
Describe whether exclusion criteria were pre-established. Report if sample or data points were omitted from analysis. If yes, report if this was due to attrition or intentional exclusion and provide justification.	Exclusion criteria were pre-established. Forecasts were omitted where they corresponded to an observed data point that was retrospectively adjusted in official reporting statistics by >5%.	

Statistics	Indicate where provided: section/figure legend	N/A
Describe statistical tests used and justify choice of tests.		N/A

Data availability	Indicate where provided: section/submission form	N/A
For newly created and reused datasets, the manuscript includes a data availability statement that provides details for access (or notes restrictions on access).	Abstract, Discussion	
When newly created datasets are publicly available, provide accession number in repository OR DOI and licensing details where available.	Code and data DOI: 10.5281/zenodo.6895700	
If reused data is publicly available provide accession number in repository OR DOI, OR URL, OR citation.	Katharine Sherratt, Hugo Gruson, Helen Johnson, Rene Niehus, Bastian Prasse, Frank Sandman, Jannik Deuschel, Daniel Wolffram, Sam Abbott, Alexander Ullrich, Graham Gibson, Evan L Ray, Nicholas G Reich, Daniel Sheldon, Yijin Wang, Nutcha Wattanachit, Lijing Wang, Jan Trnka, Guillaume Obozinski, … Sebastian Funk. (2022). European Covid-19 Forecast Hub (v2022.07.21) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6874754	

Code availability	Indicate where provided: section/figure legend	N/A

For any computer code/software/mathematical algorithms essential for replicating the main findings of the study, whether newly generated or re-used, the manuscript includes a data availability statement that provides details for access or notes restrictions.	Abstract, Discussion	
Where newly generated code is publicly available, provide accession number in repository, OR DOI OR URL and licensing details where available. State any restrictions on code availability or accessibility.	URL: https://github.com/covid19-forecast-hub-europe/euro-hub-ensemble DOI: 10.5281/zenodo.6895700 License: MIT	
If reused code is publicly available provide accession number in repository OR DOI OR URL, OR citation.		N/A

Reporting:

The MDAR framework recommends adoption of discipline-specific guidelines, established and endorsed through community initiatives.

Adherence to community standards	Indicate where provided: section/figure legend	N/A
State if relevant guidelines (e.g., ICMJE, MIBBI, ARRIVE, STRANGE) have been followed, and whether a checklist (e.g., CONSORT, PRISMA, ARRIVE) is provided with the manuscript.	EPIFORGE guidelines for reporting epidemic forecasting; checklist provided in Supplement	

---

* We provide the following guidance regarding transparent reporting and statistics; we also refer authors to [Ten common statistical mistakes to watch out for when writing or reviewing a manuscript](#).

Sample-size estimation
- You should state whether an appropriate sample size was computed when the study was being designed
- You should state the statistical method of sample size computation and any required assumptions
- If no explicit power analysis was used, you should describe how you decided what sample (replicate) size (number) to use

Replicates
- You should report how often each experiment was performed
- You should include a definition of biological versus technical replication
- The data obtained should be provided and sufficient information should be provided to indicate the number of independent biological and/or technical replicates
- If you encountered any outliers, you should describe how these were handled
- Criteria for exclusion/inclusion of data should be clearly stated
- High-throughput sequence data should be uploaded before submission, with a private link for

5

reviewers provided (these are available from both GEO and ArrayExpress)

Statistical reporting
- Statistical analysis methods should be described and justified
- Raw data should be presented in figures whenever informative to do so (typically when N per group is less than 10)
- For each experiment, you should identify the statistical tests used, exact values of N, definitions of center, methods of multiple test correction, and dispersion and precision measures (e.g., mean, median, SD, SEM, confidence intervals; and, for the major substantive results, a measure of effect size (e.g., Pearson's r, Cohen's d)
- Report exact p-values wherever possible alongside the summary statistics and 95% confidence intervals. These should be reported for all key questions and not only when the p-value is less than 0.05.

Group allocation
- Indicate how samples were allocated into experimental groups (in the case of clinical studies, please specify allocation to treatment method); if randomization was used, please also state if restricted randomization was applied
- Indicate if masking was used during group allocation, data collection and/or data analysis

Characterising information gains and losses when collecting multiple epidemic model outputs

Sherratt K, Srivastava A, Ainslie K, Singh DE, Cublier A, Marinescu MC, Carretero J, Garcia AC, Franco N, Willem L, Abrams S, Faes C, Beutels P, Hens N, Müller S, Charlton B, Ewert R, Paltra S, Rakow C, Rehmann J, Conrad T, Schütte C, Nagel K, Abbott S, Grah R, Niehus R, Prasse B, Sandmann F, Funk S. Characterising information gains and losses when collecting multiple epidemic model outputs. Epidemics. 2024 Mar 27;47:100765. doi: 10.1016/j.epidem.2024.100765.

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1701639	**Title**	Ms
First Name(s)	Katharine		
Surname/Family Name	Sherratt		
Thesis Title	Collaborative outbreak modelling for decision support: evaluating trade-offs from multi-model combination		
Primary Supervisor	Sebastian Funk		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Epidemics		
When was the work published?	2024		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	PhD by Publication		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

| For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary) | I first contributed building and maintaining the European Scenario Hub infrastructure. My work included adapting and developing software for collecting, validating, and visualising scenario projections, and producing all documentation. This was supported by Hugo Gruson with supervision from Sebastian Funk and collaboration with the US Scenario Hub. During the project, I led four rounds of collaborative scenario modelling. I facilitated stakeholder workshops to codesign scenarios, and designed procedures for collecting metadata, visualising, and narratively synthesising results from the scenario projections.

For this paper, I conceived and conducted the research questions and conducted all analysis, visualisations, and drafting, supervised by Sebastian Funk. First, I led work analysing trajectories for policy relevant characteristics, including writing code and creating visualisations, as part of the development of the Scenario Hub. This aspect of the work was also reviewed by the team at the ECDC. Second, I planned and conducted all analysis and visualisation for work comparing an ensemble from trajectories to the Vincentised ensemble from quantile summaries. In later drafts I added the Linear Opinion Pool ensemble, drawing on code developed by Emily Howerton and Evan Ray. The third aspect of this work was to ensemble scenario trajectories using a weighting procedure based on ongoing evaluation against observed data. I developed all code and visualisation for this element of the work. I led all draft writing, revisions, and paper management. We sought a first round of review from all those who had contributed to the Scenario Hub before submission for peer review. Code contributions are at: https://github.com/epiforecasts/multi-model-information |

SECTION E

Student Signature	Katharine Sherratt
Date	14 June 2024

Supervisor Signature	Sebastian Funk
Date	14 June 2024

Characterising information gains and losses when collecting multiple epidemic model outputs

Katharine Sherratt [a],[*], Ajitesh Srivastava [b], Kylie Ainslie [c],[d], David E. Singh [e], Aymar Cublier [e], Maria Cristina Marinescu [f], Jesus Carretero [e], Alberto Cascajo Garcia [e], Nicolas Franco [g], Lander Willem [h], Steven Abrams [h],[i], Christel Faes [i], Philippe Beutels [h], Niel Hens [h],[i], Sebastian Müller [j], Billy Charlton [j], Ricardo Ewert [j], Sydney Paltra [j], Christian Rakow [j], Jakob Rehmann [j], Tim Conrad [k], Christof Schütte [k], Kai Nagel [j], Sam Abbott [a], Rok Grah [l], Rene Niehus [l], Bastian Prasse [l], Frank Sandmann [l], Sebastian Funk [a]

[a] *London School of Hygiene & Tropical Medicine, London, UK*
[b] *University of Southern California, Los Angeles, USA*
[c] *Dutch National Institute of Public Health and the Environment (RIVM), Bilthoven, Netherlands*
[d] *School of Public Health, University of Hong Kong, Hong Kong Special Administrative Region*
[e] *Universidad Carlos III de Madrid, Madrid, Spain*
[f] *Barcelona Supercomputing Center, Barcelona, Spain*
[g] *University of Namur, Namur, Belgium*
[h] *University of Antwerp, Antwerp, Belgium*
[i] *UHasselt, Hasselt, Belgium*
[j] *Technische Universität Berlin, Berlin, Germany*
[k] *Zuse Institute Berlin (ZIB), Berlin, Germany*
[l] *ECDC, Stockholm, Sweden*

ARTICLE INFO

ABSTRACT

Background: Collaborative comparisons and combinations of epidemic models are used as policy-relevant evidence during epidemic outbreaks. In the process of collecting multiple model projections, such collaborations may gain or lose relevant information. Typically, modellers contribute a probabilistic summary at each time-step. We compared this to directly collecting simulated trajectories. We aimed to explore information on key epidemic quantities; ensemble uncertainty; and performance against data, investigating potential to continuously gain information from a single cross-sectional collection of model results.

Methods: We compared projections from the European COVID-19 Scenario Modelling Hub. Five teams modelled incidence in Belgium, the Netherlands, and Spain. We compared July 2022 projections by incidence, peaks, and cumulative totals. We created a probabilistic ensemble drawn from all trajectories, and compared to ensembles from a median across each model's quantiles, or a linear opinion pool. We measured the predictive accuracy of individual trajectories against observations, using this in a weighted ensemble. We repeated this sequentially against increasing weeks of observed data. We evaluated these ensembles to reflect performance with varying observed data.

Results: By collecting modelled trajectories, we showed policy-relevant epidemic characteristics. Trajectories contained a right-skewed distribution well represented by an ensemble of trajectories or a linear opinion pool, but not models' quantile intervals. Ensembles weighted by performance typically retained the range of plausible incidence over time, and in some cases narrowed this by excluding some epidemic shapes.

Conclusions: We observed several information gains from collecting modelled trajectories rather than quantile distributions, including potential for continuously updated information from a single model collection. The value of information gains and losses may vary with each collaborative effort's aims, depending on the needs of

---

projection users. Understanding the differing information potential of methods to collect model projections can support the accuracy, sustainability, and communication of collaborative infectious disease modelling efforts.

1. Background

During outbreaks of infectious disease, it is critical to account for the uncertainty of future disease incidence in order for public health decision-makers to fully evaluate risk (Zelner et al., 2021; Li et al., 2017). Infectious disease modellers use a variety of approaches to meet this demand for information. A common challenge is the representation of multiple sources of uncertainty, both within each model as well as across separate model projections (McCabe et al., 2021; Swallow et al., 2022). In recognising this challenge, infectious disease modelling has seen an increasing emphasis on both probabilistic modelling methods, together with collaborative approaches to modelling (Bracher et al., 2021a; Reich et al., 2022).

Probabilistic infectious disease models can address the challenge of uncertainty by simulating the complex and changing real-world process of disease transmission. Modellers must handle stochasticity in transmission dynamics, often using observed data to estimate model parameters and latent trajectories that are themselves uncertain. Each such model can generate any number of simulated trajectories, and modellers choose at what point to conclude there are sufficient iterations to reach a stable distribution of possible outcomes. The output of these simulations can then be summarised to calculate quantities of interest, such as weekly incidence of infections or cases.

When creating models to characterise the future, modellers have often drawn a distinction in the meaning of uncertainty between forecast compared to scenario projections (Lipsitch et al., 2011). Forecasts are predictions of future epidemic trajectories, and the probabilities assigned to different outcomes quantify the belief of the forecaster that these may or may not happen. In addition to potential fundamental limits to predictability, forecasts are usually reliable for, at best, a few generations of transmission (Sherratt et al., 2023) because of unmodelled factors affecting future transmission such as behavioural or policy changes, heterogeneity in transmission risk, or the emergence of new variants of different transmissibility or severity.

In contrast, scenarios are projections attuned to a particular context by being conditioned on specific factors whose futures may not be quantitatively predictable, such as options for policy interventions (Runge et al., 2023; Rhodes et al., 2020). Probabilities of future outcomes as stated by scenario models should be interpreted as valid only under the specific circumstances given by the scenario but not otherwise, without specifying any probability of the scenario itself occurring. Because of this difference, forecasts can be evaluated by confronting them with future data as it becomes available, while this evaluation is more challenging for scenarios where predictive performance will always depend on a combination of adequacy of the chosen assumptions (e.g. on pathogen biology, human behaviour and government policy), with adequacy of the model in reflecting these assumptions.

Infectious disease modelling collaborations aim to bring together models that project the future using diverse methods (Reich et al., 2022). Each collaboration sets a clearly defined target for projections, communicates this target to multiple independent modellers, and collects model results in a standardised format. This standardisation allows for a like-for-like comparison of varying modelling methods' results and accompanying uncertainty. Ensemble methods can then combine results across models. Typically, this creates a more comprehensive and robust projection (Ray et al., 2020) or reflection of expert judgement (Shea et al., 2020).

Formal, large-scale modelling collaborations have, so far, been used for influenza, Ebola, Zika, dengue fever, and COVID-19 (Reich et al., 2022). In the case of COVID-19, a number of policy-facing research groups have set up collaborations to collate forecasts and scenarios

(Borchering, 2021; Cramer et al., 2021; Funk et al., 2020; Sherratt et al., 2023), and there is a substantial effort towards expanding the practice of ensemble projections of infectious disease spread and burden. Ongoing work evaluating these efforts has focused on assessing the output of past and current ensemble modelling projects. This has included evaluating differing performance among individual models (Viboud et al., 2018; Bracher et al., 2021b; Cramer et al., 2022), and a variety of methods for creating ensembles from multiple models (Howerton et al., 2023; Ray et al., 2020; Sherratt et al., 2023; Taylor and Taylor, 2021).

The standardised format in which model projections are collected is key to meeting such projects' aims of comparing information from multiple models. The most common approach to this is to collect descriptive statistics from each model at each given time step. In this format, each modeller submits values across a pre-specified set of quantiles in order to represent uncertainty in their projection. The benefits of this system include that it should accurately represent an underlying distribution of outcomes while being storage-efficient (Genest, 1992), it is not restricted to probabilistic models producing simulations, and it allows for quantitative evaluation against observed data (Bracher et al., 2021a). Various methods for subsequent combination from quantiles depend on the view taken of uncertainty between and across model projections (Howerton et al., 2023).

However, using quantile intervals separately across each time step may lose information pertinent to epidemic decision making. As a quantile representation provides a summary across trajectories at each time step, it has no theoretical continuity through the time-series. This does not permit aggregation over time to calculate cumulative totals or means, and may misrepresent time-series characteristics including epidemic peak size or timing (Juul et al., 2021). Whilst some of these can be remedied by also collating quantiles of cumulative quantities, these still lose some of the temporal information contained in the full joint probability distribution across all future time points.

An alternative method of collecting output from multiple probabilistic models is to collect the individual simulated trajectories produced by each modeller. Each simulated trajectory comprises a single value for each time step, with modellers contributing some number of these trajectories. Each trajectory retains its own time-series characteristics, and these can therefore be summarised across different models. Collecting trajectories also creates potential for the analysis and combination of each trajectory independently from the originating model's total output. One option could include comparing each trajectory to observed data as it becomes available, even after the time of collecting model outputs. This would enable creating an ensemble projection that is conditioned on the observed accuracy of each individual trajectory. As further observed data become available, this ensemble could be updated to create a single combined projection that continuously reflects the changing performance of each trajectory. This would act similarly to methods of particle filtering in continuously conditioning on past behaviour.

We aim to explore aspects of information gains and losses from these two methods of collecting multiple model results. We contrast collecting a set of simulated trajectories, against collecting a summary at quantile intervals of those trajectories. We use the setting of the European COVID-19 Scenario Hub, where the use of quantile summaries was replaced in mid-2022 by collecting trajectories. These trajectories represent random samples from the collection of all possible trajectories of each model consistent with a given scenario and the data available up to the time at which the simulation was generated.

In this work, we assess the impact of the collection method when seeking information about policy-relevant epidemic characteristics, including cumulative totals, timing of peaks, and the extent of

uncertainty across multiple models. We then explore the information gained by the ability to compare modelled epidemic trajectories to observed data as this becomes available over time. We use this to create a multi-model ensemble which weights across all available trajectories by their past accuracy. This demonstrates the potential to continuously gain information from only a single cross-sectional collection of model results. Understanding the potential sources of information gains and losses when collecting multiple model projections may support improving the accuracy, reliability, and communication of collaborative infectious disease modelling efforts.

2. Methods

2.1. Study setting

In this work we use projections from Round 2 of the European COVID-19 Scenario Modelling Hub (Taylor and Taylor, 2021). The European COVID-19 Scenario Hub was launched in March 2022 to reflect demand for the ECDC to support longer term European policy planning. It used the existing US Scenario Hub (Borchering, 2021) as a basis for Hub infrastructure and methods. Modelling teams were recruited by word of mouth to join a series of collaborative workshops, approximately fortnightly from March through June 2022. In these sessions both policy-focussed colleagues from the ECDC and modelling-focussed researchers co-developed a set of four scenarios. Each scenario represented a combination of two possible epidemiological and policy changes that could impact the incidence of COVID-19 across Europe in the medium term.

Teams were asked to project the incidence of COVID-19 infections, cases, deaths, and hospitalisations in 32 European countries over the next year. To facilitate comparison across models, we identified and agreed a common set of key assumptions and parameters to be used by all models in each scenario as well as standard data sets to which to compare the model outputs where available. Modellers uploaded projections to a Github repository, and we summarised results across models, with a focus on targets with three or more different models. Over 2022 this process was repeated four times to explore a variety of different scenarios. In total nine separate teams submitted projections, with six teams contributing to each round.

Over June 2022 (Round 2), we specified four scenarios (A-D) as: an autumn second booster campaign among the population aged over 60 (scenarios A/C), or over 18 (scenarios B/D); and future vaccine effectiveness as 'optimistic' (equivalent to the effectiveness as of a booster vaccine against the Delta SARS-CoV-2 variant; scenarios A/B); or 'pessimistic' (as against variants Omicron BA.4/BA.5/BA.2.75; scenarios C/D). Modellers were asked to start their projections from 24th July 2022, meaning that even if data were available beyond this date they were not to inform calibration of the model. Modellers were asked to submit up to 100 simulations, each reflecting a trajectory of weekly incidence of reported cases and deaths over time for a given projection target. Modellers were informed that data presented on the Johns Hopkins University dashboard was to be used for future comparison to data (Dong et al., 2020). In practice some of the models were not calibrated to reported cases and therefore used symptomatic cases as a proxy (see model details in Supplement). Simulations were to represent random samples from the distribution of simulation trajectories consistent with the given scenario that each modelling team produced. We have published full scenario details including shared parameters, all teams' projections, and summary results online (European COVID-19 Scenario Hub).

This work specifically focuses on contrasting the sampled simulated trajectories with their representation in time-specific quantiles. We collected raw data in the form of up to 100 trajectories from each model for each projection target. We used these data to retrospectively create a marginal fixed-time quantile representation of results from each model and target. Following the current submission procedure across COVID-

19 Modelling Hubs for an individual model, we calculated a median and 22 further quantiles for each week using the values of the trajectories in that week, separately for each scenario. We processed all data in R with code available online (Sherratt and Funk, 2024).

2.2. Characterising potential information gains and losses

First we considered information about key epidemic characteristics. At the time the projections were in production, discussion with the ECDC modelling team led to an interest in: estimates of incidence over time; cumulative values over different periods; and number of distinct peaks, size, and timing of peak incidence over the projection period.

When projections were available, we estimated these characteristics from the simulated trajectories. We summed incidence over time to produce a cumulative total from each trajectory. We assessed the size of the expected burden of each target relative to a known threshold by comparing the cumulative projected total to the cumulative total of the preceding year. We identified peaks in each simulated trajectory as the local maxima in a sliding window of five weeks, using the ggpmisc R package (Aphalo, 2023). We chose a sliding window of five weeks to capture each distinct peak while avoiding detecting noise in each trajectory. We summarised across the individual peaks detected in each trajectory using quantiles at each weekly time-step, to produce a range indicating possible peak timing and maximum values across all trajectories. We produced a real-time report of this summary at the time that projections became available in July 2022.

In further retrospective analysis, we compared the use of a standard unweighted ensemble to express uncertainty across multiple models in the two representations. We created an ensemble projection from first combining all individual simulated trajectories with equal weight for each scenario, location, and outcome target. Next, we took model-specific quantiles from each model's distribution of trajectories at each time point, for each scenario, location, and outcome target. We used each set of quantiles to create linear opinion pool ensembles (LOP), which use linear extrapolation between the given quantiles to estimate the cumulative distribution function in order to then randomly sample trajectories to aggregate, again with equal weight; and a quantile-average ensemble, which takes the median across the different models' values at each quantile and time step. The LOP and quantile-average ensembles have both been used to produce ensemble projections across multiple epidemiological forecasts (Howerton et al., 2023; Ray et al., 2020; Sherratt et al., 2023). To assess the difference in uncertainty across the two ensembles, we compared the mean of the values at each quantile across all time points, outcomes and scenarios.

Lastly, we evaluated the performance of each simulated trajectory against proximity to observed data, and used this to weight an ensemble of trajectories (as above). To measure performance, we calculated the mean absolute error (MAE) for each trajectory, where the MAE is the average of the difference from observed data across all available time points for a single projection. We created a weighted ensemble from all trajectories for a country (not further separating by scenario or model) using the inverse MAE for each trajectory as a weight. To calculate weighted quantiles we used a Harrell Davis weighted estimator (Harrell and Davis, 1982) from the *cNORM* R package (v3.0.2) (Lenhard et al., 2018). As above, we calculated 23 quantiles including the median to express uncertainty.

We repeated this process to create a sequence of ensembles with changing weights over time. We created the first weighted ensemble after 4 weeks of observed data, and then created consecutive ensembles with weights re-calculated weekly to use up to the maximum available 29 weeks of observed data (to 11 March 2023). This showed varying lengths of projections repeatedly conditioned on simulated trajectories' performance against increasing data over time.

We evaluated the predictive performance of these sequences of weighted ensembles. We transformed forecasts and observed data to a logarithmic scale, as this allows a more consistent evaluation across

varying magnitudes and better reflects the exponential nature of epidemic processes (Bosse et al., 2023). We then calculated the weighted interval score for each forecast, as a quantitative performance measure that evaluates across both the accuracy and the dispersion of probabilistic forecasts (Bracher et al., 2021a). In the same way we evaluated the unweighted ensemble of trajectories described above, and used this as a relative baseline with which to compare the effect of weighting individual trajectories on ensemble performance.

3. Results

A total of six modelling teams contributed projections for various targets to the European COVID-19 Scenario Hub in Round 2. Here we focus on multi-model comparison and include only projection targets with three contributing models. These targets included 52 weeks' case and death incidence for the Netherlands and Belgium, and 41 weeks' case incidence for Spain.

Five teams contributed projections for these targets. Three teams used compartmental models, one an agent-based model, and one a machine learning method (see Supplement). Four models generated 100 simulated trajectories, and one 96 trajectories (implying a slightly smaller weight to this model in trajectory-based aggregates). In total, we consider 294,816 data points from 5920 trajectories, where each data point is the estimated weekly incidence in a simulated trajectory of an outcome in a target country and scenario over up to one year (Fig. 1.i).

Aggregating across simulated trajectories from multiple models

allowed access to information about various epidemic characteristics. These included cumulative totals, and peak size and timings (see contemporaneous report reproduced in the Supplement). By summarising across the peaks of each individual trajectory, we were able to create an estimate of uncertainty around the size and timing of peaks for each target. We were also able to summarise cumulative outcomes. For example, across all 5920 trajectories for all targets and scenarios, 10% saw a cumulative total exceeding the preceding year. These epidemic characteristics could not be meaningfully estimated from the same results summarised into quantiles.

We compared information loss in the aggregation of simulated trajectories into ensemble projections (Fig. 1). We compared an ensemble taken from all trajectories (Fig. 1.ii) with a linear opinion pool (not shown), and the quantile-average ensemble (Fig. 1.iii). We noted that a linear opinion pool ensemble produces near-identical results to taking an ensemble directly from trajectories. Across all projection targets, we observed substantially increased uncertainty in an ensemble that aggregated either directly from trajectories, or via linear opinion pool, compared to a quantile-average ensemble. This represented the wider variety of epidemic shapes projected by different models. For example, the credible interval of projections for Spain included high autumn-winter incidence, while for Belgium gave greater credibility to multiple peaks of incidence. These were not observed in the interval projections of an ensemble derived from models' quantiles.

We quantified the range of uncertainty between each ensemble by comparing the mean of values at each quantile across all time points and
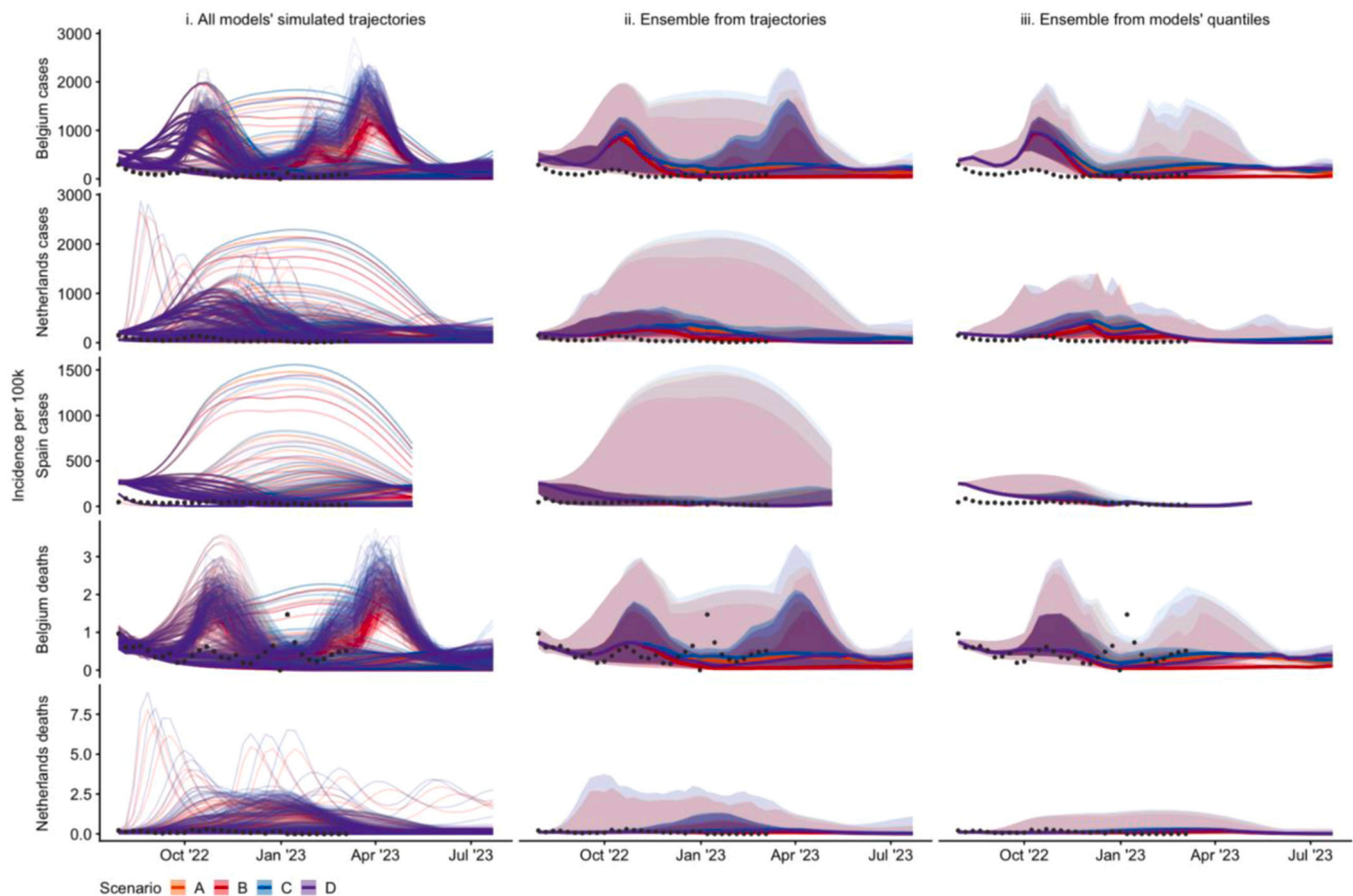
Fig. 1. Projections of incidence per 100,000 population, by country (row) and aggregation method (column) showing median, 50%, and 99% probabilistic intervals (increasingly shaded ribbons), for each scenario, using: i) no ensemble method (100 simulated trajectories per model, or 96 in case of one of the models); ii) quantile intervals of the distribution across all simulated trajectories; ii; a median across each model's projections at a given quantile interval. We do not show the linear opinion pool ensemble here as results are near-identical to the ensemble drawn directly from trajectories (ii). Scenarios included: an autumn second booster vaccine campaign among population aged 18+ (scenarios B & D) or 60+ (scenarios A & C); where vaccine effectiveness is 'optimistic' (effectiveness as of a booster vaccine against Delta; scenarios A & B) or 'pessimistic' (as against BA.4/BA.5/BA.2.75; scenarios C & D). See Supplement for further detail on individual models' trajectories.

scenarios (supplementary figure 1A). All ensembles produced similar values around the centre of the distribution, with no noticeable difference between the median values of each projection. However, across all five targets we observed that an ensemble based on either simulated trajectories, or an LOP ensemble, produced sharply increasing uncertainty between the 90–98% intervals. For example, at the upper 98% probability interval, ensemble projections for cases in Spain averaged nearly six times higher incidence when drawn directly from trajectories compared to when drawn from a median of three models' quantiles (respectively averaging 1016 and 173 weekly new cases per 100,000 population).

We then considered an ensemble of individual trajectories each weighted against a sequentially increasing amount of observed data (Fig. 2). We note that models used a variety of methods and may have been calibrated to alternative data sources (see Supplement). In comparison to the unweighted ensemble (shown in grey), we observed reduced uncertainty across weighted ensemble projections. Compared to conditioning on data up to 16 weeks before, adding 8 weeks of

additional data in weighting case projections reduced the upper 98% bound of uncertainty by at least 5% and up to 30% on average (supplementary figure 1B). The accuracy-weighted contribution of each trajectory to an ensemble varied substantially between models and targets, and over time. For example, in Spain each trajectory's weight remained stable after mid December 2022, reflecting the data by effectively downweighting those trajectories projecting sustained high incidence over winter (see Fig. 1i).

We used this information to create consecutive weekly ensembles, with weights updating as increasing observed data became available to measure trajectories' accuracies. In the combined (weighted interval) score, forecasts using weighted trajectories generally performed similarly to the unweighted equivalent, with a median relative WIS among the weighted ensembles of 0.99 (IQR: 0.89–1.05; supplementary figure 2).

When using the full 31 weeks of available data, a weighted ensemble performance improved compared to projections made without weighting on accuracy (with a median relative WIS across targets of 0.77
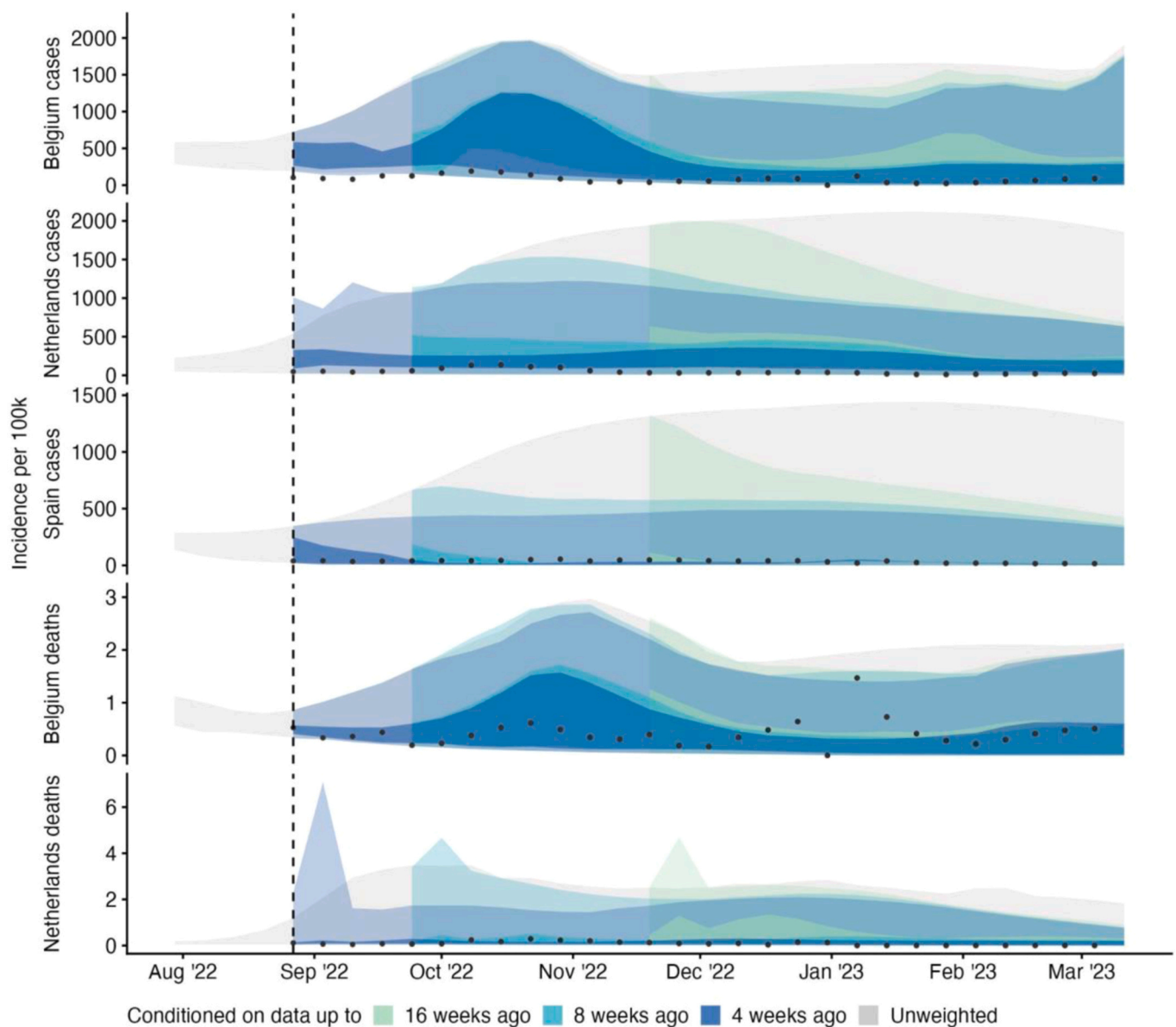
Fig. 2. Ensemble forecasts of incidence by target, using no weighting (grey ribbon), or 4, 8, and 16 weeks ahead of available data, with available data increasing weekly over time (coloured ribbons); showing 50% and 99% credible intervals. Each simulated trajectory started from 30 July 2022 and was weighted using its inverse mean absolute error against available data. We used at least 4 and up to 31 weeks of this observed accuracy data.

compared to the baseline of 1). However, this improvement was not linearly correlated with increasing data, and the relationship varied by target (Fig. 3). Weighted forecasts that used only a few data points of trajectories' accuracy performed similarly or poorly compared to the unweighted ensemble.

However, among three targets, using more accuracy data gave a stable or consistently improved performance (after 10 weeks for cases in Belgium and 17 weeks for cases and deaths in the Netherlands). This was also true of cases in Spain, with worse performance compared to the unweighted ensemble when using up to 9 weeks of data, and improving and then better relative performance after 17 weeks of data, until performance worsened once more after 27 weeks of accuracy data. In contrast, forecasts of deaths in Belgium were better with fewer weeks of accuracy data, and weighting with between 14 and 26 weeks' data produced a worse performance than the unweighted ensemble of trajectories.

4. Discussion

A significant part of the value of collaborative infectious disease modelling projects comes from the standardisation of model output across varying numbers of model teams, methods, and simulations. We compared two methods of collecting information from multiple models' projections of an epidemic. We took three scenario models for each of five projection targets, and contrasted collecting a sample of up to 100 simulated trajectories against collecting quantile intervals of those trajectories at each time step.

We found that collecting simulated trajectories enabled analysis of trajectory shapes, peaks, and cumulative total burden. We observed that trajectories contained a right-skewed probabilistic distribution, which meant that ensembles either directly from trajectories, or using a linear opinion pool method, increasingly diverged from the quantile-average ensemble in projecting the outer upper limit of the probabilistic distribution. We also found that collecting trajectories could be used to create

a competitively performing ensemble based on continuous predictive performance.

The common practice of collecting a standardised set of quantile intervals has several advantages. Firstly, combining across a set of quantiles should accurately represent the underlying distribution (Genest, 1992), and we observed that the linear opinion pool (based on a combination of quantiles) produced a near-identical ensemble as that created directly from combining individual trajectories. This suggests that the LOP ensemble may be the best choice for reflecting the widest range of uncertainty in settings where model results are only collected in quantiles, while noting that in order to create a LOP ensemble quantiles of cumulative rather than incident quantities need to be collected (Howerton et al., 2023). Furthermore, our results suggest little information about uncertainty is lost when using quantile outputs to compare the central estimates from different models. This is a useful validation for collecting multiple model results in any format when the purpose is short-term situational awareness.

Further advantages include where collecting quantile outputs also allows for a broader range of modelling methods, including quantile regression, that directly create quantile outputs rather than a joint distribution over time. Additionally, a single set of quantiles can be held in comma-separated value (csv) files of easily manageable size, requiring minimal technical knowledge of big data storage solutions or processing. This has been important in the past given a lack of readily available skills or investment in software for emergency outbreak settings. However this argument weakens with mounting evidence that this type of under-resourcing hampers outbreak response (Sherratt et al., 2024; Rivers et al., 2020).

An alternative method for multiple model collection is directly collecting models' trajectories, with the advantage of retaining each trajectory's time-dependence. We observed greater availability and flexibility of accessing information from this method in contrast to collecting quantile distributions. This was evident when comparing the tails of multiple distributions in a quantile-average ensemble, assessing
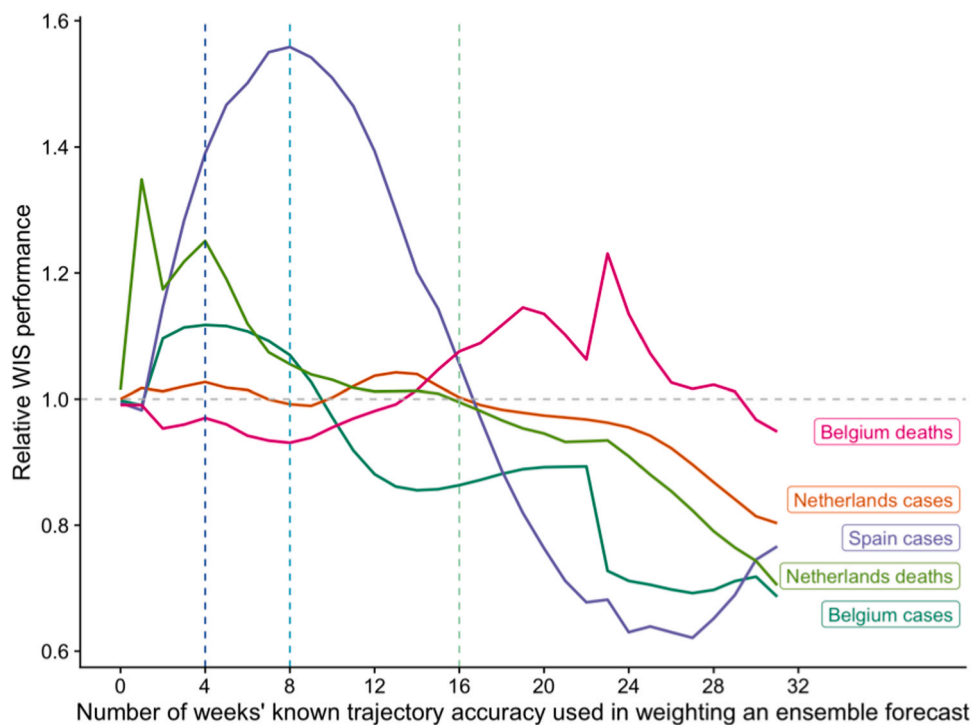
Fig. 3. Predictive performance of weighted ensembles by projection target. Weighted ensembles were created using a weighted median, where the weight of each trajectory was determined by its previous accuracy in predicting between 0 and 31 weeks of observed data (x axis). The performance of each ensemble is measured by the weighted interval score (WIS); a lower WIS score indicates better performance of the weighted ensemble than the simple unweighted median ensemble of all trajectories (reference line at 1).

the number of projected waves or the risk of crossing a specific threshold such as the burden in the preceding year, or in reevaluating projections against reported data. These analyses could also be conducted after collecting model outputs, making the method of collecting trajectories more flexible to the needs of one or multiple end-users. In particular these areas of information are more likely to be relevant to longer term preparedness and mitigation. As a result, we suggest the impact of information gains and losses from model collection may differ depending on the aim of a multi-model comparison.

Our findings comparing quantile with trajectory model outputs are compatible with ongoing work addressing issues from the loss of epidemic shape. From point forecasts, recent forecasting work has created an ensemble from multiple point forecasts in terms of similarity to canonical curve shapes (Srivastava et al., 2022). From probabilistic models it is also possible to create an ensemble of many trajectories using the centrality of each curve as a weight in a curve boxplot (Juul et al., 2021).

We have also demonstrated the potential for unique information gains when collecting simulated trajectories by assessing their performance against observed data. By conditioning the weight of each trajectory in an ensemble on subsequently observed data, we were able to create an ensemble that excluded entire trajectories, or epidemic curves, based on dependence to unrealised events. This typically either matched or reduced the uncertainty of an unweighted equivalent ensemble, and in some settings performed better overall than the unweighted equivalent.

This suggests an additional way in which collaborative modelling efforts can respond to changing outbreak dynamics and policy needs. For our setting, model results had originally been created based on a set of four scenarios relevant to policy decisions to be made in spring/ summer 2022. However, given the complex dynamics of disease transmission, no predefined future scenario is likely to accurately predict eventual reality. Among four scenarios with deliberately contrasting assumptions, most of these assumptions will be disproven by observation over time. Meanwhile, when scenario modelling outputs are collected as quantiles at each time-point, they lose their time-dependence and thus cannot be interpreted except in the light of an increasingly obsolete scenario context.

By focussing on individual model simulations in this work, we were able to abstract away from the context in which model results were created. We weighed each trajectory using only its past accuracy against observed data, regardless of the modelling technique, original scenario, or parameter values from which it arose. From this we created an ensemble that did not reflect any particular scenario assumptions, but only the time-varying accuracy of each trajectory. This meant we were able to continue to use trajectories in an ongoing evaluation, increasing the useful life of the results from a single cross-sectional collection of multiple model output.

This could be particularly useful when repeated rounds of model collection are time-intensive or computationally expensive, such as for individual-based models, or where personnel resources are constrained such as in an ongoing outbreak with potentially many competing priorities. Whilst beyond the scope of this work, future work in this area could also investigate model weights in order to rank trajectories from each scenario by proximity to observed trajectories and potentially interpret this as proximity of the given scenario assumptions to reality.

We highlight several important limitations to our comparison of information gains and losses between methods of collecting model output. In the first part of this work, we represent an analysis of trajectories that reflects our work in real-time response to the needs of policy decision-making. We did not consider alternative approaches to time-series analysis that would likely make the analysis of trajectories more robust, for example in calculations of peaks or wave durations.

In our comparison to quantile distributions, our method of collecting simulated trajectories was not specifically designed for this comparative purpose, and as a result our findings are difficult to interpret. In this

work we did not attempt to characterise how many samples might be sufficient to appropriately represent a probabilistic distribution in comparison to a quantile representation. For example, in some cases the collated trajectories were already subsampled from model runs conducted by individual teams. In contrast, in a situation with low sampling sizes of trajectories from each model, a quantile representation might provide a more stable representation.

We suggest that further work should characterise and standardise sampling techniques for model simulations in multi-model comparisons. Future study designs could focus on collating multiple representations (e.g. time-sliced quantiles and trajectories) from contributing teams directly for comparison, or collate arbitrary numbers of feasible simulated trajectories and re-weigh according to the number of simulations. Our work also demonstrates the importance of investing in and developing capacity to store and use simulation outputs rather than fixed-time quantile probabilities for well founded intercomparison modelling projects.

To conclude, we observed several information gains from collecting modelled trajectories rather than summarised quantile distributions. We highlight the potential to create continuous new information from a single collection of model output. Working from combined simulations offers the opportunity to explore creating ensembles by the shape of epidemic curve that can be updated over time, and for more detailed quantitative evaluations against observed data, such as in projected peaks or cumulative totals. We believe our findings apply whether projections are conditioned on the context of the present (as in forecasts), or on schematic futures (as in scenarios). However, the value of different information gains and losses may vary with the aims of each collaborative effort, depending on the requirements and flexibility required by projection users. Understanding potential information gains and losses when collecting model projections can support the accuracy, reliability, sustainability, and communication of collaborative infectious disease modelling efforts.

Funding declaration

Declaration of interest

CRediT authorship contribution statement

Rok Grah: Data curation, Writing – review & editing. **Ajitesh Srivastava:** Data curation, Writing – review & editing. **Kai Nagel:** Data

curation, Writing – review & editing. **Katharine Sherratt:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Christof Schütte:** Data curation, Writing – review & editing. **Tim Conrad:** Data curation, Writing – review & editing. **Sebastian Funk:** Conceptualization, Funding acquisition, Supervision, Validation, Writing – review & editing. **Maria Cristina Marinescu:** Data curation, Writing – review & editing. **Frank Sandmann:** Data curation, Writing – review & editing. **Aymar Cublier:** Data curation, Writing – review & editing. **Bastian Prasse:** Data curation, Writing – review & editing. **David E. Singh:** Data curation, Writing – review & editing. **Rene Niehus:** Data curation, Writing – review & editing. **Kylie Ainslie:** Data curation, Writing – review & editing. **Jakob Rehmann:** Data curation, Writing – review & editing. **Nicolas Franco:** Data curation, Writing – review & editing. **Alberto Cascajo Garcia:** Data curation, Writing – review & editing. **Jesus Carretero:** Data curation, Writing – review & editing. **Philippe Beutels:** Data curation, Writing – review & editing. **Christel Faes:** Data curation, Writing – review & editing. **Steven Abrams:** Data curation, Writing – review & editing. **Lander Willem:** Data curation, Writing – review & editing. **Ricardo Ewert:** Data curation, Writing – review & editing. **Billy Charlton:** Data curation, Writing – review & editing. **Sebastian Müller:** Data curation, Writing – review & editing. **Niel Hens:** Data curation, Writing – review & editing. **Sam Abbott:** Writing – review & editing. **Christian Rakow:** Data curation, Writing – review & editing. **Sydney Paltra:** Data curation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All code and data are available on Github at: epiforecasts/multi-model-information; DOI: https://doi.org/10.5281/zenodo.10891377 .

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.epidem.2024.100765.

References

Pedro J. Aphalo, ggpmisc: Miscellaneous Extensions to "ggplot2". 2023. [Online]. Available: ⟨https://docs.r4photobiology.info/ggpmisc/⟩.

Borchering, R.K., 2021. Modeling of future COVID-19 cases, hospitalizations, and deaths, by vaccination rates and nonpharmaceutical intervention scenarios — United States, April–September 2021. MMWR Morb. Mortal. Wkly. Rep. 70 https://doi.org/10.15585/mmwr.mm7019e3.

Bosse, N.I., Abbott, S., Cori, A., van Leeuwen, E., Bracher, J., Funk, S., 2023. Scoring epidemiological forecasts on transformed scales. PLoS Comput. Biol. 19 (8), e1011393 https://doi.org/10.1371/journal.pcbi.1011393.

Bracher, J., Ray, E.L., Gneiting, T., Reich, N.G., 2021a. Evaluating epidemic forecasts in an interval format. PLoS Comput. Biol. 17 (2), e1008618 https://doi.org/10.1371/journal.pcbi.1008618.

Bracher, J., et al., 2021b. A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. Nat. Commun. 12 (1), 5173. https://doi.org/10.1038/s41467-021-25207-0.

Cramer, E.Y., et al., 2021. The United States COVID-19 Forecast Hub dataset. medRxiv 2021. https://doi.org/10.1101/2021.11.04.21265886.

Cramer, E.Y., et al., 2022. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proc. Natl. Acad. Sci. USA 119 (15), e2113561119. https://doi.org/10.1073/pnas.2113561119.

Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. 20 (5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1.

European COVID-19 Scenario Hub, Round 2. [Online]. Available: ⟨https://covid19scenariohub.eu/report2.html⟩.

Funk, S., et al., 2020. Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. medRxiv 2020. https://doi.org/10.1101/2020.11.11.20220962.

Genest, C., 1992. Vincentization revisited. Ann. Stat. 20 (2), 1137–1142.

Harrell, F.E., Davis, C.E., 1982. A new distribution-free quantile estimator. Biometrika 69 (3), 635–640. https://doi.org/10.1093/biomet/69.3.635.

Howerton, E., et al., 2023. Context-dependent representation of within- and between-model uncertainty: aggregating probabilistic predictions in infectious disease epidemiology. J. R. Soc. Interface 20 (198), 20220659. https://doi.org/10.1098/rsif.2022.0659.

Juul, J.L., Græsbøll, K., Christiansen, L.E., Lehmann, S., 2021. Fixed-time descriptive statistics underestimate extremes of epidemic curve ensembles. Nat. Phys. 17 (1) https://doi.org/10.1038/s41567-020-01121-y.

Lenhard, A., Lenhard, W., Gary, S., cNORM - Generating Continuous Test Norms. 2018⟨https://doi.org/10.13140/RG.2.2.25821.26082⟩.

Li, S.-L., et al., 2017. Essential information: Uncertainty and optimal control of Ebola outbreaks. Proc. Natl. Acad. Sci. USA 114 (22), 5659–5664. https://doi.org/10.1073/pnas.1617482114.

Lipsitch, M., Finelli, L., Heffernan, R.T., Leung, G.M., Redd, S.C., 2011. Improving the evidence base for decision making during a pandemic: the example of 2009 influenza A/H1N1. Biosecurity Bioterrorism Biodefense Strategy Pract. Sci. 9 (2), 89–115. https://doi.org/10.1089/bsp.2011.0007.

McCabe, R., et al., 2021. Communicating uncertainty in epidemic models. Epidemics 37, 100520. https://doi.org/10.1016/j.epidem.2021.100520.

Ray, E.L., et al., 2020. Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. medRxiv. https://doi.org/10.1101/2020.08.19.20177493.

Reich, N.G., et al., 2022. Collaborative hubs: making the most of predictive epidemic modeling. Am. J. Public Health 112 (6), 839–842. https://doi.org/10.2105/AJPH.2022.306831.

Rhodes, T., Lancaster, K., Lees, S., Parker, M., 2020. Modelling the pandemic: attuning models to their contexts. BMJ Glob. Health 5 (6), e002914. https://doi.org/10.1136/bmjgh-2020-002914.

Rivers, C., Martin, E., Meyer, D., Inglesby, T.V., Cicero, A.J., Cizek, J. 2020. Modernizing and expanding outbreak science to support better decision making during public health crises: Lessons for COVID-19 and beyond. The Johns Hopkins Center for Health Security. Available at: https://centerforhealthsecurity.org/sites/default/files/2023-02/200324-outbreak-science.pdf . Accessed 1 March 2024.

Runge, M.C., et al., 2023. Scenario design for infectious disease projections: integrating concepts from decision analysis and experimental design. medRxiv, 2023.10.11.23296887. https://doi.org/10.1101/2023.10.11.23296887.

Shea, K., et al., 2020. Harnessing multiple models for outbreak management. Science 368 (6491), 577–579. https://doi.org/10.1126/science.abb9934.

Sherratt, K., et al., 2023. Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. eLife 12, e81916. https://doi.org/10.7554/eLife.81916.

Sherratt, K., et al., 2024. Improving modelling for epidemic responses: reflections from members of the UK infectious disease modelling community on their experiences during the COVID-19 pandemic [version 1; peer review: awaiting peer review]. Wellcome Open Res. 9 (12) https://doi.org/10.12688/wellcomeopenres.19601.1.

Sherratt, K., & Funk, S. (2024). epiforecasts/multi-model-information: Publication release (v1.1). Zenodo. https://doi.org/10.5281/zenodo.10891377.

Srivastava, A., Singh, S., Lee, F., Shape-based evaluation of epidemic forecasts. arXiv, Nov. 11, 2022⟨https://doi.org/10.48550/arXiv.2209.04035⟩.

Swallow, B., et al., 2022. Challenges in estimation, uncertainty quantification and elicitation for pandemic modelling. Epidemics 38, 100547. https://doi.org/10.1016/j.epidem.2022.100547.

Taylor, J.W., Taylor, K.S., 2021. Combining Probabilistic Forecasts of COVID-19 Mortality in the United States. Eur. J. Oper. Res. https://doi.org/10.1016/j.ejor.2021.06.044.

Viboud, C., et al., 2018. The RAPIDD ebola forecasting challenge: synthesis and lessons learnt. Epidemics 22, 13–21. https://doi.org/10.1016/j.epidem.2017.08.002.

Zelner, J., Riou, J., Etzioni, R., Gelman, A., 2021. Accounting for uncertainty during a pandemic. Patterns 2 (8). https://doi.org/10.1016/j.patter.2021.100310.

Supplementary Information

SI Figure 1
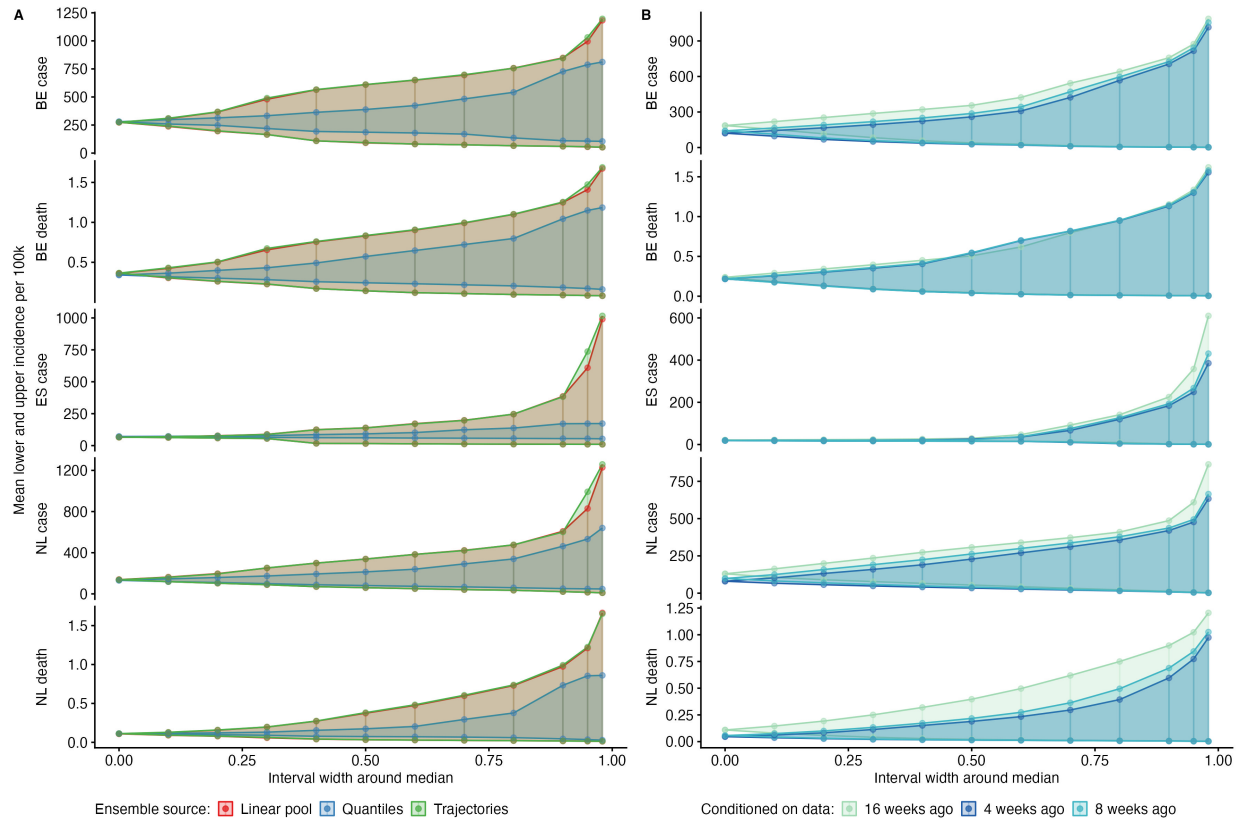
SI Figure 1. Mean central prediction intervals at increasing distances from the median. The 52-week mean of incidence per 100,000 population across all time points and scenarios, showing mean central prediction intervals at increasing distances from the median (interval width), by aggregation method (A) or weighting (B). The median estimate for each ensemble has 0 interval width (x-axis), with uncertainty increasing until an interval width at 0.98 represents the 1%-99% credibility interval around the median.

SI Figure 2
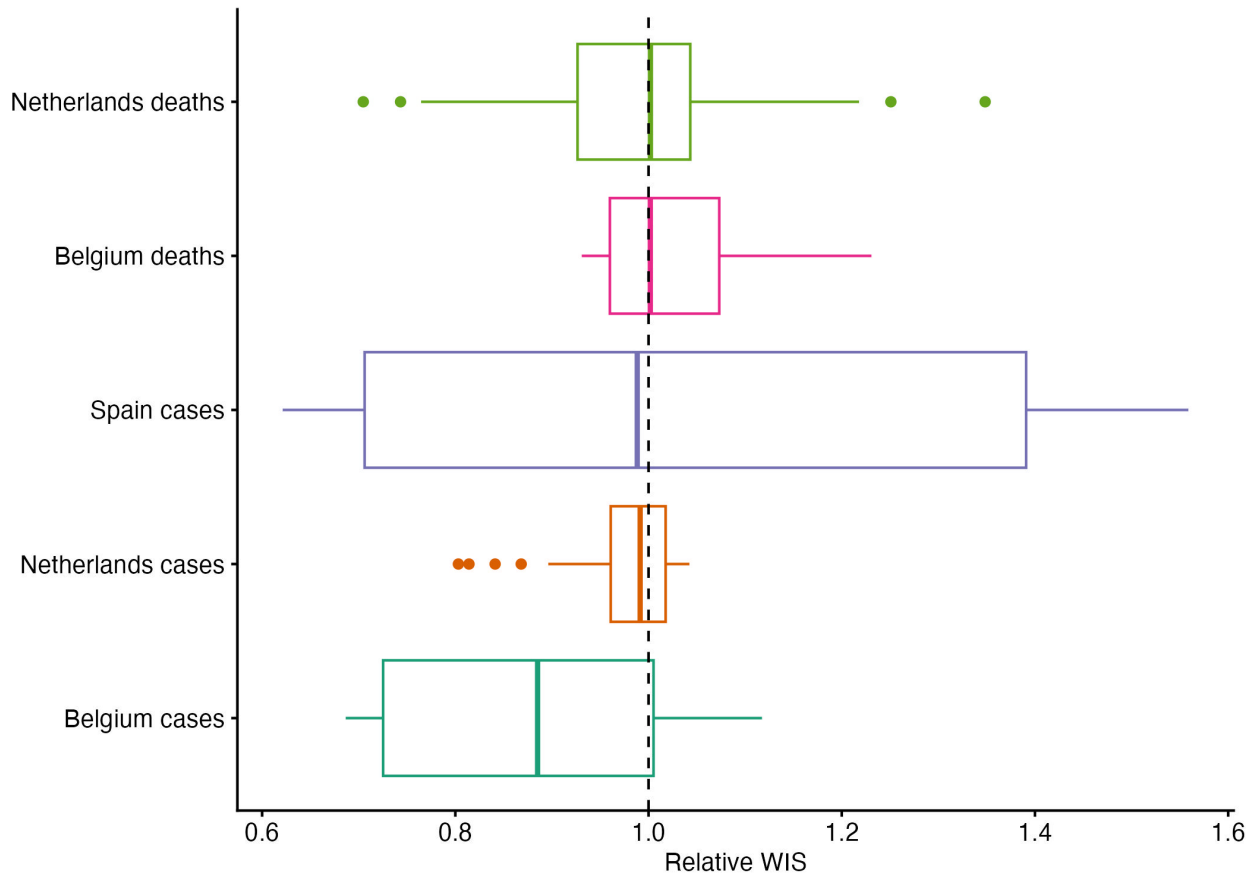
SI Figure 2. Distribution of forecast performance scores (relative WIS), of forecasts from model trajectories weighted using 0 through 31 weeks' available data. Performance is compared to an unweighted ensemble (reference line at 1).

SI Table 1

Team	Methods
ECDC	
ECDC-CM_ONE	Discrete-time, deterministic, mean-field SEIR-type compartmental model on metapopulation level. Population divided by age, vaccination status, and previous recovery; incl. seasonality, BA2 & behavior.
Dutch National Institute of Public Health and the Environment (RIVM)	
RIVM-vacamole	Deterministic, age-structured SEIR model, accounting for differences in susceptibility/infectiousness by age, seasonality, contact patterns, modes of vaccine protection, and waning immunity.
SIMID	
SIMID-SCM	Stochastic age-structured discrete time extended compartmental model
Universidad Carlos III de Madrid	

Team	Methods
UC3M-EpiGraph	Agent-based parallel simulator that models individual interactions extracted from social networks and demographical data.
University of Southern California	
USC-SIkJalpha	Uses SIKJalpha which models temporally varying infection, death, and hospitalization rates. Learning is performed by reducing the problem to multiple simple linear regression problems.

SI Table 1. Teams that contributed models to Round 2 of the European Scenario Hub, with self-described methods and links to further information. See also:

- Full model metadata, at: https://github.com/covid19-forecast-hub-europe/covid19-scenario-hub-europe/tree/main/model-metadata
- Information about each model's assumptions for Round 2, at: https://github.com/covid19-forecast-hub-europe/covid19-scenario-hub-europe/tree/main/model-abstracts/2022-07-24

Round 2 report

The following pages are the original website reporting for the European Scenario Hub Round 2 as of July 2022.

The report is currently (January 2023) available at: https://covid19scenariohub.eu/report2.html

Code to generate this report is available at: https://github.com/european-modelling-hubs/covid19-scenario-hub-europe-website/blob/main/report2.Rmd

Round 2

Scenarios

We asked teams of researchers across Europe to use quantitative models to project COVID-19 outcomes for 32 European countries over the next year. In order to explore different sets of assumptions about drivers of the pandemic, we asked teams to vary four sets of parameters. We can describe this in a 2x2 scenario specification:

	Age 60+ booster campaign • 2nd* booster recommended for 60+ • Uptake starts 15th September, and reaches 50% coverage by 15th December	**Age 18+ booster campaign** • 2nd* booster recommended for general population, ages 18+ • Uptake starts 15th September, and reaches 50% coverage by 15th December
Optimistic vaccine effectiveness • Increased booster vaccine effectiveness to that seen against **Delta variant**	Scenario A	Scenario B
Pessimistic vaccine effectiveness • Reduced booster vaccine effectiveness against infection from **BA.4/BA.5/BA.2.75 variants**	Scenario C	Scenario D

See also the full scenario details (https://github.com/covid19-forecast-hub-europe/covid19-scenario-hub-europe/wiki/Round-2) for more detail on the common set of assumptions teams used to create their models.

In Round 2, we asked modellers to start their projections from the 2022-07-24. Data after this date were not included, and as a result, model projections are unlikely to fully account for later information on the changing variants or behavioural patterns.

In this report we only show results from countries with at least 3 models.

Current situation

We consider vaccination rates in countries for which multiple teams of modellers contributed projections.

Vaccination % among 60+ population, relative to total population
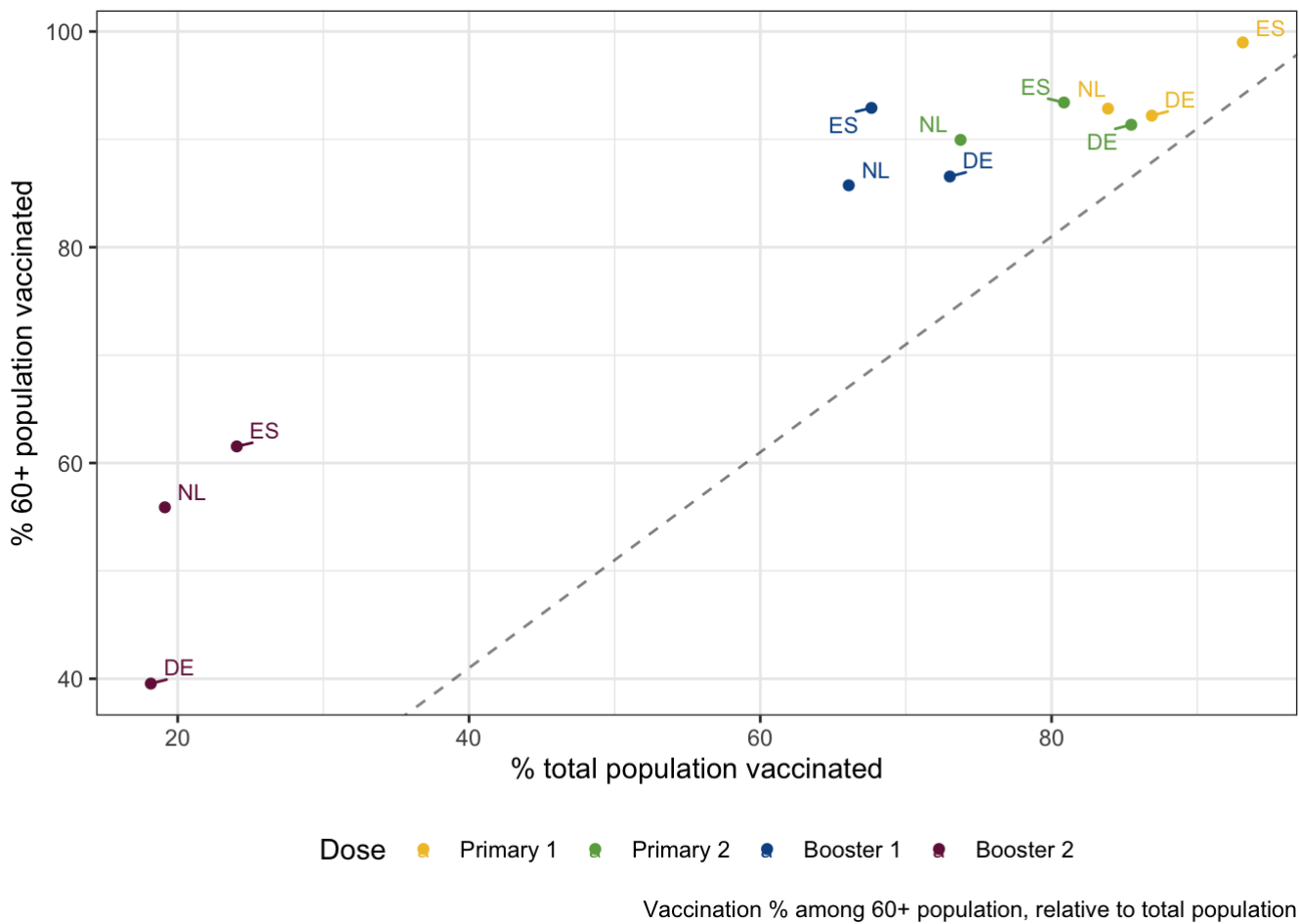
Participating teams

6 models contributed scenario projections to Round 2.

Models

Participating teams by number of countries and horizon

Team	Countries	Weeks
USC-SIkJalpha	31	52
ECDC-CM_ONE	28	53
MODUS_Covid-Episim	1	53
RIVM-vacamole	1	53
SIMID-SCM	1	52
UC3M-EpiGraph	1	41

Countries

Number of independent model projections for each target variable and location

Code	Country	Infection	Case	Hosp	Icu	Death
BE	Belgium	1	3	2	1	3
DE	Germany	1	2	2	0	1
ES	Spain	1	3	2	0	2

128

Code	Country	Infection	Case	Hosp	Icu	Death
NL	Netherlands	1	3	2	1	3

Cumulative outcomes

For each model and scenario, we compare the total number of outcomes over the entire projection period as a % of the total country population. We compared the cumulative number of projected outcomes to the cumulative total over one year before projections started (July 2021 to July 2022).

Death

Total incident death over projection period
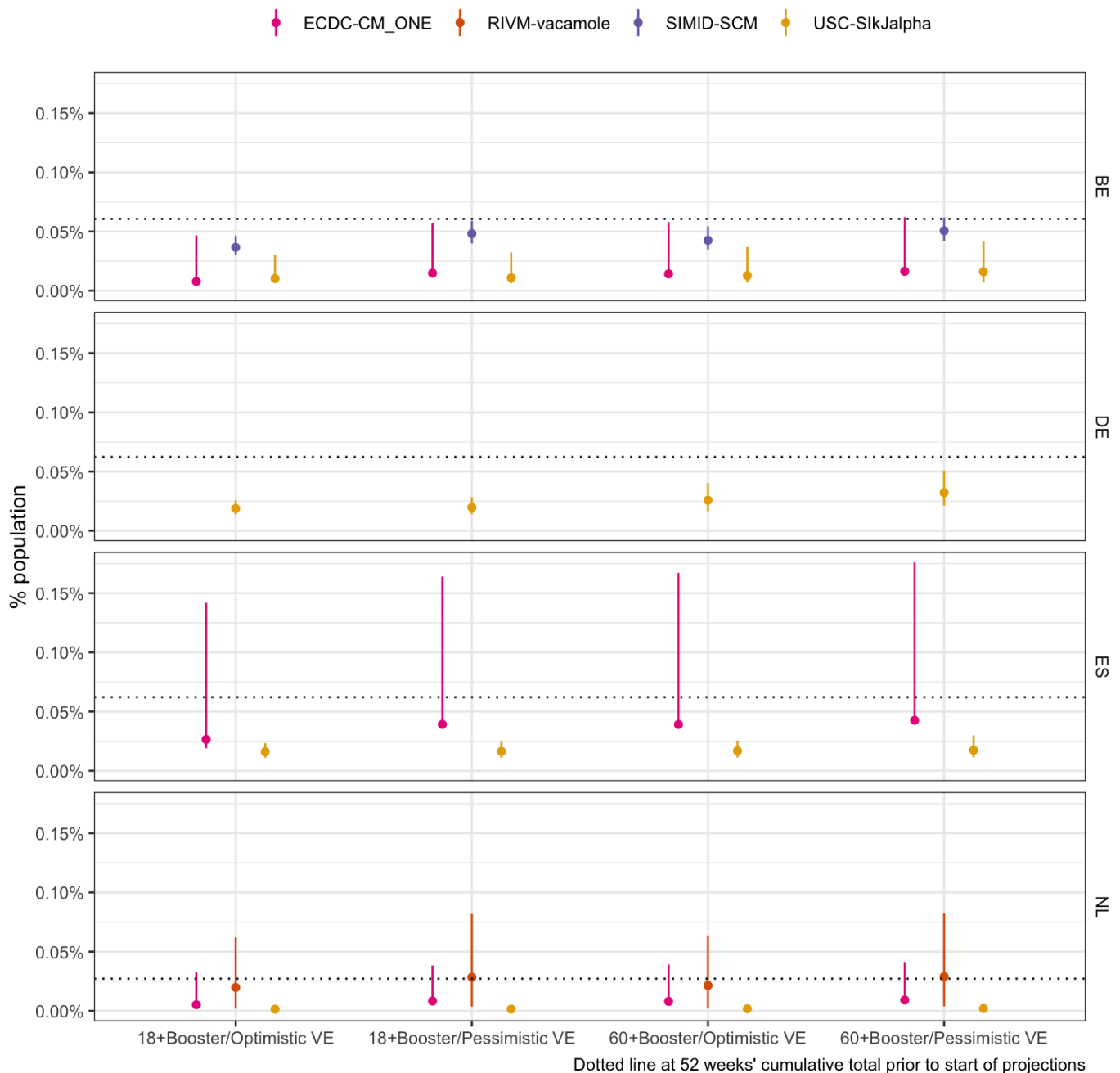
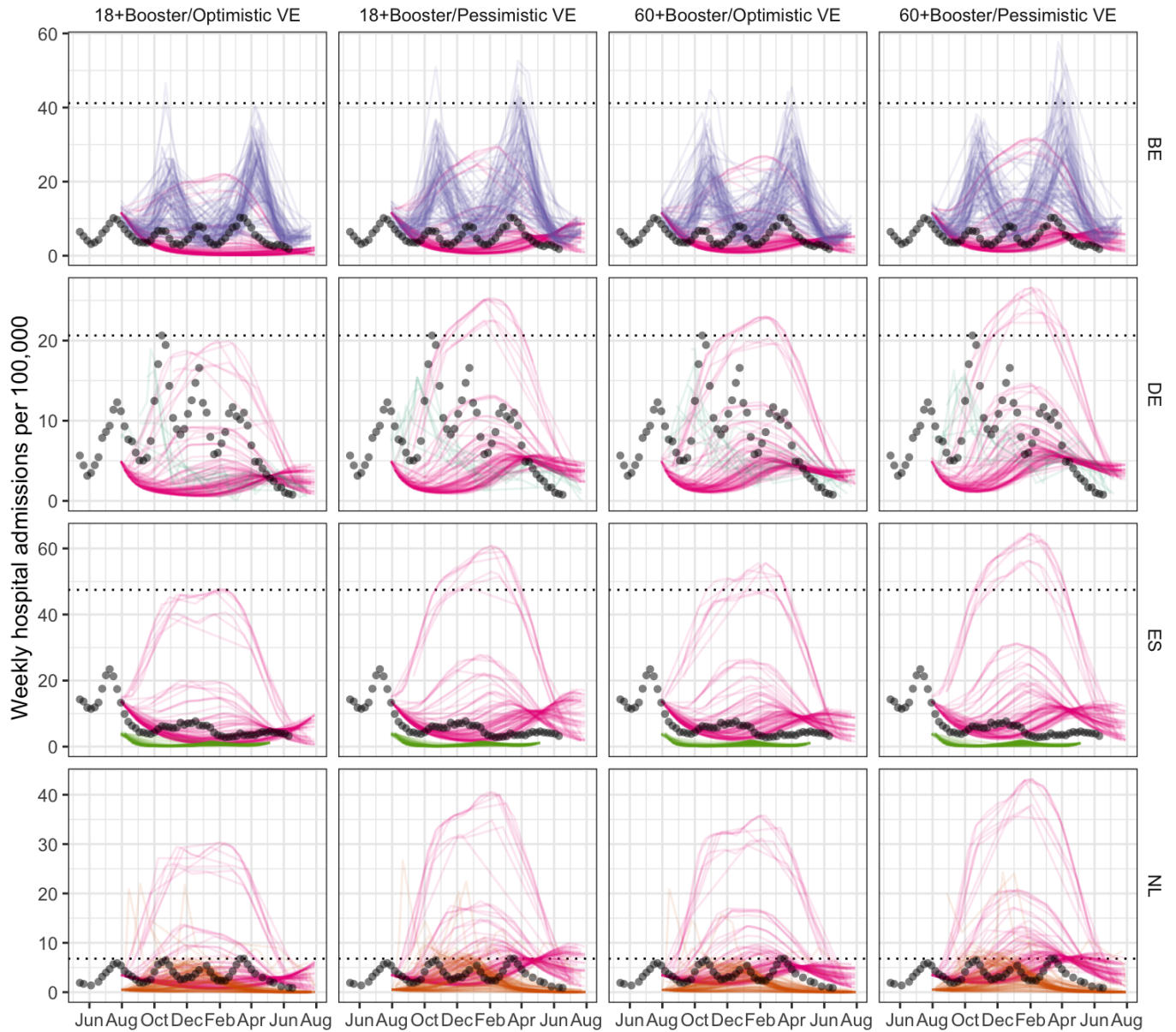Dotted line at 52 weeks' cumulative total prior to start of projections

Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)

Case

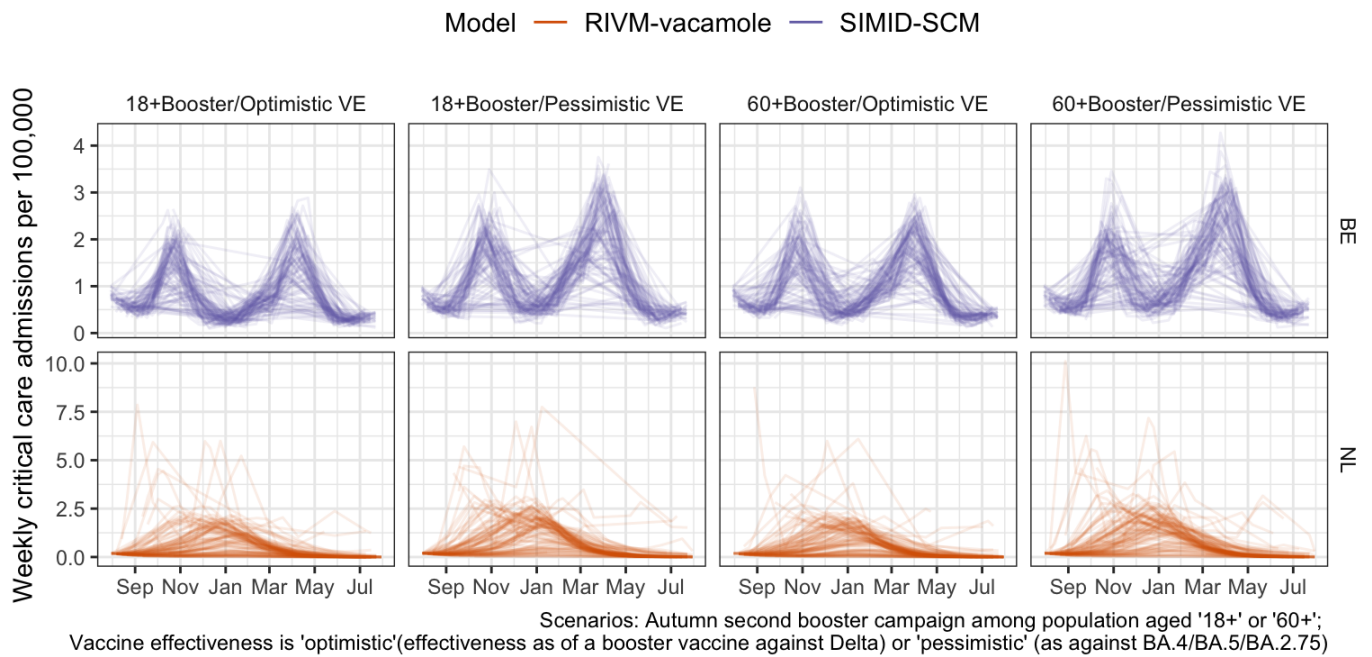Total incident case over projection period
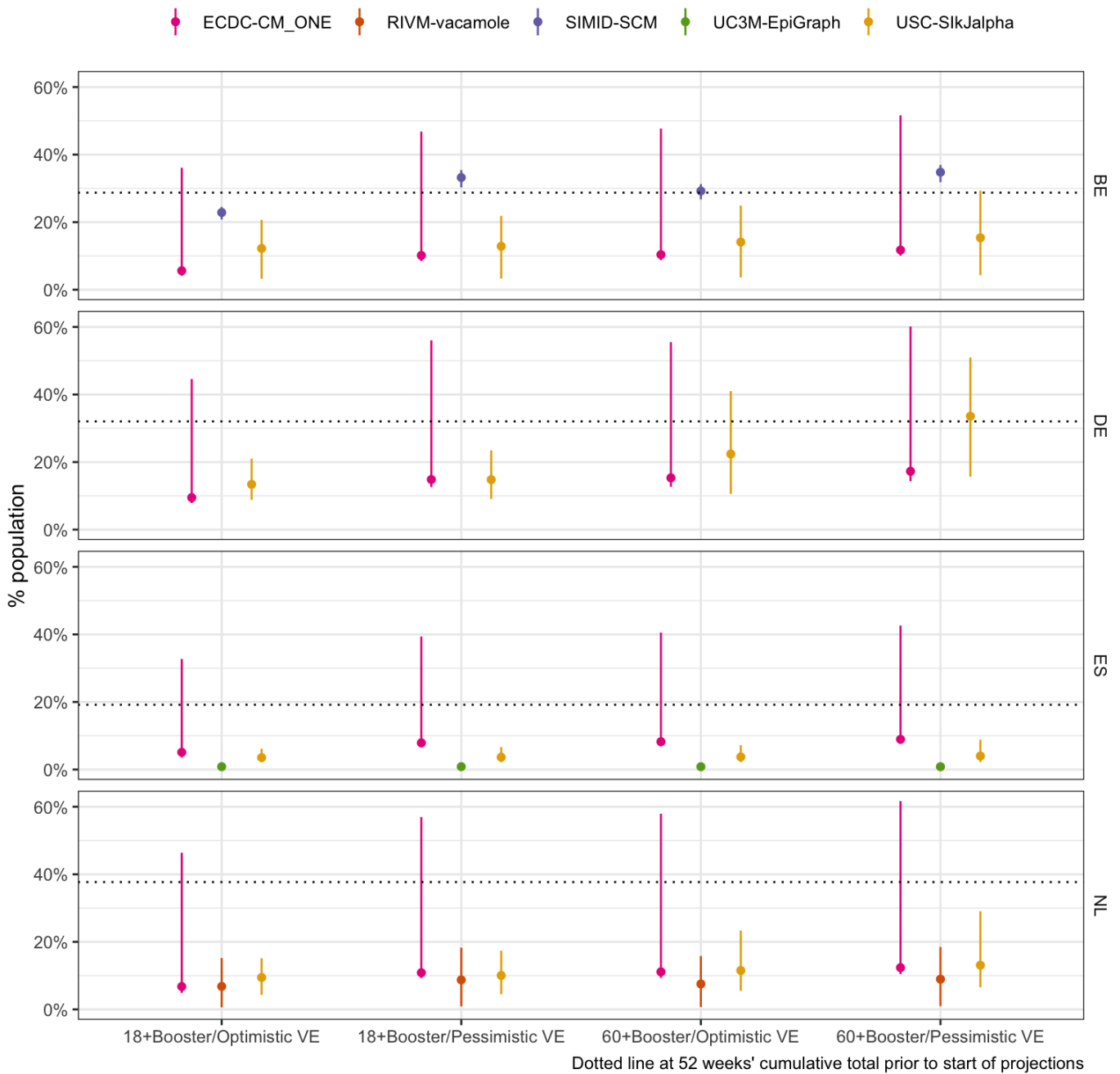
Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)
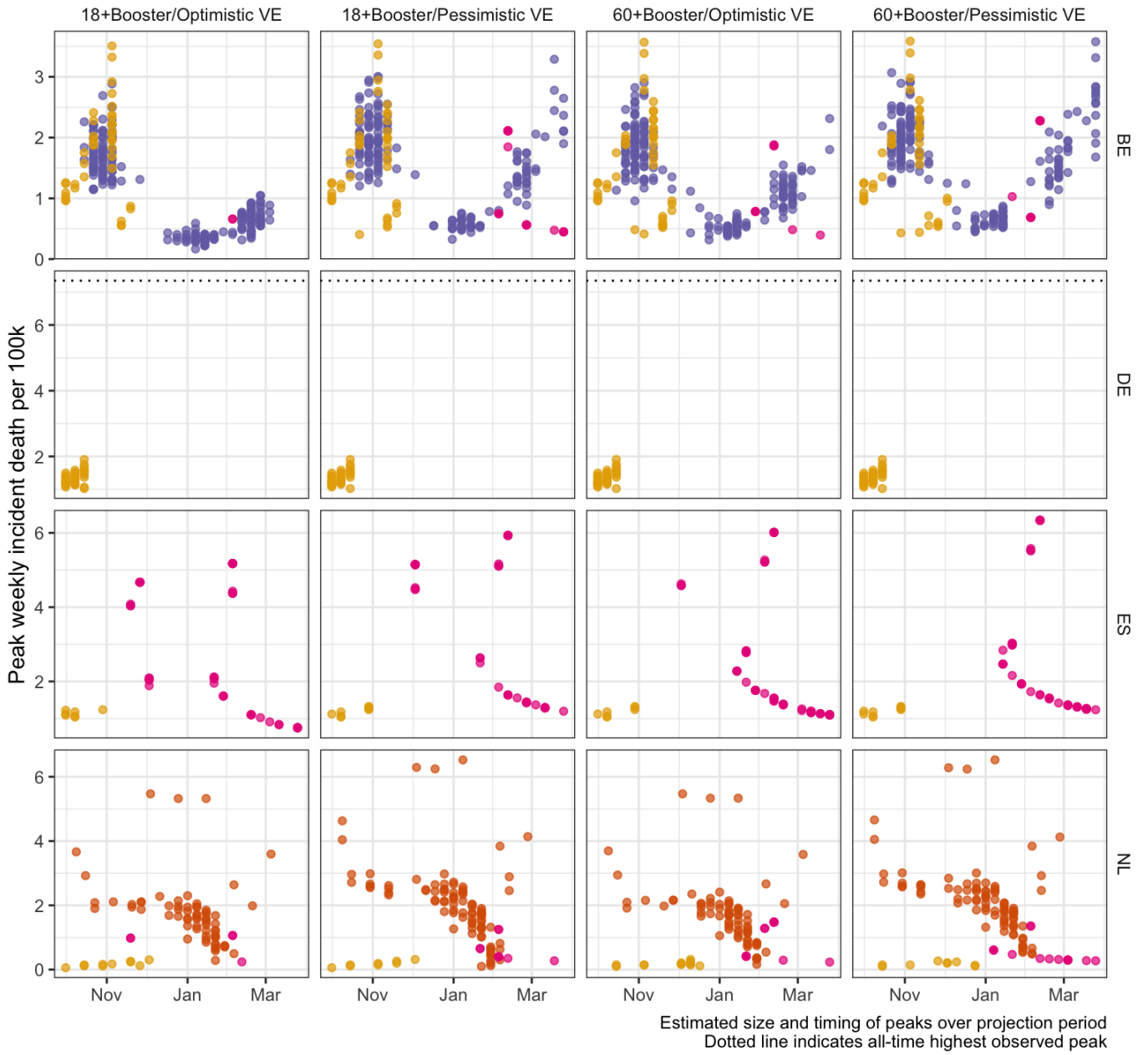
Infection

Total incident infection over projection period
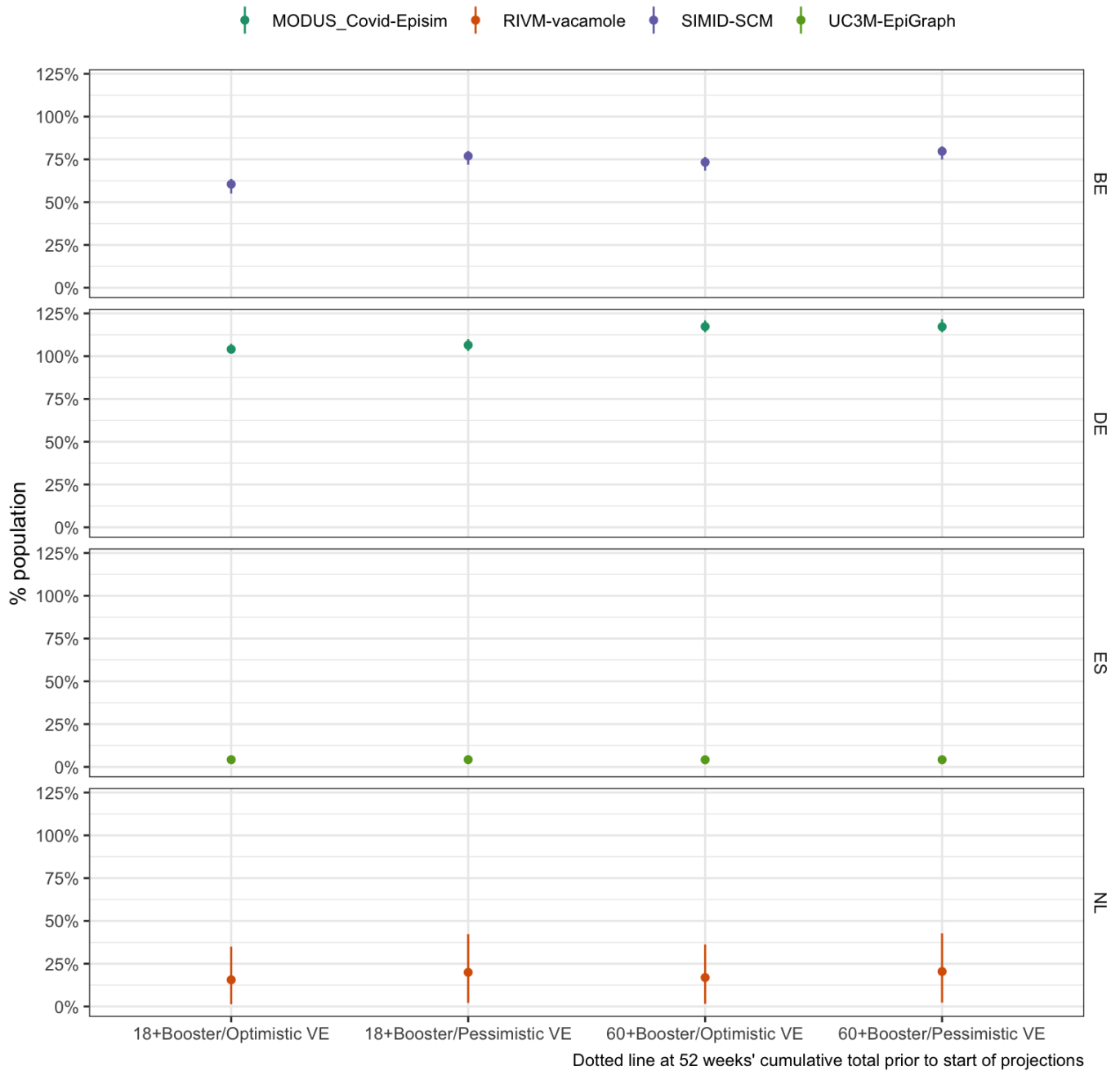
Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)

Hosp

Total incident hosp over projection period
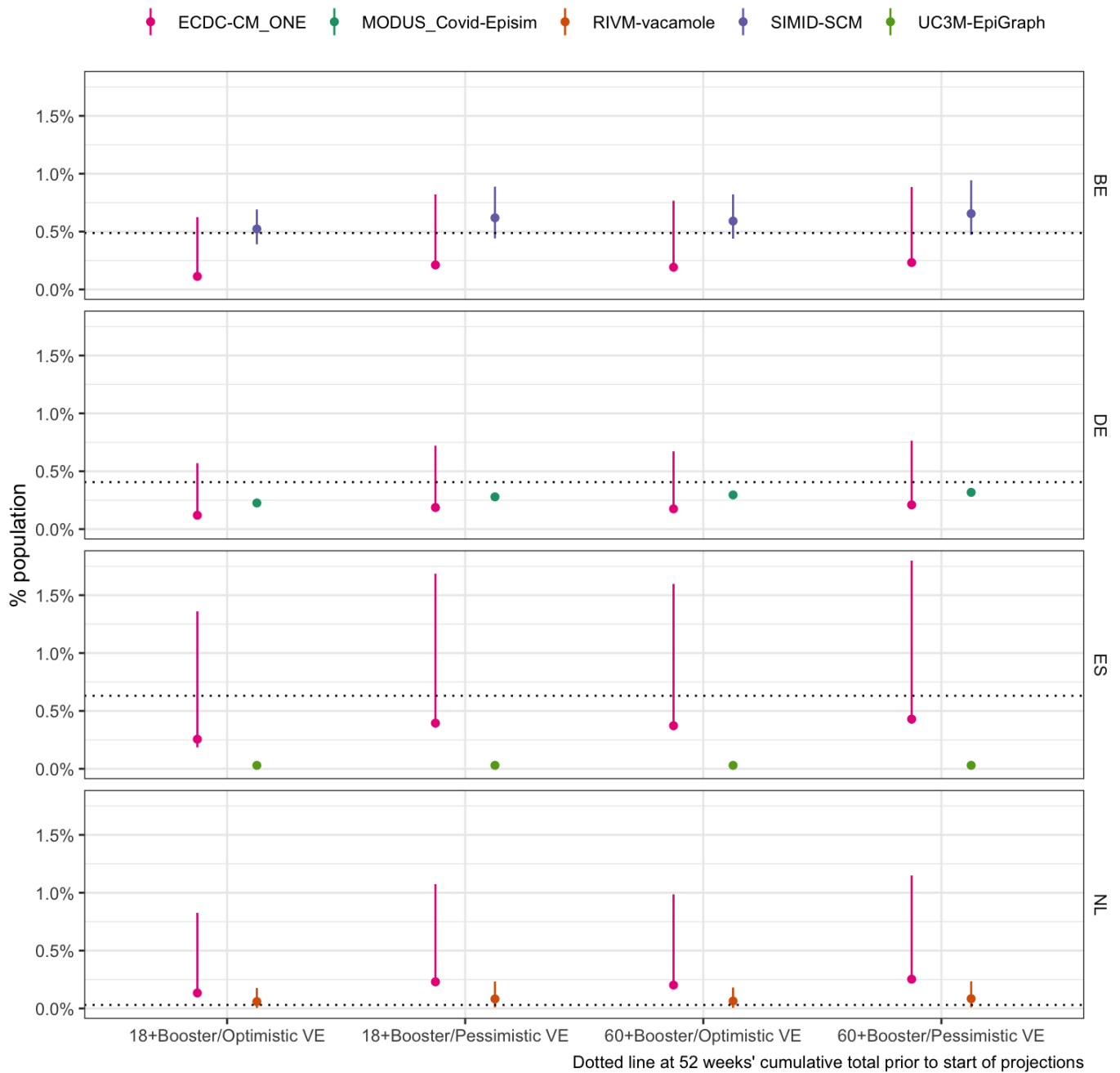
Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)

Icu

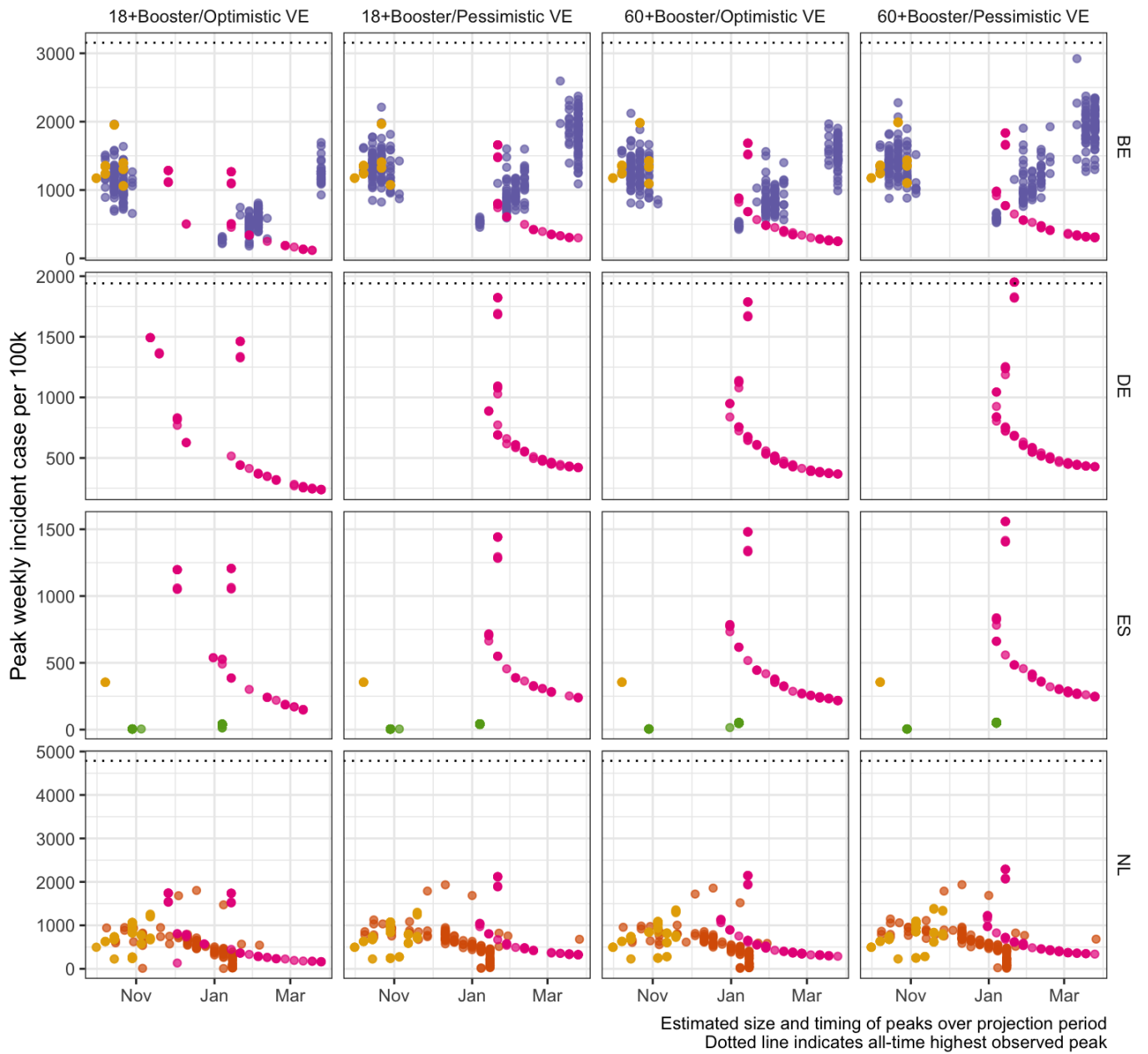Total incident icu over projection period
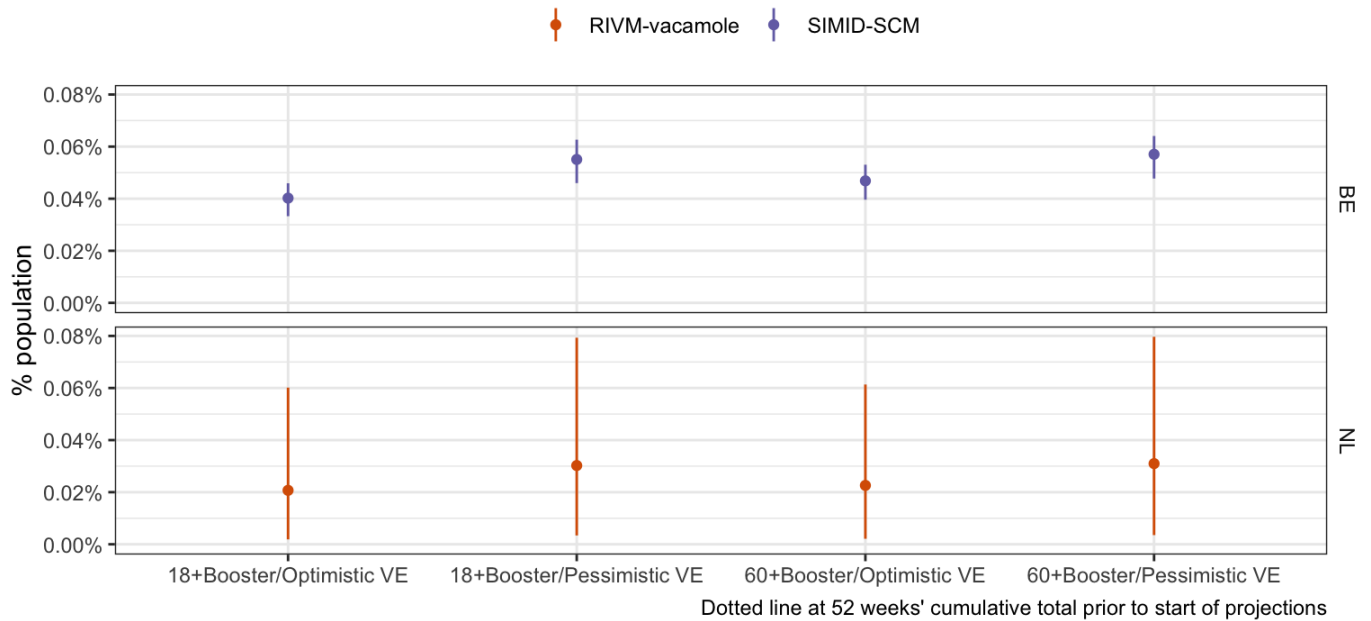
Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)
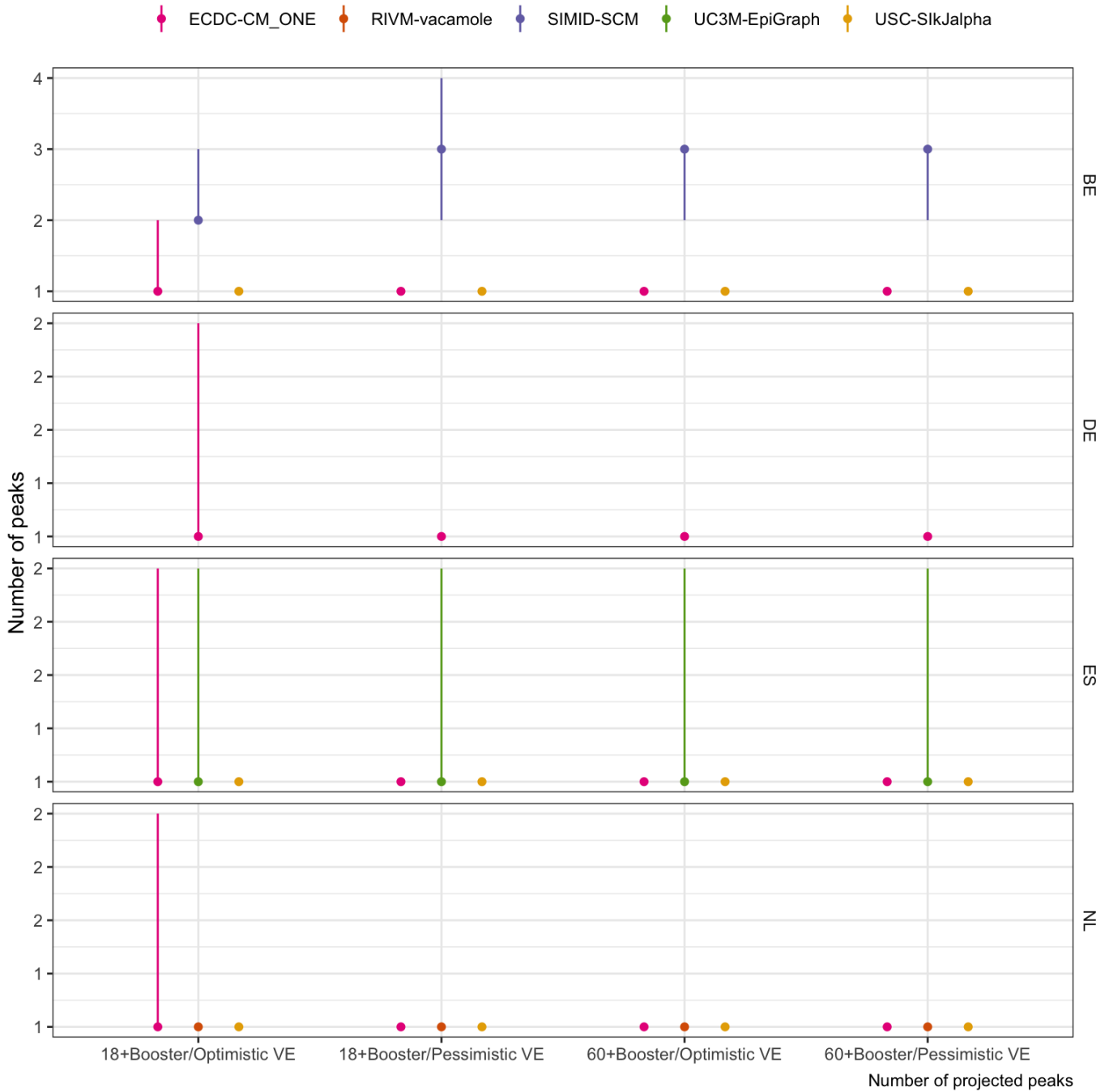
Incident outcomes

We explored the incidence of COVID-19 per 100,000 over the projection period and in terms of projected peaks in incidence. We summarised peaks both over the entire projection period, and over only the autumn-winter period (October through March); we considered (A) the timing and maximum weekly incidence of each peak, and (B) the total number of peaks.

Trajectories
Death

Scenarios: Autumn second booster campaign among population aged '18+' or '60+';
Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)

Case

Scenarios: Autumn second booster campaign among population aged '18+' or '60+';
Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)
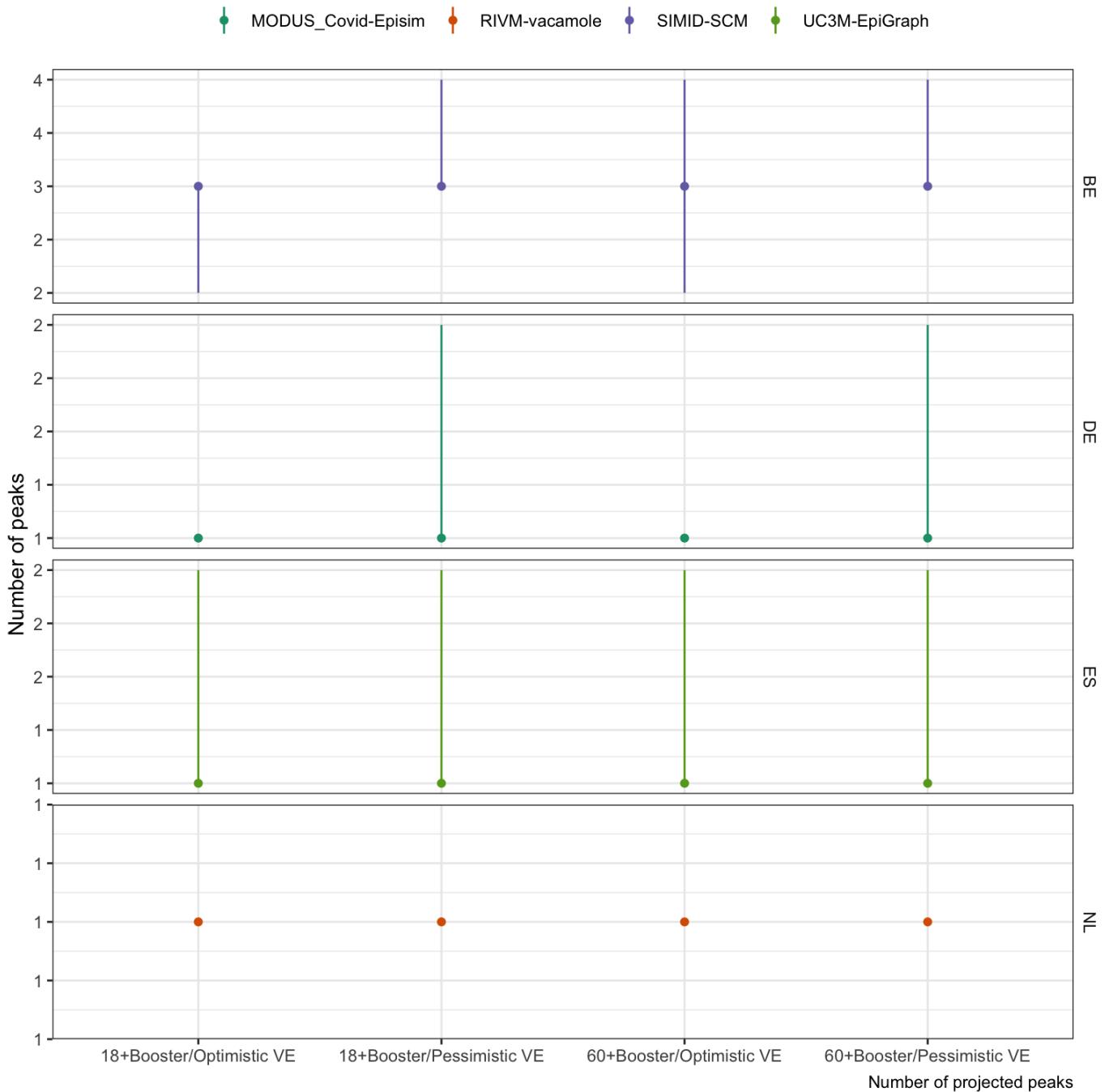
Infection

Scenarios: Autumn second booster campaign among population aged '18+' or '60+';
Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)

Hosp

Scenarios: Autumn second booster campaign among population aged '18+' or '60+';
Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)
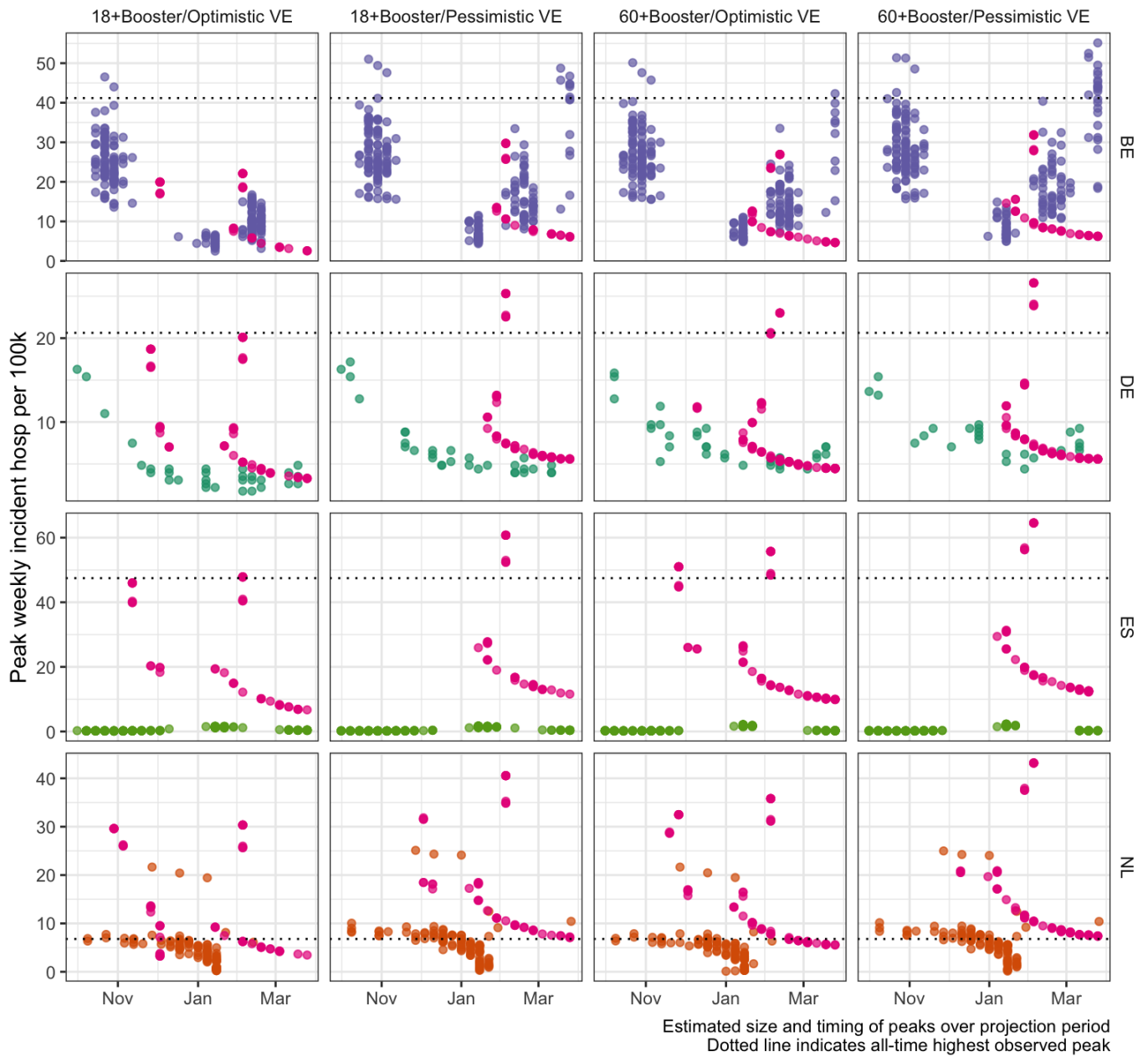
Icu

Scenarios: Autumn second booster campaign among population aged '18+' or '60+';
Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)

Peaks

Autumn-winter

Projections over October 2022 through March 2023

Death

A. Size and timing of peaks. Boxplots show summary of the likely value at peak incidence (median and interquartile range); points show timing and size of peaks from independent sample simulations

Estimated size and timing of peaks over projection period
Dotted line indicates all-time highest observed peak

B. Projected number of peaks (median with 5-95% probability)

Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)
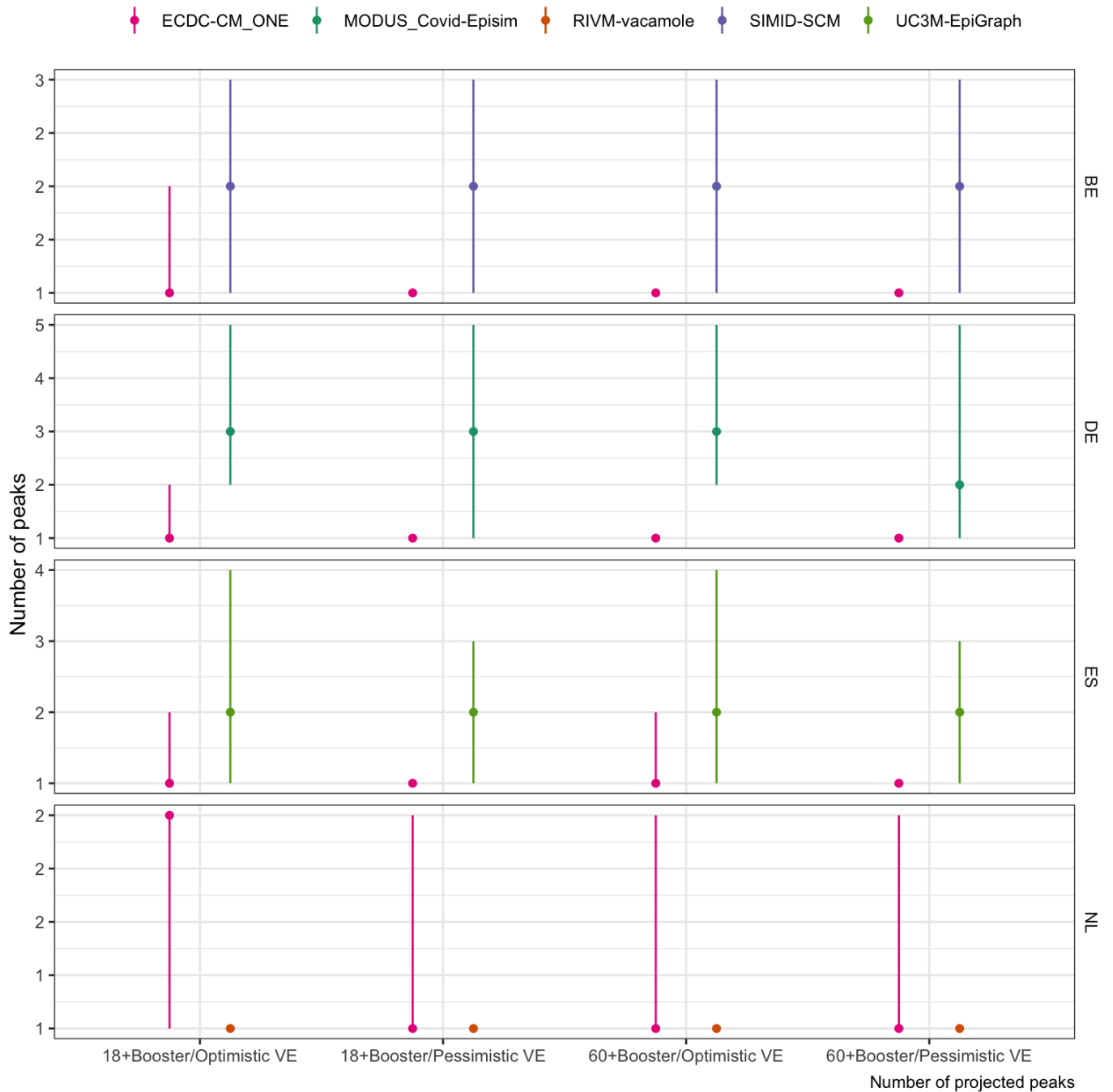
Case

A. Size and timing of peaks. Boxplots show summary of the likely value at peak incidence (median and interquartile range); points show timing and size of peaks from independent sample simulations

Estimated size and timing of peaks over projection period
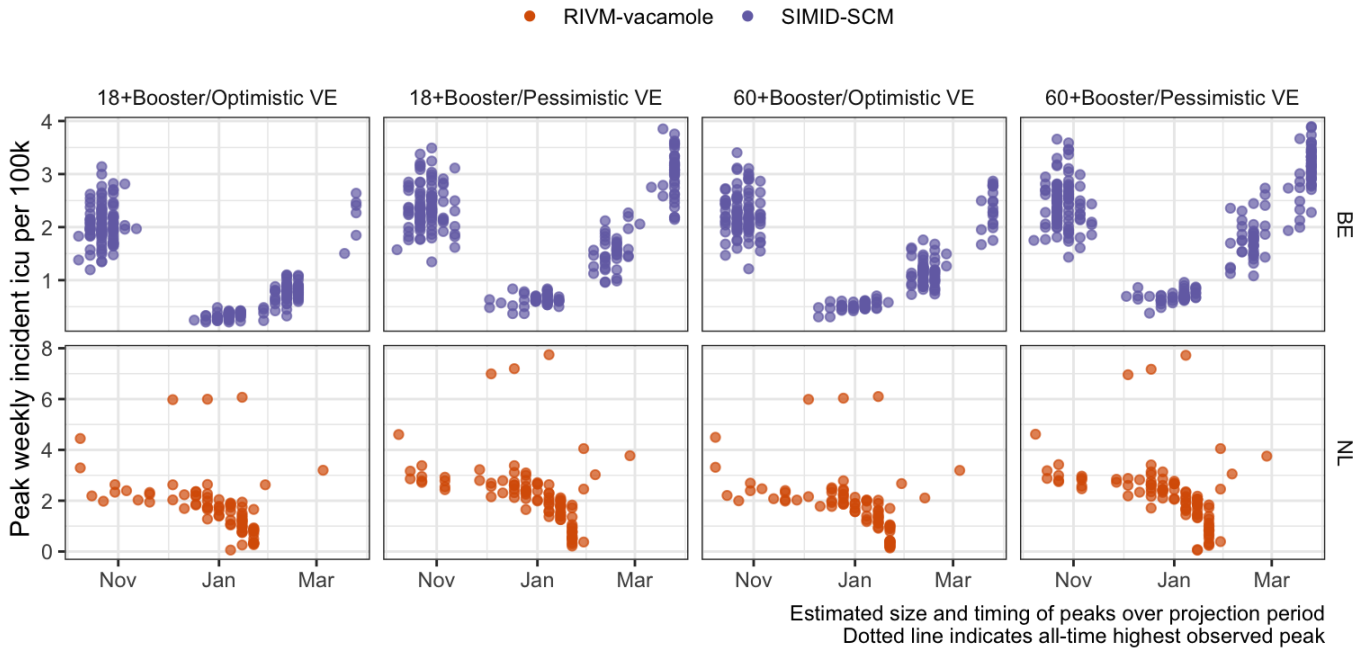Dotted line indicates all-time highest observed peak

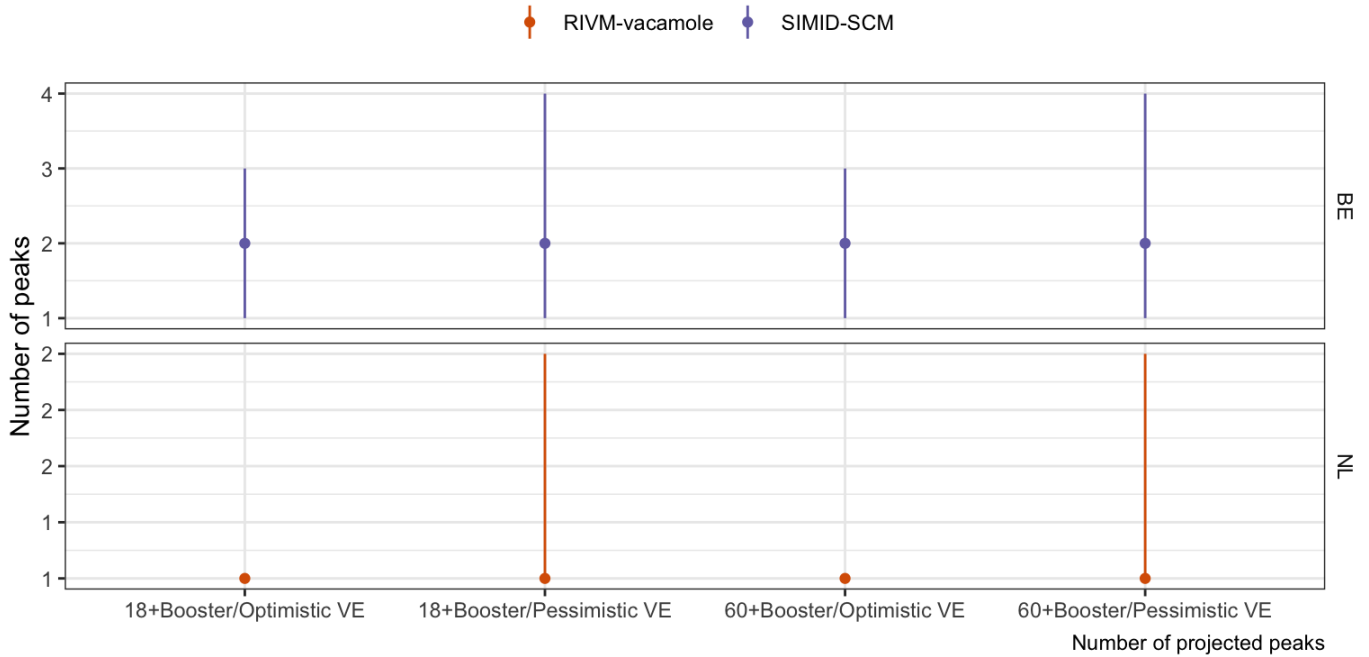B. Projected number of peaks (median with 5-95% probability)

Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)

Infection

A. Size and timing of peaks. Boxplots show summary of the likely value at peak incidence (median and interquartile range); points show timing and size of peaks from independent sample simulations

Estimated size and timing of peaks over projection period
Dotted line indicates all-time highest observed peak

B. Projected number of peaks (median with 5-95% probability)

Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)
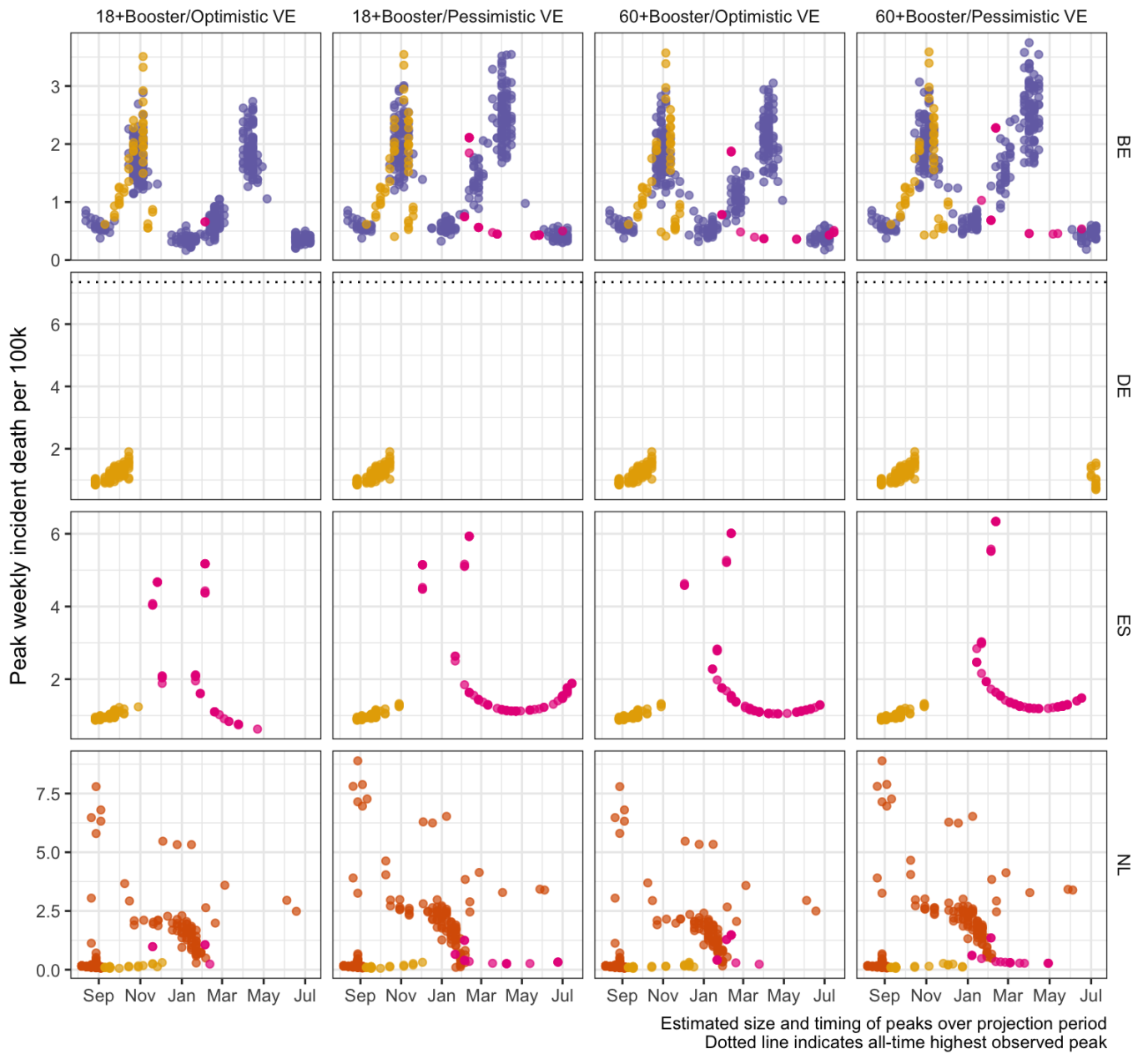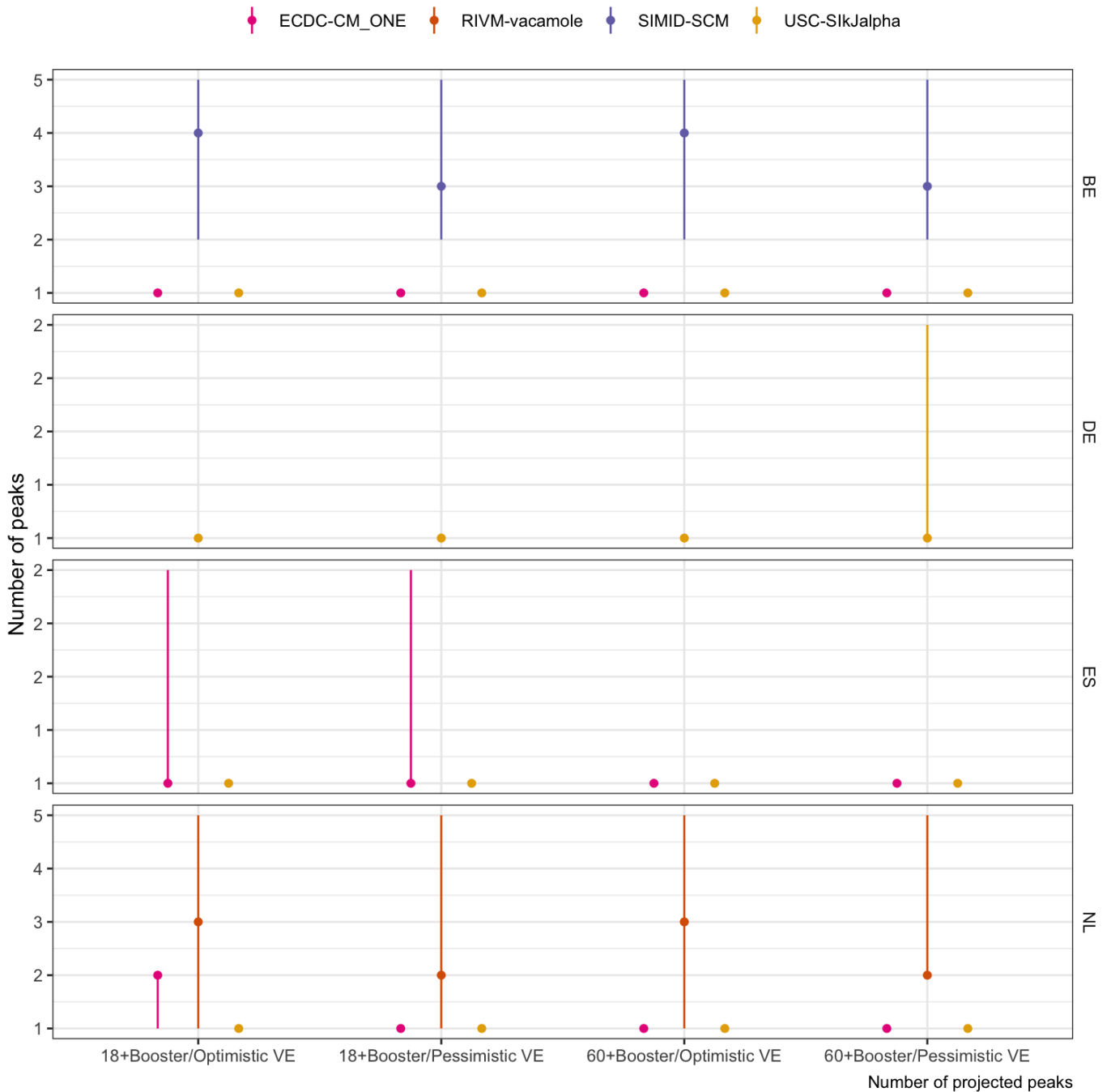
Hosp

A. Size and timing of peaks. Boxplots show summary of the likely value at peak incidence (median and interquartile range); points show timing and size of peaks from independent sample simulations

B. Projected number of peaks (median with 5-95% probability)

Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)

Icu

A. Size and timing of peaks. Boxplots show summary of the likely value at peak incidence (median and interquartile range); points show timing and size of peaks from independent sample simulations

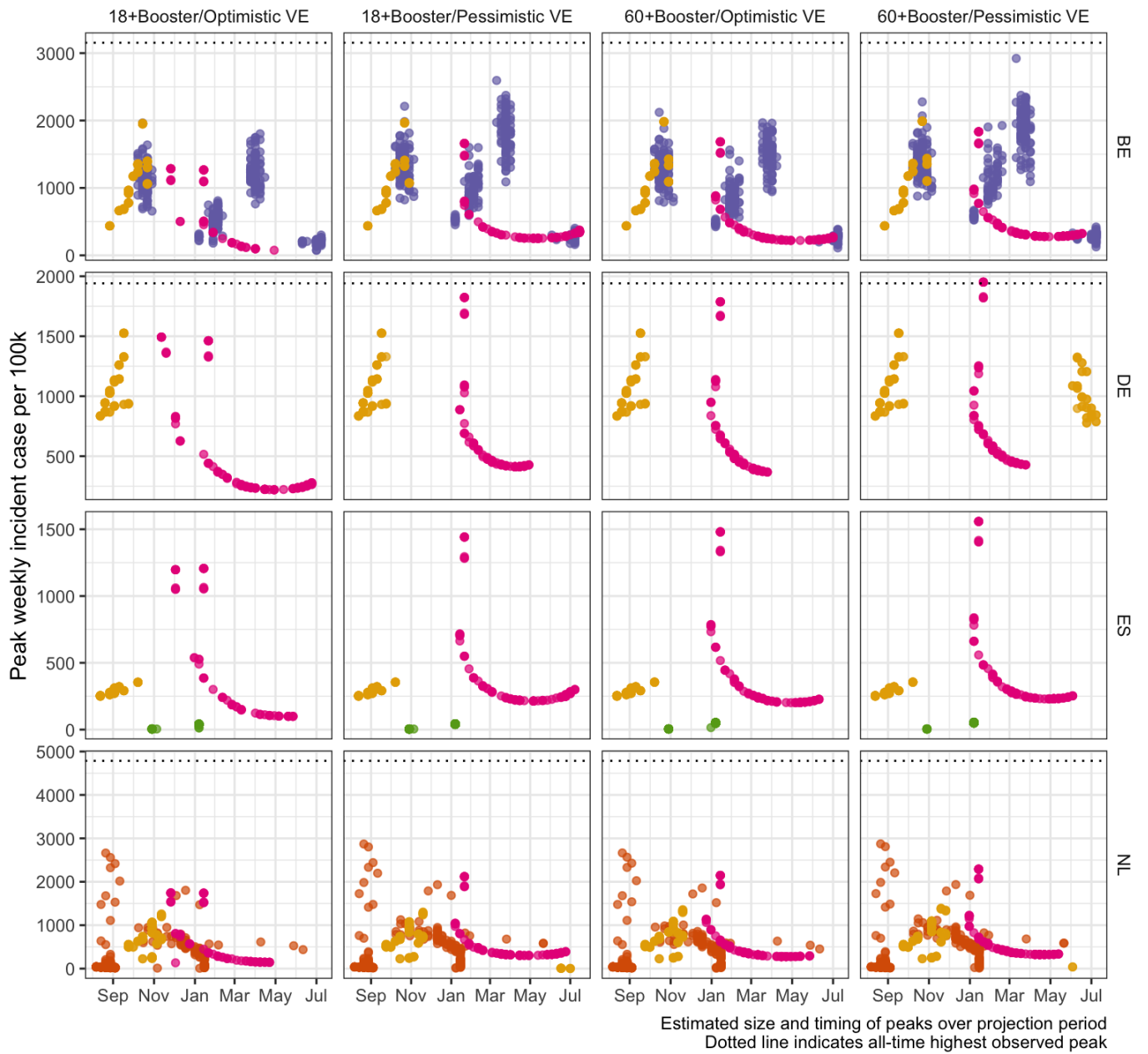B. Projected number of peaks (median with 5-95% probability)

Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)
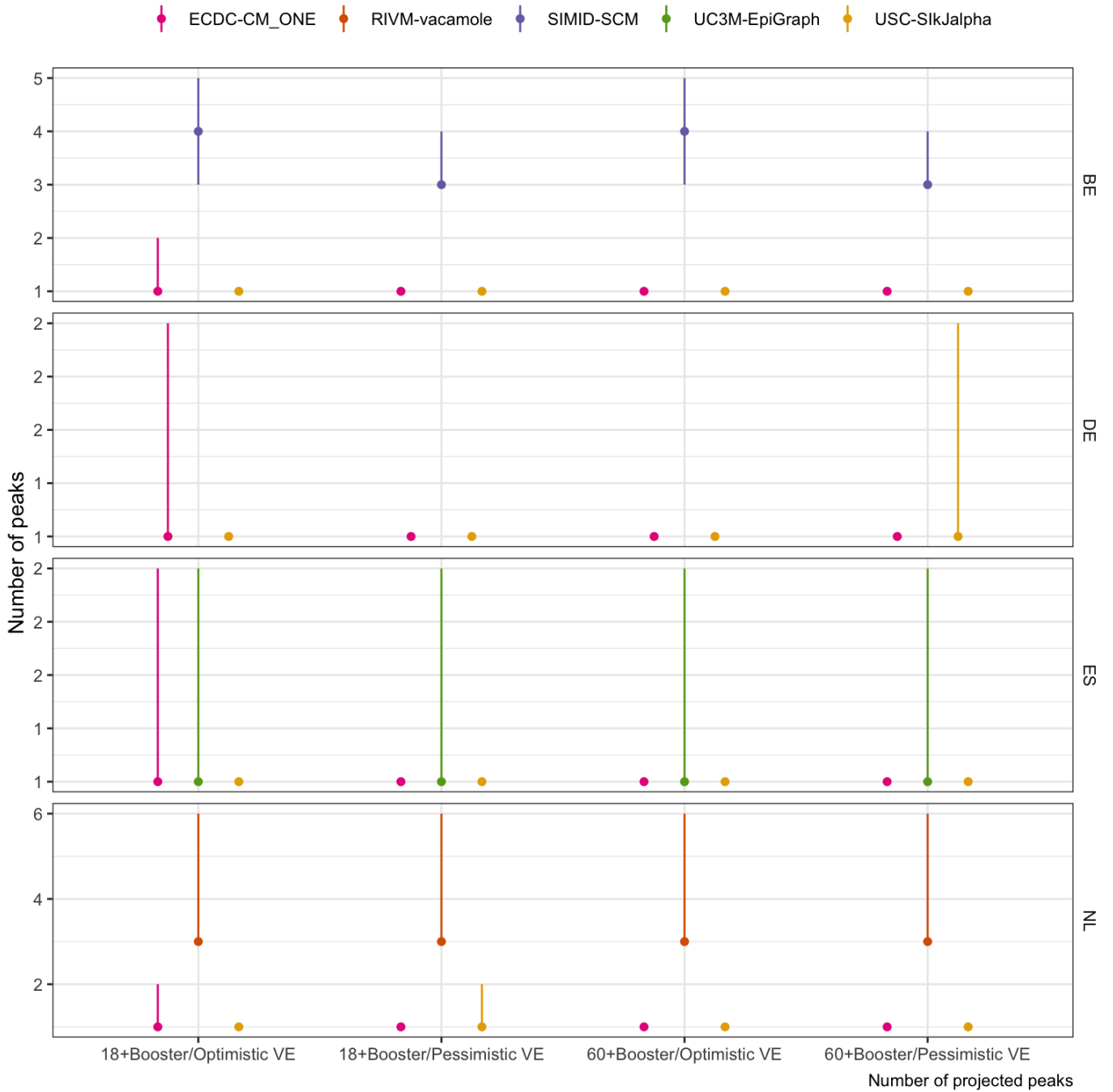
Entire projection period

Projections over June 2022 through June 2023

Death

A. Size and timing of peaks. Boxplots show summary of the likely value at peak incidence (median and interquartile range); points show timing and size of peaks from independent sample simulations

Estimated size and timing of peaks over projection period
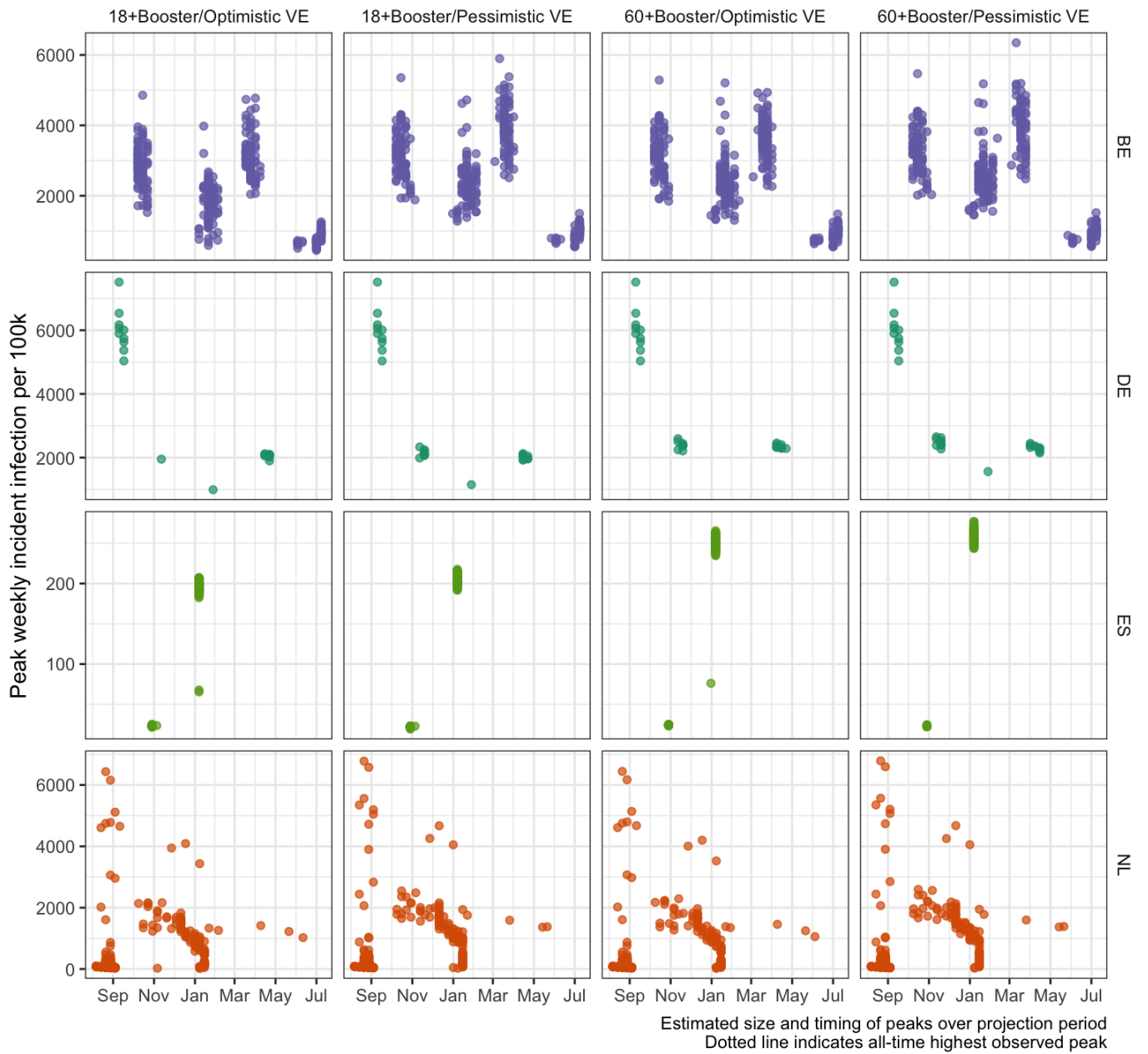Dotted line indicates all-time highest observed peak

B. Projected number of peaks (median with 5-95% probability)

Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)

Case

A. Size and timing of peaks. Boxplots show summary of the likely value at peak incidence (median and interquartile range); points show timing and size of peaks from independent sample simulations

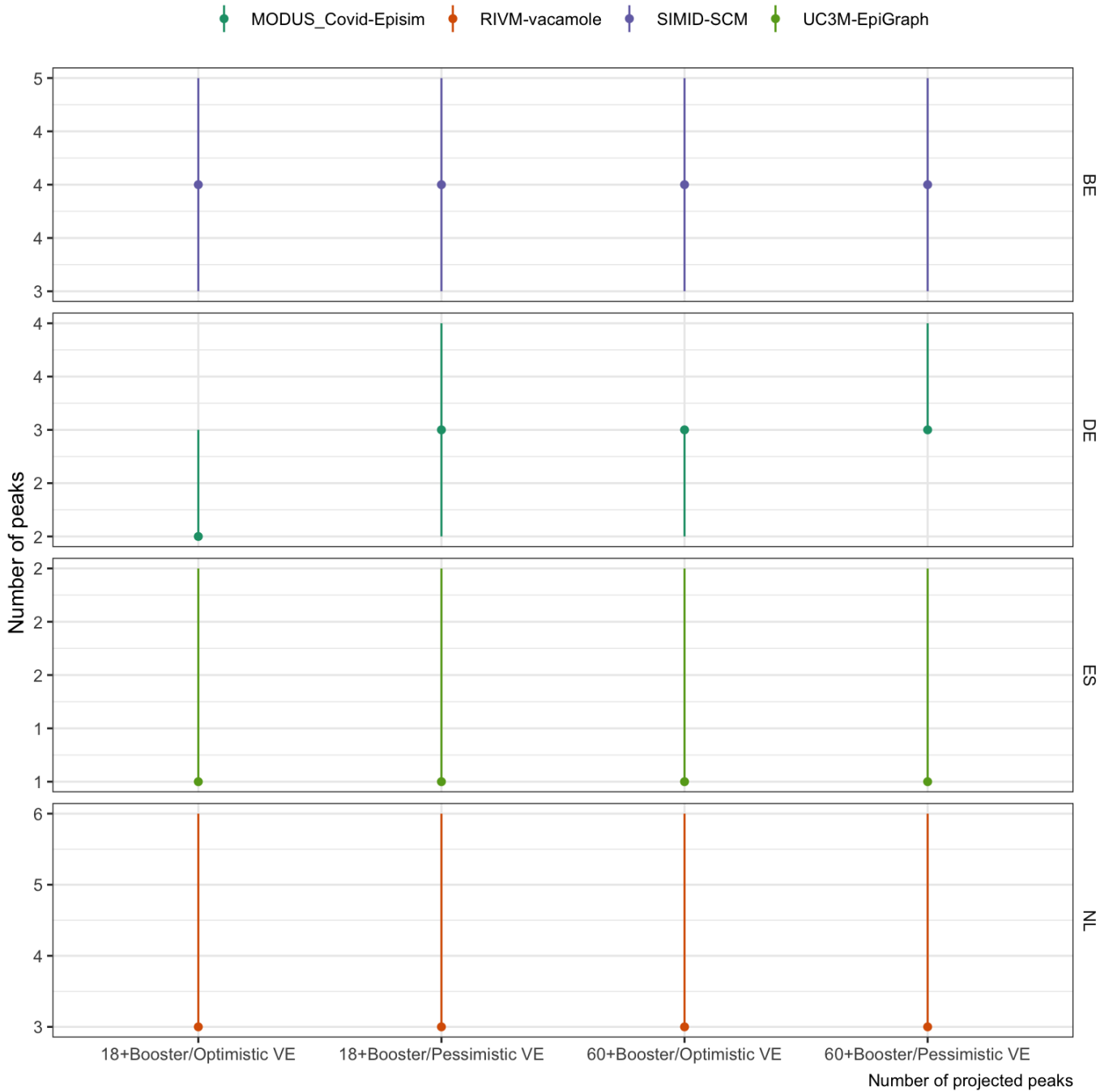B. Projected number of peaks (median with 5-95% probability)

Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)

Infection

A. Size and timing of peaks. Boxplots show summary of the likely value at peak incidence (median and interquartile range); points show timing and size of peaks from independent sample simulations

Estimated size and timing of peaks over projection period
Dotted line indicates all-time highest observed peak

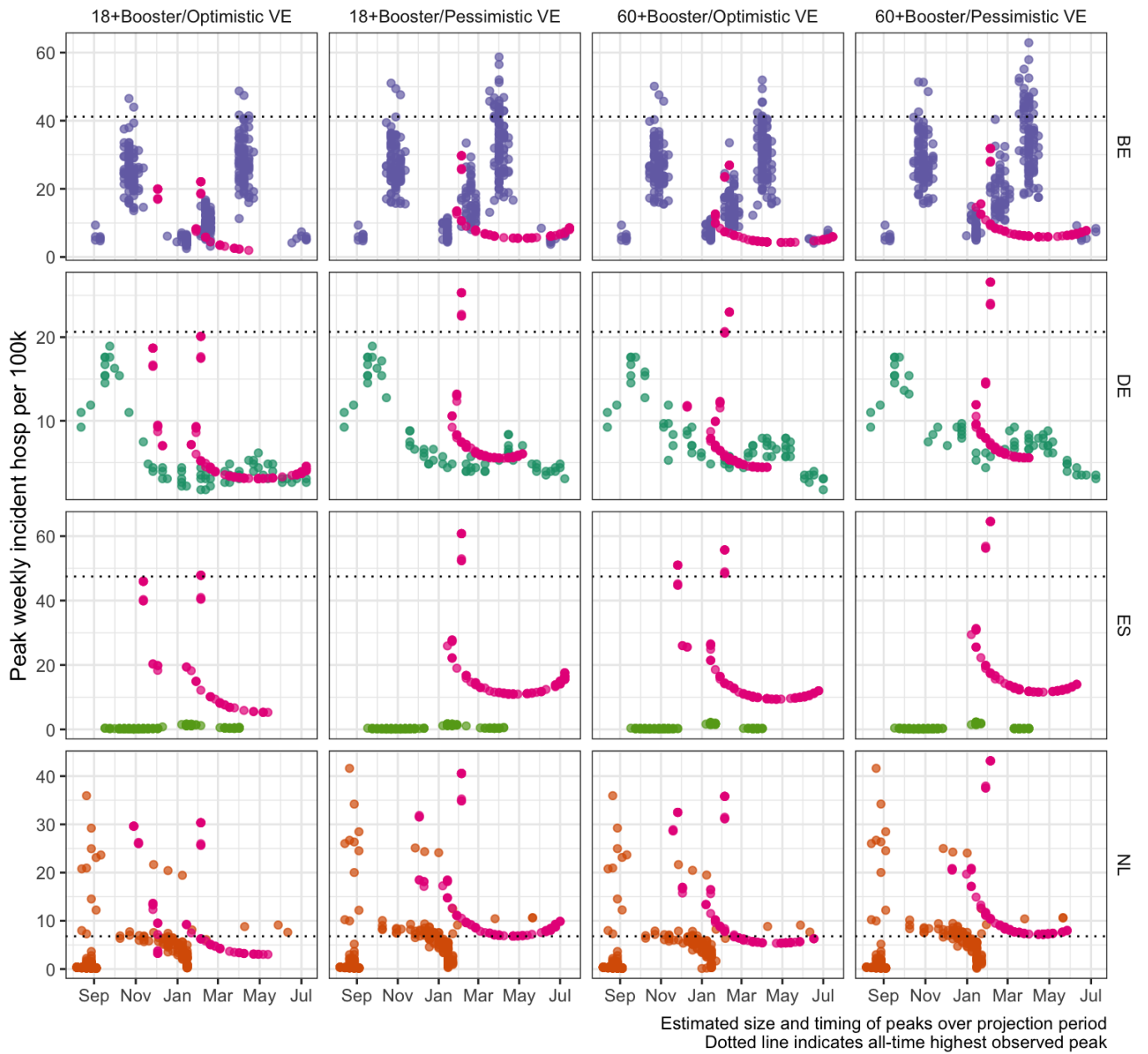B. Projected number of peaks (median with 5-95% probability)

Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)
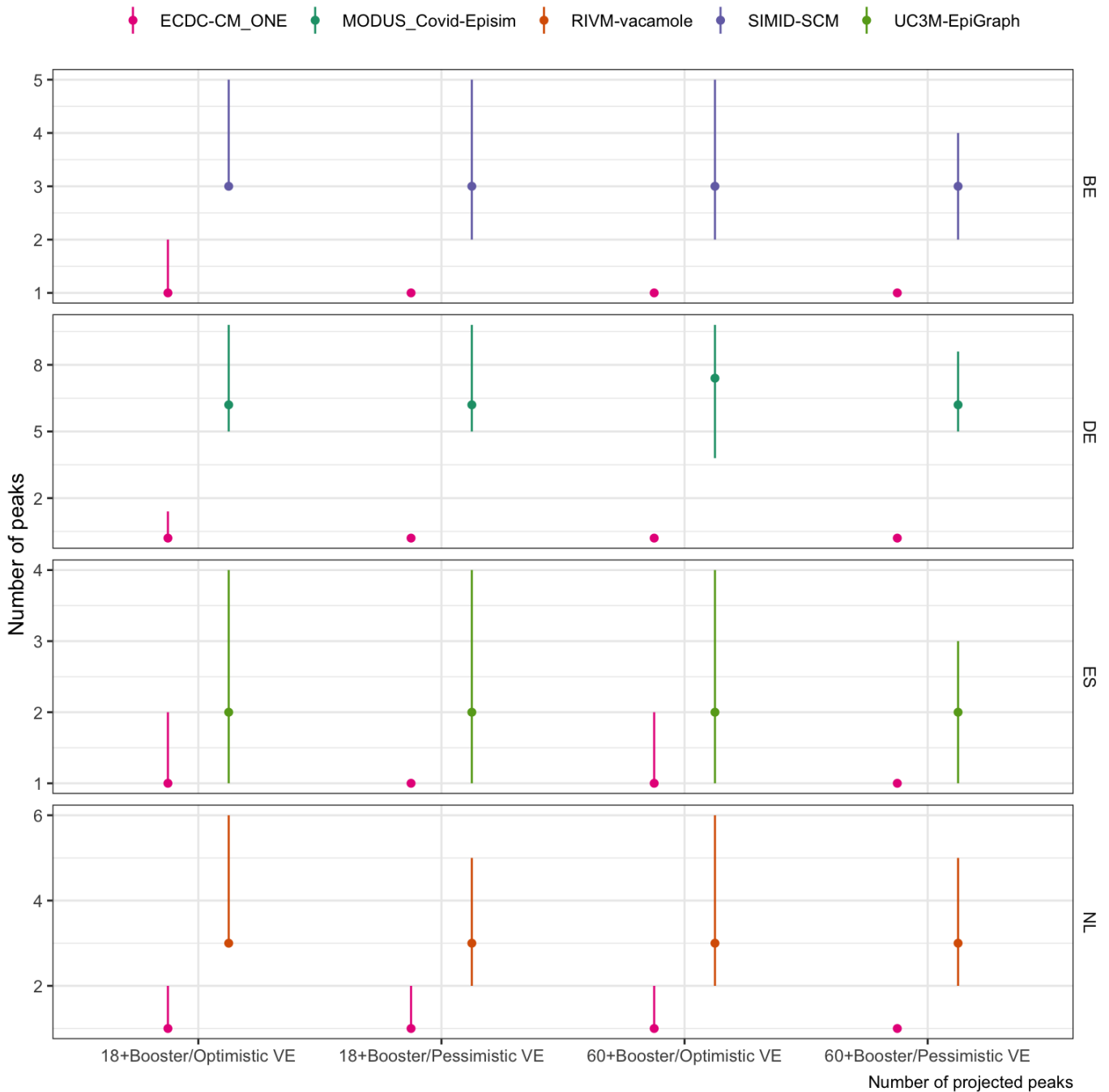
Hosp

A. Size and timing of peaks. Boxplots show summary of the likely value at peak incidence (median and interquartile range); points show timing and size of peaks from independent sample simulations

Estimated size and timing of peaks over projection period
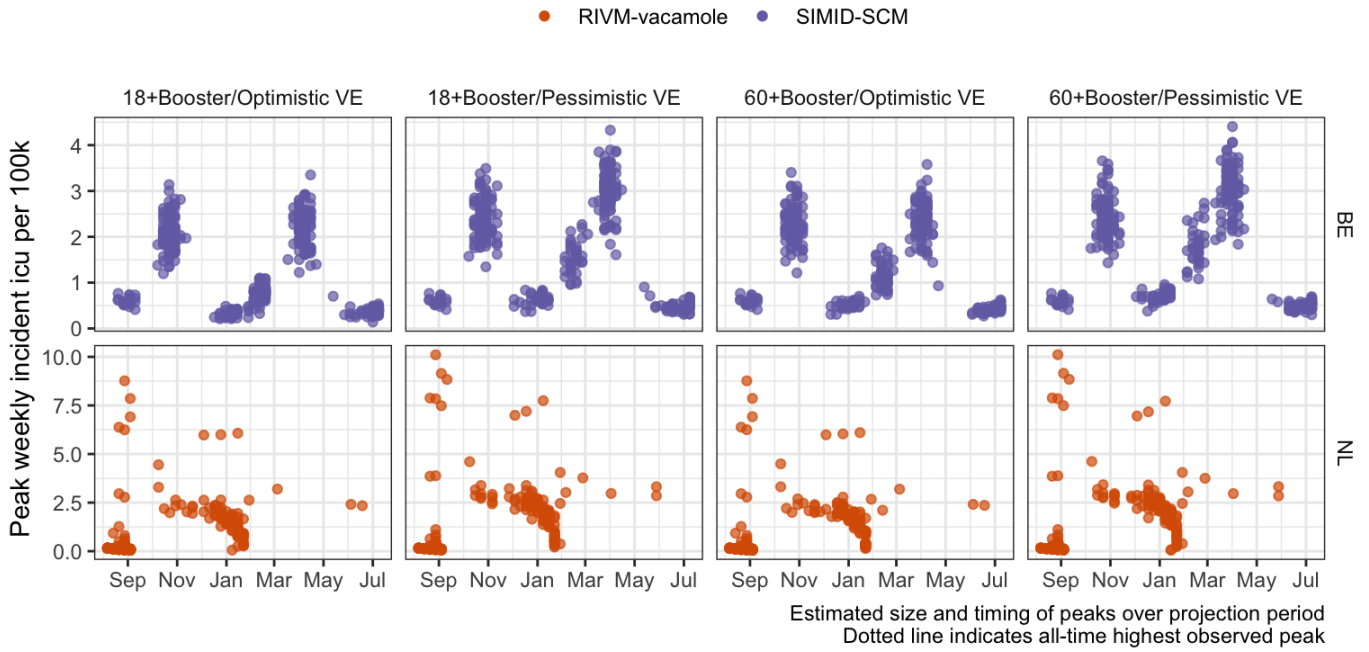Dotted line indicates all-time highest observed peak

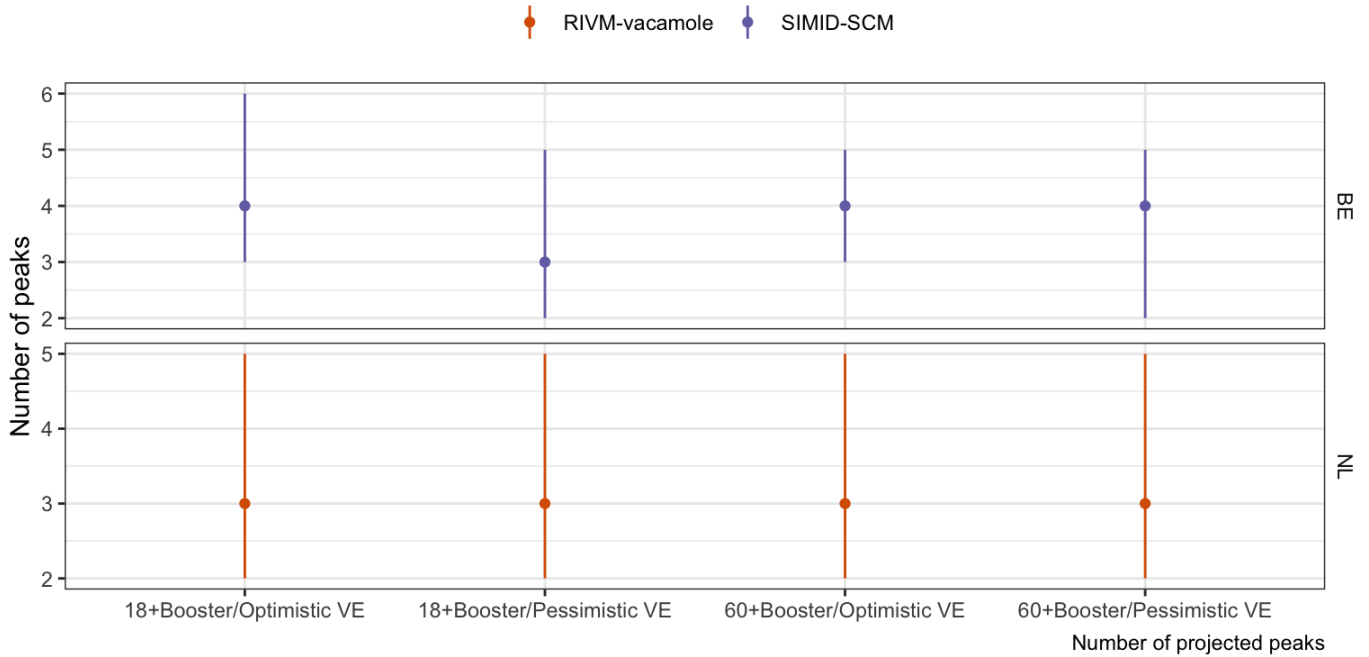B. Projected number of peaks (median with 5-95% probability)

154

Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)

Icu

A. Size and timing of peaks. Boxplots show summary of the likely value at peak incidence (median and interquartile range); points show timing and size of peaks from independent sample simulations

Estimated size and timing of peaks over projection period
Dotted line indicates all-time highest observed peak

B. Projected number of peaks (median with 5-95% probability)

Number of projected peaks

Scenarios: Autumn second booster campaign among population aged '18+' or '60+'; Vaccine effectiveness is 'optimistic'(effectiveness as of a booster vaccine against Delta) or 'pessimistic' (as against BA.4/BA.5/BA.2.75)

Contact us (contact.html)

Improving modelling for epidemic responses: reflections from members of the UK infectious disease modelling community on their experiences during the COVID-19 pandemic

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1701639	**Title**	Ms
First Name(s)	Katharine		
Surname/Family Name	Sherratt		
Thesis Title	Collaborative outbreak modelling for decision support: evaluating trade-offs from multi-model combination		
Primary Supervisor	Sebastian Funk		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Wellcome Open Research		
When was the work published?	January 2024		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	PhD by Publication		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	This work was developed in collaboration with Anna Carnegie (co-first author), Sam Abbott, and Yang Liu. The original concept of surveying modellers involved in the UK COVID-19 response was developed jointly and then led by Sam Abbott and Anna Carnegie. I provided feedback on the survey design, together with Yang Liu, and supported its promotion. We then jointly designed the workshop, with implementation led by Sam Abbott and Anna Carnegie. I reviewed this together with Yang Liu, participated in the workshop, and gathered materials after the workshop completed. I then led on synthesising the experiences and recommendations documented throughout. This was initially intended as a workshop report before developing further into the academic output presented here. In this phase of the work, I first led work gathering and digitising all workshop outputs, for example notes, mind maps, and dot plots from the day, converting these into usable data. I then led work ordering and synthesising these qualitative data into themes and recommendations, with review and feedback from Anna Carnegia, Sam Abbott and Yang Liu. Anna Carnegie then wrote an initial outline for a workshop report. I first reviewed this and then with joint agreement I led work to draft an academic paper reporting our findings. I led on the development of this write up, including the background, formalising the structure of the methods, writing the results and discussion. pI then managed several rounds of review from the core authorship group and those who participated in the workshop.

SECTION E

Student Signature	Katharine Sherratt
Date	14 June 2024

Supervisor Signature	Sebastian Funk
Date	14 June 2024

RESEARCH ARTICLE

Improving modelling for epidemic responses: reflections from members of the UK infectious disease modelling community on their experiences during the COVID-19 pandemic [version 1; peer review: 2 approved]

Katharine Sherratt [1]*, Anna C Carnegie [1]*, Adam Kucharski[1], Anne Cori[2], Carl A B Pearson [1,3], Christopher I Jarvis[1], Christopher Overton[4-6], Dale Weston[7], Edward M Hill [8,9], Edward Knock[2], Elizabeth Fearon[10], Emily Nightingale [1], Joel Hellewell[11], W John Edmunds[1], Julián Villabona Arenas[1], Kiesha Prem [1,12], Li Pi [13], Marc Baguelin[1,2], Michelle Kendall[8], Neil Ferguson[2], Nicholas Davies[1], Rosalind M Eggo[1], Sabine van Elsland [2], Timothy Russell [1,11], Sebastian Funk [1], Yang Liu [1], Sam Abbott [1]

[1]Centre for Mathematical Modelling of Infectious Disease, London School of Hygiene & Tropical Medicine, London, UK
[2]MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London, UK
[3]South African DSI-NRF Centre of Excellence in Epidemiological Modelling and Analysis (SACEMA), Stellenbosch University, Stellenbosch, Western Cape, South Africa
[4]All Hazards Intelligence, Data Analytics and Surveillance, UK Health Security Agency, London, UK
[5]Department of Mathematical Sciences, University of Liverpool, Liverpool, UK
[6]Department of Mathematics, The University of Manchester, Manchester, UK
[7]Emergency Response Department Science & Technology Behavioural Science, UK Health Security Agency, London, UK
[8]Warwick Mathematics Institute and The Zeeman Institute for Systems Biology & Infectious Disease Epidemiology Research, University of Warwick, Coventry, UK
[9]Joint UNIversities Pandemic and Epidemiological Research, JUNIPER, https://maths.org/juniper/, UK
[10]Institute for Global Health, University College London, London, UK
[11]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK
[12]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore
[13]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK

* Equal contributors

Open Peer Review

Approval Status ✔ ✔

	1	2
version 1 08 Jan 2024	✔ view	✔ view

Abstract

Background

The COVID-19 pandemic both relied and placed significant burdens on

the experts involved from research and public health sectors. The sustained high pressure of a pandemic on responders, such as healthcare workers, can lead to lasting psychological impacts including acute stress disorder, post-traumatic stress disorder, burnout, and moral injury, which can impact individual wellbeing and productivity.

Methods

As members of the infectious disease modelling community, we convened a reflective workshop to understand the professional and personal impacts of response work on our community and to propose recommendations for future epidemic responses. The attendees represented a range of career stages, institutions, and disciplines. This piece was collectively produced by those present at the session based on our collective experiences.

Results

Key issues we identified at the workshop were lack of institutional support, insecure contracts, unequal credit and recognition, and mental health impacts. Our recommendations include rewarding impactful work, fostering academia-public health collaboration, decreasing dependence on key individuals by developing teams, increasing transparency in decision-making, and implementing sustainable work practices.

Conclusions

Despite limitations in representation, this workshop provided valuable insights into the UK COVID-19 modelling experience and guidance for future public health crises. Recognising and addressing the issues highlighted is crucial, in our view, for ensuring the effectiveness of epidemic response work in the future.

Keywords
modelling, COVID-19, pandemic response
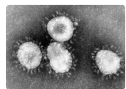
This article is included in the Coronavirus (COVID-19) collection.

1. **Srikanth Umakanthan** [iD], The University of the West Indies at Saint Augustine Faculty of Medical Sciences, Saint Augustine, Trinidad and Tobago

2. **Robert Moss** [iD], The University of Melbourne, Melbourne, Australia

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Katharine Sherratt (katharine.sherratt@lshtm.ac.uk)

Author roles: Sherratt K: Data Curation, Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Carnegie AC**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Kucharski A**: Investigation, Writing – Review & Editing; **Cori A**: Investigation, Writing – Review & Editing; **Pearson CAB**: Investigation, Writing – Review & Editing; **Jarvis CI**: Investigation, Writing – Review & Editing; **Overton C**: Investigation, Writing – Review & Editing; **Weston D**: Investigation, Writing – Review & Editing; **Hill EM**: Investigation, Writing – Review & Editing; **Knock E**: Investigation, Writing – Review & Editing; **Fearon E**: Investigation, Writing – Review & Editing; **Nightingale E**: Investigation, Writing – Review & Editing; **Hellewell J**: Investigation, Writing – Review & Editing; **Edmunds WJ**: Investigation, Writing – Review & Editing; **Villabona Arenas J**: Investigation, Writing – Review & Editing; **Prem K**: Investigation, Writing – Review & Editing; **Pi L**: Investigation, Writing – Review & Editing; **Baguelin M**: Investigation, Writing – Review & Editing; **Kendall M**: Investigation, Writing – Review & Editing; **Ferguson N**: Investigation, Writing – Review & Editing; **Davies N**: Investigation, Writing – Review & Editing; **Eggo RM**: Investigation, Writing – Review & Editing; **van Elsland S**: Investigation, Writing – Review & Editing; **Russell T**: Investigation, Writing – Review & Editing; **Funk S**: Investigation, Writing – Review & Editing; **Liu Y**: Investigation, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Abbott S**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

How to cite this article: Sherratt K, Carnegie AC, Kucharski A *et al.* **Improving modelling for epidemic responses: reflections from members of the UK infectious disease modelling community on their experiences during the COVID-19 pandemic [version 1; peer review: 2 approved]** Wellcome Open Research 2024, **9**:12 https://doi.org/10.12688/wellcomeopenres.19601.1

First published: 08 Jan 2024, **9**:12 https://doi.org/10.12688/wellcomeopenres.19601.1

Introduction

The response to the COVID-19 pandemic necessitated a multi-pronged approach, with infectious disease transmission modelling playing a key role in informing strategy and policy decisions[1,2]. Input from UK modellers was mostly channelled through weekly meetings of the Scientific Pandemic Influenza Group on Modelling, Operational subgroup (SPI-M-O) feeding into the Scientific Advisory Group for Emergencies (SAGE)[3]. This advisory group, drawing on expertise from the academic, and public health sectors, developed planning scenarios and short-to-medium term forecasts and projections, routinely estimated key parameters such as the reproduction number (a proxy for transmissibility), conducted routine data analysis, as well as authoring ad-hoc reports on modelling results relevant to the ongoing pandemic in the UK[4,5]. Some of these analyses resulted in academic papers along with those produced by the wider UK modelling community (e.g. 6,7).

The high-pressure environment and daunting responsibilities of those at the frontlines of pandemic response have been shown to exert significant psychological tolls. Notably, healthcare workers (HCWs) involved in infectious disease outbreaks, including COVID-19, have been shown to experience profound and enduring psychological impacts. These include acute stress disorder, post-traumatic stress disorder (PTSD), burnout, as well as moral injury[8–10]. Moral injury refers to a specific form of distress that stems from guilt, anxiety, and loss of trust when actions or roles conflict with one's deeply held moral beliefs. These psychological impacts not only diminish individual wellbeing but can also considerably affect worker productivity, with lasting effects that can linger for years, as exemplified by the 2002/2003 SARS epidemic[11,12].

However, the experiences and challenges faced by non-healthcare responders to the pandemic, such as those involved in modelling and research, have received comparatively less attention[8,10]. Stressors such as high workloads, long hours, tight deadlines, and harassment from the public and press during the COVID-19 response had the potential to cause both visible and invisible impacts. These include mental health impacts, exhaustion, social isolation, compromised career progression in academia, and moral injury.

The experiences of modelling responders have not been systematically discussed but are indirectly reflected in issues of staff retention and burnout across institutions. With the aim of bridging this gap, on March 28th, 2023, we organised a one-day workshop to create a space for collective reflection and strategising improvements for future epidemic responses. This paper seeks to provide an outline of the workshop proceedings, the collective themes that emerged from our discussions, and synthesise our suggested actions into a set of priority recommendations to enhance future epidemic responses.

Methods
Our approach
We employed an iterative, participatory approach to both design and run a reflective workshop with members of the UK modelling community in order to facilitate the summarisation of our collective experiences. In the interest of clearly relaying the proceedings and results emanating from the workshop, we use the term 'participants' to refer to attendees (i.e. ourselves, including the organisers) in the remainder of the methods and the results.

Workshop design
We aimed to ensure the content of the reflective session captured the needs of the individuals at the forefront of the UK modelling response. To inform the content of the workshop, the session organisers (SA and ACC), alongside two additional members of the UK modelling community, solicited informal feedback from individuals involved in the COVID-19 response. This feedback included the personal and professional ramifications of participating in COVID-19 response work, along with the obstacles to effective response work and strategies to address them.

We then engaged an external facilitator to assist in planning the agenda and guiding participants throughout the session. This aimed to ensure unbiased management of discussions and to enable participants to express themselves openly in a safe and supportive environment. To select an appropriate facilitator, we sought input from the broader scientific community and chose an individual with a track record of successfully delivering similar events.

Initial discussion topics were developed by the session organisers in consultation with the external facilitator, drawing on anecdotal evidence from conversations with other modellers who were involved in the COVID-19 response. Further feedback was solicited from two members (KS and YL) of the UK modelling community who were not directly involved in the organisation process. This resulted in a set of discussion topics that addressed the concerns and interests of the community.

Participants
We aimed to include a diverse range of participants involved in the UK COVID-19 modelling response, encompassing researchers and professional services staff. A brief expression of interest form was disseminated by the session organisers to the UK modelling community via organisation mailing lists, personal networks (aiming to also reach those who may have transitioned away from the infectious disease modelling field), and social media channels to ensure representation across different levels of seniority. We invited all those who expressed an interest to attend. We provided a small travel fund for participants on a first come first served basis for those travelling from across the UK.

Workshop structure
Participant arrival and introduction
Upon arrival, the facilitator encouraged participants to engage with flipchart papers displaying "snapshot" questions with attendees providing their responses using stickers. These were:

1. Do you think sufficient action is currently being taken to improve future outbreak responses to the standard you think is acceptable?

164

2. Who is responsible for ensuring people are supported, and appropriately credited for their work?

3. Summarise your pandemic experience in one word.

See the supplementary information for the multi-choice answers[13].

At the formal start time, the facilitator opened with an overview of the day's agenda, establishing expectations and a code of conduct for participants. The Chatham House Rule ("share the information you receive, but do not reveal the identity of who said it"[14]) were introduced to ensure that individuals would not be identified, while allowing for the synthesis of outputs. A co-organiser (SA) shared their personal pandemic timeline (see the supplementary information[13]), setting the stage for the first exercise.

Iterative discussions of experience

Participants divided themselves into pairs, after being encouraged to work with someone they would not usually interact with. They were asked to discuss their individual pandemic timelines for 15 minutes each, while the partner asked questions based on those we developed when designing the workshop, listened, and asked follow up questions. The following questions were provided:

1. What was your pandemic timeline? What were the highs and lows?

2. What was your experience of pandemic work like?

3. What were some of the things that helped assist you to do effective research during the outbreak response?

4. Do you think team science was appropriately supported over the pandemic?

5. Has your employer or the wider community taken action to help mitigate any of the personal or professional costs/challenges you identified? What more can be done?

6. Do you think there were barriers to doing effective and sustainable COVID-19 outbreak response work? If so, what were they?

7. What has been done and what more can be done to reduce any barriers to effective outbreak response work in the future?

We also provided suggested follow-on questions which are available in the supplementary information[13].

Pairs then formed groups of four to identify common themes from their one-to-one reflections using post-it notes. The group was then brought together and themes were summarised and organised into headline categories. This approach maintained anonymity for the participants, while capturing their reflections in a summarised form. As a combined group we then further discussed these topics, leading to the identification of six major themes.

Synthesising recommendations

The latter portion of the workshop focussed on pinpointing recommendations for action. Participants were presented with primary categories derived from the morning's discussions. Participants were then divided into new groups, with each group assigned a theme. Each group was tasked with developing recommendations and potential implementers. Participants could move between themes and contribute their thoughts. These recommendations, along with actionable steps and suggestions for those responsible for implementing, were then exhibited on a wall for group review. Finally, attendees used dot stickers to identify priorities, allowing a visual representation of the group's consensus.

We (ACC, KS, YL, SA) then reviewed the contents of the group discussions based on the post-it notes, whiteboards, and recommendation board created during the session. Two authors (ACC and KS) independently digitised the output, and four authors (ACC, KS, YL, SA) independently reviewed results. We then came to a consensus on the common themes of participants' experiences, using the major themes identified by participants as a guide, and priority recommendations for stakeholders. We prepared an initial draft and shared this with participants. Finally, we integrated feedback, ensuring that the insights derived from the workshop were preserved.

Results
Outputs from the workshop
Summary of attendees

The event was attended by 27 individuals, including 25 research staff and two professional services staff. Staff attended from five higher education institutions (London School of Hygiene & Tropical Medicine, Imperial College London, University of Warwick, Liverpool University, the University of Oxford), and the UK Health Security Agency (UKHSA). The majority of attendees were based in London. Participants represented various career stages, including early, mid-career, and senior academics and professionals. Among the attendees were multiple members of SPI-M-O and SAGE.

Snapshot reflections

In response to the initial snapshot question, "Do you think sufficient action is currently being taken to improve response work to a standard you think is acceptable?", the participants expressed an overwhelmingly negative view (17/18). In the second snapshot question, "Who should ensure that individuals are adequately supported and credited for their response work" (this question allowed multiple answers), participants suggested this responsibility was shared among stakeholders. Affirmative responses were more common on the panel listing smaller groups than on the panel listing larger organisations, indicating that respondents considered themselves (9/56), line managers (16/56), and research groups (12/56) more responsible for this task compared to larger organisations such as institutions (6/56), academic funders (5/56), and the "system more generally" (8/56). The second panel displaying the larger organisations was situated to the right of the first panel, which may have resulted in decreased visibility. Photos of the panels are available in the supplementary information[13].

Each attendee was asked to summarise their pandemic experience in one word using post-it notes (see supplementary information). Positive responses were: *"exciting"*, *"valuable"*, and *"engrossing"*. Neutral responses were: *"intense"*, *"unprecedented"*, *"ambiguous"*, *"focussed"*, *"hectic"*, *"repetitive"*, and *"surreal"*. Negative responses were: *"hard"*, *"austere"*, *"stressful"*, *"lonely"*, *"harrowing"*, *"frustration"*, and *"exhausting"*. Multiple participants added stickers (indicating support) to both *"stressful"* and *"exhausting"*.

Themes from paired and group discussions

In the paired and small group discussions, several topics emerged into which participants' perspectives were grouped. These included (reproduced verbatim from those listed on the day): *"societal impact"*, *"mental health"*, *"life outside"*, *"emotions"*, *"personal"*, *"team spirit"*, *"institutional structures"*, *"work process"*, *"work feeling/support"*, *"work pressure"*, *"(negativity about)"* *"positives"*, *"career direction"*, *"rewards"*, *"access and privilege"*, *"bad stuff"*, *"COVID-19 modeller-specific experiences"*, and *"general experiences"*. In the following session, participants then refined these themes to leave the following: *"institutional factors"*, *"mental health"*, *"life/personal"*, *"work process"*, *"career direction"*, and *"social impact"*. See the Supplementary Information for the full list of participants' points[13].

After the workshop, we reviewed individual post-it notes and further refined these themes to leave: *funding and institutional support*; *recognition, rewards, and access*; *team and work dynamics*; *non-academic contributions*; and *personal impacts*. The themes emerging from the group discussions are synthesised, stratified by these themes below. We indicate direct quotes from individual authors using quotation marks and italics.

Funding and institutional support

Lack of institutional support: Insufficient institutional support for those involved in the COVID-19 modelling response was a common issue among participants. Many felt that they were not protected by their institutions during the response or in its aftermath; for example, when receiving aggression from some sectors of the media and general public. Additionally, groups highlighted the lack of processes to respond in an emergency while protecting psychological safety. This included the need for training for managers and teams, and wellbeing procedures and human resources policies.

Contract insecurity and inflexible funding rules: The precarity of short-term contracts due to heavy reliance on external grant funding was highlighted, along with implicit pressures to underestimate personnel time in funding applications to meet budget thresholds, adhere to eligibility criteria and achieve cost recovery targets. The importance of providing sufficient and sustainable personnel funding was stressed, with this including academic and professional services roles such as project managers, administrators, communications professionals, technicians, and software engineers.

Recognition, rewards and access

Inadequacy of reward metrics: Credit attribution mechanisms were a recurring concern. Participants emphasised that there are currently insufficient frameworks to reward the nature of response work itself. Hurdles in receiving recognition for work included contributing to confidential reports where involvement was unable to receive external acknowledgement. In particular, it was noted that outputs such as software tools and policy reports do not fit within the traditional academic credit structure. Similarly, participants recognised that promotions, paper authorship, and grant Principal Investigator (PI) positions were not designed to promote collaborative team working. This was identified as a problem for both the general wellbeing of researchers and the quality of the science produced. The unequal, and individual-focussed, credit structures that persisted throughout the pandemic were also discussed, with senior or well-connected researchers being identified as receiving the majority of recognition. Participants noted *"rewards not attributed equally"*, and that *"institutions got awards, not individuals (not all key players)"*. This uneven reward system was seen as contributing to a competitive culture, which was identified as a problematic aspect of response work and academia more widely.

Access to decision-making spaces: Individuals had different access to policy-making spaces which did not always reflect where or how their work was used. As a result, some individuals who lacked access reported feeling left behind when it came to updates relevant to their work. There was a general consensus that there should be more transparency regarding these forums for those involved in producing the work presented.

Team and work dynamics

Insufficient capacity: Participants highlighted issues with *"not being able to say no"* and the *"pressure [that] came in waves. 'Not again...'"* These issues contributed to poor working practices within teams, including insufficient capacity and reliance on one or two individuals to perform key tasks. In turn, this made it more challenging for these individuals to maintain a work-life balance. The highly pressurised and reactive nature of response work meant that there was not always space for teams to reflect on the effectiveness of routine aspects of the response, including whether academic groups were the best placed to perform this work. In addition, despite a need for additional capacity, working in highly reactive 'response mode' made it difficult to properly onboard new starters and hand over responsibility of tasks and projects where resources were available to do this. There was reference to other professions more adapted to response work, such as the military and emergency services, suggesting there may be learning to be gained from these sectors.

Competing demands and barriers to progression: Individuals faced challenges in balancing competing demands of and distinguishing between 'response' work and research. Some individuals sacrificed otherwise beneficial opportunities, such

as teaching. Although response work created some opportunities for career progression, these were distributed unequally relative to contribution. Access to these opportunities depended on several factors, including career stage, and relative privilege (which is the differential access to resources, opportunities, and advantages some groups have compared to others). We note that privilege is often invisible to those who have it, and recognizing one's own relative privilege is a key part of understanding and addressing social inequalities.

Collaborative working: Participants cited the positive experience of collaborative working and camaraderie within teams – academic, professional, and hybrid. However, as the pandemic progressed, there was a sense that the egalitarian working structures, which some felt were put in place at the start of the pandemic, faded: *"Shift from egalitarian structure to pre-existing hierarchies"*. Meanwhile, with close working relationships and the intensely personal impact of the COVID-19 response, professional disagreements sometimes took on an unusually emotional tone.

Non-academic contributions
Role of professional services staff: Participants highlighted the significance of integrating professional services roles into research teams, mentioning that these staff played crucial roles in response-related tasks. Participants pointed out that professional services roles, especially administrative positions like project managers, are frequently deemed ineligible costs in grant applications. Similar to academic staff, many individuals in these roles work on short-term contracts. Consequently, these positions were often under-resourced and experienced high turnover.

Public health agency workers: Participants emphasised the importance of strengthening the collaboration between academics and public health agencies, with the aim of fostering knowledge and skills exchange both during, prior to, and after responses. The importance of a bidirectional exchange was highlighted, with academics having the opportunity to learn about the practical challenges faced by public health agencies, while public health staff would benefit from access to the latest research findings. Participants called for more opportunities to facilitate these exchanges, such as joint workshops, shared working spaces, and dedicated training sessions.

Personal impacts
Public recognition: The COVID-19 pandemic brought the infectious disease modelling field public recognition and scrutiny. Participants acknowledged the personal responsibility that came with this visibility, while valuing the significance of their work. While friends and family gained deeper understanding of their work, some highlighted the challenge of work and life becoming intertwined. Participants referred to the *"surreal level of public and media interest (good or bad)"* and the idea that *"work and the world were one and the same. Neither was an escape from the other."*

Mental ill health and burnout: Participants across organisations and seniority levels reported prioritising work over their health and wellbeing, leading to extreme levels of overwork,

burnout, and associated mental health effects, including depression and anxiety. The experience was common among attendees at all levels and career types, with recognition that this can creep up over time and not enough has been done to mitigate against it. Some participants expressed guilt and a sense of 'survivor bias' from being able to remain within academia, having witnessed friends and colleagues leave the field. One post-it note summed up the feeling of *"trading off career versus health and everything"*. People were reluctant to reach out to managers or colleagues for support. With close working relationships, the personal challenges faced by colleagues inevitably impacted the wider team. No strategies were identified by participants as having been in place to address these issues during the pandemic response or having been implemented more recently.

Commitments outside of work: Several participants highlighted the challenges they faced in balancing high-intensity roles with personal obligations during the pandemic response. They shared experiences of coping with loss and caregiving responsibilities, which were particularly difficult for those whose partners were also involved in the response. Certain groups faced heightened challenges; for example, women often bore a disproportionate burden of caregiving tasks, early career researchers tended to have less stable domestic situations, and non-UK nationals experienced difficulties such as visa concerns or being separated from their home countries.

Recommendations
The strategies collectively proposed at the workshop spanned societal impact, mental health, career direction, work processes, personal life, and institutional policy. Over ninety suggestions were made for possible actions by research teams, employers, and funding entities. The full list of recommendations is available in the supplementary information[13].

Priority recommendations
Participants distilled a set of priority recommendations to enhance the support and sustainability of epidemic response work. These directives tackle crucial facets affecting the well-being and efficiency of those engaged in pandemic response. Example actions for implementing these recommendations are italicised below each recommendation (see the full list of suggested actions from the workshop in the supplementary information[13]).

1. Acknowledge, and reward, impactful response work at institution, funder, and research community levels.

    Funding bodies refine impact measures to credit all forms of output produced during, and required for, response work; institutions standardise incorporating response-driven work into criteria for doctoral theses and promotion.

2. Encourage routine interaction between academia and public health agencies, including consistently reviewing the role of each during epidemic responses.

    Government bodies and research institutes create sustainable dual positions recruiting from both sectors.

167

3. Ensure response teams are well-staffed, well-resourced, stable, and provided psychological support.

Research teams establish sustainable team-building and training programmes with long term support from funders during non-response periods to ensure individuals feel equipped and supported to engage in response work.

4. Increase the transparency of the evidence pathway from scientists to decision-makers making it easier for those across the scientific community to contribute as well as making the evidence base for decisions clearer to the general public.

Government bodies standardise rapid open access to the minutes of scientific advisory meetings and encourage input from a wider range of sources.

5. Implement best practices for a sustainable work environment.

Employers promote leave-taking and respecting work hours, and clarify communication about processes and rewards across career stages, integrate support roles into research teams, and standardise the onboarding of new team members.

Discussion

This reflective workshop brought together 27 individuals from the UK infectious disease modelling community to engage in a dialogue around the personal and professional impacts of their COVID-19 response work. Participants represented various career stages, institutions, and disciplines, enabling a diverse exploration of experiences and perspectives. We identified areas of improvement in the current approach to modelling during epidemic responses, with these including greater support for responders, line managers, and research groups. Our experiences ranged from positive to negative, with stress and exhaustion being particularly prominent. Through in-depth discussions, key themes emerged, including institutional support, mental health, career direction, and social impact. Challenges such as lack of institutional backing, insecure contracts, inadequate reward systems, and personal impacts such as mental health issues were identified. The roles of professional services staff and public health agency workers were underscored. To address these issues, we identified a variety of strategies and priority recommendations, including acknowledgement and reward of impactful response work including for professional service staff, enhanced academia-public health collaboration, minimising dependence on key individuals, increased transparency in decision-making processes, and the adoption of sustainable work practices. These findings offer valuable insights for the ongoing pandemic response and future public health emergencies.

Our approach benefitted from being embedded in the experience of the UK modelling community. The session was community-driven, adopted an informal approach, and included participants from various career stages and perspectives on the response. Prior input from the community ensured the event's relevance for attendees, while employing an external facilitator helped create a safe and structured environment for discussion. We then collectively agreed on key themes and recommendations.

However, a key limitation was participant representation. This was exacerbated by it being a one-day workshop, meaning we could only represent the views of those who were available and able to attend in person on that day. Attendees were primarily from London and Southeast England, possibly due to limited support for travel costs. Additionally, despite efforts to involve individuals who had left the infectious disease modelling field, few were able to attend. Our collective experiences are therefore likely to be missing some of the most challenging experiences and perspectives of responders, and our conclusions may be more moderate than if a wider range of participants had attended. Despite this bias, we feel this provides valuable insights into the UK COVID-19 modelling experience but should be viewed as a summary of a small group's experiences and opinions, with potential differences across jurisdictions and groups. We encourage responders in other locations to conduct similar exercises and to synthesise these findings for a broader understanding.

Ongoing efforts have begun to evaluate UK modelling work during COVID-19 both in terms of modelling results (e.g. forecasts or scenario projections[6,7]), and the systems and processes enabling the response[2,15,16]. However, so far little has been done to report the experiences of responders themselves as we have done in this work. In the context of more general crisis response, more work has been done to understand the key challenges, particularly on healthcare workers (HCW). For example, hospital disaster preparedness plans may incorporate mental and behavioural health interventions (such as resource signalling, peer support, and referrals for at-risk individuals), which have proved to be effective in reducing mental health morbidities[17,18]. Lessons from previous epidemics also emphasise the importance of effective staff support and training in preparing for future outbreaks. Perceived adequacy of training and support had a protective effect on adverse outcomes in HCW responders to the SARS epidemic[19]. These approaches, which have established use in high-stress occupations, could be adapted and applied to support modellers during epidemic response situations.

The workshop identified priority recommendations aimed at enhancing support and sustainability in pandemic response work. Our discussions underscored the importance of recognising and rewarding significant contributions to public health crises at all levels. We advocated for fostering closer ties between academia and public health agencies, building well-resourced, resilient teams, and ensuring their psychological well-being. Discussions also emphasised the need for increased transparency in the evidence-to-policy pipeline, improved work-life balance, and clear institutional communication. Further suggestions included standardising onboarding procedures and integrating support roles into teams.

Whilst we identified several themes and recommendations during our workshop, we did not explicitly separate issues specific to the pandemic response from broader academic

168

challenges. Some recommendations, for example, recognising non-traditional contributions or normalising annual leave, pertain to broader issues. It is important to discern whether these concerns are long-standing systemic issues that have been simply exacerbated by the pandemic, or if they have been particularly highlighted due to the unique stressors of the pandemic.

Conclusions

As a community we want to acknowledge that the pandemic has engendered widespread hardship, stress, and ill health throughout various populations. It is crucial to reflect on and address these profound impacts as we continue to tackle the crisis and prepare for future epidemic responses.

The consequences of the COVID-19 pandemic have been profound on those at the forefront of the UK modelling response. In this work, we have summarised our experiences and whilst we recognise that many of the issues we have identified impact those in our field more generally we believe that they are particularly problematic for epidemic response work. It is evident that changes are required across multiple domains, including individual work, team dynamics, and institutional structures, to enable future effective epidemic modelling responses.

Achieving these changes necessitates investment from governments, funding bodies and institutions. The solutions needed to foster a healthy and sustainable environment for future epidemic response work will not be attainable without such investment. Additionally, there is a need for teams aiming to respond to epidemics to redefine their working methods, developing response preparedness plans that emphasise wellbeing, training, and career development. It is clear that even these localised initiatives demand time investment from those leading them, and as a result, require support.

As it stands, future epidemic responses are likely to raise similar challenges to those we have identified here, including reliance on a select number of individuals, excessive workloads and the exacerbation of systemic inequalities. It is critical we act outside of response contexts; for example, by implementing the recommendations we have outlined, to mitigate these issues and respond more effectively in future.

Consent

This work is the sole product of collaboration among the named authors. All inputs used in this work were those of the authors, with no data collection from any additional participant or data source. Therefore, all participants in this work are named authors of this manuscript and have approved both the manuscript and supplement for publication.

Data availability

Underlying data

Open Science Framework: Underlying data for 'Improving modelling for epidemic responses: reflections from members of the UK infectious disease modelling community on their experiences during the COVID-19 pandemic', https://www.doi.org/10.17605/OSF.IO/4JNCB[13].

This project contains the following underlying data:

- Data supplement.pdf

  - Survey questions

  - Snapshot questions

  - SA pandemic timeline

  - Session questions

  - Group discussion themes

  - Themes recommendations

  - Priority recommendations

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

References

1. Whitty CJM, Collet-Fenson LB: **Formal and informal science advice in emergencies: COVID-19 in the UK.** *Interface Focus.* 2021; **11**(6): 20210059.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

2. Pagel C, Yates CA: **Role of mathematical modelling in future pandemic response policy.** *BMJ.* 2022; **378**: e070615.
   PubMed Abstract | **Publisher Full Text**

3. Medley GF: **A consensus of evidence: The role of SPI-M-O in the UK COVID-19**

   response. *Adv Biol Regul.* 2022; **86**: 100918.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

4. McCabe R, Donnelly CA: **Disease transmission and control modelling at the science–policy interface.** *Interface Focus.* 2021; **11**(6): 20210013.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

5. Brooks-Pollock E, Danon L, Jombart T, *et al.*: **Modelling that shaped the early COVID-19 pandemic response in the UK.** *Philos Trans R Soc Lond B Biol Sci.*

169

2021; **376**(1829): 20210001.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

6. Funk S, Abbott S, Atkins BD, *et al.*: **Short-term forecasts to inform the response to the Covid-19 epidemic in the UK.** *medRxiv.* 2020.
    Publisher Full Text

7. Keeling MJ, Dyson L, Tildesley MJ, *et al.*: **Comparison of the 2021 COVID-19 roadmap projections against public health data in England.** *Nat Commun.* 2022; **13**(1): 4924.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

8. Greenberg N, Docherty M, Gnanapragasam S, *et al.*: **Managing mental health challenges faced by healthcare workers during covid-19 pandemic.** *BMJ.* 2020; **368**: m1211.
    PubMed Abstract | **Publisher Full Text**

9. Riedel PL, Kreh A, Kulcar V, *et al.*: **A Scoping Review of Moral Stressors, Moral Distress and Moral Injury in Healthcare Workers during COVID-19.** *Int J Environ Res Public Health.* 2022; **19**(3): 1666.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

10. Shale S: **Moral injury and the COVID-19 pandemic: reframing what it is, who it affects and how care leaders can manage it.** *BMJ Lead.* 2020; **4**(4).
    Publisher Full Text

11. Maunder RG, Lancee WJ, Balderson KE, *et al.*: **Long-term Psychological and Occupational Effects of Providing Hospital Healthcare during SARS Outbreak.** *Emerg Infect Dis.* 2006; **12**(12): 1924–1932.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

12. Chau SWH, Wong OWH, Ramakrishnan R, *et al.*: **History for some or lesson for all? A systematic review and meta-analysis on the immediate and long-term mental health impact of the 2002–2003 Severe Acute Respiratory Syndrome (SARS) outbreak.** *BMC Public Health.* 2021; **21**(1): 670.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

13. Sherratt K, Abbott S, Carnegie A, *et al.*: **Improving modelling for epidemic responses: reflections.** *Open Science Framework.* [Dataset], 2023.
    http://www.doi.org/10.17605/OSF.IO/YG46B

14. Chatham House: **Chatham House Rule.** Aug. 26, 2022.
    Reference Source

15. Covid-19 Public Inquiry: **UK Covid-19 Inquiry.** (accessed May 15, 2023).
    Reference Source

16. Royal Society: **Lessons from modelling the pandemic.** (accessed May 13, 2023).
    Reference Source

17. Walton M, Murray E, Christian MD: **Mental health care for medical staff and affiliated healthcare workers during the COVID-19 pandemic.** *Eur Heart J Acute Cardiovasc Care.* 2020; **9**(3): 241–247.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

18. Watson P: **Stress, PTSD, and COVID-19: the Utility of Disaster Mental Health Interventions During the COVID-19 Pandemic.** *Curr Treat Options Psychiatry.* 2022; **9**(1): 14–40.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

19. Brooks S, Amlôt R, Rubin GJ, *et al.*: **Psychological resilience and post-traumatic growth in disaster-exposed organisations: overview of the literature.** *BMJ Mil Health.* 2020; **166**(1): 52–56.
    PubMed Abstract | **Publisher Full Text**

Open Peer Review

Current Peer Review Status: ✓ ✓

✓   **Robert Moss** (iD)

The University of Melbourne, Melbourne, Victoria, Australia

This article reports findings arising from a one-day workshop held in London in early 2023, in which participants were asked to reflect on the personal and professional impacts of their involvement in informing strategy and policy during the COVID-19 pandemic, for the purposes of identifying key issues (and potential recommendations to mitigate these issues) and to provide guidance for future public health crises.

As the authors acknowledge in the introduction, there is a well-established body of literature concerning the impact of disasters on exposed persons, and ways to mitigate these impacts, but this literature primarily focuses on direct responders and survivors. Less attention has been given to persons who contribute indirectly to disaster response, such as the infectious disease modelling community represented in this article. I'm only aware of one paper that reflects on (professional) impacts of the COVID-19 response on modellers: "The COVID-19 response illustrates that traditional academic reward structures and metrics do not reflect crucial contributions to modern science", which was written in 2020 by three authors of this article.

The common themes and personal impacts identified here may not necessarily be surprising (many reflect broader, long-standing issues in academia) but they deserve genuine attention and reflection. Many of these issues resonate strongly with my own experiences, and those of my colleagues. I wholeheartedly agree with the authors' conclusion that it is "critical we act outside of response contexts".

The workshop was carefully planned and conducted. The supplementary materials attest to the level of engagement from the participants, and to how thoroughly the organisers have captured and reported the findings. It is unfortunate, but entirely understandable, that only a few individuals who had left the field were able to attend the event.

I only have a few minor comments regarding the article text.

1. In "Snapshot reflections", the third sentence begins "Affirmative responses were more common ...", referring to the question "Who should ensure that individuals are adequately supported and credited for their response work".
   My first thought was that this is not a question with a "yes" or "no" answer. It took me a moment to recall that this was a multiple-choice question, and so participants were selecting responsible

organisational units from a predefined list.

The first half of this sentence could potentially be removed, so that it begins "Respondents considered themselves (9/56), line managers (16/56), and research groups (12/56) more responsible ...".

2. In the "Team and work dynamics" results section, the authors report that we might learn from other professions that are better adapted to emergency response work. I think this is a great suggestion, and worth highlighting as an example action for the third priority recommendation ("Ensure response teams are well-staffed, well-resourced, stable, and provided psychological support").

3. The "Non-academic contributions" results section highlights the importance of strengthening the collaboration between academics and public health agencies. This reminds me of Pan-InfORM (Pandemic Influenza Outbreak Research Modelling), a Canadian initiative that was established in 2009 for this very purpose. The authors could cite a recent review of Pan-InfORM activities (published in 2021, doi:10.3934/publichealth.2021020) as an international example.

4. Regarding the final sentence of the discussion:

"It is important to discern whether these concerns are long-standing systemic issues that have been simply exacerbated by the pandemic, or if they have been particularly highlighted due to the unique stressors of the pandemic",

I'm not sure I fully appreciate this distinction. If a concern has been "particularly highlighted" by the pandemic, I still interpret it as meaning that the concern was relevant prior to the pandemic, and I wonder if the intended meaning is that the concern was not identified or appreciated prior to the pandemic?

This sentence also led me to expect an outline of different approaches that might be used to address long-standing systemic issues versus those were specific to the pandemic response. Otherwise I don't understand why this distinction is being made.

References

1. Tariq M, Haworth-Brockman M, Moghadas SM: Ten years of Pan-InfORM: modelling research for public health in Canada.*AIMS Public Health*. 2021; **8** (2): 265-274 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Infectious disease modelling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 03 May 2024

https://doi.org/10.21956/wellcomeopenres.21712.r76468

Srikanth Umakanthan 🆔

The University of the West Indies at Saint Augustine Faculty of Medical Sciences, Saint Augustine, Trinidad and Tobago

The authors have touched on a critical topic that was very much neglected during the COVID-19 pandemic.
This study very well highlights the role and coordination between the employer, employee and the community using the available resources in the best possible manner.
However the following comments need to be addressed:
1. The keywords should not be similar to those in the title.
2. The workshop's type, duration, and target audience should be indicated.
3. The prior aims and objectives of the workshop and the percentage of it being achieved should be included.
4. The professional details of the external facilitator who aided in developing the discussion topics need to be mentioned.
5. The link between the results and the study methods needs to draw a clear conclusion. In the methods the authors look to "identify the Key issues such as lack of institutional support, insecure contracts, unequal credit and recognition, and mental health impacts" but the results look into the "to propose recommendations for future epidemic responses".
6. The list of specialties and the topics of the keynote speakers should be included.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Pathology, infectious diseases

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Supplementary Information: Improving modelling for epidemic responses: reflections from members of the UK infectious disease modelling community on their experiences during the COVID-19 pandemic

Katharine Sherratt* (1), Anna C. Carnegie* (1), Adam Kucharski (1), Anne Cori (2), Carl A. B. Pearson (1,3), Christopher I. Jarvis (1), Christopher Overton (4, 5, 6), Dale Weston (7), Edward M. Hill (8, 9), Edward Knock (2), Elizabeth Fearon (10), Emily Nightingale (1), Joel Hellewell (11), W. John Edmunds (1), Julián Villabona Arenas (1), Kiesha Prem (1,12), Li Pi (13), Marc Baguelin (1,2), Michelle Kendall (8), Neil Ferguson (2), Nicholas Davies (1), Rosalind M. Eggo (1), Sabine van Elsland (2), Timothy Russell (1,11), Sebastian Funk (1), Yang Liu (1), Sam Abbott (1)

*contributed equally

1. Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, UK.
2. MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, UK.
3. South African DSI-NRF Centre of Excellence in Epidemiological Modelling and Analysis (SACEMA), Stellenbosch University, South Africa
4. All Hazards Intelligence, Data Analytics and Surveillance, UK Health Security Agency, UK
5. Department of Mathematical Sciences, University of Liverpool, UK
6. Department of Mathematics, University of Manchester, UK
7. Emergency Response Department Science & Technology Behavioural Science, UK Health Security Agency, UK
8. Warwick Mathematics Institute and The Zeeman Institute for Systems Biology & Infectious Disease Epidemiology Research, University of Warwick, UK
9. Joint UNIversities Pandemic and Epidemiological Research https://maths.org/juniper/
10. Institute for Global Health, University College London, UK
11. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK
12. Saw Swee Hock School of Public Health, National University of Singapore, Singapore
13. Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, UK

Supplementary Information

Survey questions
Snapshot questions
SA pandemic timeline
Session questions
Group discussion themes
Themes
Recommendations

Survey questions

How best would you describe your role in responding to COVID-19?
- Academic researcher
- Civil servant
- Professional services staff
- Other

How best would you describe the role you are doing now?
- Academic researcher
- Civil servant
- Professional services staff
- Other

Approximately how many years experience do you have in your role?
- 0-4 years experience
- 5-9 years experience
- 10+ years experience

If you feel comfortable doing so, please provide an overview of the work you did during the COVID-19 response.
Examples of this might include:
- facilitating meetings
- data management
- producing routine estimates for surveillance and submitting them to advisory bodies.
- developing tools and methods and support their use by other responders
- providing scenario estimates in response to policymakers requests
- writing reports
- academic paper writing
- attending meetings
- supporting researchers
- speaking to the media, ... etc.

On a scale from 0 to 10, how would you rate your professional experience in responding to COVID-19 in the UK?
0   - extremely negative
5   - neutral
10 - extremely positive

What professional costs, if any, did you experience from being involved in the response?

Think about both short- and long-term costs

What professional benefits, if any, did you experience from being involved in the response?
Think about both short- and long-term benefits, for example:
1. number of first author papers
2. promotions/career progression
3. successful grants

On a scale from 0 to 10, how would you rate your personal experience in responding to COVID-19 in the UK?
0 - extremely negative
5 - neutral
10 - extremely positive

What personal costs, if any, did you experience from being involved in the response?
Think about both short- and long-term costs

What personal benefits, if any, did you experience from being involved in the response?
Think about both short- and long-term benefits

Has your employer or the wider community taken action to help mitigate any of the personal or professional costs you identified?
● Yes
● No
● Somewhat

Can you identify areas where action has been taken and areas where it has not?

Do you think there were barriers to doing effective and sustainable COVID-19 outbreak response work?
● Yes
● No
● Maybe

What were some of the barriers to doing effective and sustainable outbreak response work?
For example, thinking about:

1. funders
2. employer organisations
3. supervisors
4. peers
5. computing resources
6. human resources

What can be done to reduce barriers and better support those involved in outbreak response work in the future?

What were some of the things that helped assist you to do effective research during the outbreak response?
For example, thinking about:

1. funders
2. employer organisations
3. supervisors
4. peers
5. computing resources
6. human resources

Do you think sufficient action is currently being taken to improve future outbreak responses to the standard you think is acceptable?
- Yes
- No
- Unsure

Is there anything else you'd like to add?

Snapshot questions

Whiteboards with the following questions were available for the first hour of the session for participants to add to with either post it notes or sticker dots. Numbers in brackets [#] represent the number of sticker dots placed by participants.

1. Do you think sufficient action is currently being taken to improve future outbreak responses to the standard you think is acceptable?
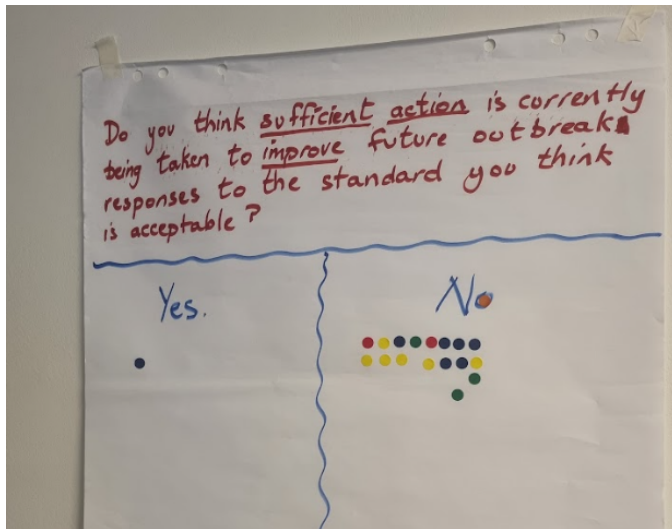
- Yes [1]
- No [18]

2. Who is responsible for ensuring people are supported, and appropriately credited for their work?
   - You [9]
   - Your line manager [16]
   - Your research team [13]
   - Your institution [6]
   - Academic funders [5]
   - The system more generally [8]
   - Other? [0]
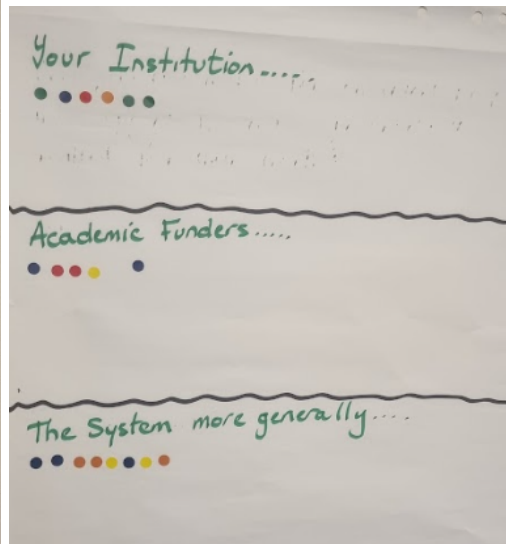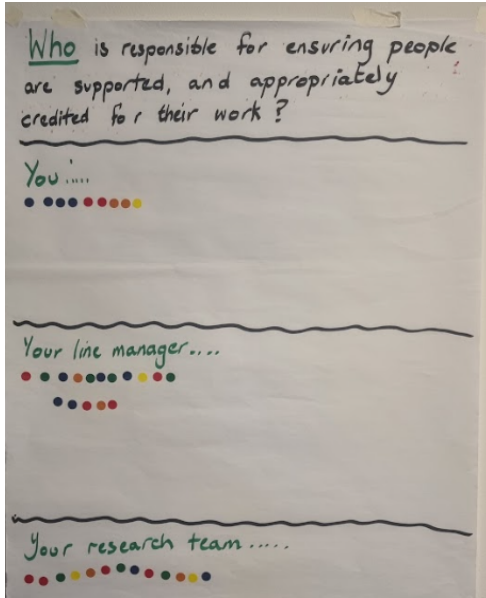
3. Summarise your pandemic experience in one word.
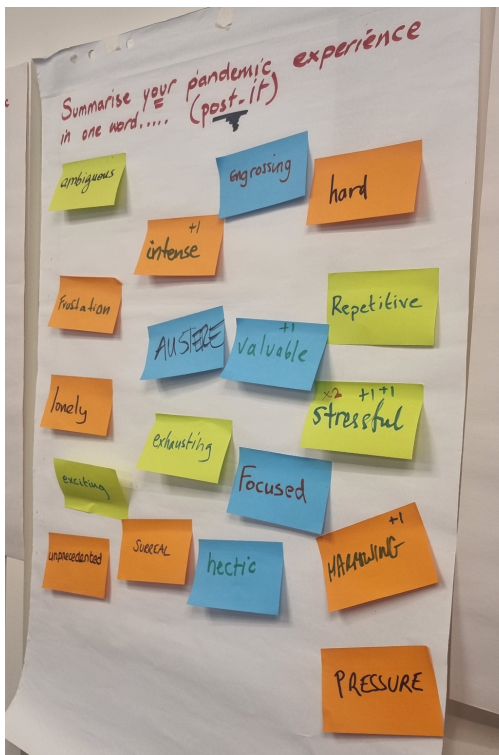
- Ambiguous
- Engrossing
- Hard
- Intense (+1)
- Frustration
- Austere
- Valuable (+1)
- Repetitive
- Lonely
- Exhausting
- Focused
- Stressful (x2) (+1) (+1)
- Exciting
- Unprecedented
- Surreal
- Hectic
- Harrowing (+1)

Session questions

1. What was your pandemic timeline? What were the highs and lows?
2. What was your experience of pandemic work like?
   Below are some suggested directions to take this:
   - What did you expect to experience when you started working on the pandemic response, and how did this differ from what you actually experienced?
   - Did you get any benefits?
   - Were there any negatives?
   - How did your professional and personal experiences differ?
   - Do the professional gains outweigh the personal costs or vice versa?
   - Do you think privilege played a role in your experience?
3. What were some of the things that helped assist you to do effective research during the outbreak response?
   Some examples could be:
   - Management practices
   - Contract duration
   - Professional services staff
   - Funders
   - Employer organisations
   - Supervisors
   - Peers
   - Computing resources
   - HR policies
4. Do you think team science was appropriately supported over the pandemic? Some potential directions:
   - What elements of team science were most important to your work?
   - Has anything changed since prior to the pandemic in terms of support for team science?
   - What role did professional services staff, research administrators and managers, press and media teams - have on your ability to do effective response work?
5. Has your employer or the wider community taken action to help mitigate any of the personal or professional costs/challenges you identified? What more can be done?
6. Do you think there were barriers to doing effective and sustainable COVID-19 outbreak response work? If so, what were they?
7. What has been done and what more can be done to reduce any barriers to effective outbreak response work in the future

Group discussion themes

Participants formed six groups and were asked to summarise their pandemic experience using post it notes, organised into self-identified themes on a whiteboard (one per group). All text is reproduced below, source data photos available on request.

Group 1

- Life outside
    - Partner in field or not
    - Personal circumstances
    - Comradeship - shared experience
    - Importance of camaraderie in fields like military response - conflicts w/ academic reward structure
- Rewards
    - Rewards not attributed equally
    - Institutions got awards, not individuals (not all key players)
    - Don't have good mechanisms for rewarding teams (as opposed to individuals)
- Work pressure
    - Work life balance
    - Overwork
    - Lack of strong management
    - Lack of leadership
    - Line managing
    - Mismatch org - task
- Bad stuff
    - Authorship issues
    - Professional conflicts
    - Professional character assassination
- Access & privileges
    - Different layers of privilege
    - Institutional size (bigger uns = easier)
    - Country of origin
    - Well connected researcher = better access to data

Group 2

- Career direction
    - Responsibility at expense of research breadth
    - Need to get back to long term planning
    - Some career opportunities (but added stress)
    - Missing out on non-publication experiences
    - Big picture vs detail
- Work process
    - Loose in a tornado
    - Lots of competing demands

- ○ Difficult to stop rapid work
  - ○ Outsourcing prioritisation
  - ○ Hybrid work
- (Negativity about) positives
  - ○ Shame associated with "doing well"
  - ○ Positive ? networking + connections
  - ○ Career advancement
  - ○ Acknowledge previous bad work habits
- Work "feeling" / support
  - ○ Not recognising MH impacts when you're in it
  - ○ Lack of understanding from non C19 colleagues
  - ○ Missing "in person" cues in interaction
  - ○ Need for more support - burnout
  - ○ Lack of support / less guidance
  - ○ Missing informal "check-ins" working remotely
  - ○ Pride in overwork
- Personal
  - ○ Better boundaries / priorities
  - ○ Paused personal commitments
  - ○ "Is it still March 2020?"

Group 3
- A
  - ○ Team spirit
    - ■ New vs existing
    - ■ Lack of support
    - ■ Individual shared
    - ■ Competitive
    - ■ Collaboration
  - ○ Institutional
    - ■ Leadership
    - ■ Protection
    - ■ Recognition
    - ■ Junior members
- B
  - ○ Mental health
    - ■ Lack of desire
    - ■ Reintegration
    - ■ Uncertainty
    - ■ Work-life balance
    - ■ Children, partners / home, family
    - ■ Travel
    - ■ Guilt
    - ■ Feeling stuck in time-space-role

- - - ■ Variable speed of time
      - ■ Dealing with losses
    - ○ Societal impact
      - ■ Responsibility
      - ■ Value of work
      - ■ Representation of work
      - ■ Sustainability
      - ■ Guilt

Group 4
- ● COVID modeller specific
  - ○ Everyone busy = harder to get support
  - ○ Talking about & acknowledge the relative difficulties of each others experiences
  - ○ Surreal level of public + media interest (good or bad)
  - ○ Differences in levels of success & contribution resulting in inequality in various forms, i.e. lots of success based partly on luck
  - ○ Hard to onboard and shift responsibility (sorry new people!)
  - ○ Difficult to balance routine modelling and interesting innovative work
  - ○ Being lumbered with "shit" jobs
  - ○ Worked too much because too much work to do
- ● More general experiences
  - ○ Worked too much because nothing else to do
  - ○ Work + the world were one and the same - neither was an escape from the other
  - ○ Pressure came in waves "not again…"
  - ○ And coincided with Xmas
  - ○ Feels like time didn't exist i.e. no memories for a lot of it
  - ○ Really hating the policies for young people

Group 5
- ● Work
  - ○ Lack of operational data
  - ○ What is "good enough"?
  - ○ Not being able to say no
  - ○ What was the point of some of the work
  - ○ Other work pressures e.g. teaching
  - ○ Waves of work
  - ○ Working overnight
  - ○ Clashes over motivation/priorities
- ● Emotions
  - ○ Powerlessness
  - ○ Sadness
  - ○ Frustration + anger
  - ○ Uncertainty
  - ○ Desensitised

- - - Pride
  - Life
    - Skewed sense of time
    - Life events
    - Child care (x2)
    - False memories
  - What to do next?

Group 6
- Structures / work
  - Shift from egalitarian structure to pre existing hierarchies
  - Collaborative work excitement
  - Team work challenges
  - Gambling / return on investment
  - All a blur
  - Survivor bias
  - Responsibility of managers to look after - who is responsible?
  - Trust
- Structures / life
  - Burn-out creeps up on you - how to avoid
  - Life events (juggling)
  - Self-care
  - Coping mechanisms
  - Personal pandemic preparedness
  - Caring for others
  - Impact of team members' personal challenges on everyone
- Emotions / work
  - Reconciling different perspectives on hierarchy & team working
  - Who gets credit
  - Space for communicating
  - Guilt
  - Comparing self with others
  - Frustration
  - Feeling useful
  - Feeling of lack of legitimacy
  - Pride
  - Self perception vs others' perception
  -
- Emotions / life
  - Trading off career vs health / everything "just 1 more"
  - Personal responsibility (feelings of)
  - Maternity leave
  - "Fresh blood" (return from leave)
  - Fear of missing out

Themes

Participants were tasked with grouping top-level themes from the group discussions into broad categories as a spider-map. This created the following:

- Work life
  - Life outside
  - Life
  - Personal
  - More general experiences
- Work - covid modeller specific
  - Work
    - Work process
    - Work pressure
  - Social impact
  - Team spirit
    - Work "feeling support"
    - Mental health
    - Emotions
- Rewards
  - Structures
  - Institutional
  - Access & priveleges
- Career direction
  - Bad stuff
  - (Negativity about) positives

Recommendations

We reproduce text from the poster boards. We include the implementor suggested by contributors at the end of each statement where this was done.

Social impact
- Educate public / policy maker of modelling knowledge for better communication
- Develop primers/training and build on existing links [*research team*]
- Ensure impactful COVID work understood at funder/institution level (esp. If less obvious/visible)
- Support for dissemination of work/case studies [*team/institution/funder*]
- Funder buy-in: creating opportunities to support consolidation of developed methods and tools
- See credit [and fund] all outputs - incl. Software and communication, media, public engagement etc. [*funders/managers/institution*]
- Formalise connections between teams/disciplines/functions - make sure we retain "what worked"
- Provide support (e.g. MH) to reduce burnout + increase retention of institutional memory
- Do[...?]ting connections [*line managers*]

Mental health
- Culture shift
    - Reframe excellence
- Raising awareness (literacy) - also of team role in crisis response
- Preparedness; peace time team building / leadership training [*wellbeing manager, centre manager*]
- Institutional safeguarding
- "We think we do this well, but we don't" needs embedding over a sustained time
- Departmental leadership buy in; lead by example (processes - not emailing at night etc) [*funders, institution, people*]
- Resilience training (peers support) [*institution*]
- Psychological first aid [*institution*]
- Individual + management training
- Adequate resourcing of central services + research teams to mitigate against burnout = funding. [*funders, institution*]

Career direction
- Now
    - Possibly greater acceptance of portfolio PHDs
    - Leeway in examiners for PhDs during covid/emergencies
    - More recognition of non-traditional outputs
- Future
    - Improve ease of movement in/out of public health agencies (eg UKHSA), eg dual positions
    - Shifting routine data analysis to public health bodies ASAP

- Stop/end routine activities as soon as they're not useful
- Find a way to credit confidential work
- How/who
    - Institutions & senior academics to write letters of support for PHD students & staff with 'non-traditional' outputs [*managers, institution*]
    - Joint appointments at public health agencies [*funders*]
    - Credit all outputs: academic, tool development, communications, public engagements, confidential reporting
    - Clear expectations for promotion [*managers, funders, institutions*]
    - Distinguish "academic" vs "emergency" response
    - Disaster roster + exercise

Work process
- Capacity for cycling between response & research
- Reduce structural reliance on 1-2 people [in] a team performing specific task
- [O]n call system?
- [A]nnual leave
- Prioritise capacity
- Project manager incorporated in team [*department, funder*]
- More transparency from government committees so groups without people on them didn't get as left behind as it seemed (to me at least)
- Transparent preparedness plan *[UKHSA, govt]*
    - Who sits on gov committees
    - What their roles/responsibilities
    - White paper [*dedicated working group incl funders*]
- Clarify roles for pandemic response: software engineering, policy-related roles
- Broader reward system, e.g. [*funder*]
    - Code/software
    - reports/briefing notes
    - "Middle" authors
    - Data collection
- Automation / routine vs. one-off - value/impact?
- Regular re-assessment of cost-effectiveness of tasks
- Ask yourself if you really need to do this so often/ so quickly
- Weekly discussions: priorities [*team leads*]

Life/personal
- Mechanism for feeding back & instituting good working practices
- "GWP" [good working practices] reps & a committee that reports to univ./institution executive board
    - Identify people to take on the role of rep
    - Regular surveys to gauge feelings & elicit suggestions
- Time off in lieu (mandatory)
- Paid overtime (capped)

- Establish working group to make recommendations at a national level [*institution/funders*]
- Realistic funding <> deliverables
- Hobby
- Normalise taking annual leave
- Respect (enforce) working hours [*individuals & institutions*]
- Role models
- Don't try very hard (Relax)
    - Active role
- Guidelines for line managers = cultural change
- Delay emails on weekends and 7pm-7am (allow control)
    - Auto-send emails to remind people to take break from work
- PDR [performance and development review] for work/life balance (part of PDR but checked by welfare manager)
    - Palm trees on slack
    - Normalise people taking holidays > holiday snaps
- "Covid impact statement" but for more general issues
- Have paid wellbeing officer(s)/manager to check in on personal issues & balance, give professional advice, etc…

Institutional
- Credit/rewards
    - Clear comms on processes
    - Meetings non-science (coping, emotion, credits etc)
        - > mandatory incorporation in protocols; normalising [*institution*]
- Starter pack
    - Processes
    - Support & social
    - Technical
    - Mentorship/buddies
    - Available training
    - survey input regular
    - share experience between institutions [*department groups*]
- Resilience
    - Firefighter mentality
    - Back-up / replaceable
        - learn from other professions
        - Structure ahead of time & practice [*dedicated working group*]
- Responsibility
    - Hierarchy
    - Map roles
    - Back-up/replaceable
    - Function of/process for support roles (software, admin/comms)
        - Starter pack

- available guidelines
- structured updating [*institution, department*]
- Clear guidance
    - Communication
    - Expectations
    - Reference resource
        - Project manager
        - career path for support roles with growth
        - comms / software / admin project management [*funder, institution*]
- Manager training
    - Niche to normal
    - Timely, frequent
    - across seniority
        - Mandatory training [*institution*]
- Lessons learned
    - Exit interview feed into updating processes
    - Left since start of pandemic
    - All staff, not only academic
        - Survey / exit interview [*manager/department*]

Additional
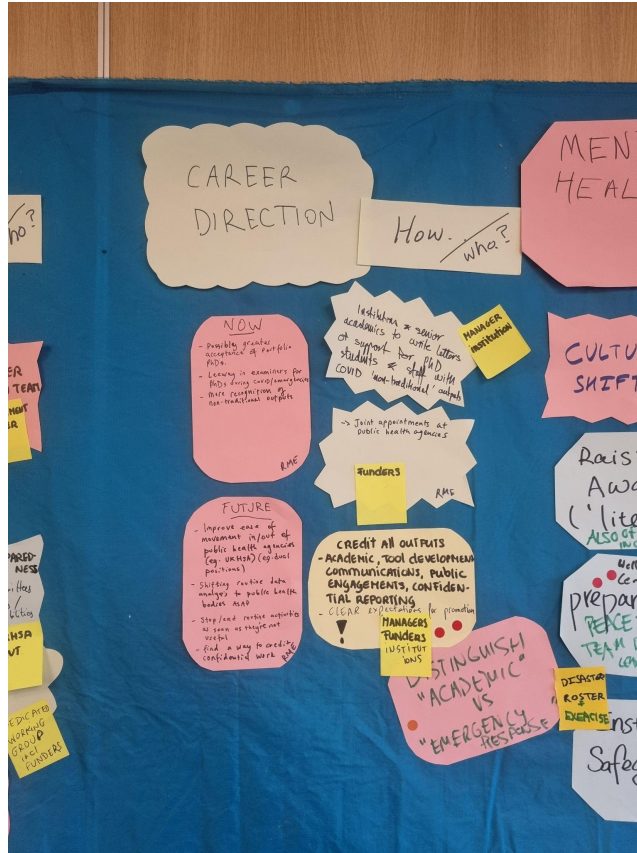- Pay rise
- Now now v just now prioritisation

Figure 1. Example of board showing recommendations

Priority recommendations

These recommendations were highlighted by at least one participant during a group discussion. Recommendations which were highlighted by two or more participants are shown in bold.

1. Research teams should ensure that impactful COVID work is understood by funders and institutions, especially if less obvious/visible.
2. **Research teams should initiate sustainable team building and training programs during non-response periods.**
3. Employers should ensure sustainable funding for academic and professional services roles to reduce burnout risks.
4. **Employers should develop processes and guidelines for career growth support for professional services staff.**
5. **Employers need to create or expand Wellbeing Officer roles to monitor work-life balance and provide guidance.**
6. Incorporation of work-life balance components in annual performance and development reviews is essential for employers.
7. Employers should implement capped paid overtime and formal Time Off In Lieu policies and routinely analyse and act upon staff survey and exit interview data.
8. Funding bodies should adjust eligibility criteria to adequately compensate non-academic staff for activities.
9. **Funding bodies should refine impact measures to acknowledge all outputs, including academic, tool development, communications, public engagement, and confidential reporting.**