Characterising information loss due to aggregating epidemic model outputs

Katharine Sherratt (1), Ajitesh Srivastava (2), Kylie Ainslie (3,4), David E. Singh (5), Aymar Cublier (5), Maria Cristina Marinescu (6), Jesus Carretero (5), Alberto Cascajo Garcia (5), Nicolas Franco (7), Lander Willem (8), Steven Abrams (8,9), Christel Faes (9), Philippe Beutels (8), Niel Hens (9), Sebastian Müller (10), Billy Charlton (10), Ricardo Ewert (10), Sydney Paltra (10), Christian Rakow (10), Jakob Rehmann (10), Tim Conrad (11), Christof Schütte (11), Kai Nagel (10), Rok Grah (12), Rene Niehus (12), Bastian Prasse (12), Frank Sandmann (12), Sebastian Funk (1)

1 London School of Hygiene and Tropical Medicine, London, UK; 2 University of Southern California, Los Angeles, USA; 3 Dutch National Institute of Public Health and the Environment (RIVM), Bilthoven, Netherlands; 4 School of Public Health, University of Hong Kong, Hong Kong SAR, China; 5 Universidad Carlos III de Madrid, Madrid, Spain; 6 Barcelona Supercomputing Center, Barcelona, Spain; 7 University of Namur, Namur, Belgium; 8 University of Antwerp, Antwerp, Belgium; 9 UHasselt, Hasselt, Belgium; 10 Technische Universität Berlin, Berlin, Germany; 11 Zuse Institute Berlin (ZIB), Berlin, Germany; 12 ECDC, Stockholm, Sweden

Funding declaration: KS, SF funded by ECDC and Wellcome (210758/Z/18/Z). AS funded by National Science Foundation Award 2135784, 2223933. KA funded by Netherlands Ministry of Health, Welfare and Sport, and European Union's Horizon 2020 research and innovation programme - project EpiPose (grant agreement number 101003688). DES, AC, MM, JC, ACG funded by U3CM, Instituto de Salud Carlos III, Gobierno de España, European Commission. NF, LW, SA, CF, PB, NH funded by European Union's Horizon 2020 research and innovation programme (grant number 101003688 – EpiPose project). SM, BC, RE, SP, CR, JR, TC, CS, KN funded by Ministry of research and education (BMBF) Germany (grants number 031L0300D, 031L0302A). RG, RN, BP, FS funded by ECDC.

Abstract

Background. Collaborative comparisons and combinations of multiple epidemic models are used as policy-relevant evidence during epidemic outbreaks. Typically, each modeller summarises their own distribution of simulated trajectories using descriptive statistics at each modelled time step. We explored information losses compared to directly collecting a sample of the simulated trajectories, in terms of key epidemic quantities, ensemble uncertainty, and performance against data.

Methods. We compared July 2022 projections from the European COVID-19 Scenario Modelling Hub. Using shared scenario assumptions, five modelling teams contributed up to 100 simulated trajectories projecting incidence in Belgium, the Netherlands, and Spain. First, we compared epidemic characteristics including incidence, peaks, and cumulative totals. Second, we drew a set of quantiles from the sampled trajectories for each model at each time step. We created an ensemble as the median across models at each quantile, and compared this to an ensemble of quantiles drawn from all available trajectories at each time step. Third, we compared each

trajectory to between 4 and 29 weeks of observed data, using the mean absolute error to weight trajectories in consecutive ensembles.

Results. We found that collecting models' simulated trajectories, as opposed to collecting models' quantiles at each time point, enabled us to show additional epidemic characteristics, a wider range of uncertainty, and performance against data. Sampled trajectories contained a right-skewed distribution which was poorly captured by an ensemble of models' quantile intervals. Ensembles weighted by predictive performance narrowed the range of plausible incidence over time, excluding some epidemic shapes altogether.

Conclusions. Understanding potential information loss when collecting model projections can support the accuracy, reliability, and communication of collaborative infectious disease modelling efforts. The importance of different information losses may vary with each collaboration's aims, with lesser impact on short term predictions compared to assessing threshold risks and longer term uncertainty.

Data availability. All code and data available on Github:
https://github.com/covid19-forecast-hub-europe/aggregation-info-loss

Background

One of the key challenges in infectious disease modelling is the representation of multiple sources of uncertainty, both within each model as well as across separate model projections [1], [2]. In public health decision-making, accounting for the full extent of uncertainty in future infectious disease incidence is critical to mitigate exposure to high impact vulnerabilities, such as health systems operating beyond surge capacity [3], [4].

A probabilistic infectious disease model characterises both epistemic and aleatoric uncertainties arising from complex and changing causes of disease transmission. In order to simulate this real-world process, modellers must handle stochasticity in transmission dynamics, typically using observed data to estimate model parameters that are themselves uncertain. Each probabilistic 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 (e.g., weekly incidence of infections or cases).

When creating models to characterise the future, modellers have often drawn a distinction between forecasts and scenarios [5]. 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 [6] 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 qualitative factors whose futures may not be quantitatively predictable, such as options for policy interventions [7], [8]. 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 multiple models for projecting the future using a variety of different methods [9]. Each collaboration attempts to standardise epistemic uncertainty across different models by setting a single common target for projections and collecting results from multiple independent models. This allows a like-for-like comparison of the stochastic uncertainty produced by varying modelling methods. Ensemble methods can then combine across models to generate a more comprehensive and robust prediction [10] or reflection of expert judgement [11].

Formal, large-scale modelling collaborations have, so far, been used for influenza, Ebola, zika, dengue fever, and COVID-19 [9]. In the case of COVID-19, a number of policy-facing research groups have set up collaborations to collate forecasts and scenarios [12]–[15], 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 evaluating the output of past and current ensemble modelling projects. This has included evaluating differing performance among individual models [16]–[18], and a variety of methods for creating ensembles from multiple models [10], [19], [15], [11].

However, to our knowledge so far no evaluation has been performed of how probabilistic models are represented before they are compared or combined in an ensemble. To comparably assess the output from multiple probabilistic models, the same set of statistics should be used to summarise each model's distribution. One approach to this uses descriptive statistics taken across all simulated trajectories from each model at each given time step. In several COVID-19 modelling hub efforts each modeller submits a common set of quantiles for each time point estimated from any number of trajectories.

This approach may lose information pertinent to epidemic decision making, such as misrepresenting the size and timing of peak incidence [20]. This loss is particularly relevant when trying to capture the full extent of variability across multiple models. For the European COVID-19 Scenario Hub, the use of quantile summaries was replaced in mid-2022 by individual model trajectories that represent random samples from the collection of all possible trajectories of the model consistent with a given scenario and the data available up to the time at which the simulation was generated.

We aim to explore three aspects of information loss associated with collecting a set of quantile intervals as a method of collecting multiple models' projections of infectious disease incidence. Specifically, we explore the impact of information loss on policy-relevant epidemic characteristics, including cumulative totals and timing of peaks; the ability to capture the full extent of uncertainty across multiple models when creating an ensemble; and the information gained by comparing modelled epidemic trajectories to observed data. Understanding the potential sources of information loss in collecting multiple model projections may support improving the accuracy, reliability, and communication of collaborative infectious disease modelling efforts.

Methods

Study setting

In this work we use projections from Round 2 of the European COVID-19 Scenario Modelling Hub [20]. The European COVID-19 Scenario Hub was launched in March 2022 to reflect demand for longer term European policy planning from the ECDC. It used the existing US Scenario Hub [14] 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 independent projections. 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 [21]. In practice some of the models were not calibrated to reported cases and therefore used symptomatic cases as a proxy. 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 [22].

Model aggregation

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 [23].

Characterising information loss

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, and cumulative 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, chosen to balance avoiding noise and capturing distinct weight. We assessed peaks both over the entire projection period, and over only the autumn-winter period (October 2022 through March 2023). 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 separately created a quantile ensemble created by first estimating model-specific quantiles from each model's distribution of trajectories at each time point, for each scenario, location, and outcome target, before calculating the median across the different models' values at each quantile and time step. This is the procedure that has been used before to produce ensemble projections across multiple epidemiological forecasts [10], [15]. 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 Harrel Davis weighted estimator [22] from the *cNORM* R package (v3.0.2) [23]. As above, we calculated 23 quantiles including the median to express uncertainty.

We repeated this process to create consecutive 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.

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 5,920 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 (figure 1.i).

Aggregating across simulated trajectories from multiple models created additional information about various epidemic characteristics that is not available from aggregating across quantile summaries from each model. Both quantile and simulated trajectories yield a probabilistic estimate of the weekly incidence of each projection target. 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, or estimating time-series characteristics including epidemic peak size or timing.

We reported on these characteristics using the simulated trajectories (see contemporaneous report reproduced in the Supplement). 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 an ensemble projection (figure 1). We compared an ensemble taken from all trajectories (figure 1.ii) with a median ensemble from models' quantiles (figure 1.iii). Across all projection targets, we observed substantially increased uncertainty in an ensemble that aggregated directly from trajectories. This represented the wider variety of epidemic shapes projected by different models, for example where the credible interval of projections included high autumn-winter incidence in Spain, and gave greater credibility to multiple peaks of incidence in Belgium. These were not observed in the interval projections of an ensemble derived from models' quantiles.

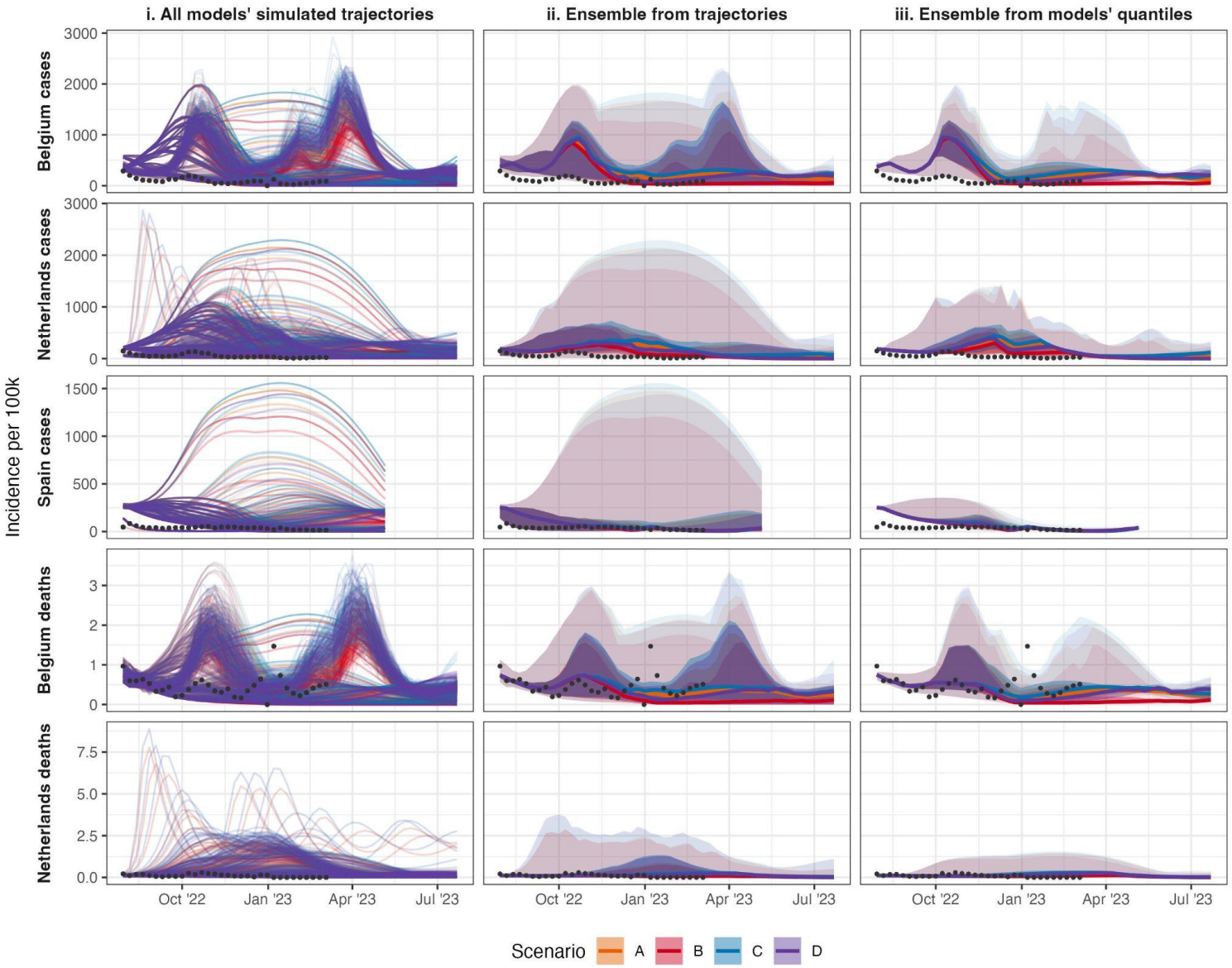

Figure 1. Projections of incidence per 100,000 population, by country (row) and aggregation method (column) showing median, 50%, and 99% probabilistic intervals (increasingly shaded ribbons), for each scenario, using: i) no ensemble method (100 simulated trajectories per model, or 96 in case of one of 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. 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).

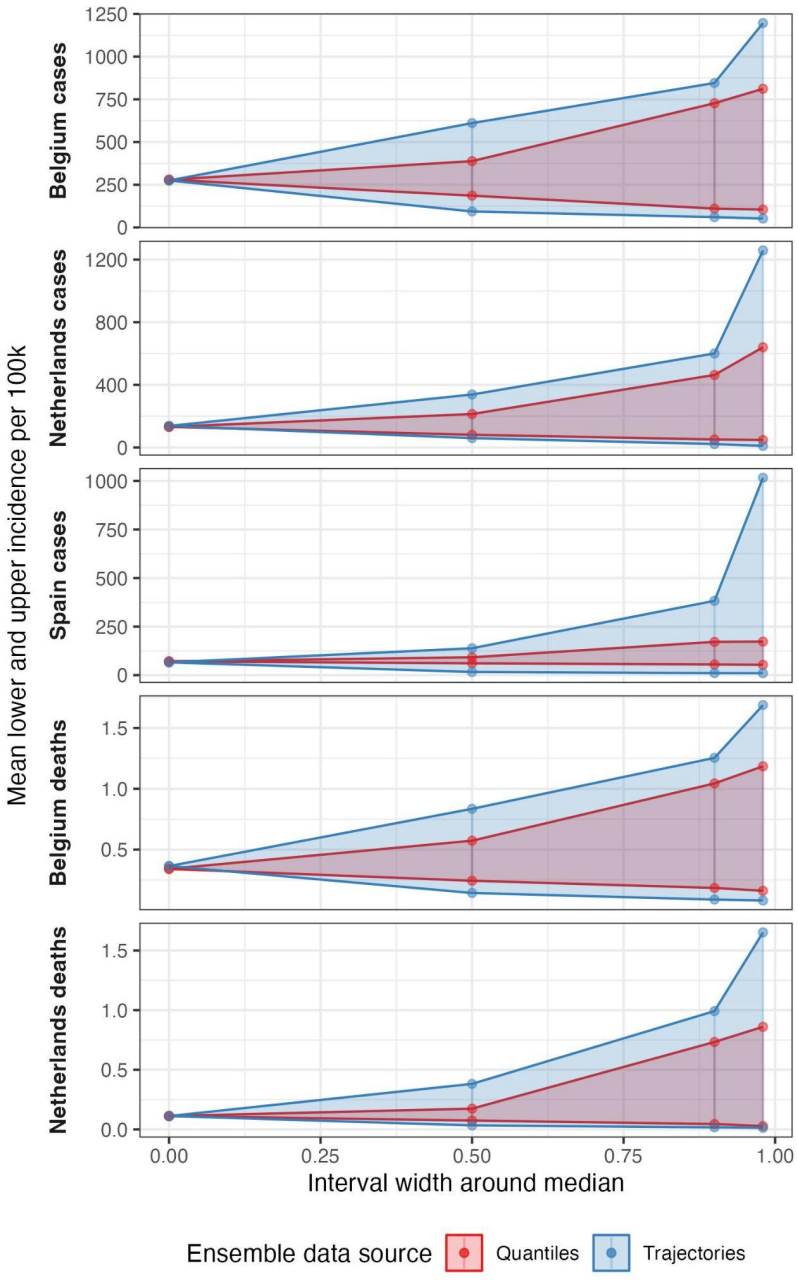

Figure 2. The 52-week mean of incidence per 100,000 population across all time points and scenarios, showing mean central prediction intervals at increasing distances from the median (interval width), by aggregation method. Methods for aggregating model outputs are: a median across individual-model quantiles (red), and quantiles from the sampled trajectories of all models at each time point (blue). 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. With increasing levels of uncertainty, the distances between the lower and upper bounds are on average reduced in an ensemble created from the median of quantile summaries of model simulations (red), compared to trajectories (blue).

To compare between the two ensembles, we considered the mean of values of these two ensembles at each quantile across all time points and scenarios (figure 2). This showed both ensembles produced similar values around the centre of the distribution, with no noticeable difference between the median values of each projection. However, the two ensembles increasingly diverged in projecting the outer upper limit of the probabilistic distribution. At the upper 98% probability interval, ensemble projections for cases in Spain averaged nearly six times higher incidence when drawn from 100 trajectories compared to when drawn from quantiles (respectively averaging 1016 and 173 weekly new cases per 100,000 population). Across all five targets, we observed the pattern that an ensemble based on simulated trajectories produced sharply increasing uncertainty between the 90% to 98% intervals. Meanwhile in an ensemble based on quantiles projected values were closer across upper bound probabilistic intervals.

We then considered an ensemble of individual trajectories each weighted against a sequentially increasing amount of observed data (figure 3). We note that models used a variety of methods and may have been calibrated to alternative data sources (see Supplement). We observed reduced uncertainty across ensemble projections when weights in an ensemble were updated over time. 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 1). When weighted by performance, the 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 figure 1i).

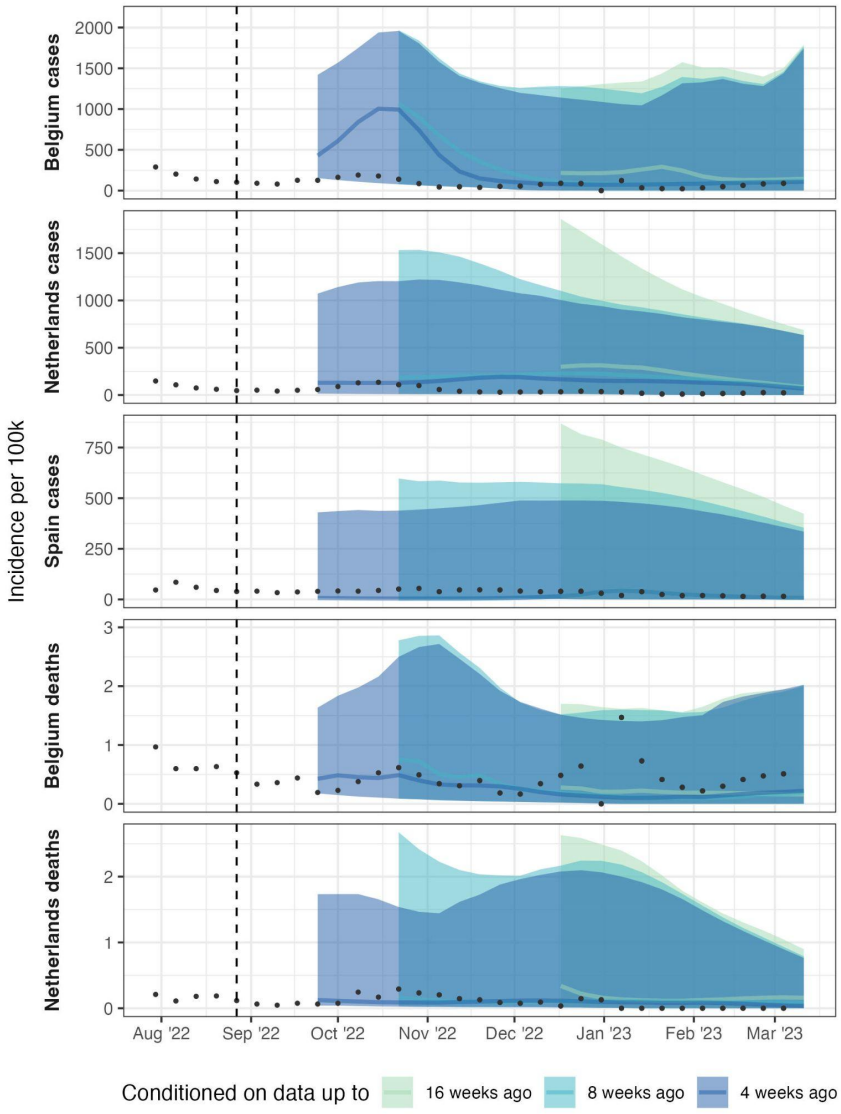

Figure 3. Weighted ensemble forecasts of incidence 4, 8, and 16 weeks ahead of available data, with available data increasing weekly over time. Showing median (lines) and 99% credible interval (shaded ribbons). Each simulated trajectory started from 30 July 2022 and was weighted using its inverse mean absolute error against available data. We used at least 4 and up to 29 weeks of observed data. This means the first ensemble is weighted by scoring trajectories against data between 30 July to the 27 August 2022 (vertical dashed line), and gives a 4 week ahead forecast for the 24 September (start of dark blue ribbon), an 8 week ahead forecast for the 22 October (start of mid blue ribbon), and 16 week forecast for the 17 December 2022 (start of light green ribbon). We created consecutive weekly ensembles, with weights updating as increasing observed data became available, shown continuously by projection horizon.

Discussion

We compared two methods of collecting data about multiple models' projections of an epidemic. We compared a sample of up to 100 simulated trajectories from three scenario models for each of five projection targets, with the representation of those trajectories by taking probabilistic quantile intervals at each time step. We compared the two results in terms of quantifiable epidemic characteristics, the range of uncertainty in an ensemble model, and the use of each trajectory's predictive performance in an ensemble. We found that collecting simulated trajectories showed trajectory shapes, peaks, and cumulative total burden; contained a right-skewed distribution which was poorly summarised by an ensemble of quantile intervals; and could be used in an ensemble based on continuous predictive performance.

Using a standardised set of quantile intervals has several advantages. Theoretically, combining across a set of quantiles should accurately represent the underlying distribution [26], with various methods for doing so depending on the view taken of uncertainty between and across model projections [11]. Second, 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. Standardisation in quantile form allows for a direct comparison between multiple models, which can be made available for modellers and decision makers to evaluate across many contributors' best efforts to express the probable range of future outcomes [27]. Third, 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, although this argument weakens with mounting evidence that this type of under-resourcing hampers outbreak response.

Collecting model projections in probabilistic quantile intervals trades the time-dependence of each simulated trajectory, for a fixed-time instantaneous summary across trajectories at each time step. We have shown several types of information loss from this trade-off, in line with ongoing work addressing similar 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 [28]. 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 [20].

We demonstrated a clear use case for 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 could be used for ongoing evaluation of scenario projections, increasing the useful life of data 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.

We highlight several important limitations to our comparison of information loss between methods of collecting model output. Our method of collecting simulated trajectories was not specifically designed for the purpose of comparing them to equivalent quantile probabilities, and as a result our findings are difficult to interpret. In some cases, the collated trajectories were already subsampled from model runs conducted by individual teams. Future studies 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.

We believe our results show potential losses across three different types of information relevant to collaborative multi-modelling projects in infectious disease. This applies to all aims and methods of combining multiple models, whether projections are conditioned on the context of the present (as in forecasts), or on schematic futures (as in scenarios). However, the impact of information loss may differ depending on the aim of a multi-model comparison. Our results suggest little information is lost in comparisons of the central estimates from different models, which is a useful validation for collecting multiple model results in any format when the purpose is short-term situational awareness. However, we observed substantial information loss when comparing the tails of multiple distributions, assessing 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, areas particularly relevant to longer term preparedness and mitigation.

We suggest that further work should characterise and standardise sampling techniques for model simulations in multi-model comparisons. Working from combined simulations then offers the opportunity to explore creating ensembles by the shape of epidemic curve, and for more detailed quantitative evaluations against observed data, such as in projected peaks or cumulative totals. This 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.

References

[1] R. McCabe *et al.*, ‘Communicating uncertainty in epidemic models’, *Epidemics*, vol. 37, p. 100520, Dec. 2021, doi: 10.1016/j.epidem.2021.100520.

[2] B. Swallow *et al.*, ‘Challenges in estimation, uncertainty quantification and elicitation for pandemic modelling’, *Epidemics*, vol. 38, p. 100547, Mar. 2022, doi: 10.1016/j.epidem.2022.100547.

[3] S.-L. Li *et al.*, ‘Essential information: Uncertainty and optimal control of Ebola outbreaks’, *Proc. Natl. Acad. Sci.*, vol. 114, no. 22, pp. 5659–5664, May 2017, doi: 10.1073/pnas.1617482114.

[4] C. S. Lutz *et al.*, ‘Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples’, *BMC Public Health*, vol. 19, no. 1, p. 1659, Dec. 2019, doi: 10.1186/s12889-019-7966-8.

[5] M. Lipsitch, L. Finelli, R. T. Heffernan, G. M. Leung, and S. C. Redd; for the 2009 H1N1 Surveillance Group, ‘Improving the Evidence Base for Decision Making During a Pandemic:

The Example of 2009 Influenza A/H1N1', *Biosecurity Bioterrorism Biodefense Strategy Pract. Sci.*, vol. 9, no. 2, pp. 89–115, Jun. 2011, doi: 10.1089/bsp.2011.0007.
[6] K. Sherratt *et al.*, 'Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations', *eLife*, vol. 12, p. e81916, Apr. 2023, doi: 10.7554/eLife.81916.
[7] K. Shea *et al.*, 'Harnessing multiple models for outbreak management', *Science*, vol. 368, no. 6491, pp. 577–579, May 2020, doi: 10.1126/science.abb9934.
[8] T. Rhodes, K. Lancaster, S. Lees, and M. Parker, 'Modelling the pandemic: attuning models to their contexts', *BMJ Glob. Health*, vol. 5, no. 6, p. e002914, Jun. 2020, doi: 10.1136/bmjgh-2020-002914.
[9] N. G. Reich *et al.*, 'Collaborative Hubs: Making the Most of Predictive Epidemic Modeling', *Am. J. Public Health*, vol. 112, no. 6, pp. 839–842, Jun. 2022, doi: 10.2105/AJPH.2022.306831.
[10] E. L. Ray *et al.*, 'Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S.', *medRxiv*, p. 2020.08.19.20177493, Aug. 2020, doi: 10.1101/2020.08.19.20177493.
[11] E. Howerton *et al.*, 'Context-dependent representation of within- and between-model uncertainty: aggregating probabilistic predictions in infectious disease epidemiology', *J. R. Soc. Interface*, vol. 20, no. 198, p. 20220659, Jan. 2023, doi: 10.1098/rsif.2022.0659.
[12] S. Funk *et al.*, 'Short-term forecasts to inform the response to the Covid-19 epidemic in the UK', *medRxiv*, p. 2020.11.11.20220962, Nov. 2020, doi: 10.1101/2020.11.11.20220962.
[13] E. Y. Cramer *et al.*, 'The United States COVID-19 Forecast Hub dataset'. medRxiv, p. 2021.11.04.21265886, Nov. 04, 2021. doi: 10.1101/2021.11.04.21265886.
[14] R. K. Borchering, '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.*, vol. 70, 2021, doi: 10.15585/mmwr.mm7019e3.
[15] K. Sherratt *et al.*, 'Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations', *eLife*, p. Forthcoming, Jun. 2022, doi: 10.1101/2022.06.16.22276024.
[16] C. Viboud *et al.*, 'The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt', *Epidemics*, vol. 22, pp. 13–21, Mar. 2018, doi: 10.1016/j.epidem.2017.08.002.
[17] J. Bracher *et al.*, 'A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave', *Nat. Commun.*, vol. 12, no. 1, p. 5173, Aug. 2021, doi: 10.1038/s41467-021-25207-0.
[18] E. Y. Cramer *et al.*, 'Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States', *Proc. Natl. Acad. Sci.*, vol. 119, no. 15, p. e2113561119, Apr. 2022, doi: 10.1073/pnas.2113561119.
[19] J. W. Taylor and K. S. Taylor, 'Combining Probabilistic Forecasts of COVID-19 Mortality in the United States', *Eur. J. Oper. Res.*, Jun. 2021, doi: 10.1016/j.ejor.2021.06.044.
[20] J. L. Juul, K. Græsbøll, L. E. Christiansen, and S. Lehmann, 'Fixed-time descriptive statistics underestimate extremes of epidemic curve ensembles', *Nat. Phys.*, vol. 17, no. 1, Art. no. 1, Jan. 2021, doi: 10.1038/s41567-020-01121-y.
[21] E. Dong, H. Du, and L. Gardner, 'An interactive web-based dashboard to track COVID-19 in real time', *Lancet Infect. Dis.*, vol. 20, no. 5, pp. 533–534, May 2020, doi: 10.1016/S1473-3099(20)30120-1.
[22] European COVID-19 Scenario Hub, 'Round 2'. https://covid19scenariohub.eu/report2.html
[23] K. Sherratt and S. Funk, 'covid19-forecast-hub-europe/aggregation-info-loss'. Accessed: Jun. 23, 2023. [Online]. Available: https://github.com/covid19-forecast-hub-europe/aggregation-info-loss/releases/tag/v0.1
[24] F. E. Harrell and C. E. Davis, 'A new distribution-free quantile estimator', *Biometrika*, vol. 69, no. 3, pp. 635–640, Dec. 1982, doi: 10.1093/biomet/69.3.635.
[25] A. Lenhard, W. Lenhard, and S. Gary, 'cNORM - Generating Continuous Test Norms'. 2018.

doi: 10.13140/RG.2.2.25821.26082.
[26] C. Genest, ‘Vincentization Revisited’, *Ann. Stat.*, vol. 20, no. 2, pp. 1137–1142, 1992.
[27] J. Bracher, E. L. Ray, T. Gneiting, and N. G. Reich, ‘Evaluating epidemic forecasts in an interval format’, *PLOS Comput. Biol.*, vol. 17, no. 2, p. e1008618, Feb. 2021, doi: 10.1371/journal.pcbi.1008618.
[28] A. Srivastava, S. Singh, and F. Lee, ‘Shape-based Evaluation of Epidemic Forecasts’. arXiv, Nov. 11, 2022. doi: 10.48550/arXiv.2209.04035.