Characterising information loss due to aggregating epidemic model outputs

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

1 London School of Hygiene and Tropical Medicine, London, UK; 2 University of Southern California, Los Angeles, USA; 3 RIVM, Bilthoven, Netherlands; 4 Universidad Carlos III de Madrid, Madrid, Spain; 5 Barcelona Supercomputing Center, Barcelona, Spain; 6 University of Namur (Belgium), Namur, Belgium; 7 University of Antwerp (Belgium), Antwerp, Belgium; 8 University of Hasselt (Belgium), Hasselt, Belgium; 9 TU Berlin, Berlin, Germany; 10 ZIB Berlin, Berlin, Germany; 11 ECDC, Stockholm, Sweden

Target journal: Epidemics, US Scenario Hub special issue

Keywords: information loss, uncertainty, aggregation, epidemic modelling

Abstract

Background. Epidemic modelling projections, and particularly comparisons between multiple models, are increasingly seen as a reliable source of policy-relevant evidence during epidemic outbreaks. Results from multiple models are typically collected by asking modellers to summarise a distribution of results using descriptive statistics. Here we look at information loss from this method, compared to collecting a subsample of underlying model simulations. We explored information losses in terms of key epidemic quantities, uncertainty in an ensemble, and evaluating performance over time.

Methods. We compared projections from Round 2 of the European COVID-19 Scenario Modelling Hub. We asked modellers to model a set of pre-specified scenarios, and from each team collected up to 100 simulations for each projection target. First, we used all simulations to compare key epidemic characteristics including peak times and cumulative totals. Second, to recreate current practice, we drew a set of standard quantiles from the submitted simulations for each model, and followed the widely used ensemble method of taking a median average across models at each quantile. We compared this to an ensemble created by drawing probabilistic quantiles from all available simulations at each time step. Third, we compared each simulation to observed data using mean average error. We used this to weight each simulation in a weighted median ensemble of all simulations.

Results. We found that by collecting simulations we were able to show trajectory shapes, peaks, and cumulative total burden. The sample of simulations contained a right-skewed distribution which was poorly summarised by an ensemble of quantile intervals. As expected, we observed wide variation in the forecasting performance of each simulation. An ensemble weighted by

predictive performance substantially narrowed the range of plausible future incidence and excluded some epidemic shapes altogether.

Conclusions. 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.

Data availability. All code and data available on Github: covid19-forecast-hub-europe/covid19-scenario-hub-europe/tree/analysis/analysis

Background

One of the key challenges in infectious disease modelling is the representation and communication of the multiple sources of uncertainty within as well as variability across model projections [1], [2]. Understanding the full extent of uncertainty in future infectious disease incidence is crucial for decision-making to account for exposure to high impact vulnerabilities in public health, such as health systems operating beyond surge capacity [3], [4].

A probabilistic infectious disease model characterises the fundamental epistemic and aleatoric uncertainties arising from complex and changing causes of disease transmission. In order to simulate this real-world process, modellers must at the same time handle stochasticity in transmission dynamics, typically using observed data to estimate model parameters that are themselves uncertain. Each probabilistic model can be run over any number of simulations, 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.

When creating models to characterise the future, modellers have often drawn a distinction between forecasts and scenarios [5]. Forecasts are unconditional 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. These can usually be reliably made only for at best a few generations of transmission [6] because of unmodelled factors affecting future transmission such as behavioural or policy changes or heterogeneity in transmission risk. In contrast, scenario projections are predictions 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 the 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.

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] and reflection of expert judgement [11].

Formal, large-scale modelling collaborations have so far been used for influenza, zika and 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]–[14], [6], and there is a substantial effort towards expanding the practice of ensemble projections of infectious disease. 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 [15]–[17], and a variety of methods for creating ensembles from multiple models [10], [18], [6], [11].

However, so far no evaluation has been performed of the method used for capturing output across multiple probabilistic models. 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 simulations from each model at each given time step. In several COVID-19 modelling hub efforts each modeller submits a common set of 23 quantiles drawn from any number of simulations. This approach may lose information pertinent to epidemic decision making, such as misrepresenting the size and timing of peak incidence [19]. 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 of equal probability from each of the models.

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. We started the European Scenario Hub in March 2022 to reflect demand for longer term European policy planning from the ECDC. We used the existing US Scenario Hub [14] as a basis for Hub infrastructure and methods. We recruited modelling teams by word of mouth to join a series of collaborative workshops, approximately fortnightly 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 political changes that would impact the incidence of COVID-19 across Europe in the medium term.

We asked teams 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 we repeated this process 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 BA.4/BA.5/BA.2.75; scenarios C/D). We asked modellers to start their projections from 24th July 2022, excluding any data after this date. Modellers were asked to submit up to 100 sample simulations, each showing a trajectory of weekly incidence over time for a given projection target. We requested that simulations were of equal probability and randomly sampled from any total number of simulations each modelling team produced. We have published full scenario details including shared parameters, all teams' projections, and summary results online [20].

Model aggregation

This work specifically focuses on contrasting the sampled simulations with their representation in quantiles. We collected raw data in the form of up to 100 sample simulations from each model for each projection target. We used this data to retrospectively create a fixed-time quantile representation of results from each model and target. For each model, we took the available sample simulations for each projection target at each week of results. Following the current submission procedure across COVID-19 Modelling Hubs for an individual model, we drew a median and 22 further quantiles to represent a range of up to 99% credibility in the probable incidence of each weekly projection target. We processed all data in R with code available online [21]. Each week of each projection target for each of four scenarios was then represented in six ways: by three models' up to 100 sample simulations, and by each of those three models' sample simulations summarised as a set of quantile probabilities.

Characterising information loss

First we considered information about key epidemic characteristics from analysis of either sample simulations or their quantile representation. Discussion with the ECDC modelling team led to an interest in estimates of incidence over time, and cumulative outbreak number, size,

and timing of peak incidence over the projection period. Both representations yield a probabilistic estimate of the weekly incidence of each projection target. As the quantile representation was a summary across samples at each time step it did not permit aggregating across time for cumulative totals or means, or estimates of epidemic peak size or timing. We therefore estimated the remaining characteristics only using summaries across the sample simulations. We summed incidence over time to produce a cumulative total from each simulation, and identified peaks in each simulated trajectory as the local maxima in a sliding window of five weeks.

Secondly, we compared the use of a standard unweighted ensemble to express uncertainty across multiple models in the two representations. From multiple quantile predictions, an ensemble projection can be aggregated and combined in various ways, including simple or weighted averaging, or a linear opinion pool that makes different assumptions about between-prediction uncertainty. Here we use only an equally weighted median across all models' estimates for each value. The unweighted median is similarly used to produce ensemble projections across multiple Hubs [6], [10].

We explored differences created in ensemble projections expressed as a standard set of quantiles. First we created an ensemble projection summarised across all individual sample simulations. We took a set of quantiles from across all models' submitted values for each interval in the time-series, for each scenario, location, and outcome target. Next, we separately created a quantile ensemble created from each model's quantile summary. For this, we took all submitted samples from each model and drew quantiles from that model's sample distribution, for each scenario, location, and outcome target. We then took an equally weighted median across multiple models' values at each quantile interval, for each target at each time step. To assess the difference between uncertainty across the two ensembles, we compared the average of values at each quantile across all time points and scenarios

Lastly, we conditioned sampled simulations on observed data by weighting samples by proximity to the observations in an early part of the period covered by the projection. We then created an ensemble projection for the period for which no observations were yet available. To measure performance, we took the mean absolute error (MAE) for each sample, 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 using the inverse MAE for each sample as a weight. To calculate weighted quantiles we used a Harrel Davis weighted estimator [22] in the cNORM R package (v3.0.2) [23]. As above, we took 23 quantiles including the median to express uncertainty. Differences between scenarios and models were therefore ignored and the ensemble only reflected samples' comparison to data.

Results *[also in html for readability: link]*

A total of six modelling teams submitted projections to the European COVID-19 Scenario Hub in Round 2. All teams contributed projections for all four scenarios. Two teams submitted projections for nearly all countries and targets. Here we focus on multi-model comparison and use only projection targets with three contributing models. These targets included 4 scenarios

for 52 weeks' case and death incidence for the Netherlands and Belgium, and 41 weeks' case incidence for Spain. Of three models for each of these targets, two sampled 100 simulations and one sampled 96 simulations. In total we consider 294816 data points from 5920 simulations, where each data point is the estimated weekly incidence at one time step in a simulated trajectory of an outcome in a target country and scenario over up to one year.

First we explore epidemic characteristics using sample trajectories. We summarised information about cumulative outbreak size over the projection period. For each target, we compared the cumulative number of projected outcomes to a threshold of the cumulative total over the one year before projections started (to July 2022). Across all 5920 simulations for all targets, 10% saw a cumulative total exceeding the relevant threshold. This varied widely, for example in Belgium where 25% and 2.5% of all trajectories would see a cumulative total exceeding the previous year's total number of cases and deaths respectively.

Sample trajectories also allowed us to explore projected peaks in incidence. We looked at peaks both over the entire projection period, and over only the autumn-winter period (October through March). In summarising peak characteristics, we considered both the timing and maximum weekly incidence of each peak, and the total number of peaks, representing distinct epidemic waves and the timing of their turning points. These epidemic characteristics could not be meaningfully estimated from the same results summarised into quantiles, as this sequence of summaries has no theoretical continuity through the time-series.

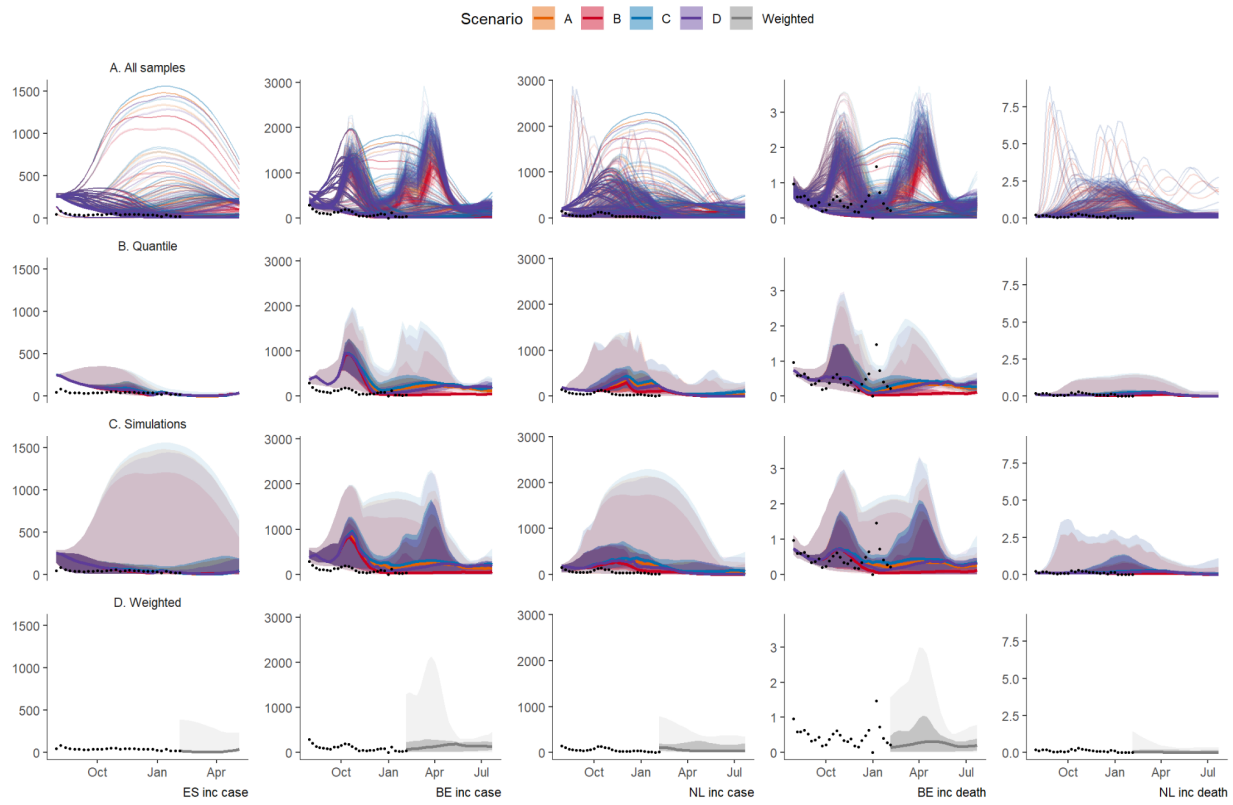

Figure 1. Projections of incidence per 100,000 population, showing median, 50%, and 99% probabilistic intervals, using: A) no ensemble method (100 samples per model); B) a median across each model's projections at a given quantile interval; C) quantile intervals of the distribution across all samples; D) a weighted median across samples with each sample weighted by its own performance against observed data.

Next we took a set of 23 quantiles from the distribution of samples provided by each model for each target. We created an ensemble using a median average at each quantile interval. We compared this ensemble (figure 1B), to taking the same set of quantiles directly from the entire set of samples provided by all models (figure 1C). To compare between the two ensembles, we took the average of values at each quantile across all time points and scenarios (figure 2).

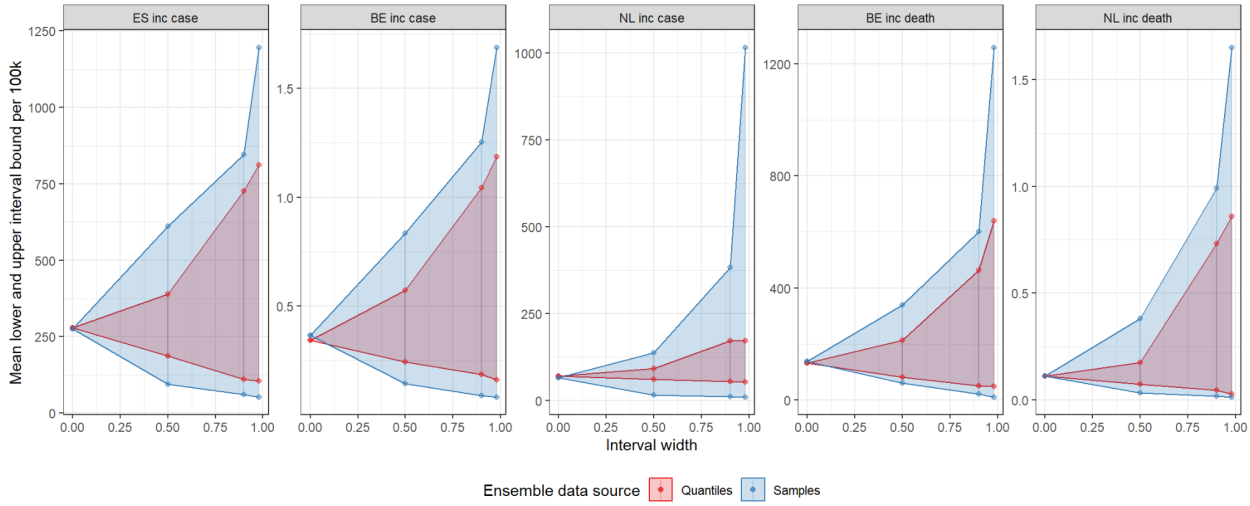

Figure 2. Mean central prediction intervals across time and scenarios. Mean lower and upper interval bounds: 52 week mean of quantiles in the sample and quantile ensembles

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 samples compared to when drawn from quantiles (respectively averaging 1016 and 173 weekly new cases per 100,000 population). Across all five targets, the pattern held that an ensemble based on samples 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.

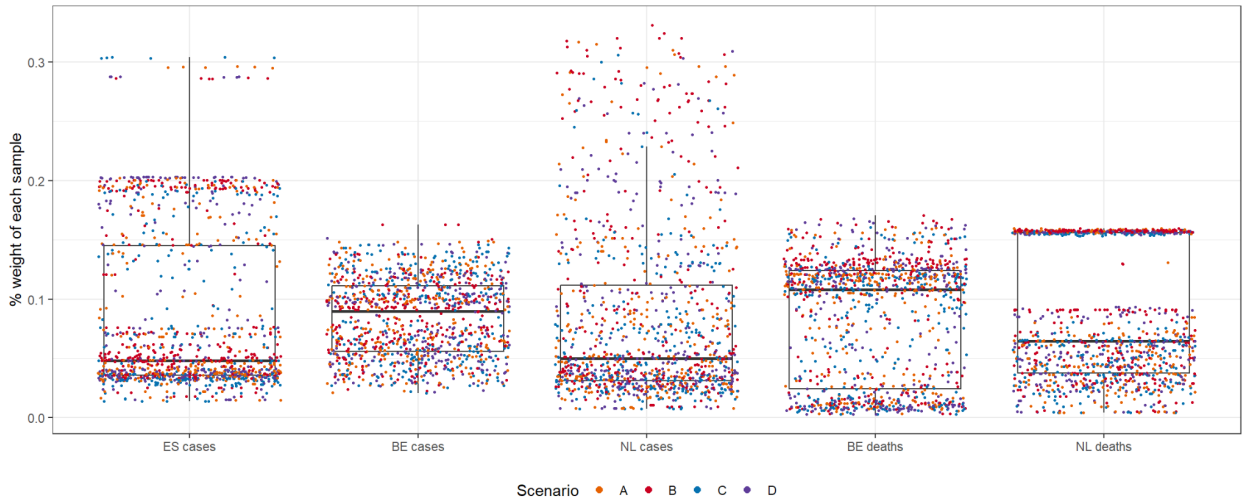

Figure 3. Weight of individual samples in an ensemble across all available samples for each of five projection targets. Samples were weighted by the inverse of mean absolute error against 28 weeks' observed data, meaning higher weight reflects better forecasting performance.

We then considered the forecasting performance of individual samples against 28 weeks' observed data (figure 3). The performance of each sample varied substantially between both models and targets. When weighted by inverse MAE, no sample received more than 0.33% weight (among n=1184 samples for each target). Weighting was heavily skewed across samples for some targets, including cases in Spain and Belgium with outlying clusters of highly weighted predictive samples from one model. Weights were more uniform across samples' forecast performance for deaths in the Netherlands and Belgium.

Creating an ensemble using the weighted performance of samples reduced the uncertainty of the future projection (figure 1D). In the Netherlands in the final week of projections for deaths, the most extreme single trajectory projected an incident 2.8 per 100,000 in the final week of projections. However when we weighted all trajectories by performance and took quantiles from this distribution, the upper limit of 99% probability suggested an incidence over 8 times lower, at 0.33 per 100,000.

Discussion

We compared two methods of collecting data about multiple models' projections of an epidemic. We compared a sample of up to 100 simulations from three scenario models for each of five projection targets, with the representation of those samples 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 samples' predictive performance in an ensemble. We found that collecting sampled simulations 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 [24], while 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 [25]. Third, a single set of quantiles can be held in csv files of 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 simulation, for a fixed-time instantaneous summary across simulations 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 work has created an ensemble from multiple models in terms of similarity to canonical curve shapes [26]. 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 [19].

We demonstrated a clear use case for collecting simulations by assessing their performance against observed data. By conditioning each simulation's weight in an ensemble on 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.

We highlight some significant limitations to our comparison of information loss between methods of collecting model output. Our method of collecting simulations was not specifically designed for the purpose of comparing them to equivalent quantile probabilities, and as a result our findings are difficult to interpret. We requested 100 simulations of equal probability to be selected at random from each model output, but this sampling strategy may have been insufficient and/or unrepresentative of the total set each modeller produced. We did not ask modelling teams how many simulations they had produced to characterise the posterior distribution and whether this was more than the submitted 100 samples. If modellers produced more than 100 simulations, then randomly selected samples could be unevenly spaced across the distribution of total simulations and therefore unrepresentative.

We calculated quantiles from the submitted subsample, rather than asking modellers to submit their own quantile probabilities from the full set of simulations they had access to. This means we cannot interpret our results as meaningfully representing a true difference between collecting sampled simulations and quantile probabilities. We could have addressed some of these issues

by asking modellers to submit all simulations, and/or estimating quantile probabilities based on all simulations.

We believe these 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 risk of crossing a specific epidemic threshold, 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 simulations rather than fixed-time quantile probabilities for reliable 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*, p. Forthcoming, Jun. 2022, doi: 10.1101/2022.06.16.22276024.

[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] 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.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] European COVID-19 Scenario Hub, 'Round 2'. https://covid19scenariohub.eu/report2.html
[21] European Covid-19 Scenario Hub, 'covid19-forecast-hub-europe/covid19-scenario-hub-europe at analysis'. https://github.com/covid19-forecast-hub-europe/covid19-scenario-hub-europe/tree/analysis
[22] 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.
[23] A. Lenhard, W. Lenhard, and S. Gary, 'cNORM - Generating Continuous Test Norms'. 2018. doi: 10.13140/RG.2.2.25821.26082.
[24] C. Genest, 'Vincentization Revisited', *Ann. Stat.*, vol. 20, no. 2, pp. 1137–1142, 1992.
[25] 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.
[26] A. Srivastava, S. Singh, and F. Lee, 'Shape-based Evaluation of Epidemic Forecasts'. arXiv, Nov. 11, 2022. doi: 10.48550/arXiv.2209.04035.