Check for updates

Fixed-time descriptive statistics underestimate extremes of epidemic curve ensembles

The uncertainty associated with epidemic forecasts is often simulated with ensembles of epidemic trajectories based on combinations of parameters. We show that the standard approach for summarizing such ensembles systematically suppresses critical epidemiological information.

Jonas L. Juul, Kaare Græsbøll, Lasse Engbo Christiansen and Sune Lehmann

Accurately communicating uncertainty is essential when forecasting. We are currently witnessing an unprecedented effort of scholars across countries and fields, competing to predict the trajectory of the novel coronavirus, COVID-19, using a plethora of approaches. These epidemic forecasts are used by governments aiming to mitigate the enormous consequences for the economy and health worldwide. Particularly crucial to decision-makers is information about the overall severity of the epidemic and, in particular, whether local hospitals will be overwhelmed. Political decisions to reopen countries and borders represent calculated risks. For that reason, it is of utmost importance that uncertainties and confidence intervals associated with forecasted epidemic trajectories are reliable and well communicated.

To forecast the trajectory of the novel virus, many researchers simulate the spread of the epidemic using mathematical models. Different classes of models are used for this purpose, including deterministic or stochastic compartmental epidemiological models[1,2] and individual-based models[3]. Given initial conditions of the epidemic prevalence in the population and a number of epidemiological and non-epidemiological parameters, these models can simulate hypothetical spreading scenarios. In reality, the initial conditions and parameters are not known exactly. For this reason, forecasts are often made by simulating a large number of plausible combinations of these inputs, and then summarizing the resulting ensemble of epidemic curves using descriptive statistics.

Regardless of the kind of model that produced the ensemble of epidemic trajectories, the ensemble is usually summarized using fixed-time descriptive statistics (see, for example, refs. [3–7]). For each time step, the instantaneous values of curves are ranked from smallest to largest and (possibly weighted) percentiles are computed. These percentiles are then used to produce confidence intervals for the forecast on the given time step — for example, the 50% least extreme values could be shown by marking the interval making out the 25th to 75th percentiles. Below, we show that in the context of forecasting trajectory extremes, however, such fixed-time approaches to producing confidence intervals suffer from a serious deficiency. This type of fixed-time statistics is biased against showing the projected peaks of the curves, and thus could obscure the part of the forecast most essential to decision-makers.

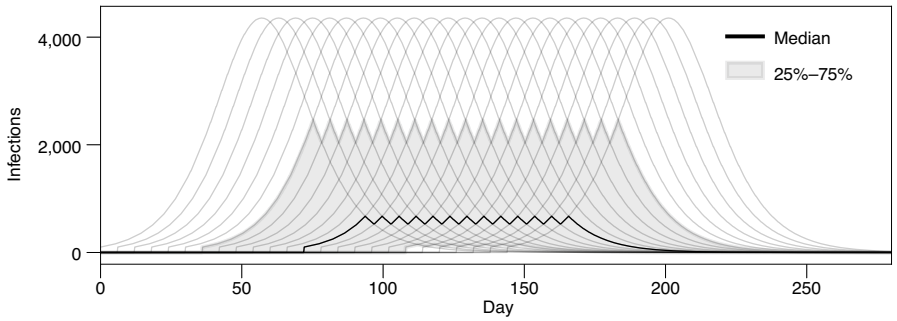

Fig. 1 | Pitfalls in using fixed-time descriptive statistics to summarize ensembles of epidemic curves. Simulations of the outbreak on the island Transmithaca (created using a deterministic compartmental model). Grey curves show individual simulations. Median and confidence intervals calculated using fixed-time statistics are defined in the legend. Simulations are identical except for the date on which the outbreak starts. The fixed-time descriptive statistics do not capture peak numbers of infections.

To illustrate the pitfalls of using fixed-time descriptive statistics to summarize ensembles of simulated epidemic trajectories, let us recount the fictional tale of the inhabitants of 'Transmithaca'.

On the island of Transmithaca, one million people lived in complete isolation from the rest of the world. A virus had ravaged the outside world, and, in the process, all viral parameters had become known with perfect precision. As Transmithaca slowly opened up for outside visitors, the inhabitants knew everything about the virus — except when it would arrive. The leaders of Transmithaca asked their epidemiologists to estimate how the disease would impact society. The epidemiologists simulated a number of scenarios, all with perfect choices of parameters, but different starting dates for the epidemic. Their simulations produced an ensemble of epidemic curves and, thinking that the individual simulated epidemic trajectories might clutter the picture, they presented the fixed-time summary statistics shown in grey and black in Fig. 1. Thus, the islanders prepared for an outbreak that might infect between 2,000 and 3,000 individuals at peak impact. As we can inspect, however, from the ensemble of time-displaced curves, the actual peak impact in every single case is more than 4,000 cases.

The tragic tale of Transmithaca is, of course, a caricature, but it illustrates a fundamental problem in the way ensembles of simulated epidemic trajectories are currently summarized. The future course of each curve is decided by the parameters and, crucially, the entire past of the curve. Models constrain trajectories to have certain forms, that is, they impose relationships among trajectory points at different times. The resulting long-term correlations between time points imply that basing summary

statistics on single points in time separately can be misleading. Therefore, the fixed-time descriptive statistics cannot be expected to capture what a real trajectory might look like. In the case of Transmithaca, the summary underestimates the infected count at peak impact, even though this quantity is identical across all curves. This is particularly unfortunate given that — if you are in charge of pandemic response — a reliable understanding of peak impact is essential.

This naturally leads us to the question: Do temporal relationships between trajectory forms pose a problem when summarizing real-world simulated epidemic trajectories? We believe that the answer to this question is: in many instances, yes. We have been involved in the task of forecasting the epidemic in Denmark, and our sampling of parameters produced an ensemble of highly time-displaced trajectories[8]. In Fig. 2a, we show part of an ensemble of epidemic trajectories produced as part of the Danish COVID-19 forecasting effort. This particular ensemble shows projected daily hospitalizations in the period 5 May to 1 October and was produced under a worst-case assumption that Danes would stop practising social distancing. For this ensemble, 67% of curve peaks lie above the 75th fixed-time percentile on the given day. For two other curve ensembles — which were produced with best-case and moderate assumptions on social distancing, respectively — 38% and 54% of peaks lie above the 75th percentile. It is clear that percentiles calculated using fixed-time statistics (shown using solid black lines) systematically underestimate peak values in these examples.

As an alternative to using fixed-time statistics, one could simply plot select curves from the ensemble as in Fig. 2a. However, to recapitulate the desired features of curve ensembles accurately, we

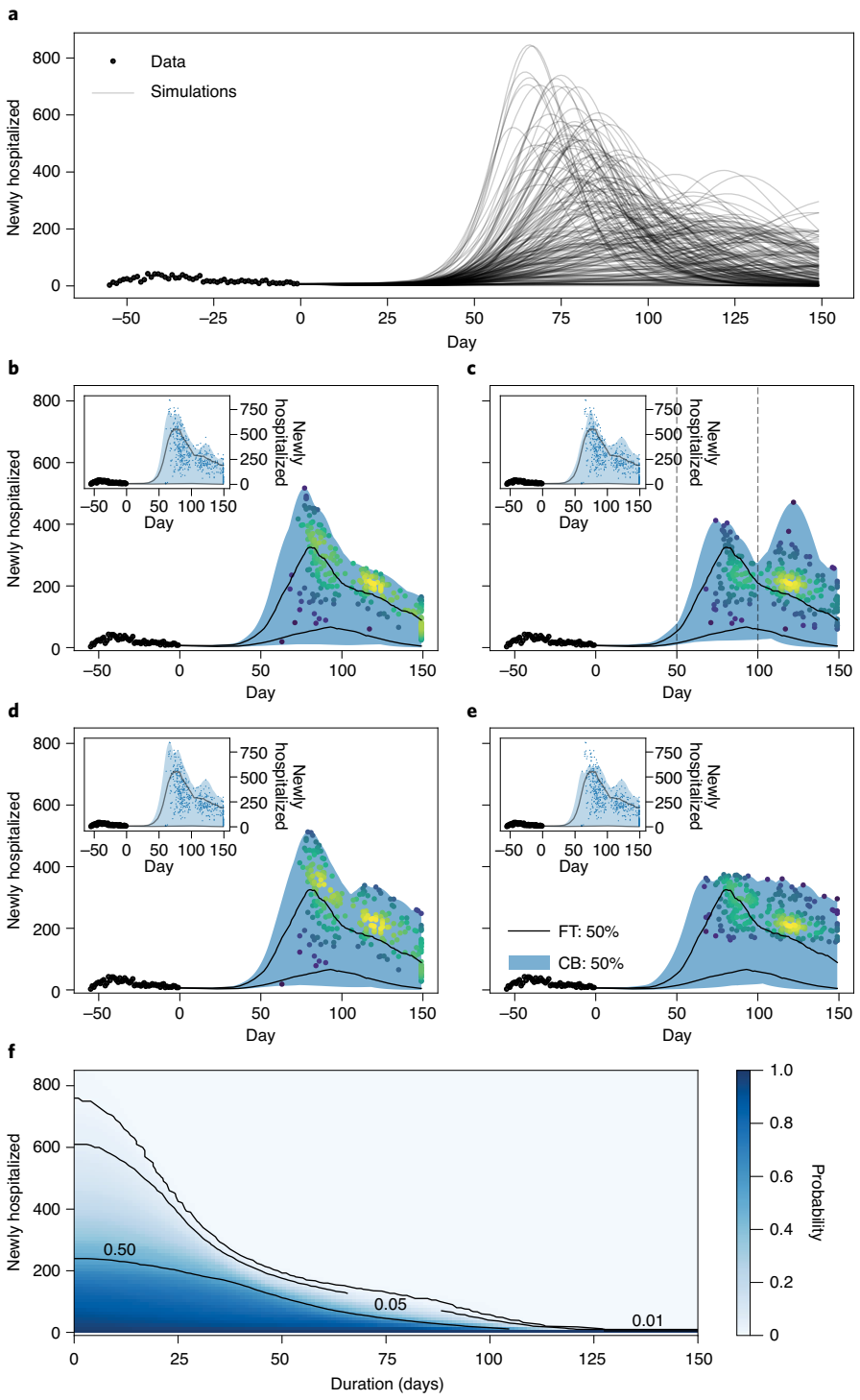

Fig. 2 | Curve-based descriptive statistics to summarize curve ensembles. a, Individual curves from an ensemble of 500 simulated epidemic trajectories produced as part of the Danish COVID-19 forecasting effort. **b**–**e**, Curve boxplot (CB) of the most central 50% (blue) curves plotted alongside the 25th–75th percentiles computed as fixed-time (FT) descriptive statistics (black lines). Dots coloured using a Gaussian kernel density estimator[15] show the position and density of peaks of the 50% most central trajectories, with yellow points indicating high density of peaks and violet indicating low density. Insets show the curve boxplot of the most central 90% (light blue) curves plotted alongside the 5th–95th percentiles computed as fixed-time descriptive statistics (grey lines) and a scatterplot showing all trajectory peaks in the ensemble. In **b**–**e**, the centrality of curves was ranked in different ways: all-or-nothing ranking for the full predicted time interval, $N_{curves} = 50$ and $N_{samples} = 100$ (**b**); all-or-nothing ranking using only the part of curves between day 50 and day 100, $N_{curves} = 10$ and $N_{samples} = 100$ (**c**); weighted ranking with a reward $f(t) = e^{-(t-t_0)\ln(2)/7}$ for all curves, t_0 being the current date; in this case, the reward for being contained in a sampled envelope gets half as big for every seven days (**d**); curves ranked according to their predicted peak number of newly hospitalized patients (**e**). The median prediction was deemed most central. **f**, Summarizing the likelihood of certain scenarios as predicted by the curve ensemble. The heatmap colour on the point (x, y) indicates the fraction of ensemble curves that at some point predict at least x consecutive days with at least y newly hospitalized patients on each day.

propose two alternative summary statistics: (1) curve-based descriptive statistics and (2) summarizing estimated likelihoods of specific scenarios of interest.

Curve-based descriptive statistics

Whereas fixed-time descriptive statistics separately evaluate the centrality of instantaneous curve values, curve-based descriptive statistics rank and visualize the centrality of entire curves. We suggest using curve boxplots to visualize trajectory confidence intervals. Curve boxplots are sometimes used for summarizing functional ensembles in simulation sciences[9,10] and the procedure for constructing a curve boxplot is straightforward:

(1) Rank curves from more central to less central.
(2) Plot the envelope containing the most central curves.

Here we define the envelope, $E(S)$, of a set of curves, S, as the area spanned by the curves in S. More precisely, a point in time (t', y) is contained by the envelope of S if there exist two curves $c_i(t), c_j(t) \in S$ such that $c_i(t') \leq y \leq c_j(t')$, where t and t' denote time values, i and j are curve indices, and y is a value on the vertical axis.

There are different ways one can rank the centrality of curves, each having its merits. With an all-or-nothing ranking method, all curves start with a centrality score of 0. N_{curves} curves are drawn uniformly at random from the ensemble and their envelope, E_{sample}, is constructed. We then check which ensemble curves are entirely contained by E_{sample}: curve c_i gets $s(c_i)$ added to its centrality score only if all points constituting the curve are contained in E_{sample}. The curve-dependent score $s(c_i)$ allows prior information to inform the ranking, for example, a curve's fit to existing data. Uniformly random samples like this are drawn N_{samples} times in total, and each time, the centrality scores of all curves are updated. In the end, curves with more centrality points are more central in the ensemble. In Fig. 2b,c, we show curve boxplots created using all-or-nothing rankings. We also show the fixed-time descriptive statistics for the ensemble.

An alternative to the all-or-nothing ranking is what we will call a weighted ranking: rewarding curves for each time step they are contained in E_{sample} (Fig. 2d). Again, one draws N_{curves} uniformly randomly from the ensemble. Now, however, we add a time- and curve-dependent value $s(c_i)f(t)$ to the centrality score of c_i if c_i is contained by E_{sample} at time t, where $f(t)$ is some function of time. The time dependency of the reward can reflect, for example, that some forecasts are expected to be very accurate in the near future but decrease in accuracy with time.

In addition to the ranking methods mentioned above (both adapted from ref. [9] with the additional curve-dependent score), one can rank the curves according to some feature of interest. In Fig. 2e, we show the curve boxplots obtained when we rank curves according to their projected maximum values of newly hospitalized cases in a single day; in other words, the median projected peak value received the highest centrality.

Likelihoods of scenarios of interest

The curve boxplots introduced above each visualize an area that contains a fraction of the ensemble curves. Sometimes, however, we may want to go beyond rough estimates of the temporal course of trajectories. It might be more interesting to quantify the risk of certain scenarios happening. As a more general approach to the issues associated with fixed-time statistics as a summary of ensembles, we note that the expectation and probability distribution of any scenario can be explicitly evaluated by counting how often they occur in the ensemble of curves. For example, consistent large numbers of hospitalized patients for long periods places a serious burden on healthcare systems[11,12]. If half of the simulated curves predict that hospitals will get at least 300 new patients every day for at least 20 consecutive days, and all curves are considered equally likely, the probability of this scenario is estimated to be 50%.

Figure 2f shows a heatmap communicating this type of risk. From Fig. 2f, we directly read off that given this model, the risk of receiving at least 200 new patients for at least 1 days in a row is higher than 50% and that there is less than 1% risk of receiving at least 400 new hospitalizations each day for at least 40 consecutive days. This plot is easily extended to ensembles with unequal curve weights.

In summary, when making forecasts of epidemic trajectories, it is important to represent the resulting curve ensembles in a way that captures the quantities of interest in an intuitive way. Here we have argued that computing confidence intervals using fixed-time descriptive statistics systematically suppresses trajectory extremes. This is natural. Fixed-time descriptive statistics are designed to show the least-extreme predictions on a given date, not to take entire curves into account. In a situation where the projected peak numbers of hospitalized patients are of the utmost importance to decision-makers and the public, however, this is unfortunate. We hope that this Comment raises awareness of this pitfall of fixed-time descriptive statistics for summarizing ensembles of epidemiological trajectories.

Here we have not discussed the ensemble generation process and have focused strictly on the uncertainty associated with a given ensemble. Depending on the assumptions behind each individual model, however, even more uncertainty may be present. For example, some models may take different approaches to uncertainty due to process noise and measurement error[13].

In addition to highlighting the shortcomings of fixed-time statistics (which have been identified before; see, for example, Appendix 2 in ref. [14]), we have suggested how curve ensemble extremes can be summarized and visualized instead.

Figure 2b–e shows that different curve rankings will result in different descriptive statistics. For the Transmithaca curves in Fig. 1, only the weighted ranking (with identical weights for all time steps) ranks curves with less extreme epidemic start dates as more central.

Further research is needed to clarify the advantages and disadvantages of the proposed methods. In the meantime, we encourage researchers to be creative and mindful about the problems of their chosen statistical method; and to communicate these openly to decision-makers.

Code availability

The code to reproduce all figures is available at www.github.com/jonassjuul/curvestat. A Python package, 'curvestat', to produce the curve-based descriptive statistics used in this Comment can be cloned from www.github.com/jonassjuul/curvestat. Ensembles in Fig. 2 were produced using a deterministic compartmental model with sampling of parameters based on literature and expert opinions, which is described in detail at https://files.ssi.dk/teknisk-gennemgang-af-modellerne-10062020. The R code used to produce the ensemble is available at https://github.com/laecdtu/C19DK. ❐

Jonas L. Juul[1,3]✉, Kaare Græsbøll[1], Lasse Engbo Christiansen[1] and Sune Lehmann[1,2]✉

[1]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark.* [2]*Copenhagen Center for Social Data Science, University of Copenhagen, Copenhagen, Denmark.* [3]*Present address: Center for Applied Mathematics, Cornell University, Ithaca, NY, USA.*

✉*e-mail: jjuul@cornell.edu; sljo@dtu.dk*

Published online: 8 December 2020
https://doi.org/10.1038/s41567-020-01121-y

References

1. Holmdahl, I. & Buckee, C. *N. Engl. J. Med.* **383**, 303–305 (2020).
2. Diekmann, O. & Heesterbeek, J. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation* Wiley Series in Mathematical and Computational Biology (Wiley, 2000).
3. Ferguson, N. M. et al. *Nature* **437**, 209–214 (2005).
4. Chinazzi, M. et al. *Science* **368**, 395–400 (2020).
5. Yang, W., Kandula, S. & Shaman, J. *Eight-week Model Projections of COVID-19 in New York City* (Columbia Univ., 2020); http://www.columbia.edu/~jls106/summary_ nyc. projection200329v1.pdf
6. Los Alamos National Laboratory COVID-19 team *COVID-19 Confirmed and Forecasted Case Data* (Los Alamos National Laboratory, accessed 28 June 2020); https://covid-19.bsvgateway.org/
7. Johns Hopkins University Infectious Disease Dynamics COVID-19 Working Group COVID scenario pipeline. *GitHub* https://github.com/HopkinsIDD/ COVIDScenarioPipeline (2020).
8. *Tillægsrapport af den 20. Maj 2020* (Statens Serum Institut, 2020).
9. Mirzargar, M., Whitaker, R. T. & Kirby, R. M. *IEEE Trans. Vis. Comput. Graph.* **20**, 2654–2663 (2014).
10. Sun, Y. & Genton, M. G. *J. Comput. Graph. Stat.* **20**, 316–334 (2011).
11. Srinivasan, S. S. et al. *Sci. Transl. Med.* **12**, eabb9401 (2020).
12. Yang, X. et al. *Lancet Respir. Med.* **8**, P475–481 (2020).
13. King, A. A., Domenech de Cellès, M., Magpantay, F. M. & Rohani, P. *Proc. R. Soc. B* **282**, 20150347 (2015).
14. Kiss, I. Z., Miller, J. C. & Simon, P. L. *Mathematics of Epidemics on Networks* Vol. 598 (Springer, 2017).
15. Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization* (John Wiley & Sons, 2015).

Acknowledgements

We thank the members of the SSI COVID-19 modelling group for an excellent collaboration and C. T. Bergstrom for comments on an early version of the manuscript. J.L.J. and S.L. received additional funding through the HOPE project (Carlsberg Foundation).

Author contributions

J.L.J. and S.L. conceived the idea. J.L.J. performed simulations, analysis and calculations. K.G. and L.E.C. devised and performed epidemiological simulations. All authors contributed to discussions and wrote the manuscript.

Competing interests

The authors declare no competing interests.