DRAFT: 12 April 2023

Methodological factors in forecast performance: the influence of model structure and target specificity on the performance of real time COVID-19 forecasts [provisional title]

Kath Sherratt*, others [TBC], and Sebastian Funk*

*London School of Hygiene and Tropical Medicine

Abstract

Performance of forecasts in capturing observed data varies in time and space with no overall "best" forecast. Two varying features of forecasting methods include the forecaster's approach to model structure; and whether the forecaster tunes their model to each target. We investigate forecasts of weekly incident deaths for 32 countries submitted to the European COVID-19 Forecast Hub between March 2021 and March 2023.

We use teams' provided metadata to categorise models by their structure (mechanistic, semi-mechanistic, statistical), and by their specificity (the number of locations each team targets; and whether the target location is the same as the team's institutional location, as a proxy for adaptation of model parameters to local conditions). We evaluate forecasts using the weighted interval score, scaled relative to a baseline of the median ensemble of all models.

Background

<table>
<tr><td>Aim</td><td>To evaluate the performance of standardised short-term forecasts of COVID-19 deaths in Europe, with respect to the long-term factors of model structure and target specificity.</td></tr>
<tr><td>Problem</td><td> Forecast performance is variable between models. Variable performance among forecasts is useful for both interpreting forecast outputs among different models, and learning how to improve forecast models and interpretation in the future.
However, it is unclear how to prioritise learning from forecast outputs in the short term (for both forecast creators and interpreters), without an understanding of whether and which features of the underlying forecast models may be associated with variable performance.
Two long term (persistent) features of forecasting methods include (in a very simplified way) the forecaster’s approach to:

Model structure – on a spectrum of deterministic, to probabilistic
Target specificity - fitting or adapting model parameters to the specific conditions of each target, on a spectrum from using only pre-specified data as published, to using only individual judgement

We investigate whether these factors make some contribution to variance in performance.
We focus only on deaths to allow for the aim of exploring target specificity as variation across countries, rather than countries-by-variable. Deaths is the more reliable forecast target between countries (re. data) and where the individual models make the most difference (vs. ensemble). </td></tr>
<tr><td>Existing analysis</td><td>
Model structure: some recent COVID-19 hub comparisons - Ray, Bracher; previous work
Target specificity: builds on expert judgement work
</td></tr>
</table>

Relevance of new work	● Need for this work
	    ○ Large scale of hub; using standardised forecasts - only possible to explore variation among teams and their models while keeping methods, targets, horizons, variables constant
	    ○ Target specificity - able to assess across many countries with comparable data
	● Relevance
	    ○ To individual modellers: areas of focus for methodological development
	    ○ To future large multi-model comparison projects: encouraging methodological diversity, use of methodological background in communicating results

Research questions	● Assess the range of performance by individual models over time ● Classify models' methodological structure and investigate whether this influences forecast performance ● Explore differences in performance by whether teams are forecasting for one or more countries (is a model that's "good/bad" overall good/bad everywhere?); including performance by whether teams are forecasting the country in which they are located

Methods

Setting	Euro hub
Unit of analysis	Quantile forecasts of weekly incident counts of COVID-19 deaths per country, for any combination of givens: location (32), target date (1-4 week horizon by week)

Sample	● Inclusion
	    o Submitted quantile forecasts over period 8 March 2021 - 10 March 2023
	    o Model metadata includes methods or linked citation

	● Exclusions
	  o ? Point-only forecasts without quantiles
	  o Anomalies: models designated “other”; forecasts made after major data anomalies

Procedure (codebase)	1. Classifying forecasts
	
	● Model structure
	  o Qualitatively define from teams’ own descriptions of methods section in metadata or citation
	  o Each model = one of {mechanistic, semi-mechanistic, statistical}. Possibly: spatial; ML; “other”
	● Target specificity
	  o Count total (ever) target locations per model
	  o Each model = single-target or multiple-target. Possibly: few-target i.e. <=3 targets
	  o Categorise team location: qualitatively categorise “home” location using location of team institution, either as given or if missing, the institution of the first contributor in metadata
	  o Each model-target pair = “home” / “foreign”
	2. Forecast evaluation
	● Scoring relative to baseline as unweighted median ensemble of all models
	● Take all included forecasts, add variables for model structure and location of team relative to target location
	● Use log scoring to help account for varying importance of error over time

Results

Setting	● Shape of epidemic across Europe over time ● Performance of baseline ensemble ● Overall trends in forecast performance by horizon / target

Sample size & characteristics	- Number of models - Model structure - Target specificity: multi-target models, single-target models; models targeting home/foreign location - Overlap between structure & specificity: e.g. perhaps all statistical models are single-target models, etc
Forecast evaluation	- Focus on relative WIS, scaled against baseline; include MAE for point forecasts - Analysis grouped by (at minimum) horizon, the category of model structure and target specificity. Possibly by: variable, time period (epidemic phase / variant) - Simple comparisons among scores by group (possibly, model: WIS ~ structure + location ; possibly with interaction)

Discussion

Summary of key results	
Methods: issues, mitigations	**Sampling** ● Model structure o Model structure as continuum of structured/unstructured, iterative (mathematical-statistical) o Methods classified into model structures as described in metadata, not verified against code o No independent secondary verification of model structure / location o No account of changing methods (leading to shift in model structure) over time ● Target specificity o Team institute location a poor proxy for individual model contributors' knowledge of epidemic in target location

<table>
<tr><td></td><td>
Location sample size – 34 teams but relatively few location/target pairs (7 Euro countries, not all teams forecasting within/outside team location)
Evaluation

Use of relative scoring to an ensemble, rather than alternative baseline of flat-line forecast as in previous work. (But could be inappropriate to use a statistical baseline to evaluate a mathematical model?)
No uncertainty in model scoring
Data inputs – evaluation against JHU data, but teams may use different data for model inputs, meaning evaluation against JHU data may be inappropriate (?). Data inputs likely affected by team location. Mitigation: teams knew evaluation would be against JHU data

Confounders

Modellers' ability to adapt parameters to changing local conditions interacts with model structure. But we only have very limited information on either factor of structure or parameters - so likely difficult to explore this in much depth.

</td></tr>
<tr><td>Conclusions / weight of evidence</td><td></td></tr>
<tr><td>Recommendations</td><td>Improvements / future work

Better classification of qualitative adaptation of model to targets
Interaction of structural (long term) factors with stochastic (short term/real time, external) events during epidemic: Identifying model responsiveness to data issues; or during epidemic phase (rising, steady, decline)

</td></tr>
</table>