SWIM Topic Meeting: Evaluating Epidemic Forecasts

Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University.
Im Neuenheimer Feld 205, 69120 Heidelberg.

Tuesday 21 April 2026, 12:00 - 17:00.

`http://swim-workshop.de/evaluating_epidemic_forecasts.html`

Programme

12:00 – 13:00 Lunch Common Room, 5th floor.

13:00 – 14:20 Session 1: Specificities & Challenges Seminar Room 4.414, 4th floor.

Johannes Bracher (Karlsruhe Institute of Technology) and Sebastian Funk (London School of Hygiene and Tropical Medicine): *Evaluating epidemic forecasts: current practices and open questions.*

Kaitlyn Johnson (London School of Hygiene and Tropical Medicine): *Evaluating forecasts for public health utility: experiences from operational forecasting.*

Cathal Mills (University of Oxford): *From metric to action: The decision value of infectious disease forecasts.*

14:20 – 15:20 Group photo / coffee break Common Room, 5th floor.

15:20 – 16:50 Session 2: Theory and Methods Seminar Room 4.414, 4th floor. Chair: Friederike Becker (Karlsruhe Institute of Technology).

Jonas Wallin (Lund University): *Local scale invariance and robustness of proper scoring rules.*

Johannes Resin (Goethe University Frankfurt): *Relative scoring rules favour baseline models and distort forecasters' incentives.*

Georgios Gavrilopoulos (ETH Zurich): *Sequential model confidence sets.*

16:50 – 17:10 Wrap-up

Practical Information

- Internet access via eduroam is available throughout the building.
- Lunch and coffee breaks are free of charge for registered participants and take place in the Common Room.
- We would like to take a group picture at the beginning of the afternoon break. This picture shall be published on the workshop website after the event. Joining the group photo is of course voluntary.

Organizers: Johannes Bracher (`johannes.bracher@kit.edu`), Sebastian Funk (`sebastian.funk@lshtm.ac.uk`), supported by Julian Heidecke (`julian.heidecke@iwr.uni-heidelberg.de`).

Sponsors: We are grateful for support by the German Research Foundation and the KIT Center MathSEE.

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

math.SEE

Abstracts

Session 1: Specifics & Challlenges.

Evaluating forecasts for public health utility: experiences from operational forecasting.
Kaitlyn Johnson (London School of Hygiene and Tropical Medicine)

Collaborative forecasting hubs have emerged as tools for coordinating, evaluating, and communicating short-term infectious disease forecasts, with the goal of supporting public health decision-making. In response to requests for local-level situational awareness in the United States, we developed the Flu MetroCast Hub, which solicits sub-state and state-level forecasts of the percent of ED visits due to influenza at 0 to 3 week horizons. We originally planned to evaluate these forecasts using proper scoring rules, comparing local to state-level performance. However, feedback from public health practitioners has revealed a consistent misalignment: technically robust metrics such as WIS relative to a baseline do not translate into an interpretable assessment of forecast reliability for decision-makers. In this talk, I share examples of this misalignment from the 2025-2026 season alongside direct feedback from public health audiences about what they want to be evaluated from forecasts. Based on the idea that the reliability of forecasts depends on the decision being made from the forecasts, I propose additionally evaluating forecasts around decision similarity: how close are the decisions a forecast generates to those that would have been made under a retrospectively defined ideal forecast? I close with open questions about how to operationalise this framing when decision landscapes vary across jurisdictions and are not known in advance.

From metric to action: The decision value of infectious disease forecasts.
Cathal Mills (University of Oxford)

Decisions for infectious disease outbreaks are difficult and consequential, required to be made in the face of considerable time and societal pressure and great uncertainty. Public health decisions can be supported by probabilistic forecasts – predictions of the future value of an epidemiological quantity, together with uncertainty. Forecasting is challenged by noisy, incomplete, and delayed data alongside non-linear and changing dynamics. Evaluation metrics for infectious disease forecasts often focus on the forecaster's perspective; improving calibration and sharpness of forecasts. Currently, there are no systematic evaluation protocols to explicitly measure a forecast's "value" – its ability to provide actionable insights for decision-makers. We here develop a systematic forecast evaluation framework for informing epidemic decision-making, focusing on three aspects: i) translating forecasts and popular evaluation metrics into public-health-relevant quantities; ii) defining evaluation metrics for decision-makers; iii) linking predictability of an epidemic to the value of forecasts for decision-making. By melding concepts from weather forecasting, information theory, and decision theory, our framework bridges conceptual gaps between forecasters and decision-makers. We illustrate the framework with an application to forecasts of weekly incident COVID-19 cases and find that ensemble models often provided the most value for decision-makers with varying levels of risk appetite. Focusing forecast evaluations on the decision-maker provides a new perspective for infectious disease modellers with the hope for improved public health decision-making in future outbreaks and epidemics.

Session 2: Theory & Methods.

Local scale invariance and robustness of proper scoring rules
Jonas Wallin (Lund University)

Jonas Wallin (Lund University): Local scale invariance and robustness of proper scoring rules (paper). Averages of proper scoring rules are often used to rank probabilistic forecasts. In many cases, the individual terms in these averages are based on observations and forecasts from different distributions. We show that some of the most popular proper scoring rules, such as the continuous ranked probability score (CRPS), give more importance to observations with large uncertainty, which can lead to unintuitive rankings. To describe this issue, we define the concept of local scale invariance for scoring rules. A new class of generalized proper kernel scoring rules is derived and as a member of this class we propose the scaled CRPS (SCRPS). This new proper scoring rule is locally scale invariant and, therefore, works in the case of varying uncertainty. Like the CRPS, it is computationally available for output from ensemble forecasts, and does not require the ability to evaluate densities of forecasts. We further define robustness of scoring rules, show why this also can be an important concept for average scores unless one is specifically interested in extremes, and derive new proper scoring rules that are robust against outliers. The theoretical findings are illustrated in three different applications from spatial statistics, stochastic volatility models and regression for count data.

Relative scoring rules favour baseline models and distort forecasters' incentives.
Johannes Resin (Goethe University Frankfurt)

Relative scores of the form $S(x, y)/S(b, y)$, where x is a prediction, b is a baseline forecast, and y is the observation, are widely used in applied forecast evaluation. With respect to the employed baseline model, (mean) relative scores below and above one then represent improved and decreased predictive performance, respectively. In this paper we provide a detailed analysis of two major drawbacks of such relative scores, considering both probabilistic and point forecast settings. Firstly, relative scores give an unfair advantage to the model serving as the baseline, which may make it hard to achieve a mean relative score below one even for highly skilled models. Secondly, they distort the incentives for forecasters and encourage hedging, meaning that forecasters aiming to minimize their evaluation scores will not report their forecasts truthfully. Geometric rather than arithmetic averaging largely avoids favouring the baseline model, but likewise leads to undesirable incentives. So-called collective relative scores are more suitable in both respects and are recommended for practical use, even though open questions remain on the exact incentives they create. We illustrate our findings in an application to daily log-returns of the German stock market index DAX.

Sequential model confidence sets.
Georgios Gavrilopoulos (ETH Zurich)

In most prediction and estimation situations, scientists consider various statistical models for the same problem, and naturally want to select amongst the best. Hansen et al. (2011) provide a powerful solution to this problem by the so-called model confidence set, a subset of the original set of available models that contains the best models with a given level of confidence. Importantly, model confidence sets respect the underlying selection uncertainty by being flexible in size. However, they presuppose a fixed sample size which stands in contrast to the fact that model selection and forecast evaluation are inherently sequential tasks where we successively collect new data and where the decision to continue or conclude a study may depend on the previous outcomes. In this article, we extend model confidence sets sequentially over time by relying on sequential testing methods. Recently, e-processes and confidence sequences have been introduced as new, safe methods for assessing statistical evidence. Sequential model confidence sets allow to continuously monitor the models' performances and come with time-uniform, nonasymptotic coverage guarantees.