SWIM Topic Meeting: Evaluating Epidemic Forecasts
Heidelberg, 21 April 2026

Housekeeping and acknowledgements

- **Wifi:** eduroam
- **Coffee break:** 14:20-15:20, Common Room (5th floor)
- **Pictures:** We will take a group picture + pictures of talks and intend to put them on the workshop website. Let us know in case you don't want to appear.
- **Dinner:** we reserved a table at *Uuuhmami* (19:00). Please note: each participant (or their institution) need to cover their own bill. **Who wants to join?**
- **Thanks to:**
  - IWR / Uni Heidelberg for hosting us, especially Julian Heidecke and Joacim Rocklöv.
  - Sponsors:

math.SEE

The SWIM Workshop Series

- **Annual workshops** (every December) cover all aspects of statistical and dynamic modelling of infectious diseases.

Karlsruhe 2024:

Heidelberg 2025:

- **Topic meetings** are focused events with one specific topic.

Upcoming events

10 June 2026, Freiburg: Topic Meeting Mathematical Modelling of Antimicrobial Resistance.

Gwen Knight (LSHTM)

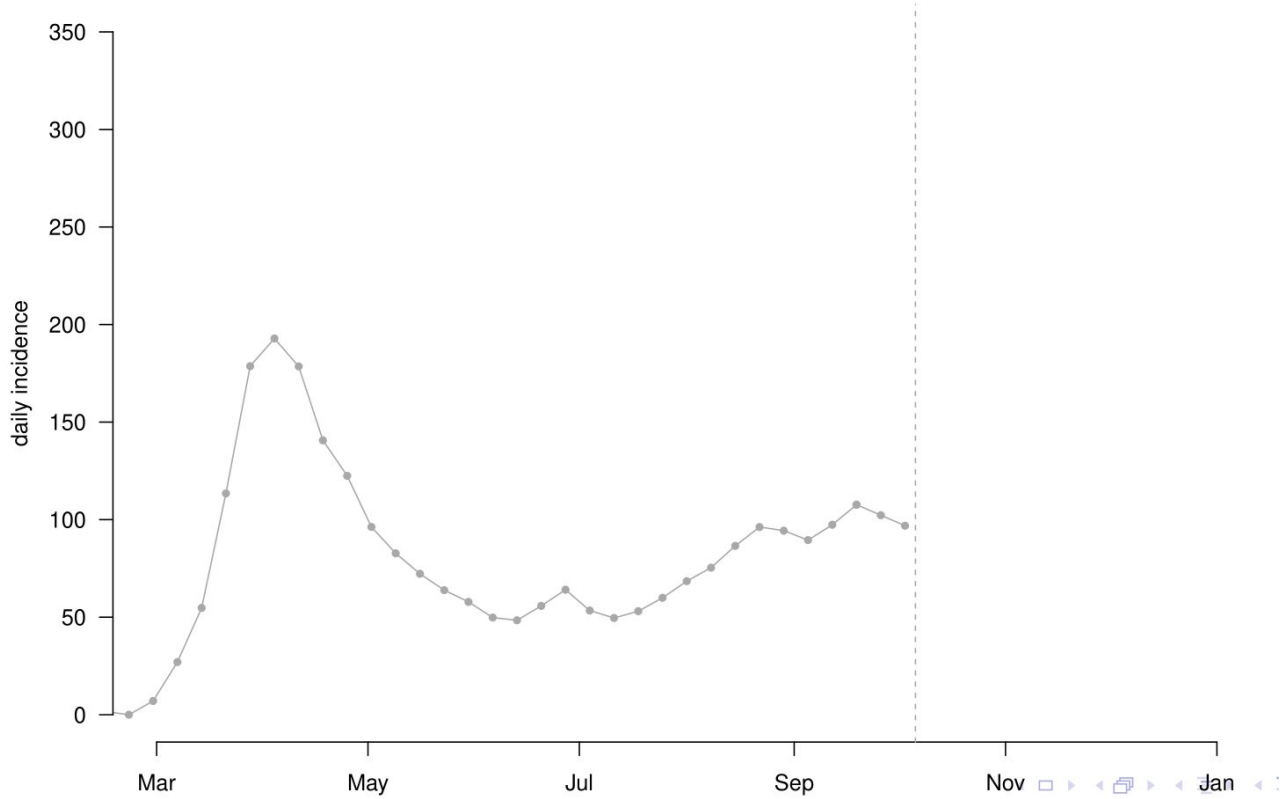
Laura Temime (CNAM)

9 December 2026, Karlsruhe: Annual Workshop.

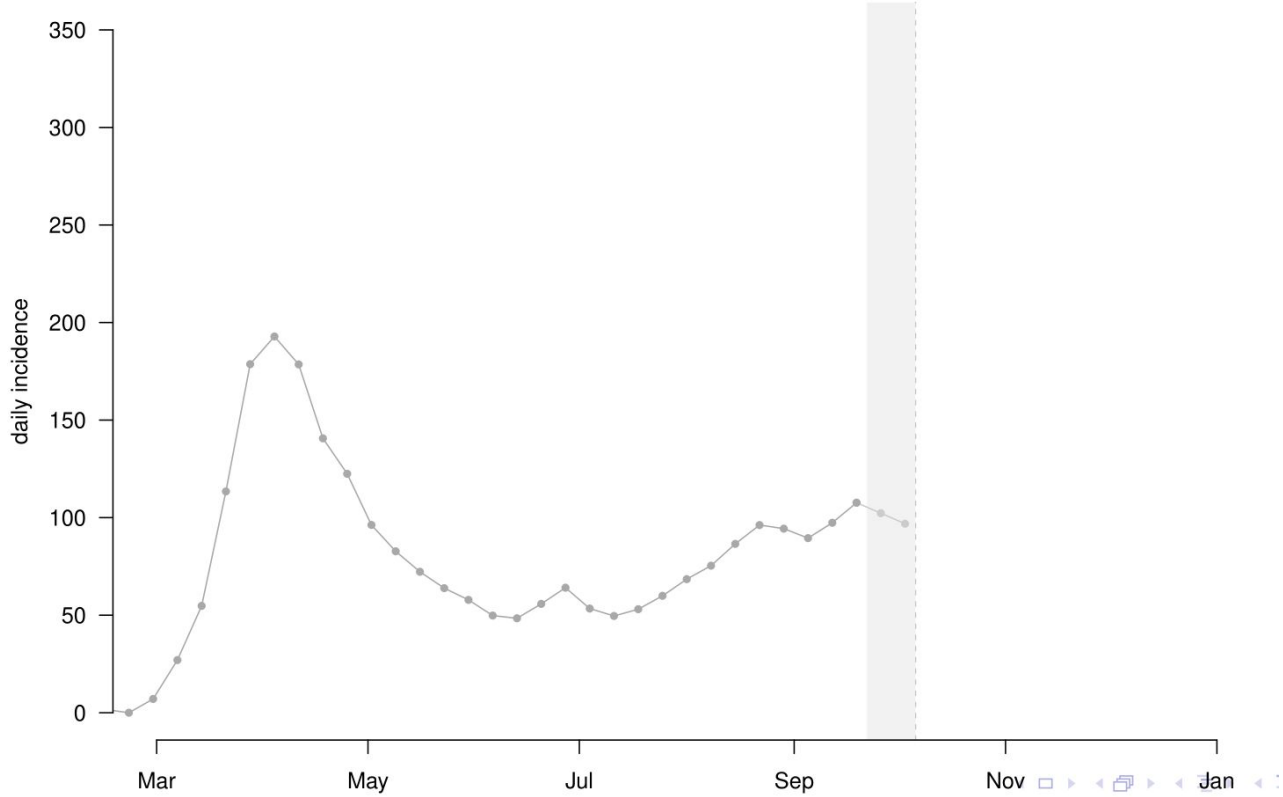
Anne Cori (Imperial)

Christian Althaus (Bern)

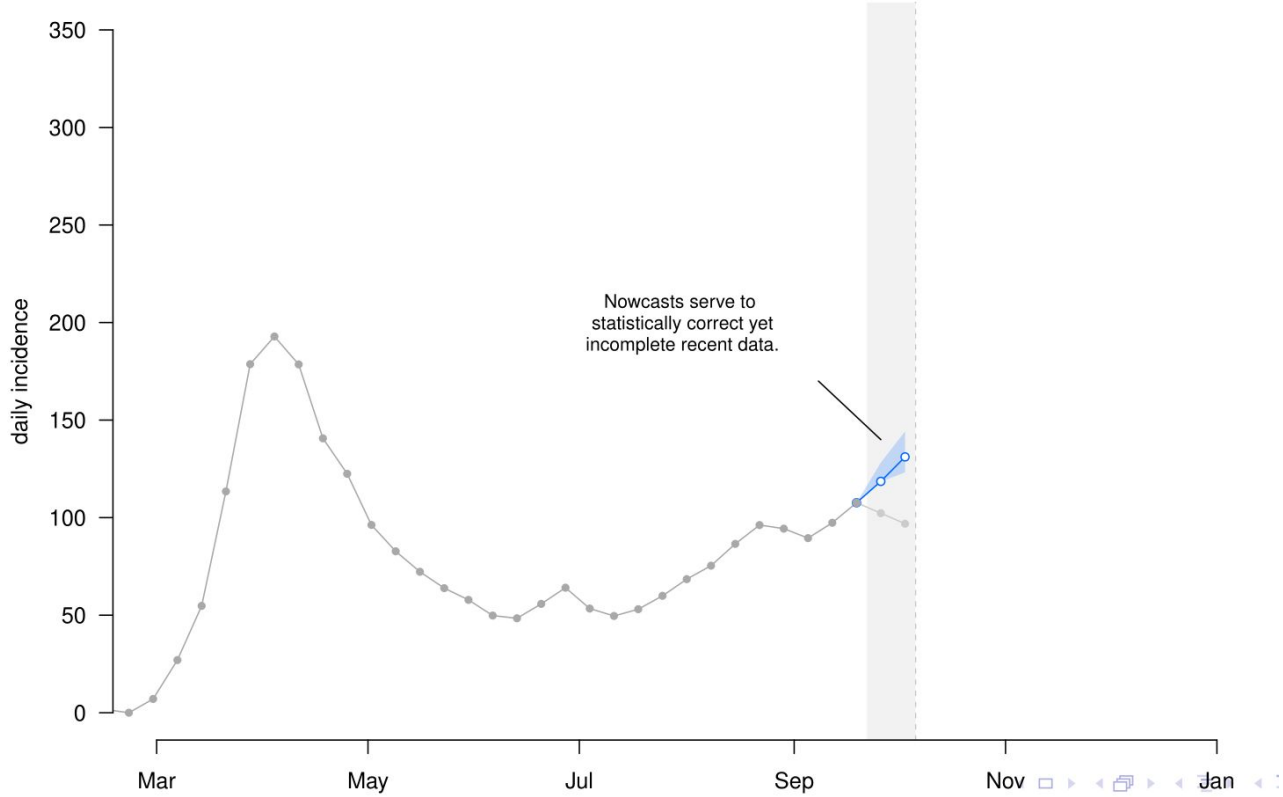
Predictive epidemic modelling



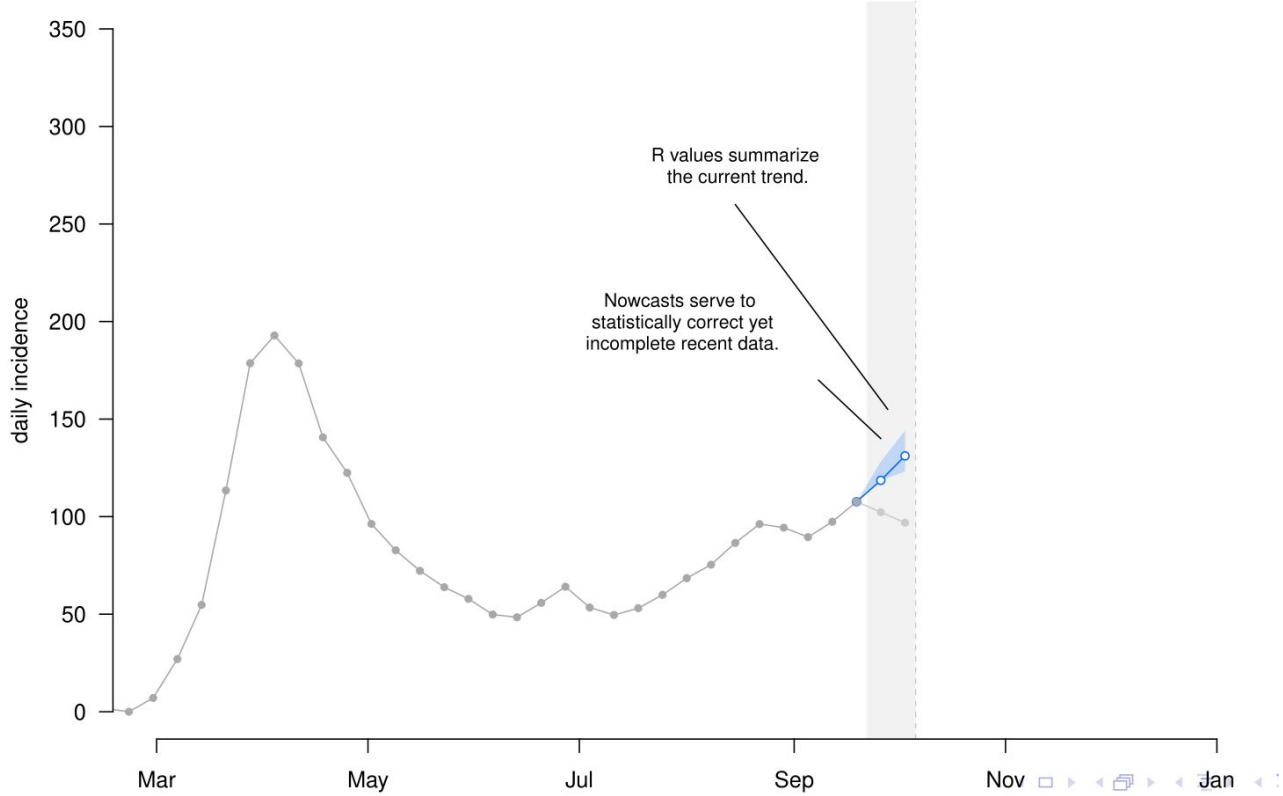
Predictive epidemic modelling



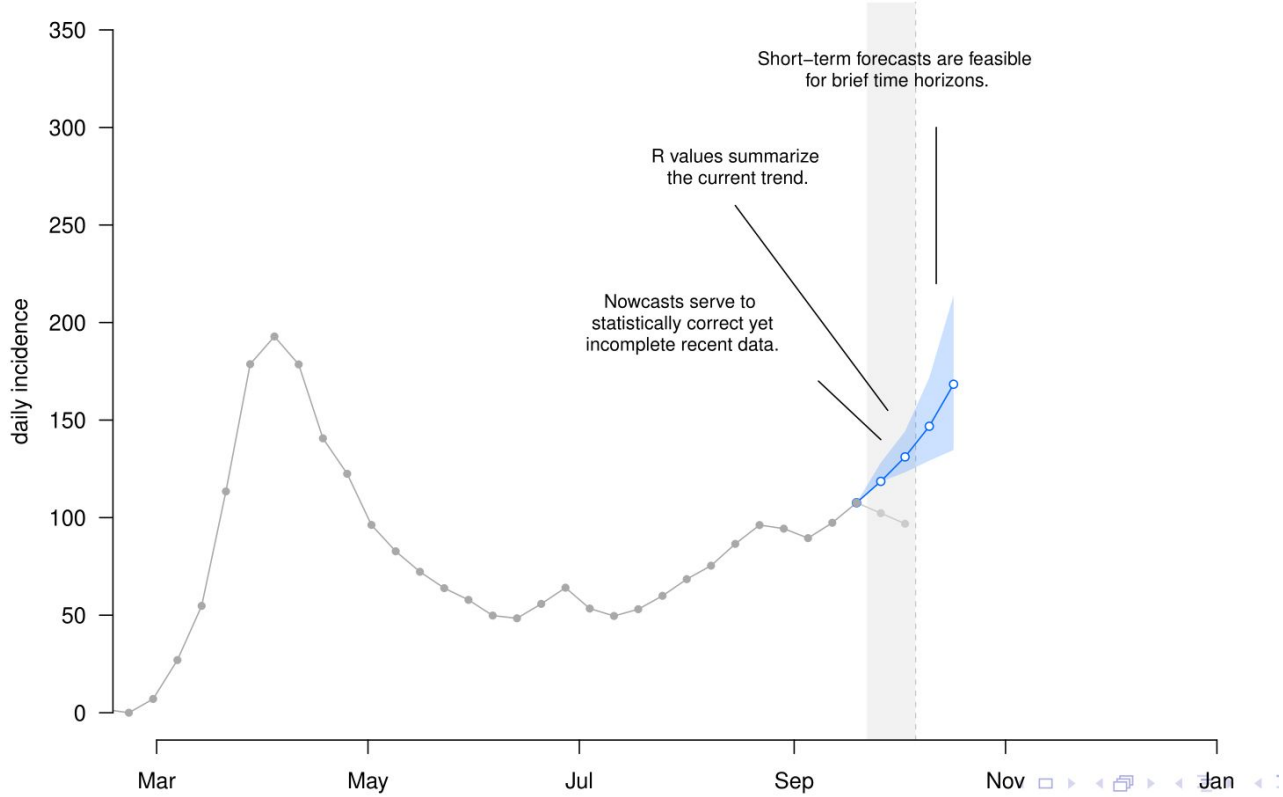
Predictive epidemic modelling



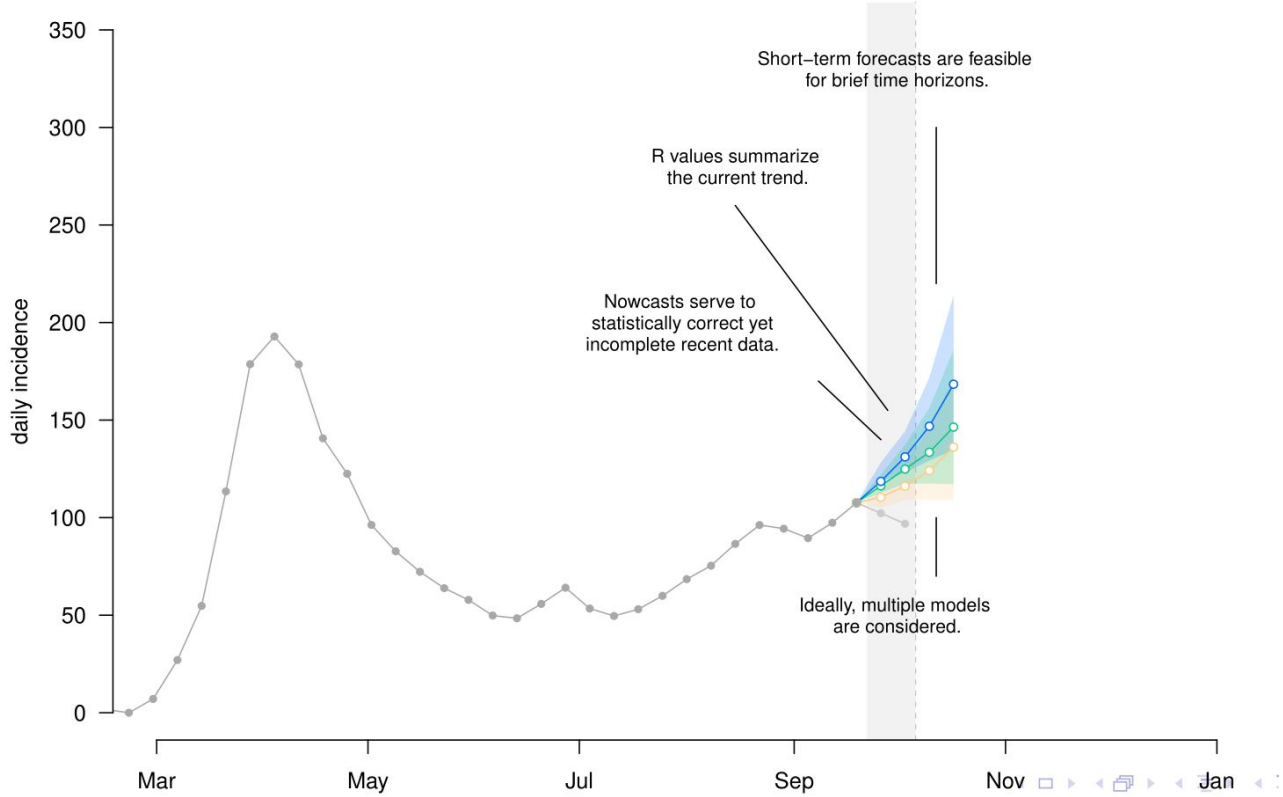
Predictive epidemic modelling



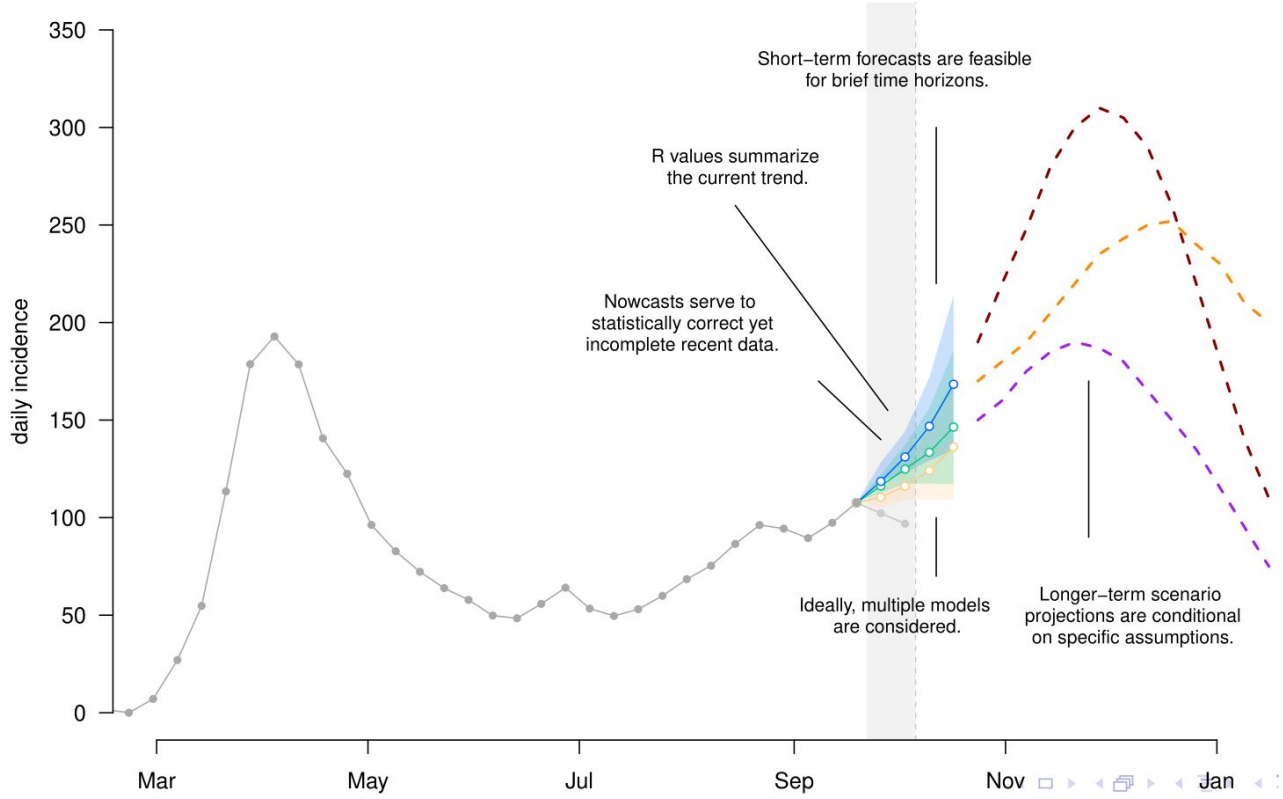
Predictive epidemic modelling



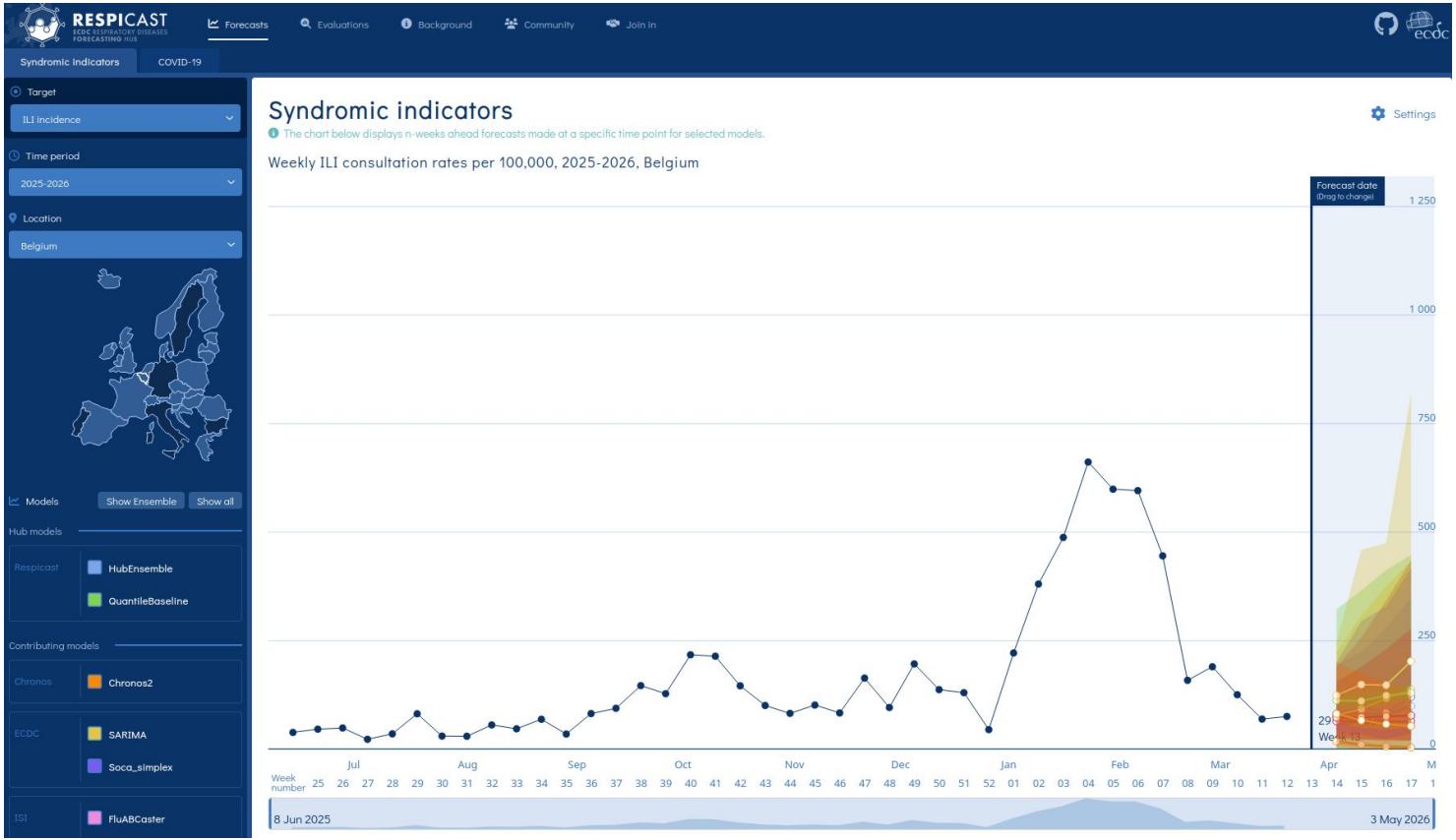
Predictive epidemic modelling



Predictive epidemic modelling



Forecast Hubs



Why evaluate forecasts?

- Can forecasts be trusted to be reliable?
- Are available forecast models better than simple heuristics?
- Which models worked best?
- Did forecasts improve decision making?

Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States

Estee Y. Cramer, Evan L. Ray, Velma K. Lopez, +291, and Nicholas G. Reich Authors Info & Affiliations

Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations

Katharine Sherratt, Hugo Gruson, Rok Grah, Helen Johnson, Rene Niehus, Bastian Prasse, Frank Sandmann, Jannik Deuschel, Daniel Wolffram ... Sebastian Funk see all »

How well did experts and laypeople forecast the size of the COVID-19 pandemic?

Gabriel Recchia, Alexandra L. J. Freeman, David Spiegelhalter

Forecasting for COVID-19 has failed

John P.A. Ioannidis [a], Sally Cripps [b], Martin A. Tanner [c]

Statistical basics

- Good **probabilistic forecasts** unite two properties:
  - **Calibration:** outcomes behave roughly as if they were coming from the predictive distributions.
  - **Sharpness:** forecasts are informative.

- **Calibration** can be assessed e.g., by checking prediction interval coverage.

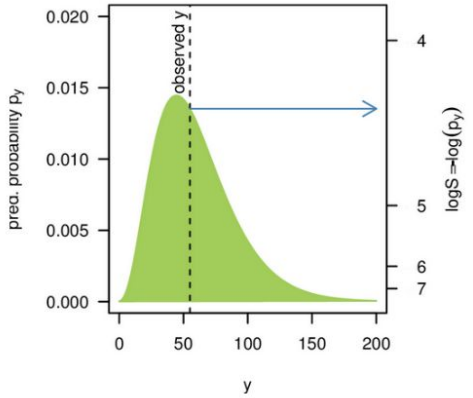
- **Proper scoring** rules assess both simultaneously and are constructed such that *honest forecasting* is encouraged.

Common proper scoring rules

- The **log score** is the predictive log-likelihood,

$$\log S(F, y) = f(y),$$

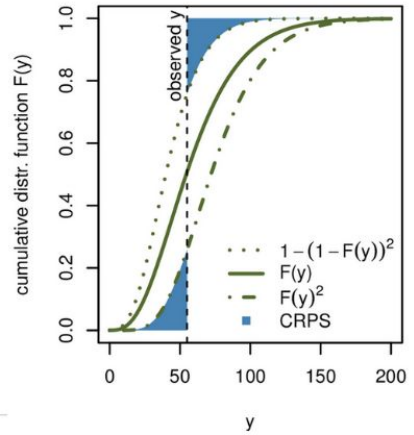
with f the predictive density or probability mass function.



- The **continuous ranked probability score**,

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} [F(x) - \mathbf{1}(x \geq y)]^2 \mathrm{d}x$$

is a probabilistic generalization of the absolute error. WIS is a quantile/interval-based approximation.



Statistical challenges*

- In many ways, ID forecasting resembles forecasting in weather, economics etc.
- Some particularities include:
  - **Strong variation of target variables:** average scores may be dominated by performance during high season.
  - Lack of clear reference standards (“climatologies”)
  - Limited number of forecast / observation pairs: comparisons often have low power.
  - Data revisions: it may be unclear which truth values to use for evaluation.
  - Opportunistic data collection: data may have gaps and reporting artifacts.

Scoring epidemiological forecasts on transformed scales

Nikos I. Bosse[1,2,3]*, Sam Abbott[1,2], Anne Cori[4], Edwin van Leeuwen[1,3,5], Johannes Bracher[6,7], Sebastian Funk[1,2,3]

Local scale invariance and robustness of proper scoring rules

David Bolin, Jonas Wallin

* This is the part where we plug our own stuff.

Statistical challenges*

- In many ways, ID forecasting resembles forecasting in weather, economics etc.
- Some particularities include:
  - Strong variation of target variables: average scores may be dominated by performance during high season.
  - **Lack of clear reference standards** (“climatologies”)
  - Low number of forecast / observation pairs: comparisons often have limited power.
  - Data revisions: it may be unclear which truth values to use for evaluation.
  - Opportunistic data collection: data may have gaps and reporting artifacts.

Mind the Baseline: The Hidden Impact of Reference Model Selection on Forecast Assessment

Manuel Stapper, Sebastian Funk

* This is the part where we plug our own stuff.

Statistical challenges*

- In many ways, ID forecasting resembles forecasting in weather, economics etc.
- Some particularities include:
  - Strong variation of target variables: average scores may be dominated by performance during high season.
  - Lack of clear reference standards ("climatologies")
  - **Low number of forecast / observation pairs:** comparisons often have limited power.
  - Data revisions: it may be unclear which truth values to use for evaluation.
  - Opportunistic data collection: data may have gaps and reporting artifacts.

Sequential model confidence sets

Sebastian Arnold, Georgios Gavrilopoulos, Benedikt Schulz, Johanna Ziegel

* This is the part where we plug our own stuff.

Statistical challenges*

- In many ways, ID forecasting resembles forecasting in weather, economics etc.
- Some particularities include:
  - Strong variation of target variables: average scores may be dominated by performance during high season.
  - Lack of clear reference standards ("climatologies")
  - Low number of forecast / observation pairs: comparisons often have limited power.
  - **Data revisions:** it may be unclear which truth values to use for evaluation.
  - Opportunistic data collection: data may have gaps and reporting artifacts.

Collaborative nowcasting of COVID-19 hospitalization incidences in Germany

Daniel Wolffram[1,2]*, Sam Abbott[3,4], Matthias an der Heiden[5], Sebastian Funk[3,4], Felix Günther[6], Davide Hailer[1], Stefan Heyder[7], Thomas Hotz[7], Jan van de Kassteele[8], Helmut Küchenhoff[9,10], Sören Müller-Hansen[11], Diellë Syliqi[9], Alexander Ullrich[5], Maximilian Weigert[9,10], Melanie Schienle[1,2], Johannes Bracher[1,2]

* This is the part where we plug our own stuff.

Statistical challenges*

- In many ways, ID forecasting resembles forecasting in weather, economics etc.
- Some particularities include:
  - Strong variation of target variables: average scores may be dominated by performance during high season.
  - Lack of clear reference standards ("climatologies")
  - Low number of forecast / observation pairs: comparisons often have limited power.
  - Data revisions: it may be unclear which truth values to use for evaluation.
  - **Opportunistic data collection:** data may have gaps and reporting artifacts.

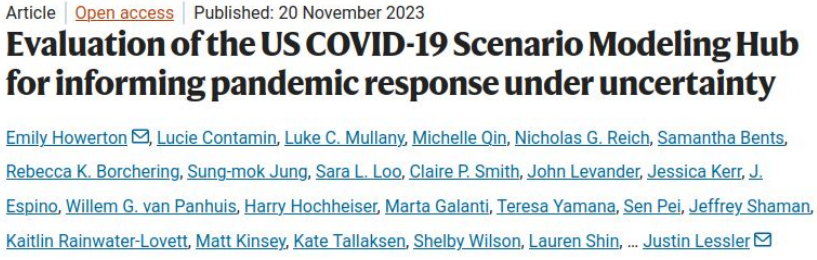
Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations

Katharine Sherratt[1]*, Hugo Gruson[1], Rok Grah[2], Helen Johnson[2], Rene Niehus[2], Bastian Prasse[2], Frank Sandmann[2], Jannik Deuschel[3], Daniel Wolffram[3],

* This is the part where we plug our own stuff.

Applied challenges

- Applied challenges in the evaluation of epidemic forecasts include:
  - **Counterfactuals:** forecasts are often meant to inform decisions which aim to change future trajectories.
  - Overall complexity: biology, behaviour, reporting etc, all need to be predicted correctly
  - Quantification of utility:
    - The costs of forecast errors are often asymmetric and hard to quantify exactly.
    - Public health experts often feel that statistical performance measures only partly reflect the utility of forecasts.



Article | Open access | Published: 20 November 2023

Evaluation of the US COVID-19 Scenario Modeling Hub for informing pandemic response under uncertainty

Emily Howerton, Lucie Contamin, Luke C. Mullany, Michelle Qin, Nicholas G. Reich, Samantha Bents, Rebecca K. Borchering, Sung-mok Jung, Sara L. Loo, Claire P. Smith, John Levander, Jessica Kerr, J. Espino, Willem G. van Panhuis, Harry Hochheiser, Marta Galanti, Teresa Yamana, Sen Pei, Jeffrey Shaman, Kaitlin Rainwater-Lovett, Matt Kinsey, Kate Tallaksen, Shelby Wilson, Lauren Shin, ... Justin Lessler

Applied challenges

- Applied challenges in the evaluation of epidemic forecasts include:

- Counterfactuals: forecasts are often meant to inform decisions which aim to change future trajectories.
- **Overall complexity:** biology, behaviour, reporting etc, all need to be predicted correctly
- Quantification of utility:
  - The costs of forecast errors are often asymmetric and hard to quantify exactly.
  - Public health experts often feel that statistical performance measures only partly reflect the utility of forecasts.

Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast FREE

Kelly R. Moran, Geoffrey Fairchild, Nicholas Generous, Kyle Hickmann, Dave Osthus, Reid Priedhorsky, James Hyman, Sara Y. Del Valle

Applied challenges

- Applied challenges in the evaluation of epidemic forecasts include:
  - Counterfactuals: forecasts are often meant to inform decisions which aim to change future trajectories.
  - Overall complexity: biology, behaviour, reporting etc, all need to be predicted correctly
  - **Quantification of utility:**
    - The costs of forecast errors are often asymmetric and hard to quantify exactly.
    - Public health experts often feel that statistical performance measures only partly reflect the utility of forecasts.

Evaluating infectious disease forecasts with allocation scoring rules FREE

Aaron Gerding ✉, Nicholas G Reich, Benjamin Rogers, Evan L Ray Author

From metric to action: The decision value of infectious disease forecasts

Cathal Mills, Nicholas J. Irons, Joseph L.-H. Tsui, Sarah Sparrow, Luiz M. Carvalho, Adam J. Kucharski, Oliver Ratmann, Ben Lambert, Christl A. Donnelly, Moritz U. G. Kraemer

Agenda

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