Evaluating forecasts for public health utility: experiences from operational forecasting

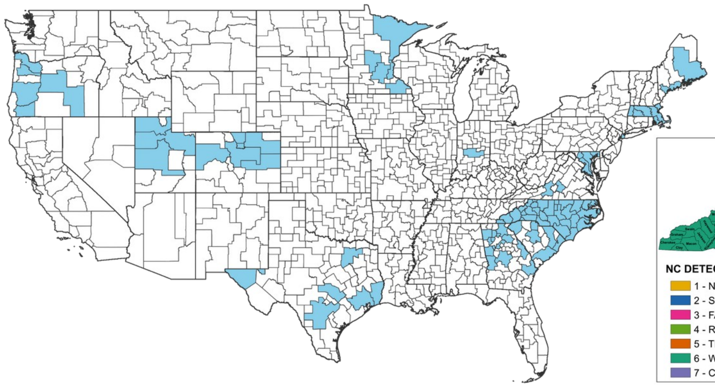
Kaitlyn Johnson
April 21st, 2026

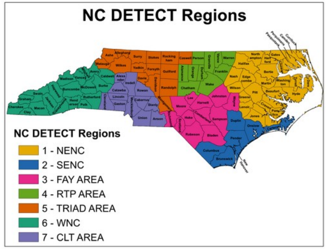
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Flu MetroCast Hub

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

- Goal: produce and evaluate local (sub-state) short-term forecasts (0-3 weeks ahead) of the percent of emergency department (ED) visits due to flu in the U.S.
    - Audience: public health practitioners who have asked for more local forecasts
- 2025-2026: first full season, increasing participation throughout season (3 → 13 models)
- Interested in: how models perform compared to state-level forecasts, which models perform best, and feedback from public health officials regarding their utility





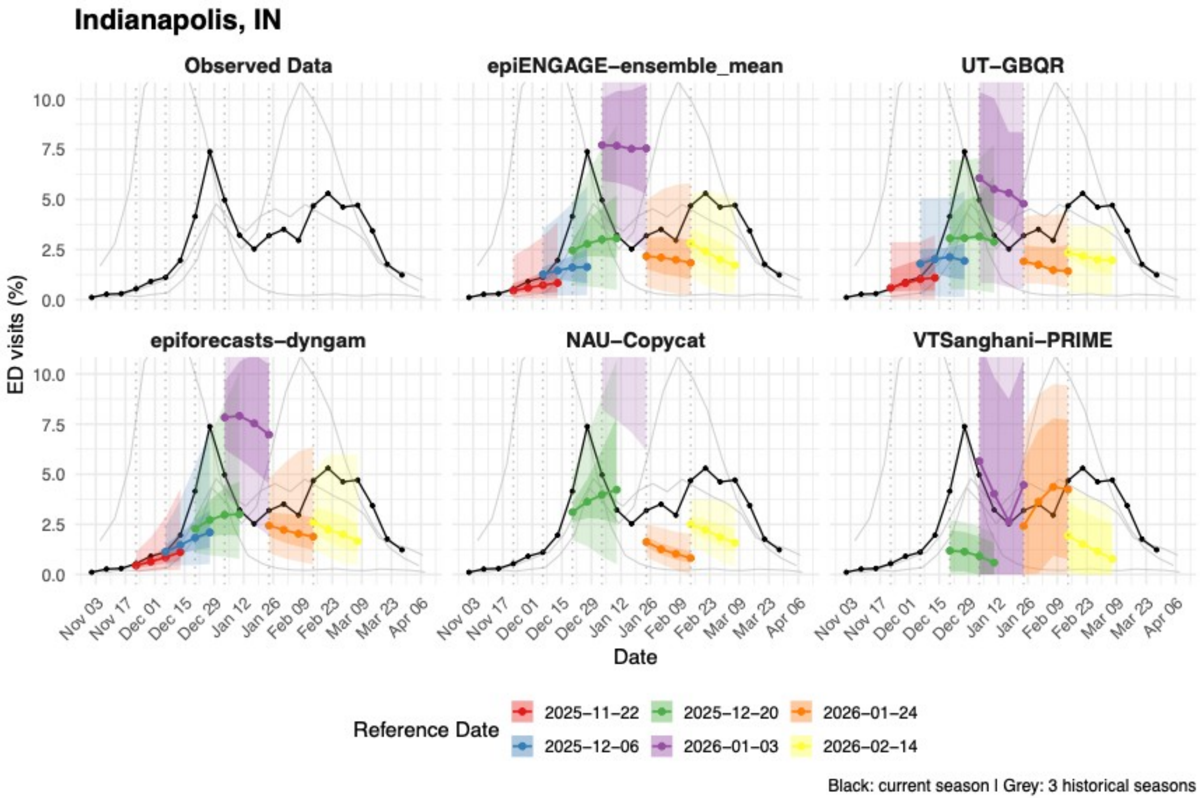
Preregistration: Evaluation of real-time performance of local-level flu forecasts for the 2025-2026 season in a subset of locations in the United States

Authors: Dongah Kim[1], Remy Pasco[1], Spencer J. Fox[2], Lauren Ancel Meyers[1,3], Nicholas G. Reich[4], Sam Abbott[5], Becky Wilson[1], Melissa Kerr[4], Katharine Sherratt[5], Sebastian Funk[5], Kaitlyn E. Johnson[5]

https://github.com/reichlab/flu-metrocast/

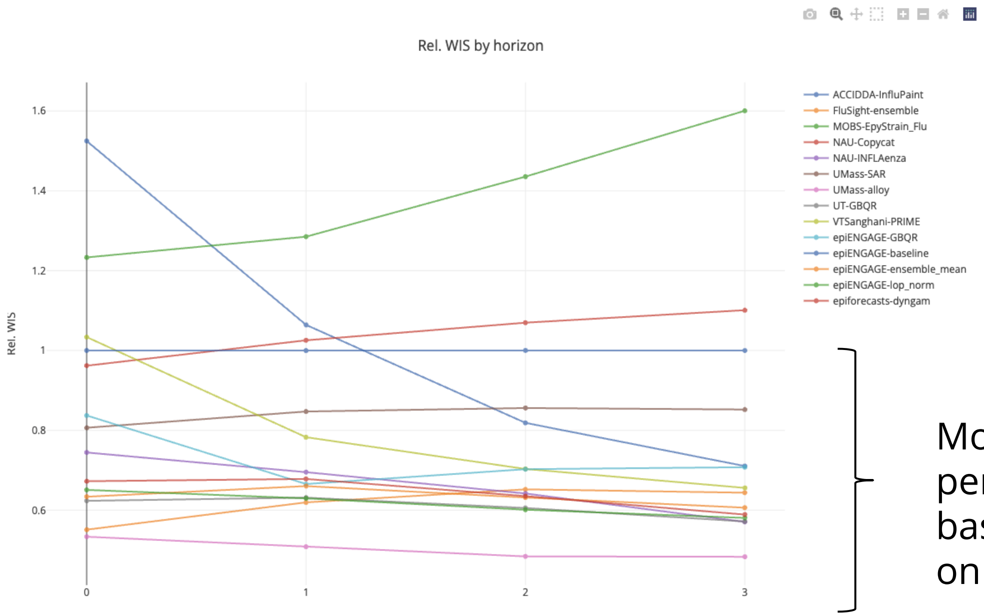
https://reichlab.io/metrocast-dashboard/

Shared consensus: models not performing well



- Models missed early and fast season onset
- Most missed peak
- Most models missed second peak (when present)

Scores do not fully capture poor performance



Most models performing better than baseline (flat baseline) on average

Feedback from public health: evaluation metrics (WIS/rWIS) aren't telling them enough about the features of a forecast that they care about

What do public health practitioners tell us they care about when evaluating forecasts?

What public health audiences care about:

- Season onset
- Peak timing
- Peak magnitude
- Trend (increasing, decreasing, stable)
- Probability of multiple peaks

What they don't find useful: WIS relative to baseline

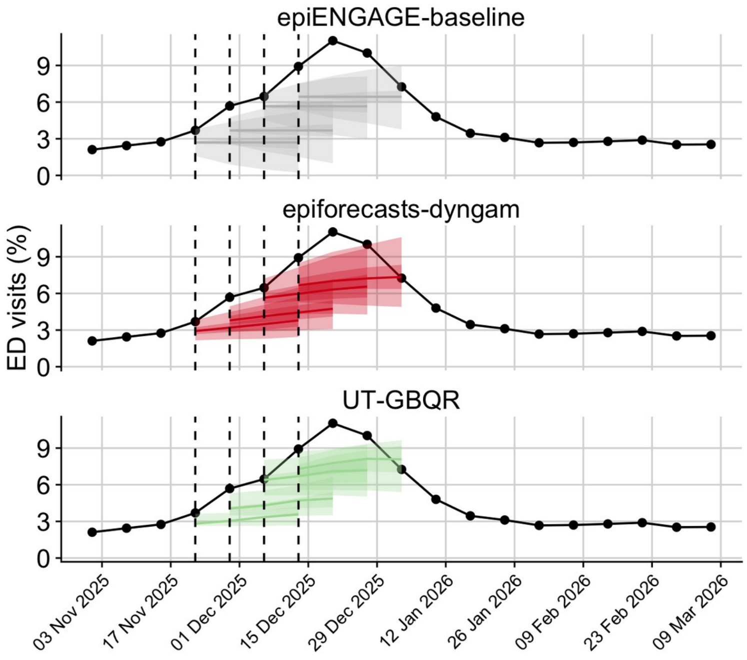
What forecasts are/could be used for:

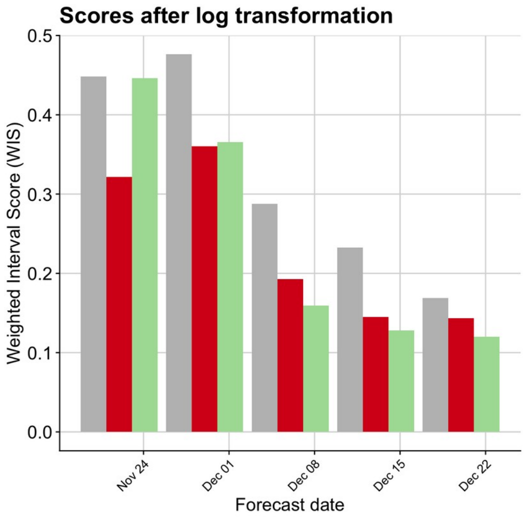
- Communicating to leadership/media/public
- Recommending vaccines/masks
- Allocating/requesting additional staff/PPE

Apparent misalignment between proper scoring metrics and what is interpretable or useful for public health

How did models perform? Missed early increase

Visual comparison of forecasts in New York City during beginning of surge

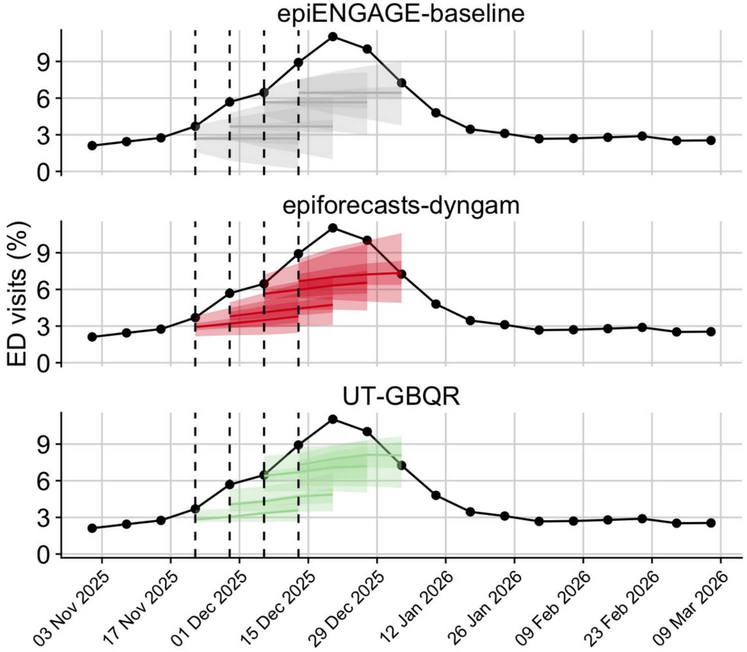


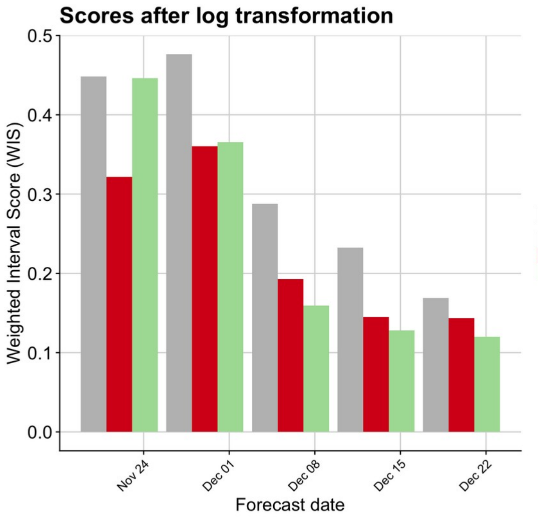


Feedback from public health officials:
These forecasts all performed poorly during the period of increase

How did models perform? Missed early increase

Visual comparison of forecasts in New York City during beginning of surge



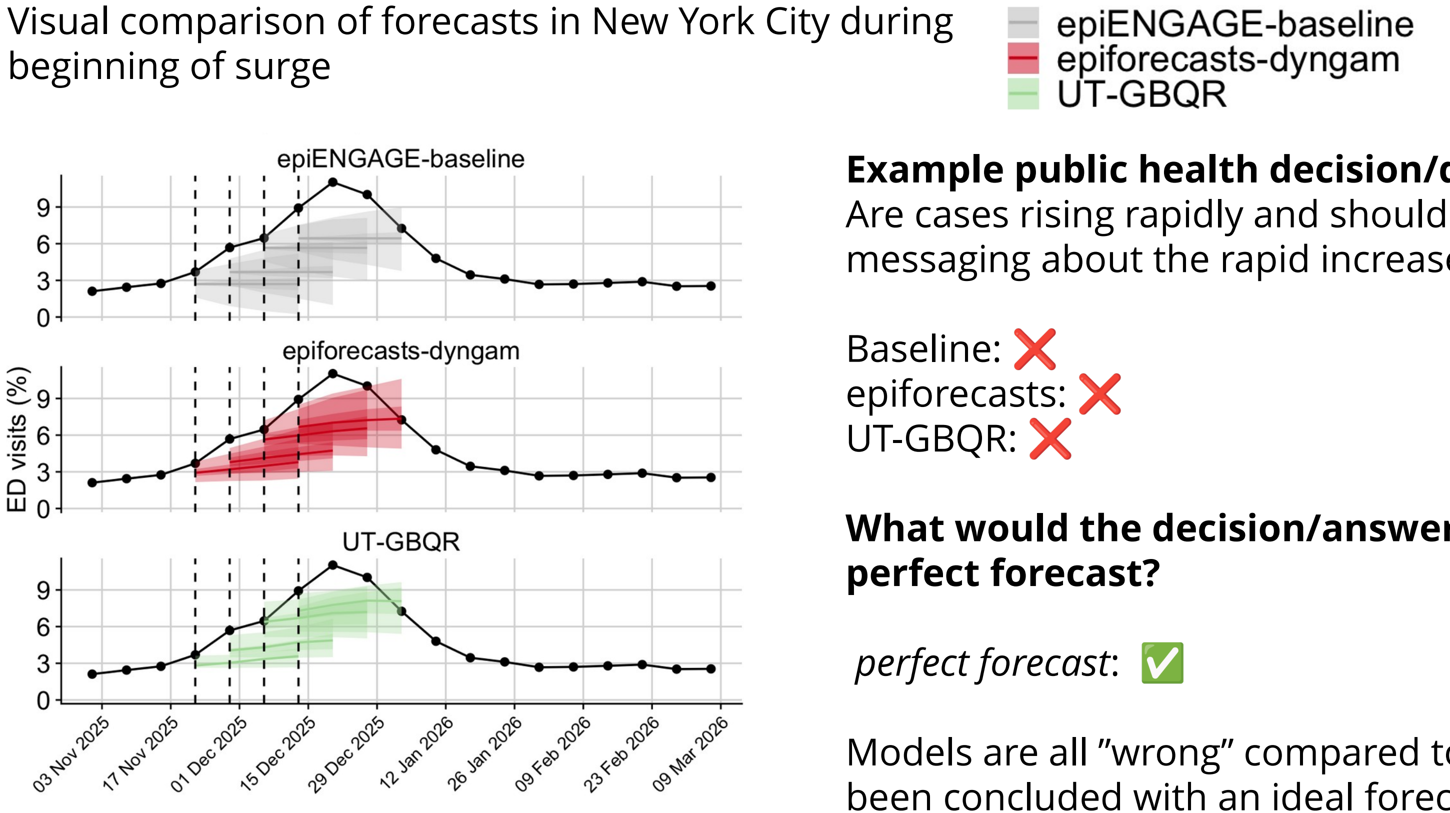


Feedback from public health officials:
These forecasts all performed poorly during the period of increase

My hypothesis from PH feedback: forecasts are being implicitly evaluated against a specific question/decision and compared to the decision that would have been made with complete information

How did models perform? Missed early increase

Visual comparison of forecasts in New York City during beginning of surge



Example public health decision/question:
Are cases rising rapidly and should we consider public messaging about the rapid increase in flu activity?

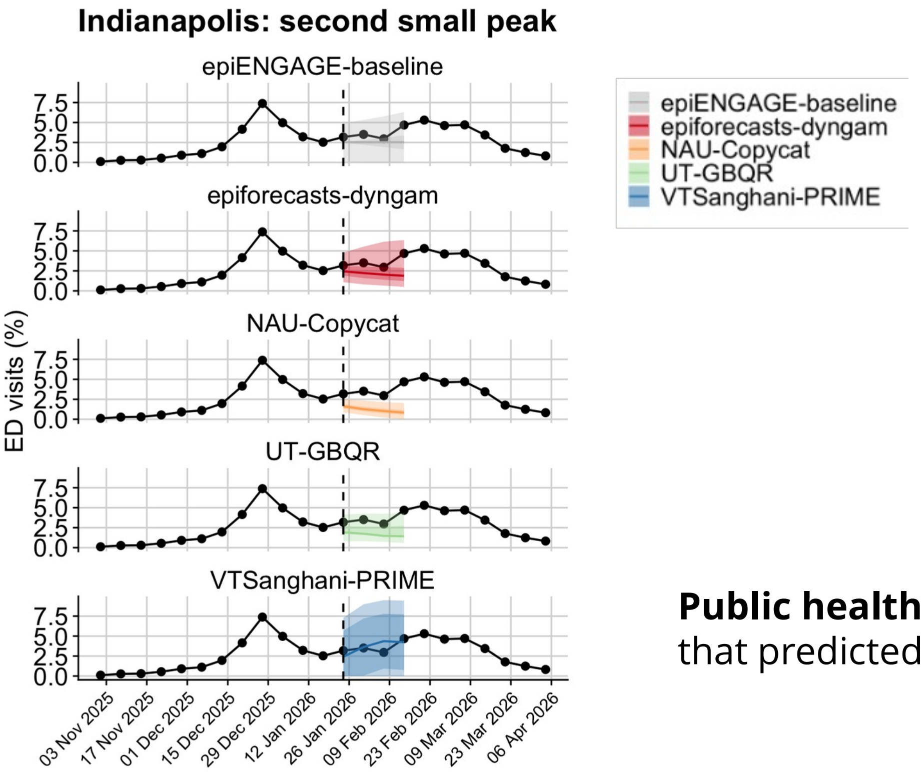
Baseline: ❌
epiforecasts: ❌
UT-GBQR: ❌

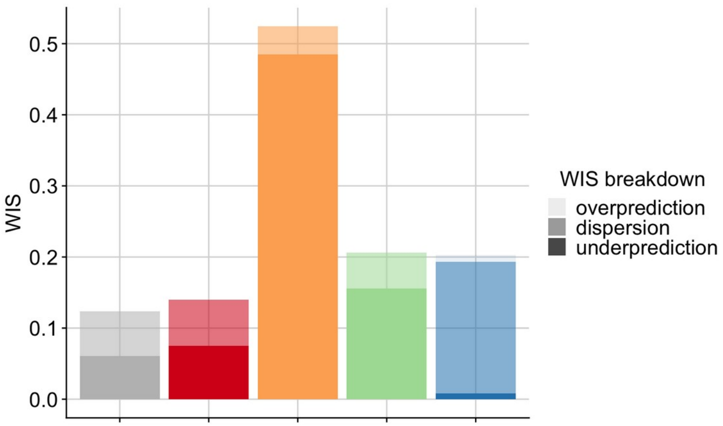
What would the decision/answer be from a perfect forecast?

perfect forecast: ✅

Models are all ”wrong” compared to what would have been concluded with an ideal forecast

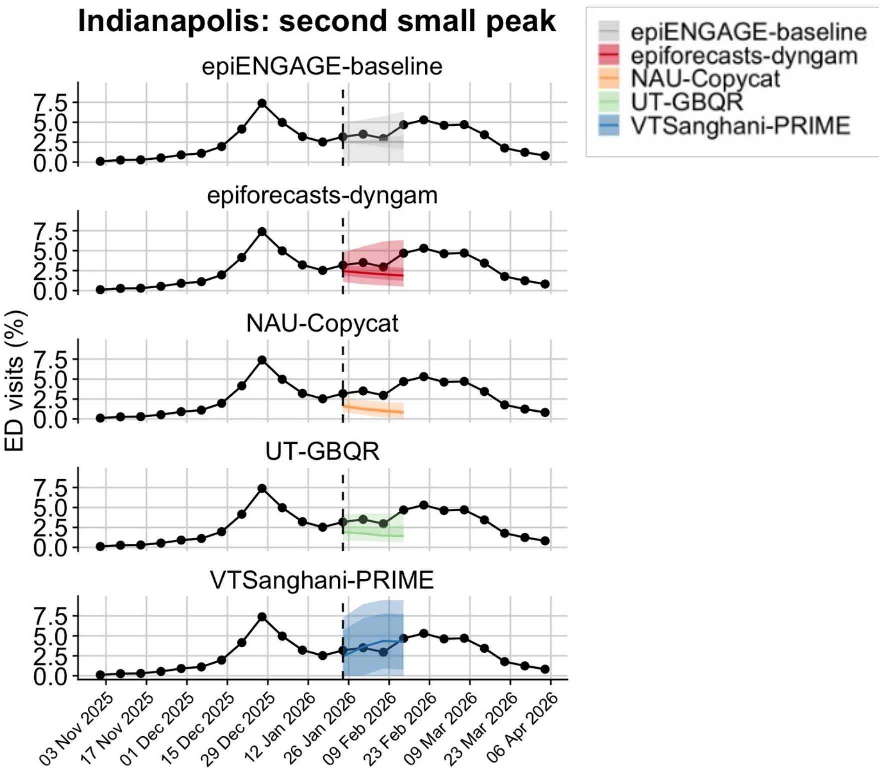
How well did models predict second smaller peak?





Public health perspective: VTSanghani-PRIME only model that predicted uptick in Feb/Mar, making it the best forecast

Only one model predicts late season uptick

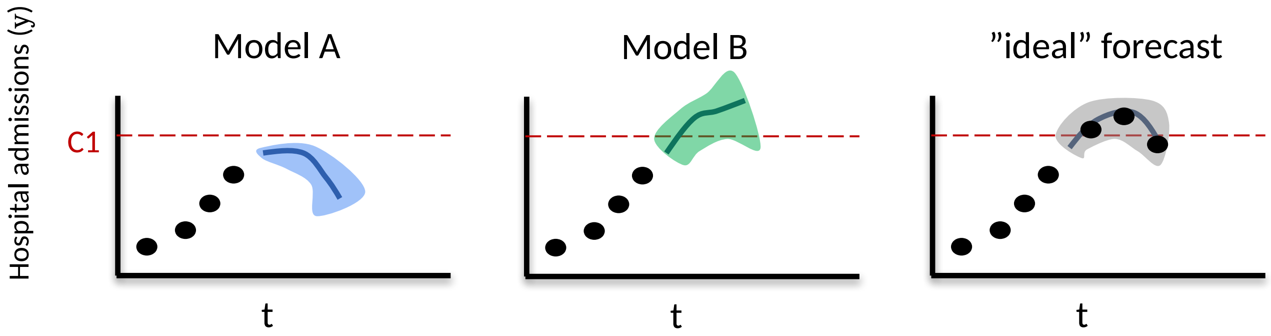


Example public health question/decision:
Will there be a second peak?

Baseline: ❌
Epiforecasts: ❌
NAU-Copycat: ❌
UT-GBQR: ❌
VTSanghani-PRIME: ✅

What would the decision/answer be from a perfect forecast?
perfect forecast: ✅

Decision/question: Is healthcare capacity (C1) likely to be exceeded?



$P(y > C1) = 0.05$

$P(y > C1) = 0.85$

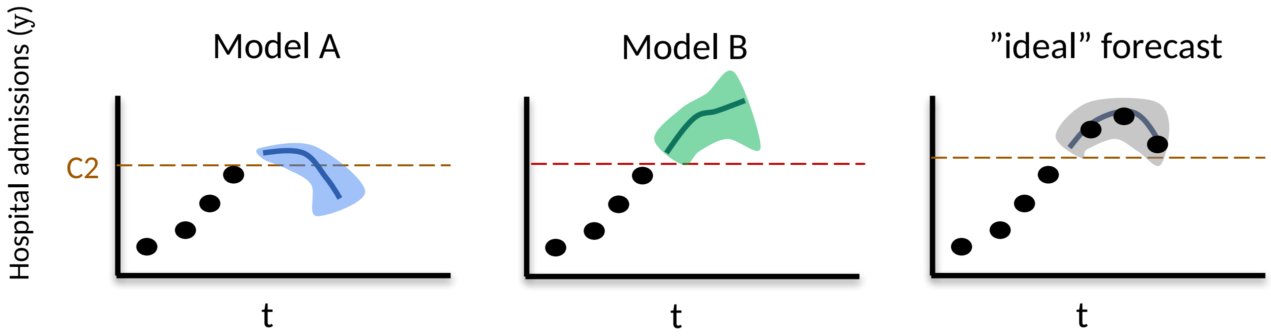
$P(y > C1) = 0.75$

Decisions

Forecasts

	C1		
Model A	❌		
Model B	✅		
“ideal” forecast	✅		

Decision/question: Is healthcare capacity (C2) likely to be exceeded?



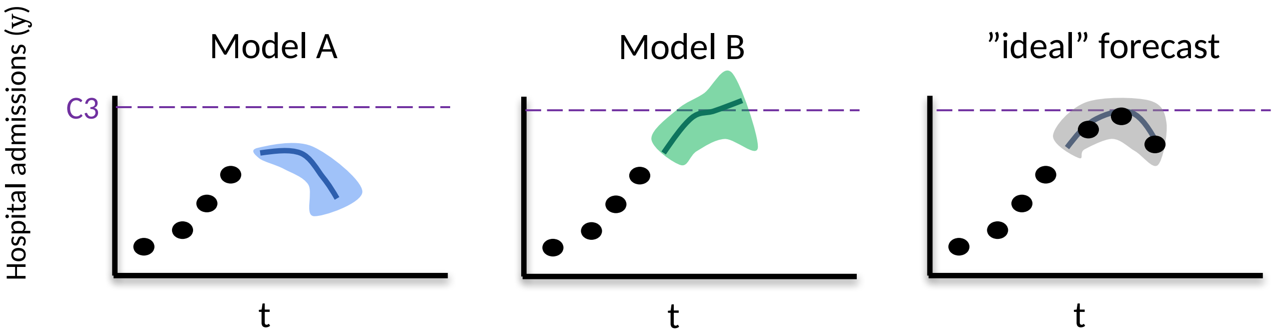
$P(y > C2) = 0.4$ $P(y > C1) = 0.97$ $P(y > C1) = 0.9$

Decision table

Forecasts

	C1	C2	
Model A	❌	✅	
Model B	✅	✅	
"ideal" forecast	✅	✅	

Decision/question: Is healthcare capacity (C3) likely to be exceeded?



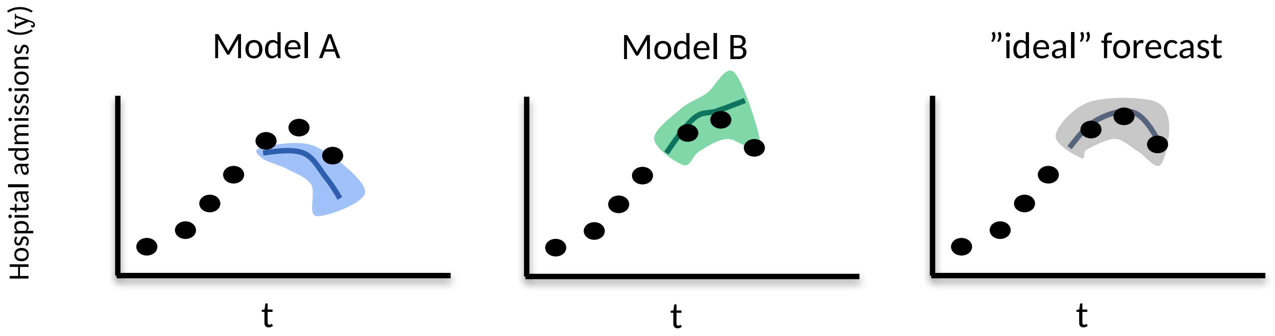
$P(y > C3) = 0$ | $P(y > C3) = 0.6$ | $P(y > C3) = 0.2$

Decision table

Forecasts

	C1	C2	C3
Model A	❌	✅	❌
Model B	✅	✅	✅
“ideal” forecast	✅	✅	❌

Takeaway: which forecast is most ”useful” depends on the decision/question



Decision table

Forecasts

	C1	C2	C3	Proportion similar to ideal
Model A	❌	✅	❌	2/3
Model B	✅	✅	✅	2/3
“ideal” forecast	✅	✅	❌	

Could also frame with a continuous decision/question e.g. number of additional staff needed and similarity to number arrived at with ideal forecast

What to do about it?

- Use weighted scoring rules (Allen. 2024): incorporate a weight function into conventional scoring rules, but do so in such a way that the resulting score remains proper
- Fix the baseline – a better baseline would make the relative scores more meaningful (Stapper & Funk 2025)
- Use allocation scoring rules (Gerding et al 2024)
- Develop a framework to evaluate for PH decisions (Mills et al 2026)
- Cater the proper scoring rules you choose to the question/decision of interest

Weighted scoringRules: Emphasizing Particular Outcomes When Evaluating Probabilistic Forecasts

Sam Allen
ETH Zürich

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

Manuel Stapper, Sebastian Funk
doi: https://doi.org/10.1101/2025.08.01.25332807

JOURNAL ARTICLE

Evaluating infectious disease forecasts with allocation scoring rules FREE

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

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
doi: https://doi.org/10.1101/2025.07.20.25331802

Open questions

- How would we as forecasters optimise our forecasts without knowing what decisions they are being used for/the context?
- If different locations operate under fundamentally different decision landscapes, how do we build evaluation frameworks that are both principled and actually useful to public health?
- If we *did* have a mapping from forecast distribution to decisions, how should we evaluate models against it?
- How to reconcile with the fact that a decision-based approach ”throws away” information on the full predictive distribution?

Acknowledgments

MetroCast Hub "Admin team"

Dongah Kim, Remy Pasco, Melissa Kerr, Lauren Meyers, Becky Wilson, Haerin Kim, Spencer Fox, Nick Reich

Thanks to epiforecasts team, Sebastian Funk

and many collaborators:

LSHTM Centre for Mathematical Modelling of Infectious Disease