Evaluating probabilistic COVID19 forecasts under partial missingness: A pairwise comparison approach

Johannes Bracher

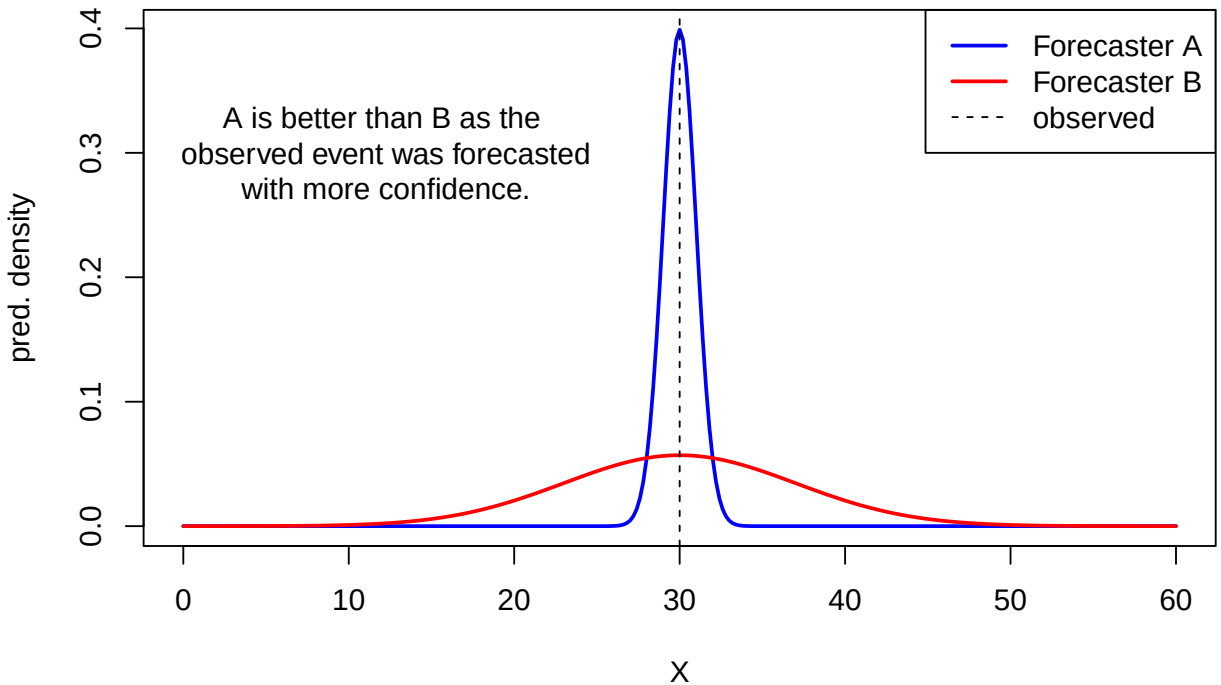
Karlsruhe Institute of Technology
Heidelberg Institute for Theoretical Studies

Thanks to J. Bien, E. Cramer, T. Gneiting, E. Ray, N. Reich, R. Tibshirani for helpful discussions

October 27, 2020

Why take into account uncertainty?

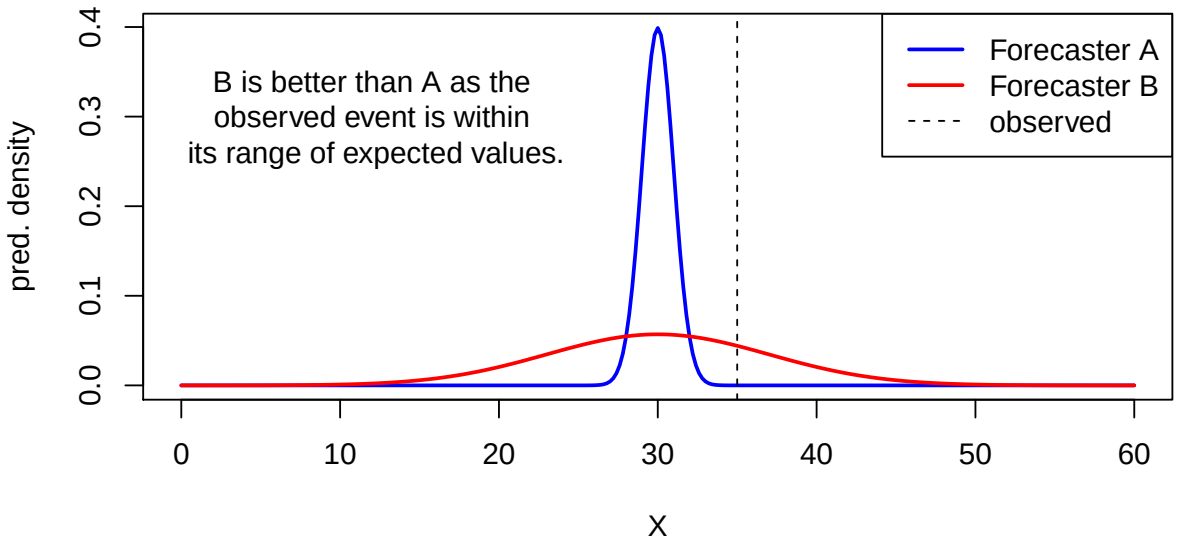
- Forecast quality cannot be fully described considering only the central tendency:



- Good forecasts "maximize sharpness subject to calibration"
- Proper scoring rules (Gneiting and Raftery 2007) allow us to compare probabilistic forecasts

Why take into account uncertainty?

- Forecast quality cannot be fully described considering only the central tendency:

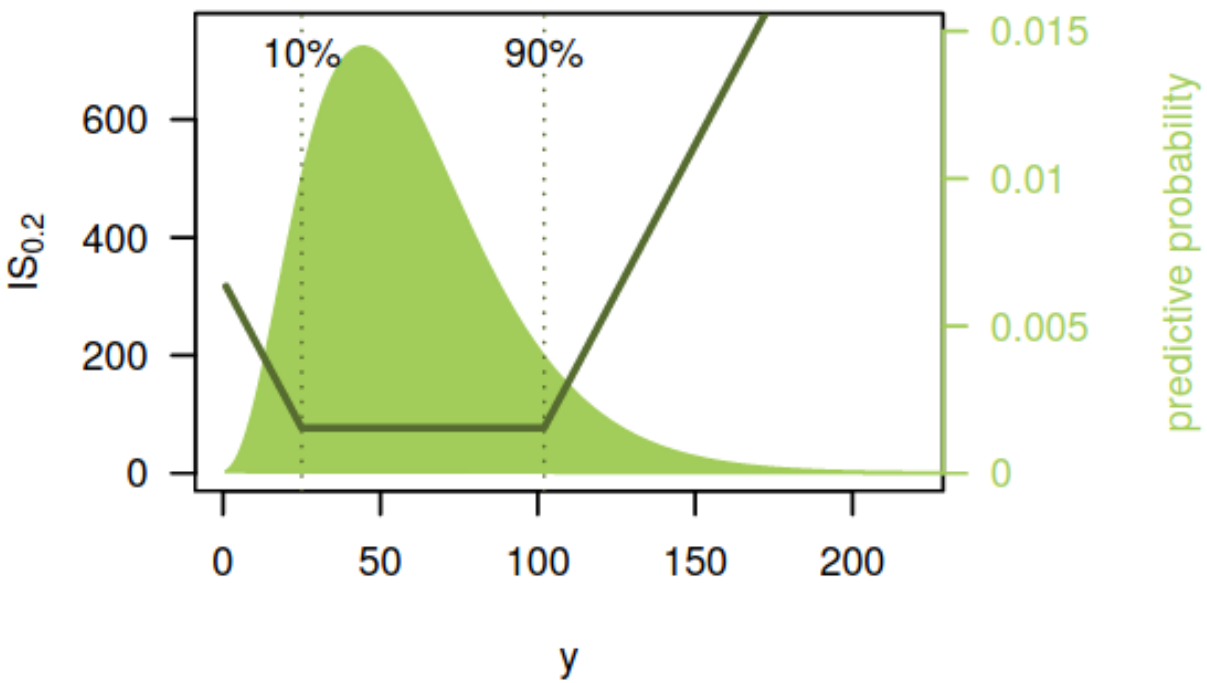


- Good forecasts "maximize sharpness subject to calibration"
- Proper scoring rules (Gneiting and Raftery 2007) allow us to compare probabilistic forecasts

The interval score

Consider a central $(1-\alpha)\times 100\%$ prediction interval $[l, u]$ and observation y. The **interval score** is given by

$$\mathrm{IS}_\alpha(F, y) = \underbrace{(u-l)}_{\text{spread}} + \underbrace{\frac{2}{\alpha}(l-y)\mathbf{1}(y<l)}_{\text{penalty for underprediction}} + \underbrace{\frac{2}{\alpha}(y-u)\mathbf{1}(y>u)}_{\text{penalty for overprediction}},$$



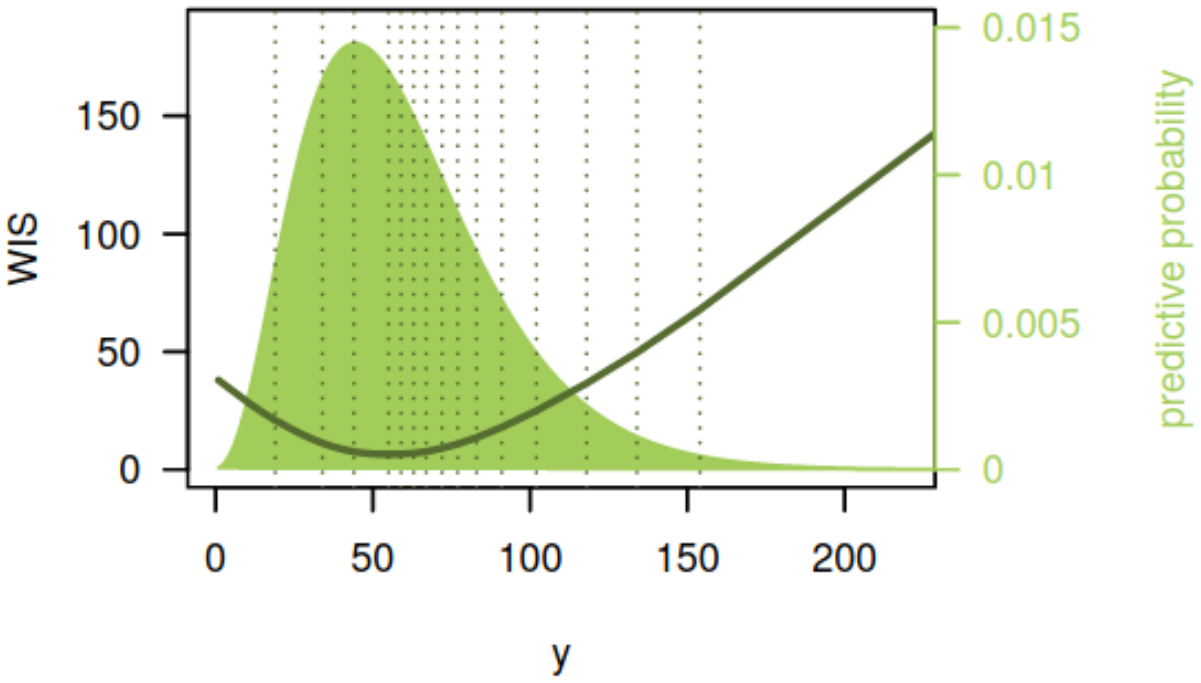
The weighted interval score

Bracher, Ray, Gneiting, Reich (2020), https://arxiv.org/abs/2005.12881

To assess prediction intervals at levels $(1-\alpha_0, \ldots, 1-\alpha_K)$ simultaneously we can use the **weighted interval score**:

$$\text{WIS}_{\alpha_{0:K}}(F, y) = \frac{1}{K+1} \times \sum_{k=0}^{K} \frac{\alpha}{2} \times \text{IS}_{\alpha_k}(F, y).$$

which approximates the CRPS and generalizes the AE.



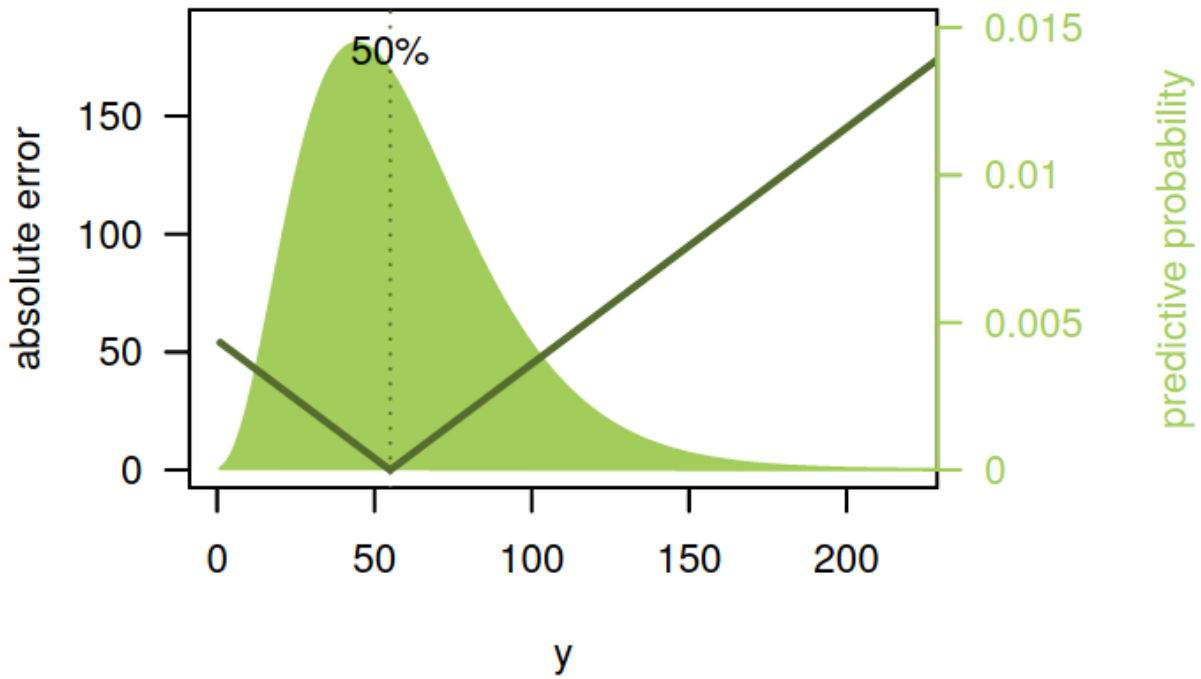
The weighted interval score

Bracher, Ray, Gneiting, Reich (2020), https://arxiv.org/abs/2005.12881

To assess prediction intervals at levels $(1-\alpha_0, \ldots, 1-\alpha_K)$ simultaneously we can use the **weighted interval score**:

$$\mathrm{WIS}_{\alpha_{0:K}}(F, y) = \frac{1}{K+1} \times \sum_{k=0}^{K} \frac{\alpha}{2} \times \mathrm{IS}_{\alpha_k}(F, y).$$

which approximates the CRPS and generalizes the AE.



Systematic comparison of different forecasters

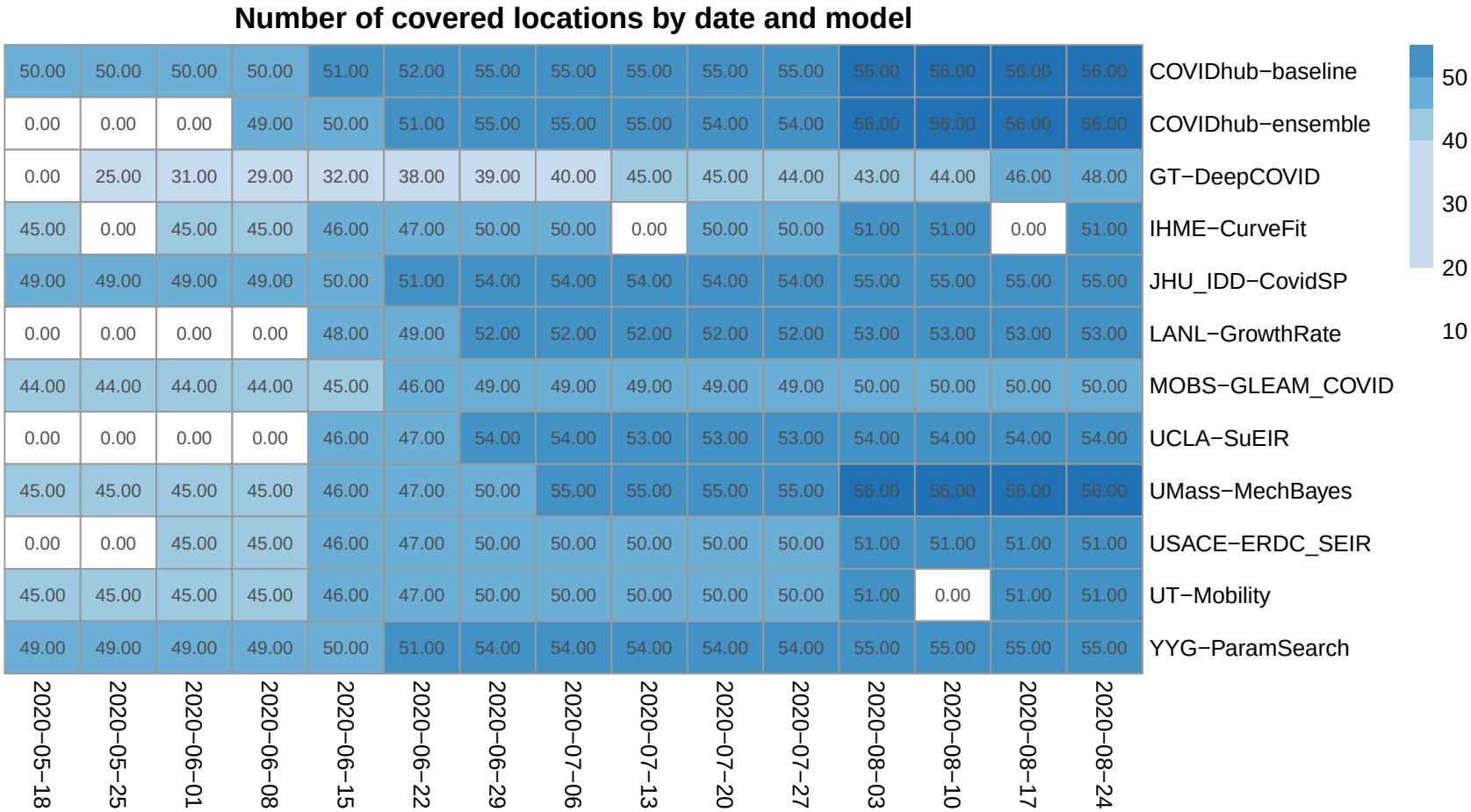
- **Mean weighted interval scores** for aggregation across space, time, target
  - preserve propriety (i.e. don't create an incentive to cheat)
  - scores from larger states and weeks with high incidence are influential

- **Restriction** to models fulfilling (E. Cramer / N. Reich):
  - covering at least half of all states
  - covering at least 12 out of 15 wks between 2020-05-18 and 2020-08-24
  - forecast targets with major revisions of truth data removed

- **Difficulty: Different forecasters cover different subsets of weeks and locations**.
  - could be dealt with mixed effects regression, but challenging
  - my suggestion: use **pairwise** comparisons
  - main assumption: "non-informative missingness"

Availability of forecasts from different models

After removal of weeks/targets with major revisions of truth data



Ratios of mean WIS

If all forecasts were available from all forecasters the following would work:

$$\theta_{ij} = \frac{\text{mean WIS model } i}{\text{mean WIS model } j} = \begin{cases} < 1 \text{ if } i \text{ better than j} \\ > 1 \text{ if } i \text{ worse than j} \end{cases}$$

$$\theta_i = \frac{\text{mean WIS model } i}{\left(\prod_{m=1}^{M} \text{mean WIS model } m\right)^{1/M}} = \begin{cases} < 1 \text{ if } i \text{ better than avg} \\ > 1 \text{ if } i \text{ worse than avg} \end{cases}$$

$$\Rightarrow \theta_{ij} = \frac{\theta_i}{\theta_j}$$

θ_i is a scale-free measure of relative performance (**relative WIS skill** wrt models $1, \ldots, M$). Ratios of average scores are easier to interpret that differences.

Problem: The θ_i can only be evaluated if all forecasters cover all prediction tasks.

Comparisons under partial missingness

If all forecasts were available from all forecasters then

$$\theta_i = \left(\prod_{m=1}^{M} \theta_{im} \right)^{1/M}.$$

If some forecasts are missing we can still compute

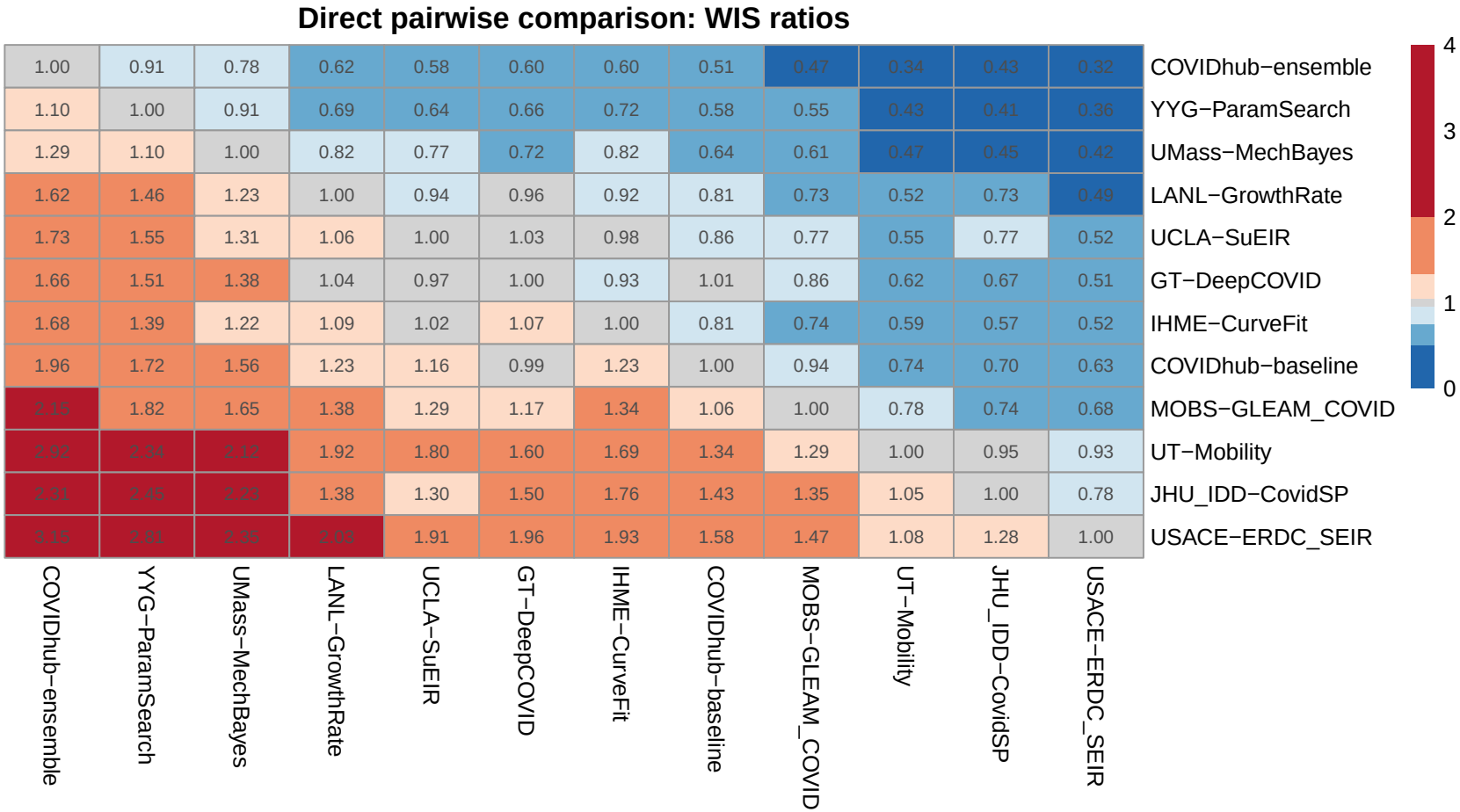
$$\tilde{\theta}_{ij} = \frac{\text{mean WIS model } i \text{ on } \mathcal{A}_{ij}}{\text{mean WIS model } j \text{ on } \mathcal{A}_{ij}}$$

with $\mathcal{A}_{ij}$ as the overlap of available forecasts by i and j and

$$\tilde{\theta}_i = \left(\prod_{m=1}^{M} \tilde{\theta}_{im} \right)^{1/M}.$$

Now generally $\tilde{\theta}_{ij} \neq \tilde{\theta}_i / \tilde{\theta}_j$, but we hope that $\tilde{\theta}_{ij} \approx \tilde{\theta}_i / \tilde{\theta}_j$

Observed WIS ratios $\tilde{\theta}_{ij}$



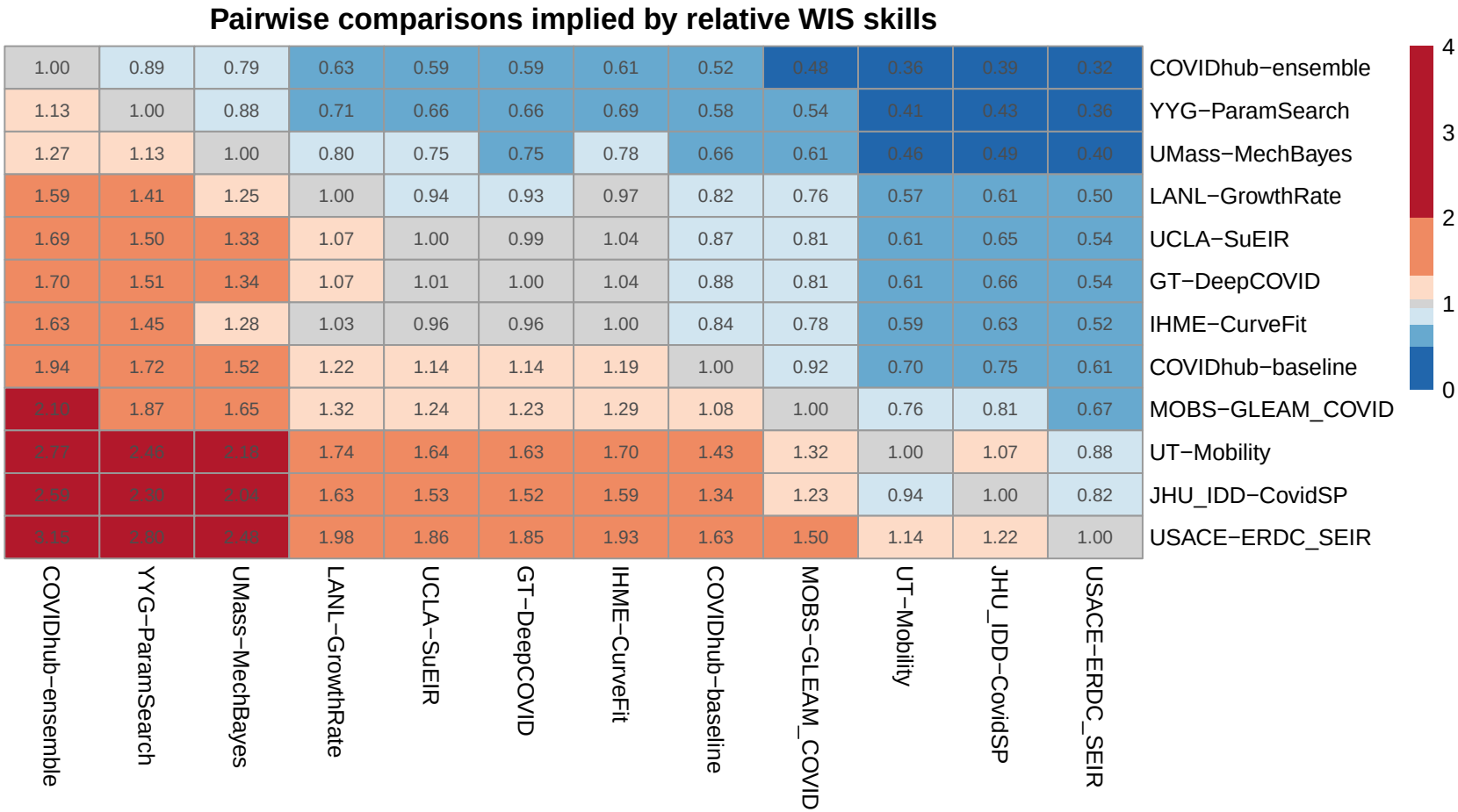
blue: row model better; red: column model better

Pairwise comparisons are *almost* transitive!

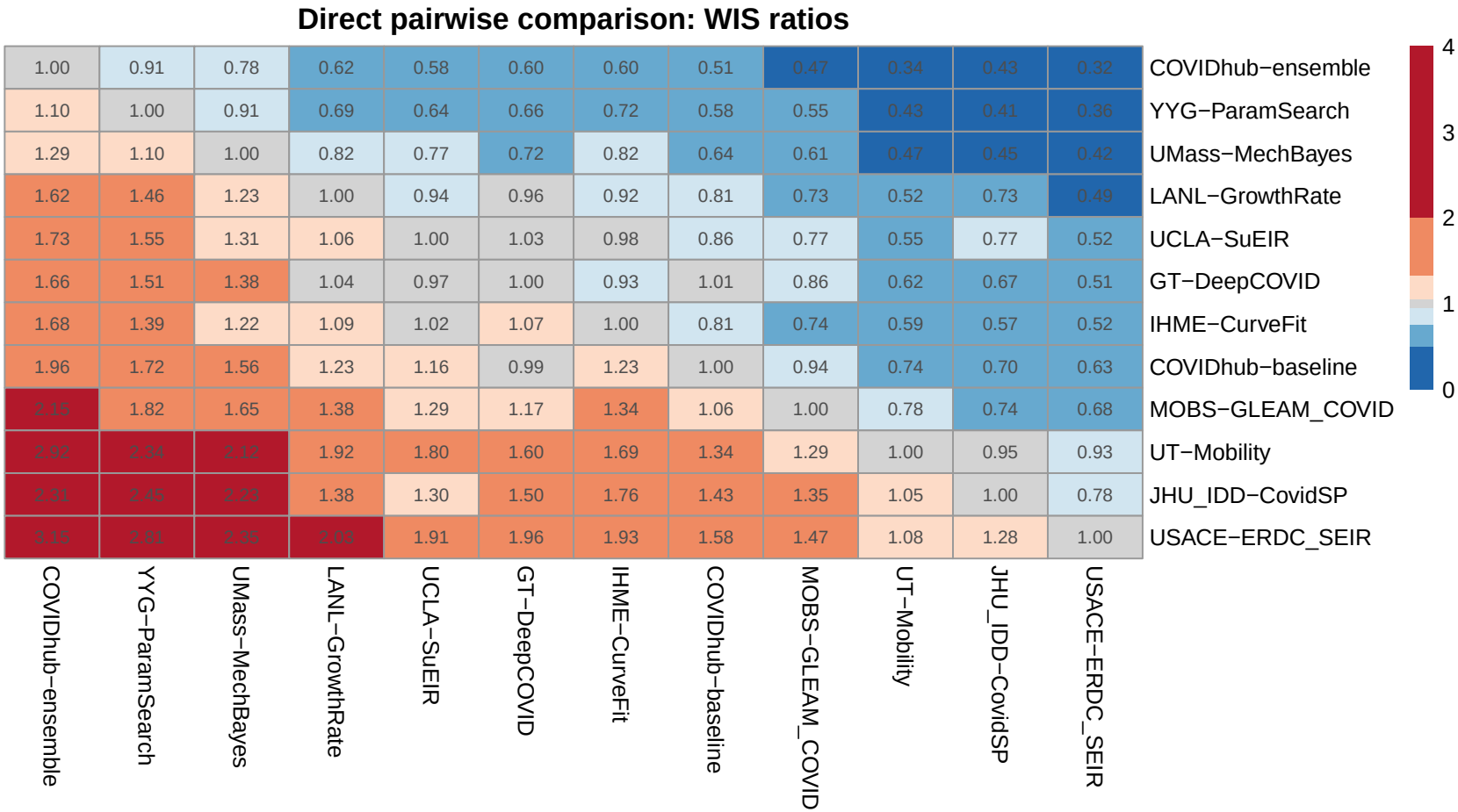
Relative WIS skills $\tilde{\theta}_i$

model	$\tilde{\theta}_i$
COVIDhub-ensemble	0.56
YYG-ParamSearch	0.63
UMass-MechBayes	0.72
LANL-GrowthRate	0.89
IHME-CurveFit	0.92
UCLA-SuEIR	0.95
GT-DeepCOVID	0.96
COVIDhub-baseline	1.09
MOBS-GLEAM_COVID	1.18
JHU_IDD-CovidSP	1.46
UT-Mobility	1.56
USACE-ERDC_SEIR	1.78

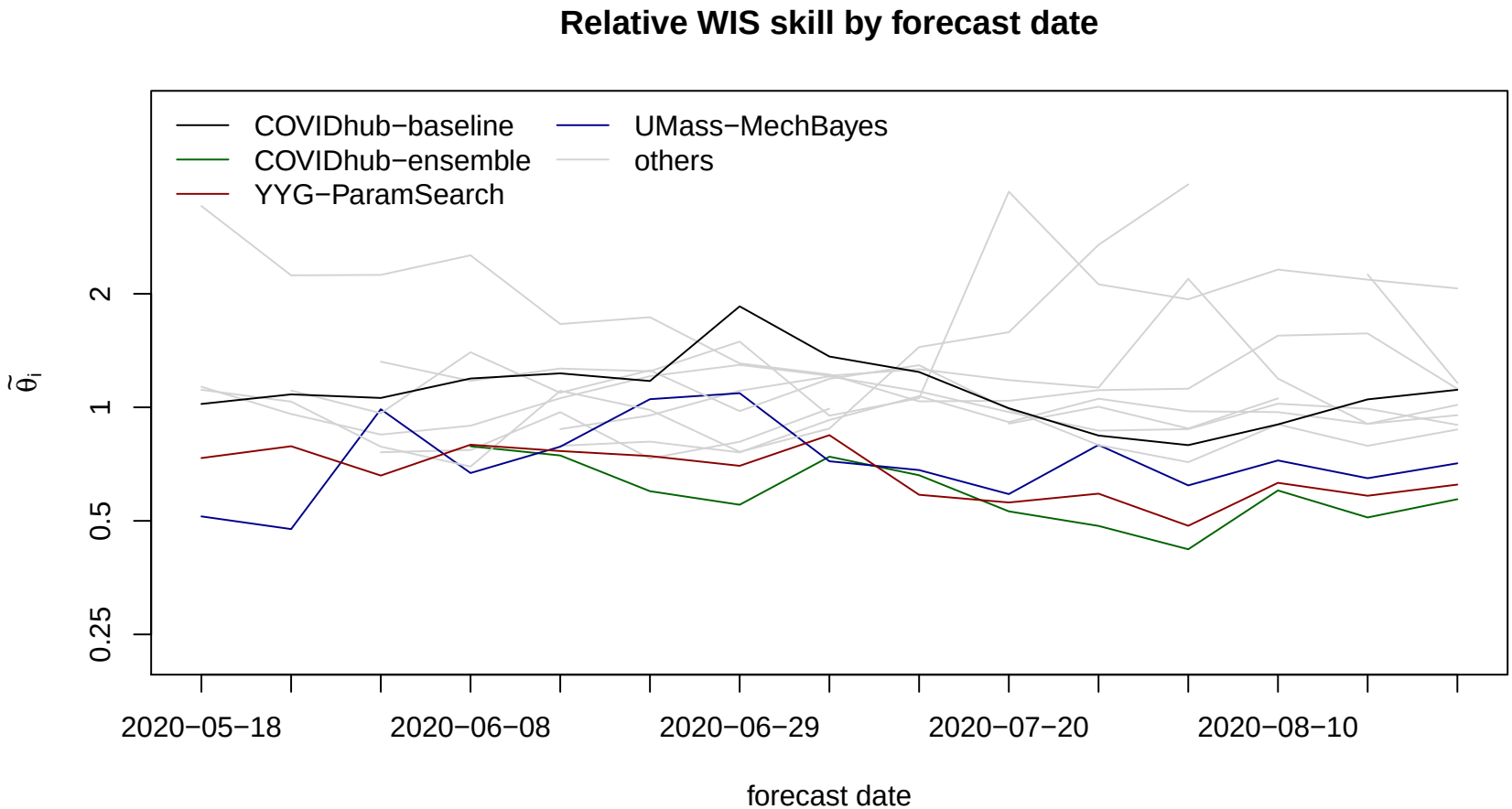
Checking agreement between $\tilde{\theta}_{ij}$ and $\tilde{\theta}_i/\tilde{\theta}_j$



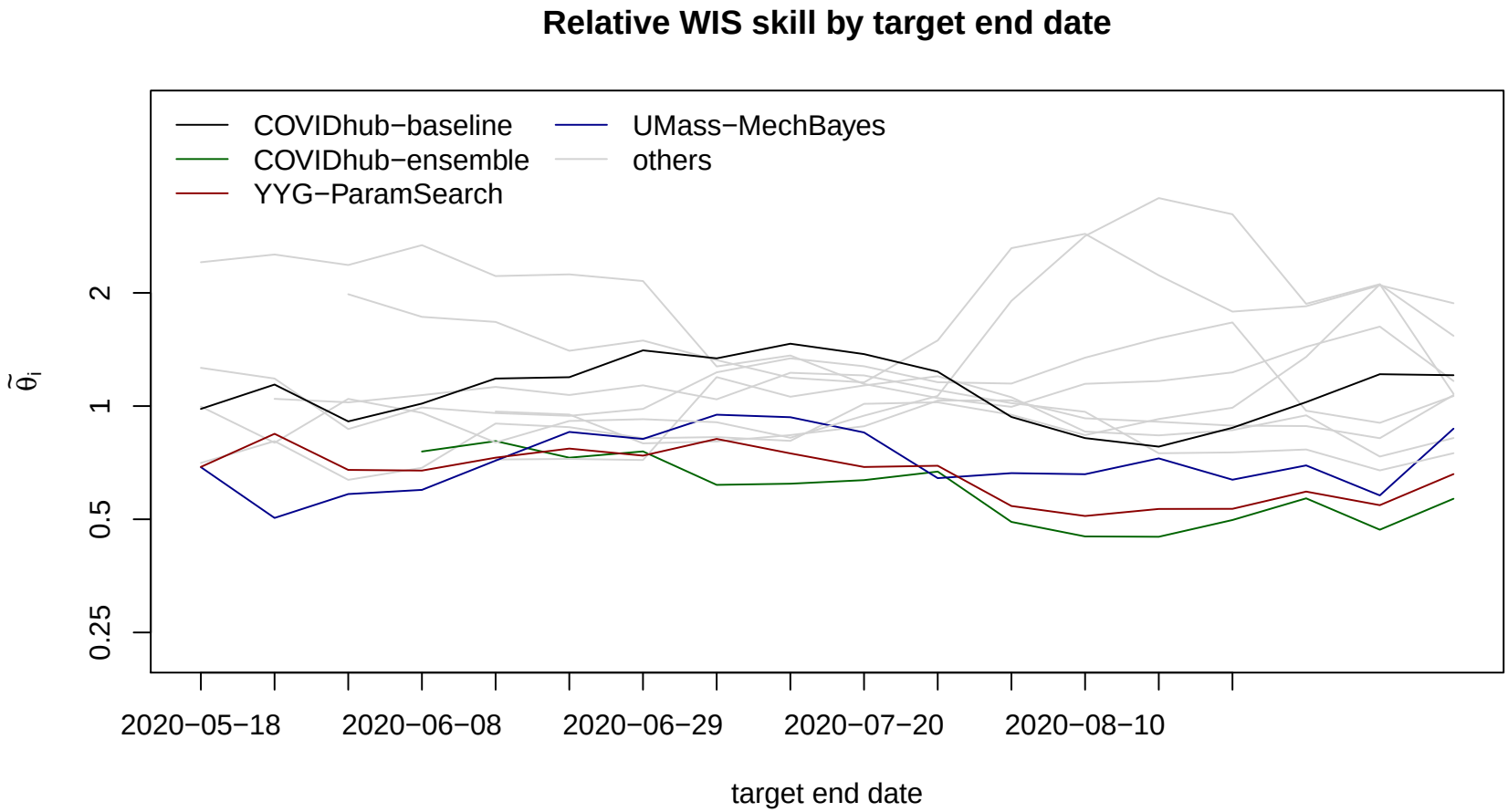
Checking agreement between $\tilde{\theta}_{ij}$ and $\tilde{\theta}_i/\tilde{\theta}_j$



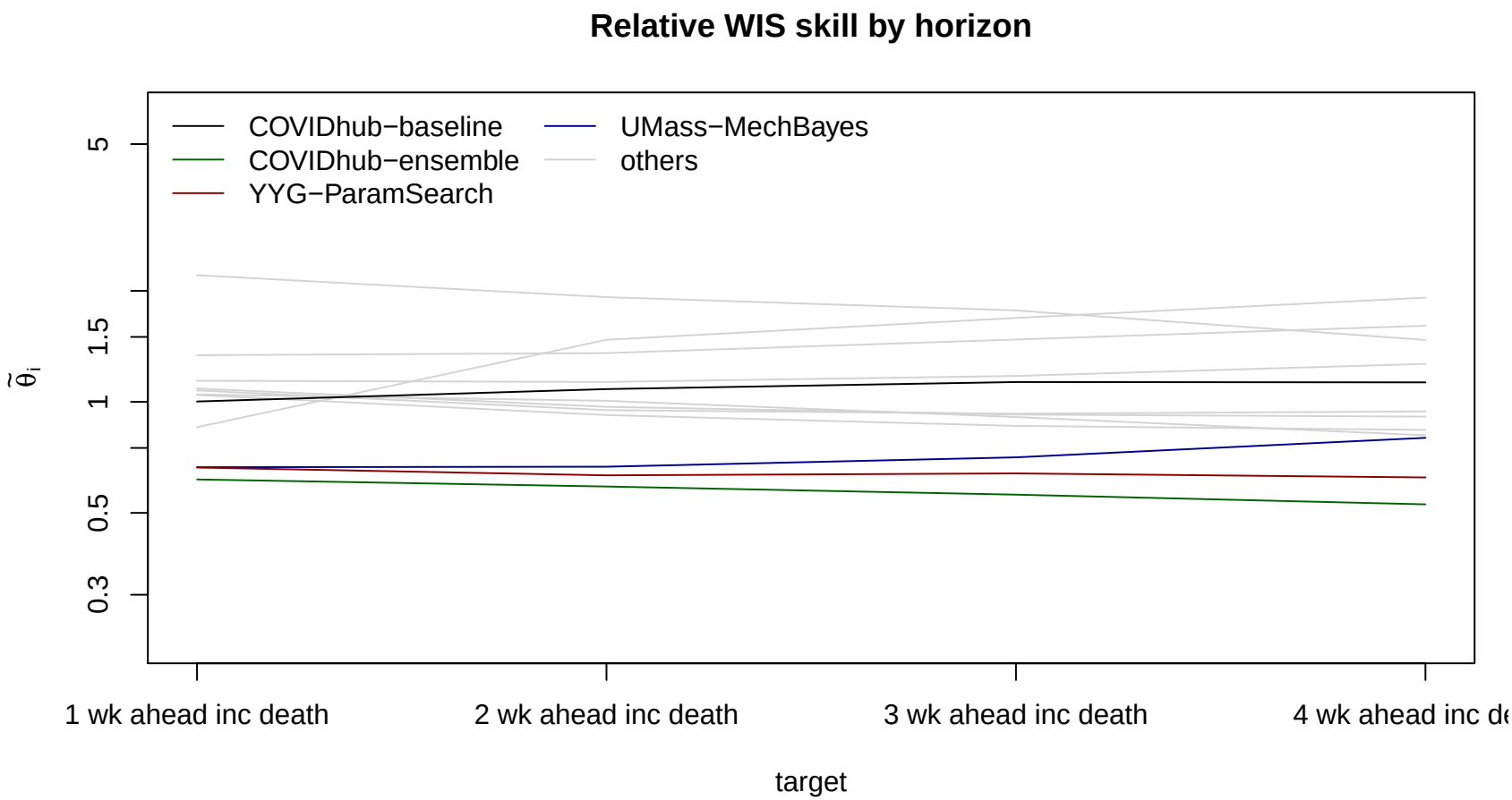
Stratified relative WIS skill



Stratified relative WIS skill



Stratified relative WIS skill



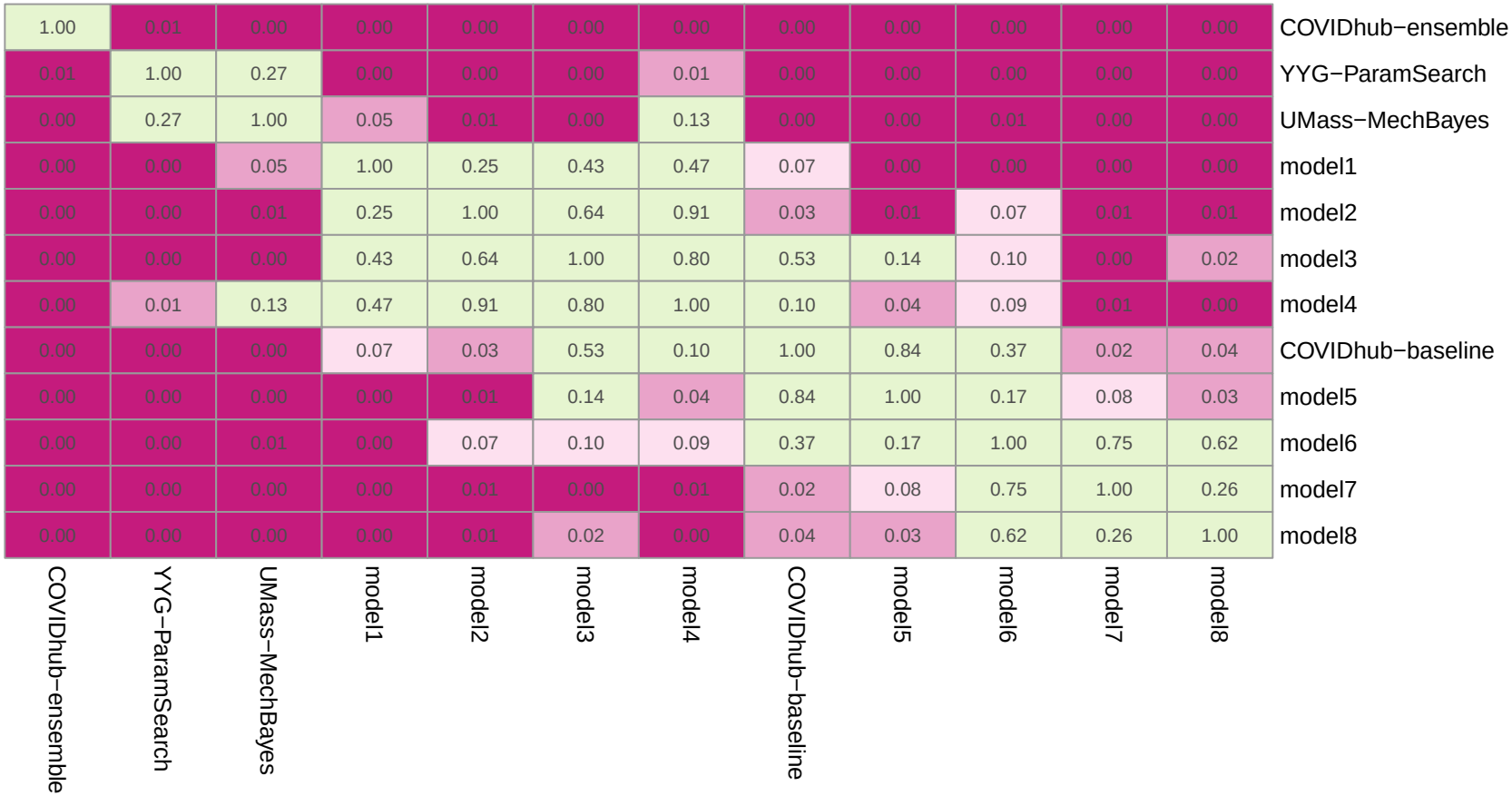
Testing for difference in forecast performance

To test for a difference in mean WIS between models i and j:

- test statistic: $\tilde{\theta}_{ij}$ ($=$ ratio of mean WIS of i and j for $\mathcal{A}_{ij}$)
- generation of reference distribution:
  - blockwise permutation of pairs of scores between i and j
  - to account for dependence between locations and horizons: all forecasts made at one forecast date treated as one block
  - p-value is given by the proportion of sampled WIS ratios exceeding the observed ratio

Results of performance tests

Permutation tests (applied to mean WIS per forecast date)



References

- Johannes Bracher, Evan L. Ray, Tilmann Gneiting, Nicholas G. Reich (2020): Evaluating epidemic forecasts in an interval format. https://arxiv.org/abs/2005.12881
- Tilmann Gneiting and Adrian Raftery (2007): Strictly Proper Scoring Rules, Prediction,and Estimation. JASA, DOI 10.1198/016214506000001437