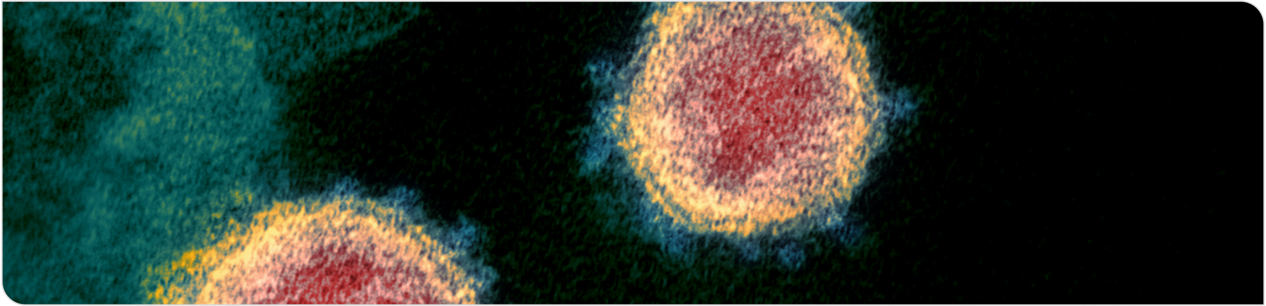
Karlsruher Institut für Technologie

Score-based evaluation of epidemic forecasts

Forecasting Infectious Disease Incidence for Public Health

Johannes Bracher | Karlsruhe Institute of Technology / Heidelberg Institute for Theoretical Studies



Evaluation and incentives: window-based taxation

Score-based forecast evaluation

Gneiting and Raftery 2007, Gneiting 2011

- A **scoring rule** s maps a prediction (distribution or point) and an observation y_{obs} to $\mathbb{R}$. Convention: lower scores are better.

  Example: absolute error

  $$\text{AE}(\hat{y}, y_{\text{obs}}) = |\hat{y} - y_{\text{obs}}|.$$

  .

- The **Bayes act** is the optimal choice (in expectation) under a given score and the forecaster's predictive distribution F.

  Under the absolute error:

  $$\hat{y}_{\text{Bayes}} = \text{argmin}_{\hat{y} \in \mathbb{R}} \mathbb{E}_F|\hat{y} - Y| = \text{med}(F).$$

  .

Score-based forecast evaluation

Gneiting and Raftery 2007, Gneiting 2011

- A **scoring rule** s maps a prediction (distribution or point) and an observation y_{obs} to $\mathbb{R}$. Convention: lower scores are better.

  Example: absolute percentage error

  $$\text{APE}(\hat{y}, y_{\text{obs}}) = |\hat{y} - y_{\text{obs}}| \;/\; y_{\text{obs}}.$$

  .

- The **Bayes act** is the optimal choice (in expectation) under a given score and the forecaster's predictive distribution F.

  Under the absolute percentage error

  $$\hat{y}_{\text{Bayes}} \;=\; \text{med}^{(-1)}(F).$$

  .

Score-based forecast evaluation

Gneiting and Raftery 2007, Gneiting 2011

- A **scoring rule** s maps a prediction (distribution or point) and an observation y_{obs} to $\mathbb{R}$. Convention: lower scores are better.

  Example: absolute percentage error

  $$\text{APE}(\hat{y}, y_{\text{obs}}) = |\hat{y} - y_{\text{obs}}| \;/\; y_{\text{obs}}.$$

  This should reflect the utility of the forecast $\hat{y}$.

- The **Bayes act** is the optimal choice (in expectation) under a given score and the forecaster's predictive distribution F.

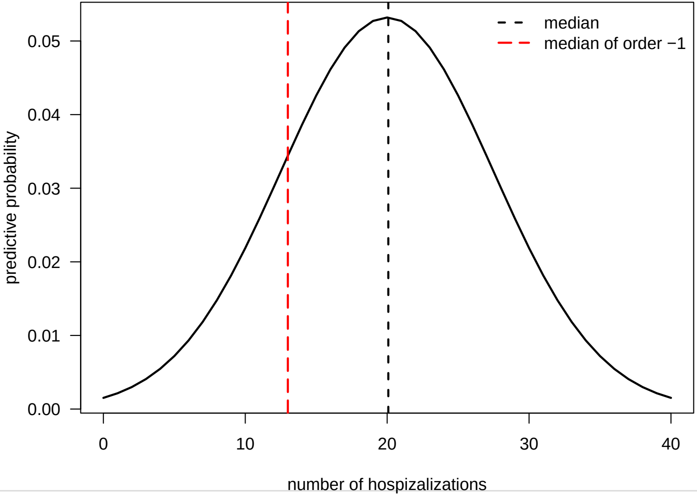
  Under the absolute percentage error

  $$\hat{y}_{\text{Bayes}} = \text{med}^{(-1)}(F).$$

  This should be a useful quantity.

APE incentivizes lower forecasts than AE

Gneiting (2011)



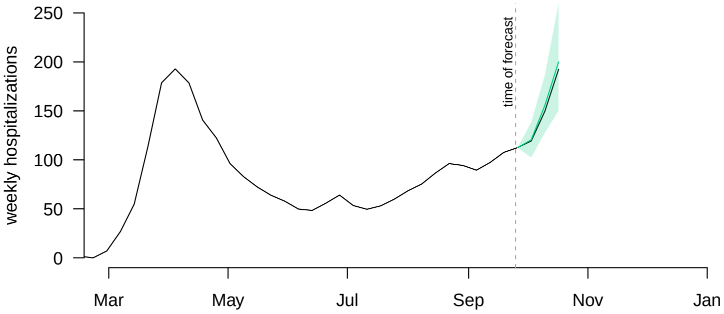
Proper scoring rules for probabilistic forecasts

Gneiting and Raftery 2007

- Epidemiological forecasts should ideally be **probabilistic**.
- A scoring rule is **stricty proper** if the Bayes act (relative to a class of distributions $\mathcal{F}$) is the forecaster's true belief F:
$$\text{argmin}_{G \in \mathcal{F}} \mathbb{E}_F[s(G, Y)] = F.$$
- Proper scores thus incentivize **honest forecasting**.

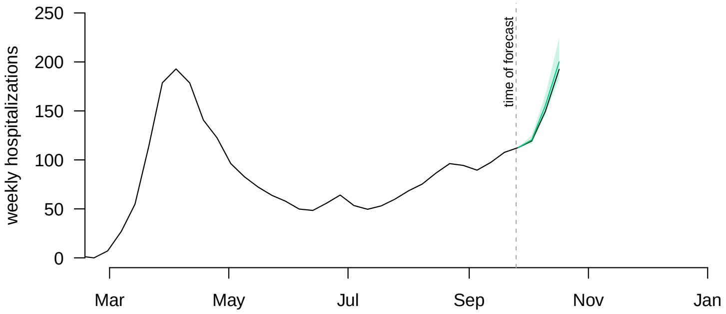
What makes a good forecast? Sharpness and calibration

- Proper scoring rules reward **sharpness** subject to **calibration**.
- Calibration: consistency of forecasts and observations.
  - can be assessed e.g., using PIT histograms.
- Sharpness: informativeness of forecasts.



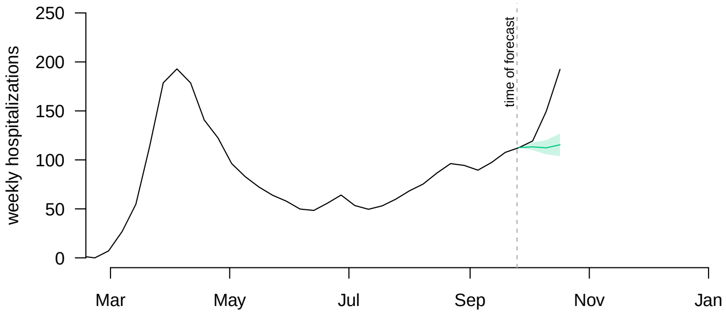
What makes a good forecast? Sharpness and calibration

- Proper scoring rules reward **sharpness** subject to **calibration**.
- Calibration: consistency of forecasts and observations.
  - can be assessed e.g., using PIT histograms.
- Sharpness: informativeness of forecasts.



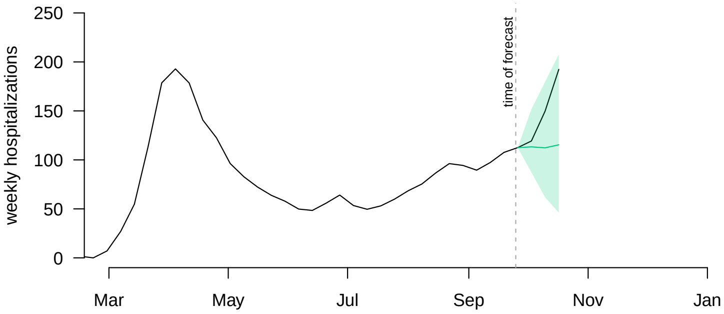
What makes a good forecast? Sharpness and calibration

- Proper scoring rules reward **sharpness** subject to **calibration**.
- Calibration: consistency of forecasts and observations.
  - can be assessed e.g., using PIT histograms.
- Sharpness: informativeness of forecasts.



What makes a good forecast? Sharpness and calibration

- Proper scoring rules reward **sharpness** subject to **calibration**.
- Calibration: consistency of forecasts and observations.
  - can be assessed e.g., using PIT histograms.
- Sharpness: informativeness of forecasts.

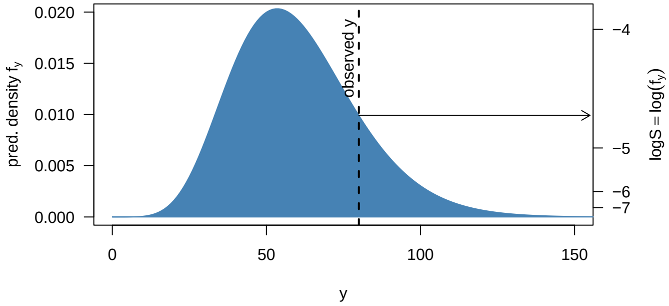


Popular proper scoring rules: logarithmic score

Gneiting and Raftery 2007

- logarithmic score:

$$\text{logS}(F, y_{\text{obs}}) = \log \text{f}(Y = y_{\text{obs}})$$



- the logarithmic score is *local*.

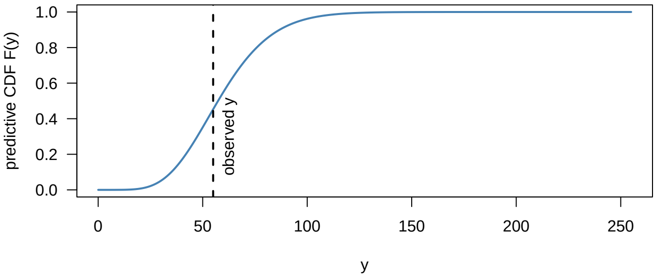
Popular proper scoring rules: CRPS

Gneiting and Raftery 2007

- Continuous ranked probability score:

$$\text{CRPS}(F, y_{\text{obs}}) = \int_{-\infty}^{\infty} \{F(x) - \mathbb{I}(y_{\text{obs}} \geq x)\}^2 \, \mathrm{d}x$$

- CRPS is *distance sensitive* and generalizes the absolute error.



- The *weighted interval score* (WIS, Bracher et al 2021) is a quantile-based approximation of the CRPS.

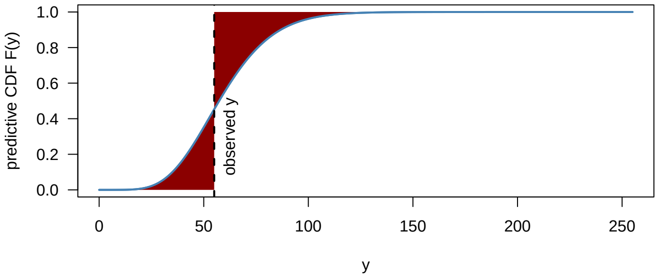
Popular proper scoring rules: CRPS

Gneiting and Raftery 2007

- Continuous ranked probability score:

$$\text{CRPS}(F, y_{\text{obs}}) = \int_{-\infty}^{\infty} \{F(x) - \mathbb{I}(y_{\text{obs}} \geq x)\}^2 \, \mathrm{d}x$$

- CRPS is *distance sensitive* and generalizes the absolute error.



- The *weighted interval score* (WIS, Bracher et al 2021) is a quantile-based approximation of the CRPS.

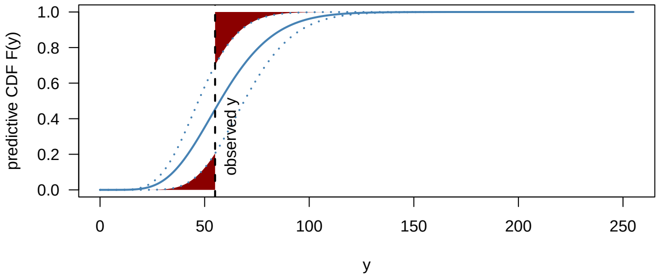
Popular proper scoring rules: CRPS

Gneiting and Raftery 2007

- Continuous ranked probability score:

$$\text{CRPS}(F, y_{\text{obs}}) = \int_{-\infty}^{\infty} \{F(x) - \mathbb{I}(y_{\text{obs}} \geq x)\}^2 \, \mathrm{d}x$$

- CRPS is *distance sensitive* and generalizes the absolute error.



- The *weighted interval score* (WIS, Bracher et al 2021) is a quantile-based approximation of the CRPS.

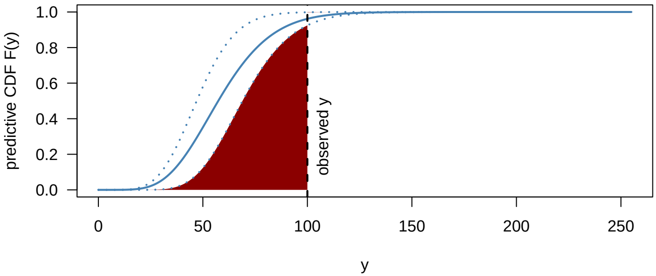
Popular proper scoring rules: CRPS

Gneiting and Raftery 2007

- Continuous ranked probability score:

$$\text{CRPS}(F, y_{\text{obs}}) = \int_{-\infty}^{\infty} \{F(x) - \mathbb{I}(y_{\text{obs}} \geq x)\}^2 \, \mathrm{d}x$$

- CRPS is *distance sensitive* and generalizes the absolute error.



- The *weighted interval score* (WIS, Bracher et al 2021) is a quantile-based approximation of the CRPS.

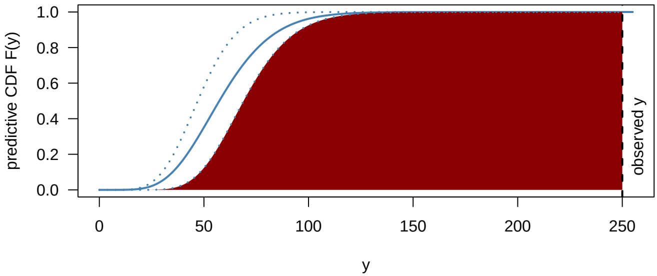
Popular proper scoring rules: CRPS

Gneiting and Raftery 2007

- Continuous ranked probability score:

$$\text{CRPS}(F, y_{\text{obs}}) = \int_{-\infty}^{\infty} \{F(x) - \mathbb{I}(y_{\text{obs}} \geq x)\}^2 \, \mathrm{d}x$$

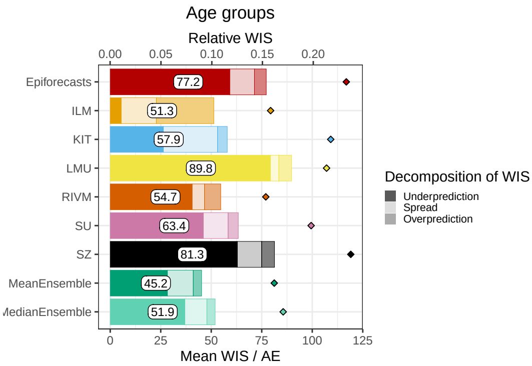
- CRPS is *distance sensitive* and generalizes the absolute error.



- The *weighted interval score* (WIS, Bracher et al 2021) is a quantile-based approximation of the CRPS.

Score decompositions

- CRPS / WIS can be decomposed into dispersion, overprediction and underprediciton (Bracher et al 2021).



Check out Daniel Wolffram's poster!

- Other decompositions exist, e.g., *miscalibration, discrimination, uncertainty* (Gneiting et al 2023).

How to choose a proper scoring rule?

- Hard to tell 🤷‍♂️
- **Applicability:**
  - CRPS requires at least an interval scale.
  - logS easy to use for bins, CRPS for quantiles and samples, Dawid-Sebastiani score for moments.
- **Purpose** (Winkler 1996):
  - For inference, the logS is generally most powerful.
  - “Distance” can be relevant in decision making, favouring CRPS.
- **Robustness:** logS can diverge to ∞, CRPS is more forgiving (to a point where it may seem lenient).
- **Scale-invariance:** logS is invariant to transformations of the target (up to a constant)
- Where feasible, several metrics should be considered **and complemented with visual inspection**.

Some more practical aspects

Karlsruher Institut für Technologie

- Purely reporting average scores is usually not very informative.
- **Visual inspection** of forecasts and observations is an important step.
- **Calibration** of forecasts should be assessed separately (e.g., via PIT histograms).
- Inclusion of **baseline models** elucidates whether models have non-trivial predictive ability.
  - It’s not totally clear what these should be...

What happens when using an improper score?

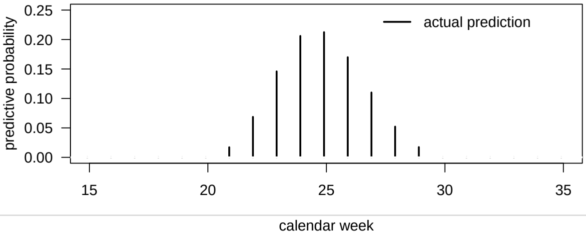
Bracher (2019), Reich et al (2019)

- "Multi-bin log score" for discrete target $Y \in \{1, \dots, N\}$

$$\text{MBlogS}(F, y_{\text{obs}}) = \underbrace{\log\left(\sum_{i=-d}^{d} \text{Prob}_F(Y = y_{\text{obs}} + i)\right)}_{\text{log-probability assigned to observation} \pm d},$$

with tolerance d.

- Example: predicting flu peak week with $d = 1$:



What happens when using an improper score?

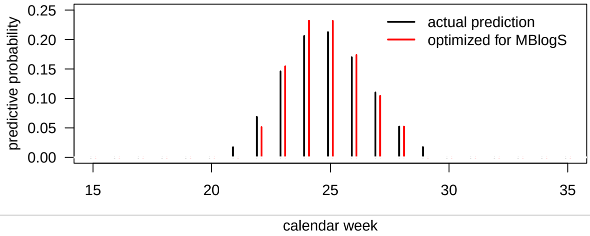
Bracher (2019), Reich et al (2019)

- "Multi-bin log score" for discrete target $Y \in \{1, \dots, N\}$

$$\text{MBlogS}(F, y_{\text{obs}}) = \underbrace{\log\left(\sum_{i=-d}^{d} \text{Prob}_F(Y = y_{\text{obs}} + i)\right)}_{\text{log-probability assigned to observation } \pm\, d},$$

with tolerance d.

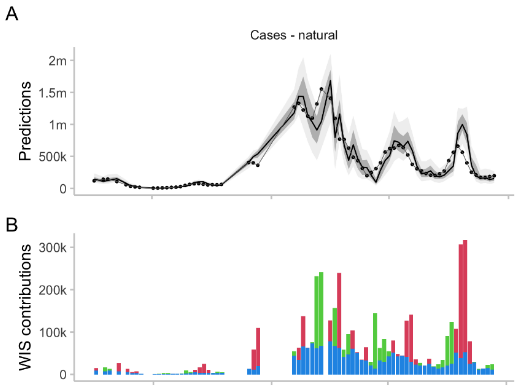
- Example: predicting flu peak week with $d = 1$:



On which scale to evaluate forecasts?

Bosse et al (2023)

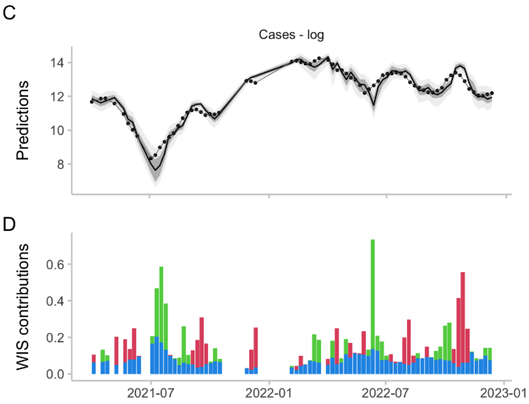
- CRPS changes when transforming forecasts and observations.
- CRPS for $\log(\text{weekly counts})$ can be interpreted as
  - a “probabilistic relative error”.
  - an assessment how well the growth rate was predicted.
  - a “variance-stabilized” score.



On which scale to evaluate forecasts?

Bosse et al (2023)

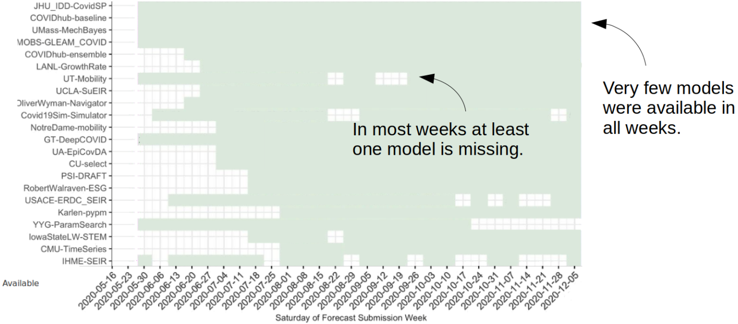
- CRPS changes when transforming forecasts and observations.
- CRPS for $\log(\text{weekly counts})$ can be interpreted as
  - a “probabilistic relative error”.
  - an assessment how well the growth rate was predicted.
  - a “variance-stabilized” score.



How to handle incongruent sets of forecasts?

Cramer et al (2022)

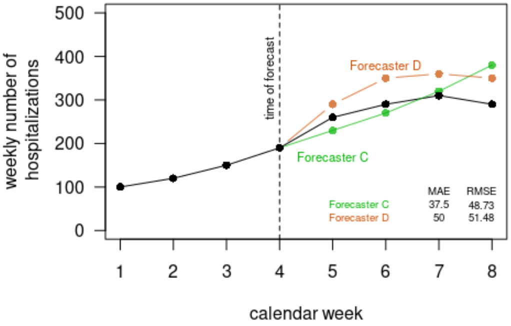
- Typically not all models provide forecasts for all targets.
- Example: COVID mortality forecasts (Cramer et al 2022):



- Heuristic solution: “pairwise tournament” approach leading to “relative WIS”.

How to align evaluation with public health utility?

- Statistical evaluation may be at odds with perceived utility.
- Example: shapes matter.



- Could likely be accounted for by multivariate scoring.

References

- Bosse, Abbott, Cori, van Leeuwen, Bracher, Funk (2022): Transformation of forecasts for evaluating predictive performance in an epidemiological context. Preprint, medRxiv.
- Bracher (2019): On the multi-bin logarithmic score used in the FluSight competitions. PNAS.
- Bracher, Ray, Gneiting, Reich (2021): Evaluating epidemic forecasts in an interval format. PLOS Computational Biology.
- Bracher, Wolffram et al (2022): National and subnational short-term forecasting of COVID-19 in Germany and Poland during early 2021. Communications Medicine.
- Cramer et al (2022): Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. PNAS.

References

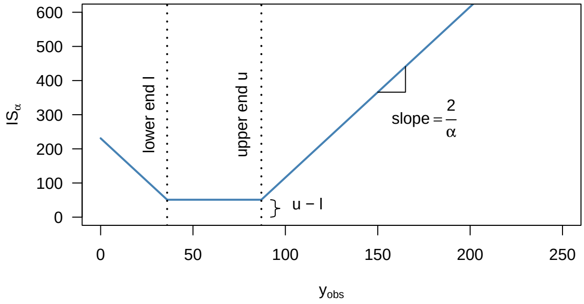
- Gneiting (2011): Making and Evaluating Point Forecasts. Journal of the American Statistical Association.
- Gneiting, Lerch, Schulz (2023): Probabilistic solar forecasting: Benchmarks, post-processing, verification. Solar Energy.
- Gneiting, Raftery (2007): Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association.
- Pollett et al (2021): Recommended reporting items for epidemic forecasting and prediction research: The EPIFORGE 2020 guidelines. PLOS Medicine.
- Winkler (1996): Scoring Rules and the Evaluation of Probabilities. Test.

Proper scoring rules: the (weighted) interval score

Bracher, Ray, Gneiting, Reich (2021)

- For a central $(1-\alpha)$ prediction interval $[l, u]$:

$$\text{IS} = \underbrace{(u - l)}_{\text{spread}} + \underbrace{\frac{2}{\alpha} \times (l - y_{\text{obs}}) \times \mathrm{I}(y_{\text{obs}} < l)}_{\text{underprediction penalty}} + \underbrace{\frac{2}{\alpha} \times (y_{\text{obs}} - u) \times \mathrm{I}(y_{\text{obs}} > u)}_{\text{overprediction penalty}}$$



- Via a weighted sum of interval scores (WIS) at different levels we can approximate the CRPS.