On the multibin logarithmic score used in the FluSight competitions

Johannes Bracher[a,1]

The FluSight challenges (1) represent an outstanding collaborative effort and have "pioneered infectious disease forecasting in a formal way" (ref. 2, p. 2803). However, I wish to initiate a discussion about the employed evaluation measure.

The competitions feature discrete or discretized targets related to the US influenza season. E.g., for the peak timing Y, a forecast distribution F consists of probabilities $p_1, \ldots, p_T$ for the $T = 33$ wk of the season. Such forecasts can be evaluated using the log score (3, 4)

$$\mathrm{logS}(F, y_{\mathrm{obs}}) = \log(p_{y_{\mathrm{obs}}}),$$

where y_{obs} is the observed value. This score is strictly proper; i.e., its expectation is uniquely maximized by the true distribution of Y. In the FluSight competitions the logS is applied in a multibin version,

$$\mathrm{MBlogS}(F, y_{\mathrm{obs}}) = \log\left(\sum_{i=-d}^{d} p_{y_{\mathrm{obs}}+i}\right),$$

to measure "accuracy of practical significance" (ref. 1, p. 3153). Depending on the target, d is either 1 or 5. Following the competitions, this score has become widely used (5–10), even though as also mentioned in ref. 1, it is improper. This may be problematic as improper scores incentivize dishonest forecasts. Assume $T > 2d$ and

$$p_1 = \cdots = p_d = p_{T-d+1} = \cdots = p_T = 0, \quad [1]$$

i.e., probability 0 for the $2d$ extreme categories. Now define a "blurred" distribution $\tilde{F}$ with

$$\tilde{p}_t = \frac{\sum_{i=-d}^{d} p_{t+i}}{2d+1}, \; t = 1, \ldots, T, \quad [2]$$

where $p_t = 0$ for $t < 1$ and $t > T$ and Eq. **1** ensures $\sum_{t=1}^{T} \tilde{p}_t = 1$. This implies

$$\mathrm{MBlogS}(F, y_{\mathrm{obs}}) = \mathrm{logS}(\tilde{F}, y_{\mathrm{obs}}) + \log(2d+1);$$

i.e., the MBlogS is essentially the logS applied to a blurred version of F. To optimize the expected MBlogS under their true belief F, forecasters should therefore not report F, but a sharper forecast G so that the blurred version $\tilde{G}$ (with $\tilde{p}_{G,1}, \ldots, \tilde{p}_{G,T}$ derived from $p_{G,1}, \ldots, p_{G,T}$ as in Eq. **2**) is close or equal to F. This follows from the propriety of the logS. An optimal G is found by maximizing $\sum_{t=1}^{T} p_t \cdot \log(\tilde{p}_{G,t})$ with respect to $p_{G,1}, \ldots, p_{G,T}$.

This optimal G can differ considerably from the original F, as Fig. 1 shows for forecasts of the 2016 to 2017 national-level peak timing by the Los Alamos National Laboratory (LANL) team (9) (downloaded from https://github.com/FluSightNetwork/cdc-flusight-ensemble/). The optimized Gs (with $d = 1$) often have their mode shifted by 1 wk and tend to be multimodal, even for unimodal F. Averaged over the 2016 to 2017 season they yield improved MBlogS for the peak timing (−0.434 vs. −0.484). This illustrates that the MBlogS may be gamed, even though we strongly doubt participants have tried to. The logS, like any other proper score, could avoid such pitfalls.

Acknowledgments

I thank T. Gneiting for helpful discussions and the FluSight Collaboration for making its forecasts publicly available.

[a]Epidemiology, Biostatistics and Prevention Institute, University of Zurich, 8001 Zurich, Switzerland

Author contributions: J.B. designed research, performed research, analyzed data, and wrote the paper.

The author declares no conflict of interest.

[1]Email: johannes.bracher@uzh.ch.

First published September 26, 2019.

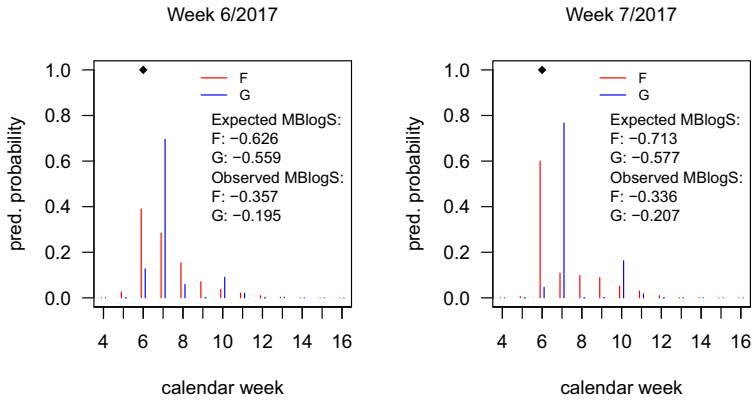

Fig. 1. Forecasts *F* for the peak week, submitted by the LANL team in weeks 6 to 7, 2017, and optimized versions *G*. Diamonds mark the observed peak week. Expected scores are computed under *F*.

1 N. G. Reich *et al.*, A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 3146–3154 (2019).
2 C. Viboud, A. Vespignani, The future of influenza forecasts. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2802–2804 (2019).
3 T. Gneiting, A. E. Raftery, Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
4 L. Held, S. Meyer, J. Bracher, Probabilistic forecasting in infectious disease epidemiology: The 13th Armitage lecture. *Stat. Med.* **36**, 3443–3460 (2017).
5 L. C. Brooks, D. C. Farrow, S. Hyun, R. J. Tibshirani, R. Rosenfeld, Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS Comput. Biol.* **14**, 1–29 (2018).
6 S. Kandula *et al.*, Evaluation of mechanistic and statistical methods in forecasting influenza-like illness. *J. R. Soc. Interface* **15**, 20180174 (2018).
7 S. Kandula, J. Shaman, Near-term forecasts of influenza-like illness: An evaluation of autoregressive time series approaches. *Epidemics* **27**, 41–51 (2019).
8 C. J. McGowan *et al.*, Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Sci. Rep.* **9**, 683 (2019).
9 D. Osthus, J. Gattiker, R. Priedhorsky, S. Y. Del Valle, Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy (with discussion). *Bayesian Anal.* **14**, 261–312 (2019).
10 D. Osthus, A. R. Daughton, R. Priedhorsky, Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited. *PLoS Comput. Biol.* **15**, 1–19 (2019).