PNAS
www.pnas.org

Supplementary Information for

Can Auxiliary Indicators Improve COVID-19 Forecasting and Hotspot Prediction?

Daniel J. McDonald, Jacob Bien, Alden Green, Addison Hu, Nat DeFries, Sangwon Hyun, Natalia L. Oliveira, James Sharpnack, Jingjing Tang, Robert Tibshirani, Valerie Ventura, Larry Wasserman, and Ryan J. Tibshirani

Daniel J. McDonald.
E-mail: daniel@stat.ubc.ca

This PDF file includes:

Supplementary text
Figs. S1 to S19 (not allowed for Brief Reports)
Legends for Dataset S1 to S2

Other supplementary materials for this manuscript include the following:

Datasets S1 to S2

Supporting Information Text

1. Examining the relative advantage of using finalized rather than vintage data

The goal of this section is to quantify the effect of not properly accounting for the question of "what was known when" in performing retrospective evaluations of forecasters. Figures S1 and S2 show what Figures 3 and 4 in the main paper would have looked like if we had simply trained all models using the finalized data rather than using vintage data. This comparison can be seen more straightforwardly in Figures S3 and S4, which show the ratio in performance between the vintage and finalized versions. When methods are given the finalized version of the data rather than the version available at the time that the forecast would have been made, all methods appear (misleadingly) to have better performance than they would have had if run prospectively. For example, for forecasting case rates 7-days ahead, the WIS of all methods is at least 8% larger than what would have been achieved using finalized data. This effect diminishes as the forecasting horizon increases, reflecting the fact that longer-horizon forecasters rely less heavily on recent data than very short-horizon forecasters. Crucially, some methods are "helped" more than others by the less scrupulous retrospective evaluation, underscoring the difficulty of avoiding misleading conclusions when performing retrospective evaluations of forecasters.

CHNG-CLI (and, to a lesser extent, the other claims-based signals) is the most affected by this distinction, reflecting the latency in claims-based reporting. This underscores the importance of efforts to provide "nowcasts" for claims signals (which corresponds to a 0-ahead forecast of what the claims signal's value will be once all data has been collected). Looking at the CHNG-CLI and DV-CLI curves in Figure S1, we can see that they perform very similarly when trained on the finalized data. This is reassuring because they are, in principle, measuring the same thing (namely, the percentage of outpatient visits that are primarily about COVID-related symptoms). The substantial difference in their curves in Figure 3 of the main paper must therefore reflect their having very different backfill profiles.

While using finalized rather than vintage data affects DV-CLI the least for forecasting, it is one of the most affected methods for the hotspot problem. This is a reminder that the forecasting and hotspot problems are fundamentally distinct. For example, the hotspot problem does not measure the ability to distinguish between flat and downward trends.

Even the AR model is affected by this distinction, reflecting the fact that the case rates themselves (i.e., the response values) are also subject to revision. The forecasters based on indicators are thus affected both by revisions to the indicators and by revisions to the case rates. In the case of the Google-AA model, in which we only used finalized values for the Google-AA indicator, the difference in performance can be wholly attributed to revisions of case rates.

2. Aggregating with geometric mean

In this section, we consider using the geometric mean instead of the arithmetic mean when aggregating the weighted interval score (WIS) across location-time pairs. There are three reasons why using the geometric mean may be desirable.

1. WIS is right-skewed, being bounded below by zero and having occasional very large values. Figure S5 illustrates that the densities appear roughly log-Gaussian. The geometric mean is a natural choice in such a context since the relative ordering of forecasters is determined by the arithmetic mean of the *logarithm* of their WIS values.

2. In the main paper, we report the ratio of the mean WIS of a forecaster to the mean WIS of the baseline forecaster. Another choice could be to take the mean of the ratio of WIS values for the two methods. This latter choice would penalize a method less for doing poorly where the baseline forecaster also does poorly. Using instead the geometric mean makes the order of aggregation and scaling immaterial since the ratio of geometric means is the same as the geometric mean of ratios.

3. If one imagines that a forecaster's WIS is composed of multiplicative space-time effects $S_{\ell,t}$ shared across all forecasters, i.e. $\text{WIS}(F_{\ell,t,f}, Y_{\ell,t}) = S_{\ell,t}E_{f,t}$ with $E_{F,t}$ a forecaster-specific error, then taking the ratio of two forecasters' geometric mean WIS values will effectively cancel these space-time effects.

Figure S6 uses the geometric mean for aggregation. Comparing this with Figure 3 of the main paper, we see that the main conclusions are largely unchanged; however, CHNG-CLI now appears better than AR. This behavior would be expected if CHNG-CLI's poor performance is attributable to a relatively small number of large errors (as opposed to a large number of moderate errors). Indeed, Figure 5 of the main paper further corroborates this, in which we see the heaviest left tails occuring for CHNG-CLI.

3. Bootstrap results

As explained in Section 2.B. of the main paper, a (somewhat cynical) hypothesis for why we see benefits in forecasting and hotspot prediction is that the indicators are not actually providing useful information but they are instead acting as a sort of "implicit regularization," leading to shrinkage on the autoregressive coefficients and therefore to less volatile predictions. To investigate this hypothesis, we consider fitting "noise features" that in truth should have zero coefficients. Recall (from the main paper) that at each forecast date, we train a model on 6,426 location-time pairs. Indicator models are based on six features, corresponding to the three autoregressive terms and the three lagged indicator values. To form noise indicator features, we replace their values with those from a randomly chosen time-space pair (while keeping the autoregressive features fixed). In particular, at each location ℓ and time t, for the forecasting task we replace the triplet $(X_{\ell,t}, X_{\ell,t-7}, X_{\ell,t-14})$ in Eq. (3) of

the main paper with the triplet $(X_{\ell^*,t^*}, X_{\ell^*,t^*-7}, X_{\ell^*,t^*-14})$, where (ℓ^*, t^*) is a location-time pair sampled with replacement from the 6,426 location-time pairs. Likewise in the hotspot prediction task, we replace the triplet $(X^{\Delta}_{\ell,t}, X^{\Delta}_{\ell,t-7}, X^{\Delta}_{\ell,t-14})$ in Eq. (5) of the main paper with $(X^{\Delta}_{\ell^*,t^*}, X^{\Delta}_{\ell^*,t^*-7}, X^{\Delta}_{\ell^*,t^*-14})$. Figures S7–S9 show the results. No method exhibits a noticeable performance gain over the AR method, leading us to dismiss the implicit regularization hypothesis.

4. Upswings and Downswings

In this section we provide extra details about the upswing / flat / downswing analysis described in the main text. Figure S10 shows the overall results, examining the average difference $\mathrm{WIS}(AR) - \mathrm{WIS}(F)$ in period. Figure S11 shows the same information for the hotspot task. On average, during downswings and flat periods, the indicator-assisted models have lower classification error and higher log likelihood than the AR model. For hotspots, both Google-AA and CTIS-CLIIC perform better than the AR model during upswings, in contrast to the forecasting task, where only Google-AA improves. For a related analysis, Figure S12 shows histograms of the Spearman correlation (Spearman's ρ, a rank-based measure of association) between the $\mathrm{WIS}(F)/\mathrm{WIS}(AR)$ and the magnitude of the swing. Again we see that case rate increases are positively related to diminished performance of the indicator models.

One hypothesis for diminished relative performance during upswings is that the AR model tends to overpredict downswings and underpredict upswings. Adding indicators appears to help avoid this behavior on the downswing but not as much on upswings. Figure S13 shows the correlation between between $\mathrm{WIS}(AR) - \mathrm{WIS}(F)$ and the difference of their median forecasts. During downswings, this correlation is large, implying that improved relative performance of F is related to making lower forecasts than the AR model. The opposite is true during upswings. This is largely to be expected. However, the relationship attenuates in flat periods and during upswings. That is, when performance is better in those cases, it may be due to other factors than simply making predictions in the correct direction, for example, narrower confidence intervals.

5. Leadingness and laggingness

In Section 2.D of the main text, we discuss the extent to which the indicators are leading or lagging case rates during different periods. To define the amount of leadingness or laggingness, we use the cross correlation function (CCF) between the two time series. The CCF is of an indicator $X_{\ell,t}$ and case rates $Y_{\ell,t}$ is defined as

$$\mathrm{CCF}_{\ell,t}(a) = \frac{1}{n_a} \sum_{i=1}^{n_a} \left((X_{\ell,t,i+a} - \overline{X}_{\ell,t})/s^X_{\ell,t} \right) \left((Y_{\ell,t,i} - \overline{Y}_{\ell,t})/s^Y_{\ell,t} \right),$$

where n_a is the number of available time points when $X_{\ell,t}$ has been shifted in time by a days, and $\overline{X}_{\ell,t}$, $s^X_{\ell,t}$ are the sample mean and standard deviation of $X_{\ell,t}$ (respectively $Y_{\ell,t}$). The result is a sequence of Pearson correlations for each selected a, where we align $X_{\ell,t}$ with the values of $Y_{\ell,t}$ that occurred a days earlier. Thus, for $a > 0$, $\mathrm{CCF}_{\ell,t}(a) > 0$ indicates that $Y_{\ell,t}$ is moving together with the values of $X_{\ell,t}$ that occur a days in the future. In this case we say that $X_{\ell,t}$ is lagging $Y_{\ell,t}$. For $a < 0$, $\mathrm{CCF}_{\ell,t}(a) > 0$ means that $Y_{\ell,t}$ is positively correlated with $X_{\ell,t}$ shifted earlier, so we say that $X_{\ell,t}$ leads $Y_{\ell,t}$.

Figure S14 shows the standardized signals for the HRR containing Charlotte, North Carolina, from August 1, 2020 until the end of September. These are the same signals shown in Figure 1 in the manuscript. To define "leadingness" we compute $\mathrm{CCF}_{\ell,t}(a)$ for each $a \in \{-15, \ldots, 15\}$ using the 56 days leading up to time t (this is the same amount of data used to train the forecasters: 21 days of training data, 21 days to get the response at $a = 21$, and 14 days for the longest lagged value). The orange dashed horizontal line represents the 97.5% significance threshold for correlations based on 56 observations. Any correlations larger in magnitude than this value are considered statistically significant under the null hypothesis of no relationship. We define leadingness to be the sum of the significant correlations that are leading (those above the dashed line with $a < 0$) while laggingness is the same but for $a > 0$. In this case, there are no significant correlations, so both scores will be 0: on September 28, DV-CLI is neither leading nor lagging cases.

Figure S15 shows the correlation between laggingness and the difference in indicator WIS and AR WIS. Unlike leadingness (Figure 5 in the manuscript) there is no obvious relationship. This is heartening as laggingness should not aid forecasting performance. On the other hand, if an indicator is more lagging than it is leading, this may suggest diminished performance. Figure S16 shows the correlation of the difference in leadingness and laggingness with the difference in WIS. The pattern here is largely similar to the pattern in leadingness described in the manuscript: the relationship is strongest in down periods and weakest in up periods with the strength diminishing as we move from down to flat to up for all indicators.

In calculating the CCF and the associated leadingness and laggingness scores, we have used the finalized data, and we look at the behavior when t is the target date of the forecast. That is we are using the same data to evaluate predictive accuracy as to determine leadingness and laggingness. It should be noted that the leadingness of the indicator at the time the model is trained may also be important. Thus, we could calculate separate leadningness and laggingness scores for the trained model and for the evaluation data and examine their combination in some way. We do not pursue this combination further and leave this question for future work.

6. Examining data in 2021

In this section, we investigate the sensitivity of the results to the period over which we train and evaluate the models. In the main paper, we end all evaluation on December 31, 2020. Figures S17–S19 show how the results would differ if we extended this

analysis through March 31, 2021. Comparing Figure S17 to Figure 3 of the main paper, one sees that as ahead increases most methods now improve relative to the baseline forecaster. When compared to other methods, CHNG-CLI appears much better than it had previously; however, all forecasters other than CHNG-COVID and DV-CLI are performing less well relative to the baseline than before. These changes are likely due to the differing nature of the pandemic in 2021, with flat and downward trends much more common than upward trajectories. Indeed, the nature of the hotspot prediction problem is quite different in this period. With a 21-day training window, it is common for there to be many fewer hotspots in training.

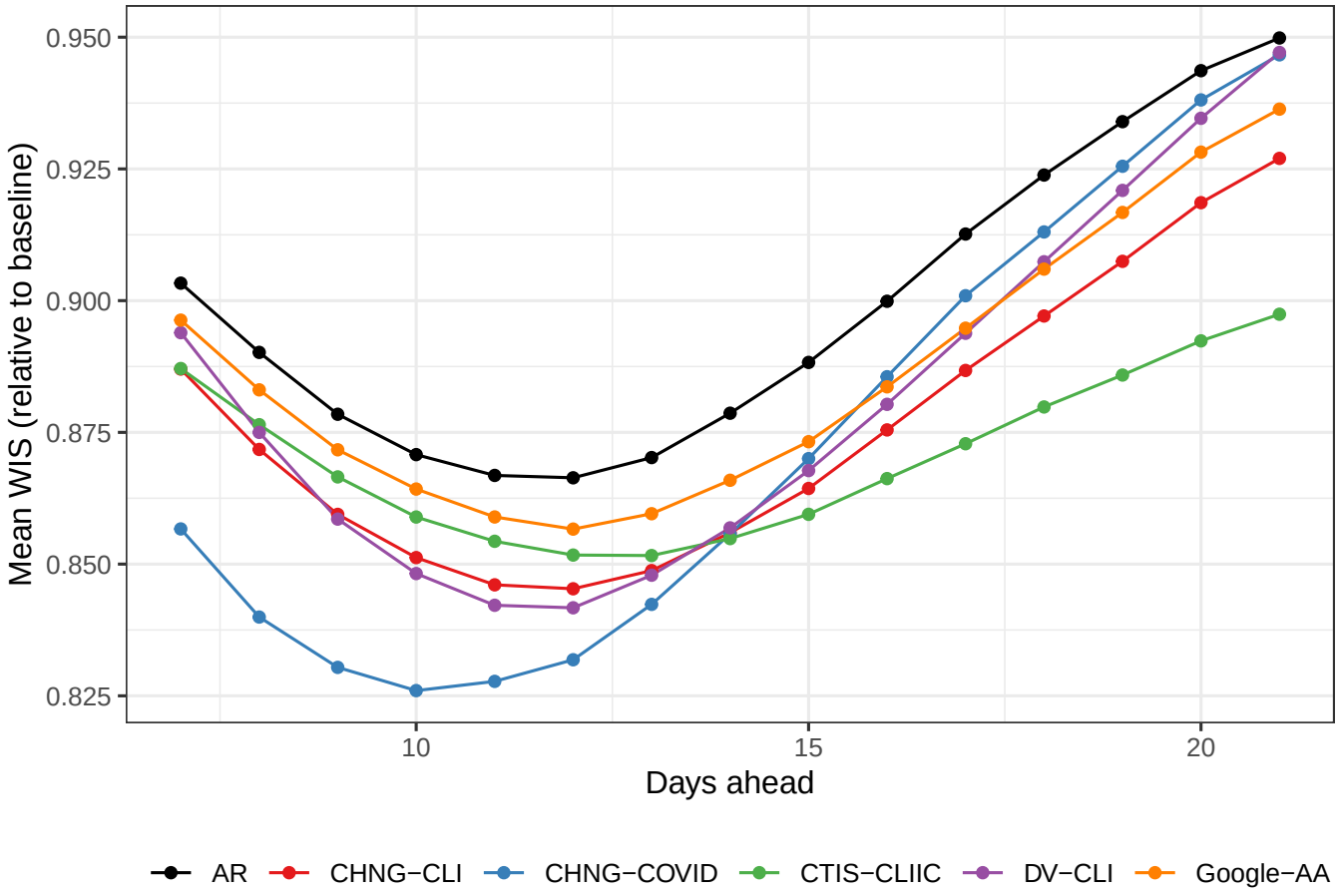

Fig. S1. Forecasting performance using finalized data. Compare to Figure 3 in the manuscript.

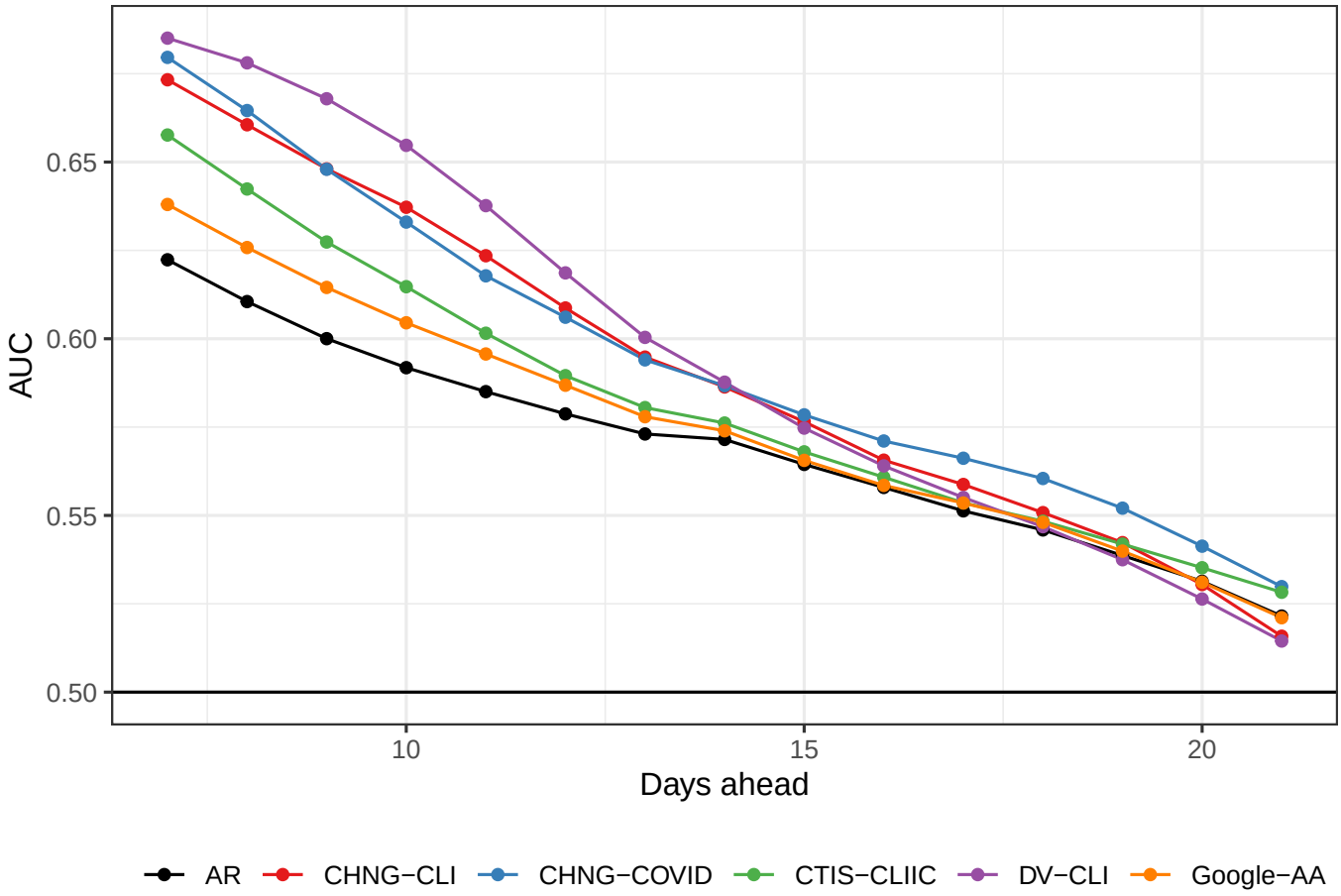

Fig. S2. Hotspot prediction performance using finalized data. Compare to Figure 4 in the manuscript.

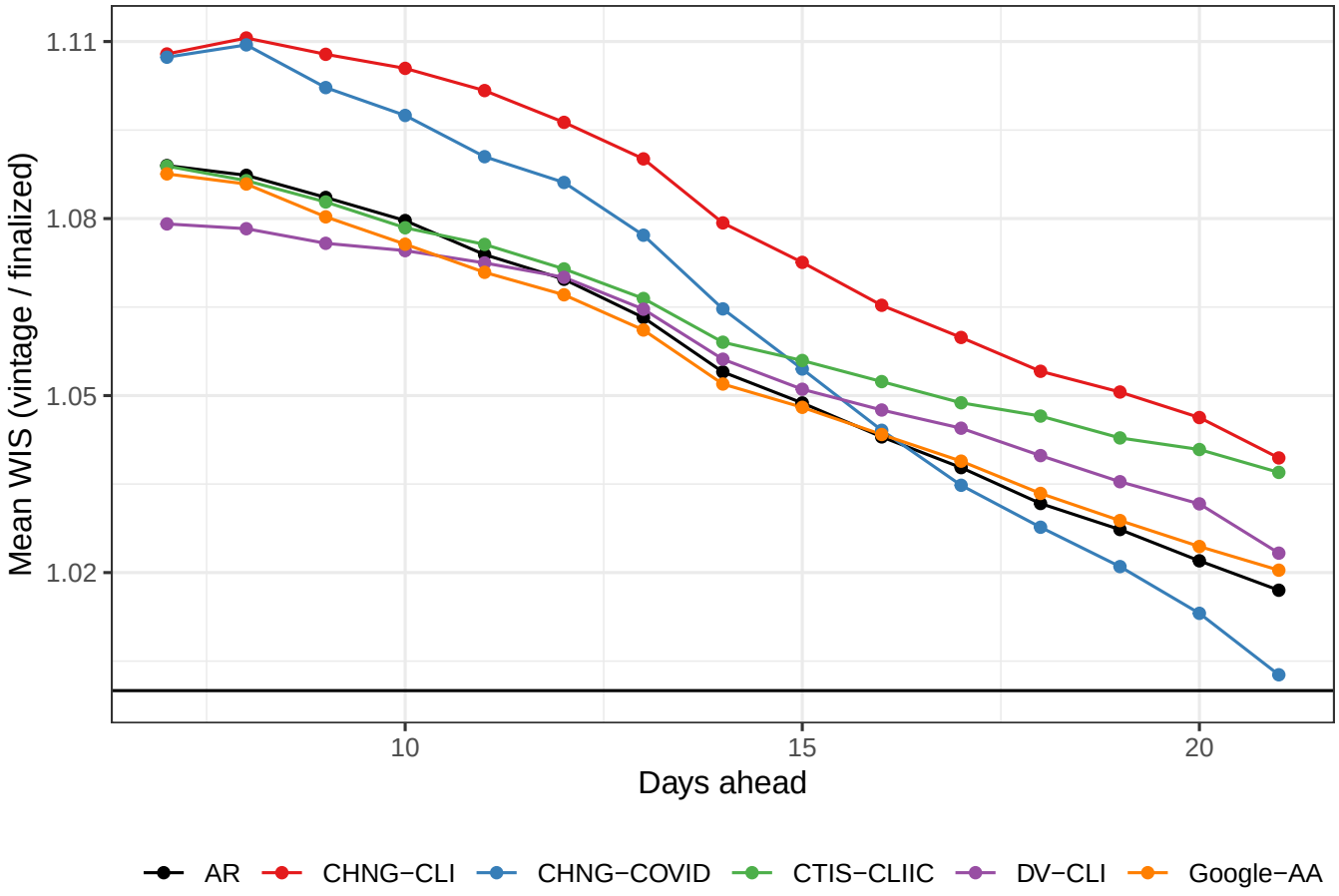

Fig. S3. Relative forecast WIS with vintage compared to finalized data. Using finalized data leads to overly optimistic performance.

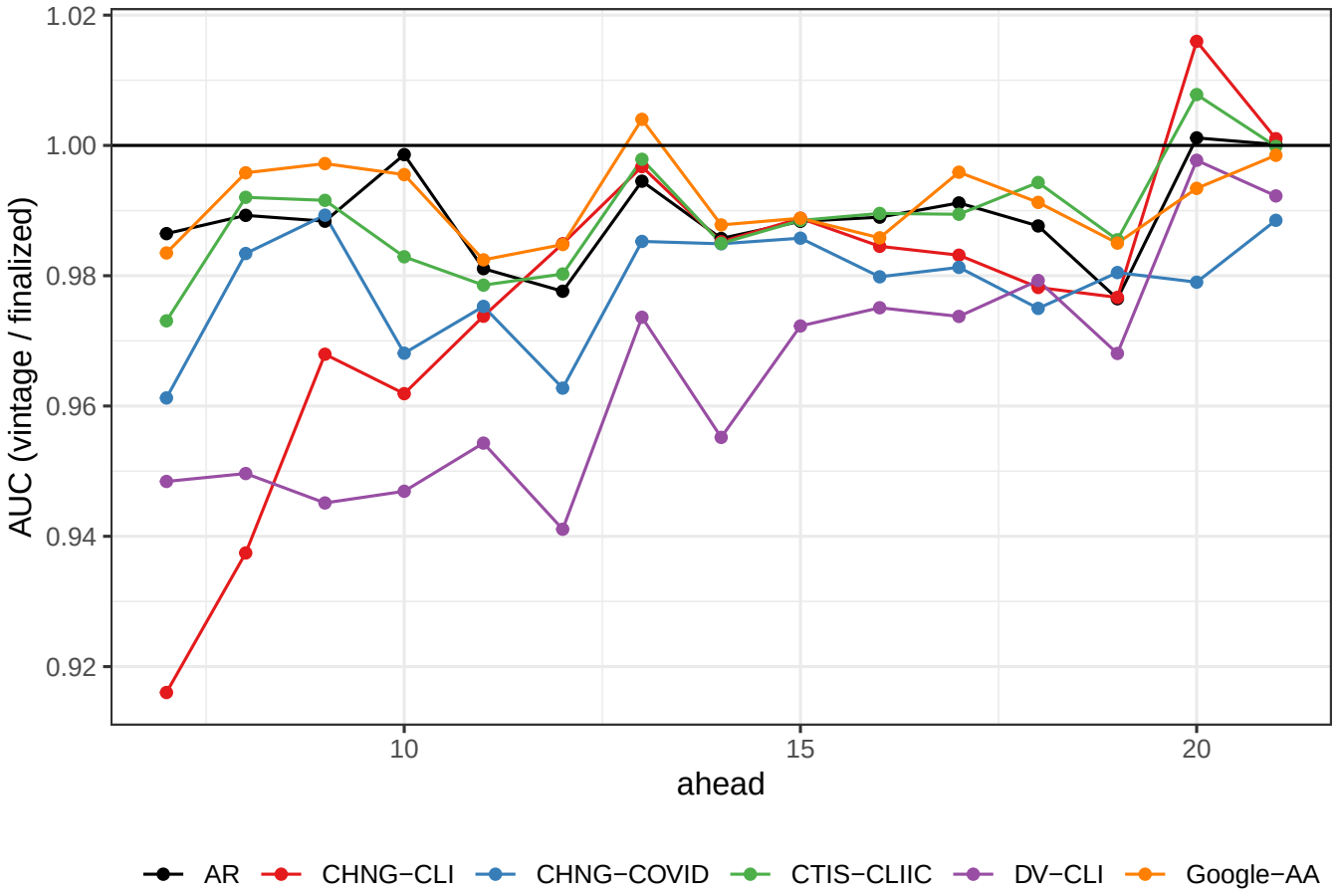

Fig. S4. Relative AUC with vintage compared to finalized data. Using finalized data leads to overly optimistic hotspot performance.

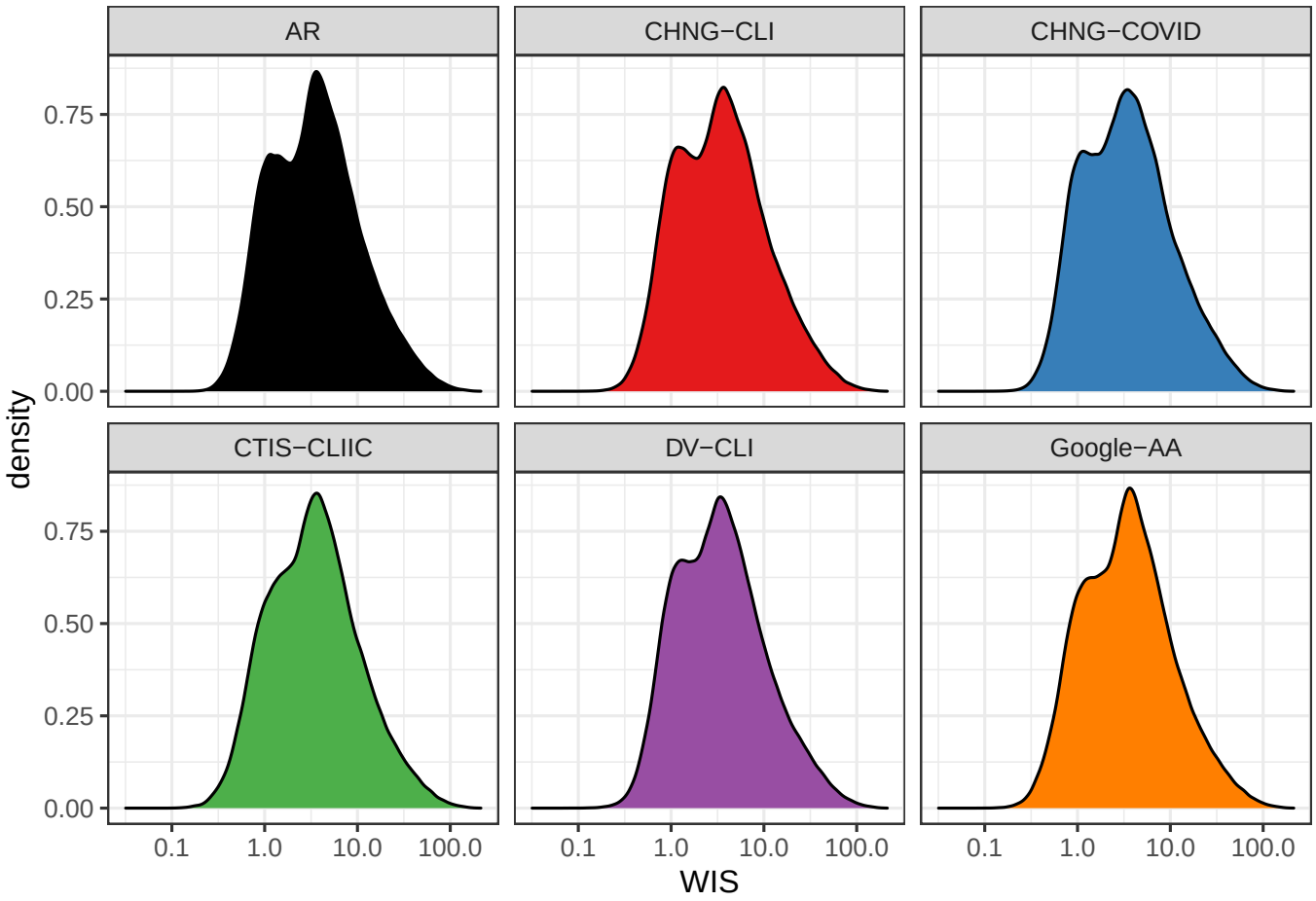

Fig. S5. Weighted interval score appears to more closely resemble a log-Gaussian distribution.

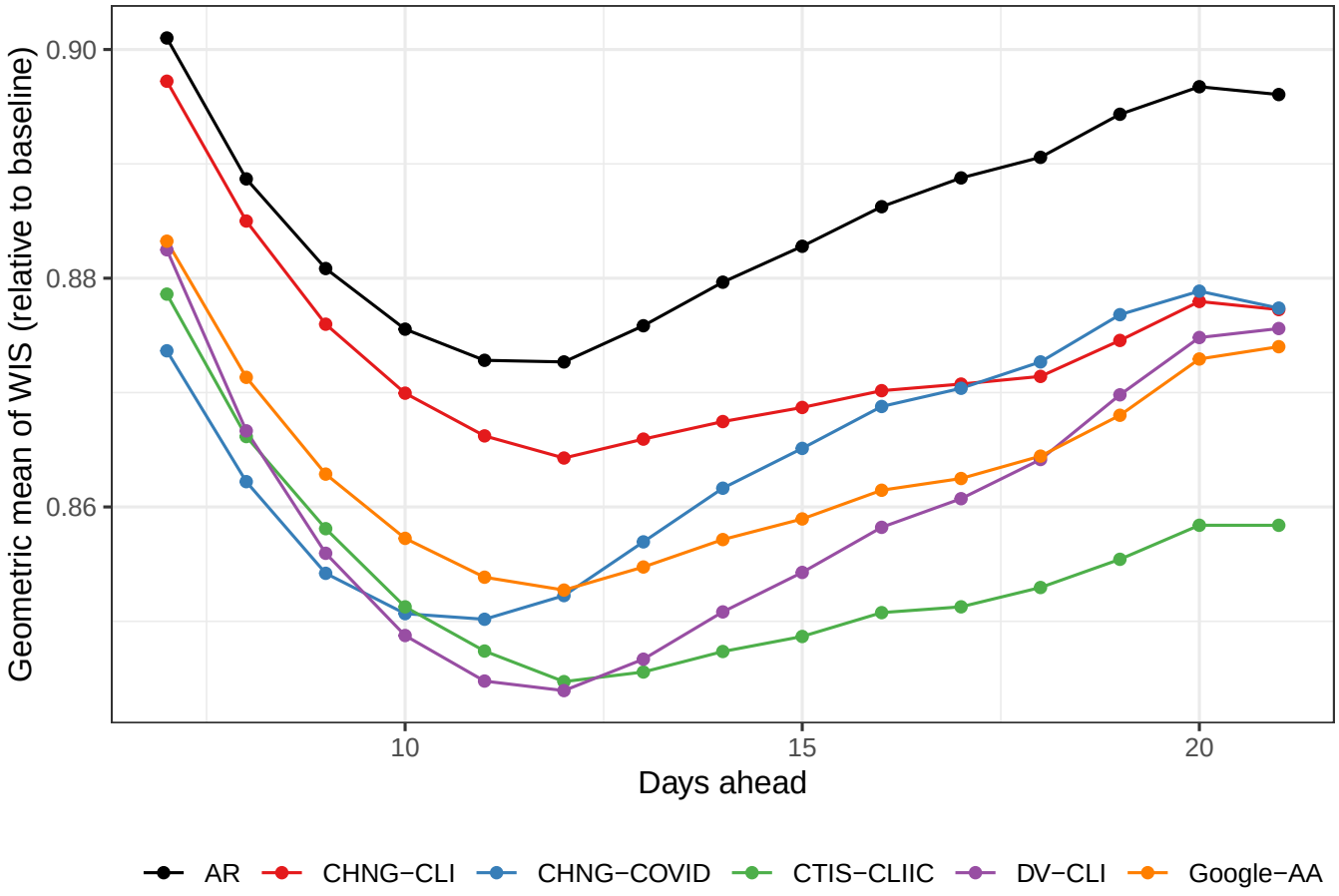

Fig. S6. Relative forecast performance using vintage data and summarizing with the more robust geometric mean.

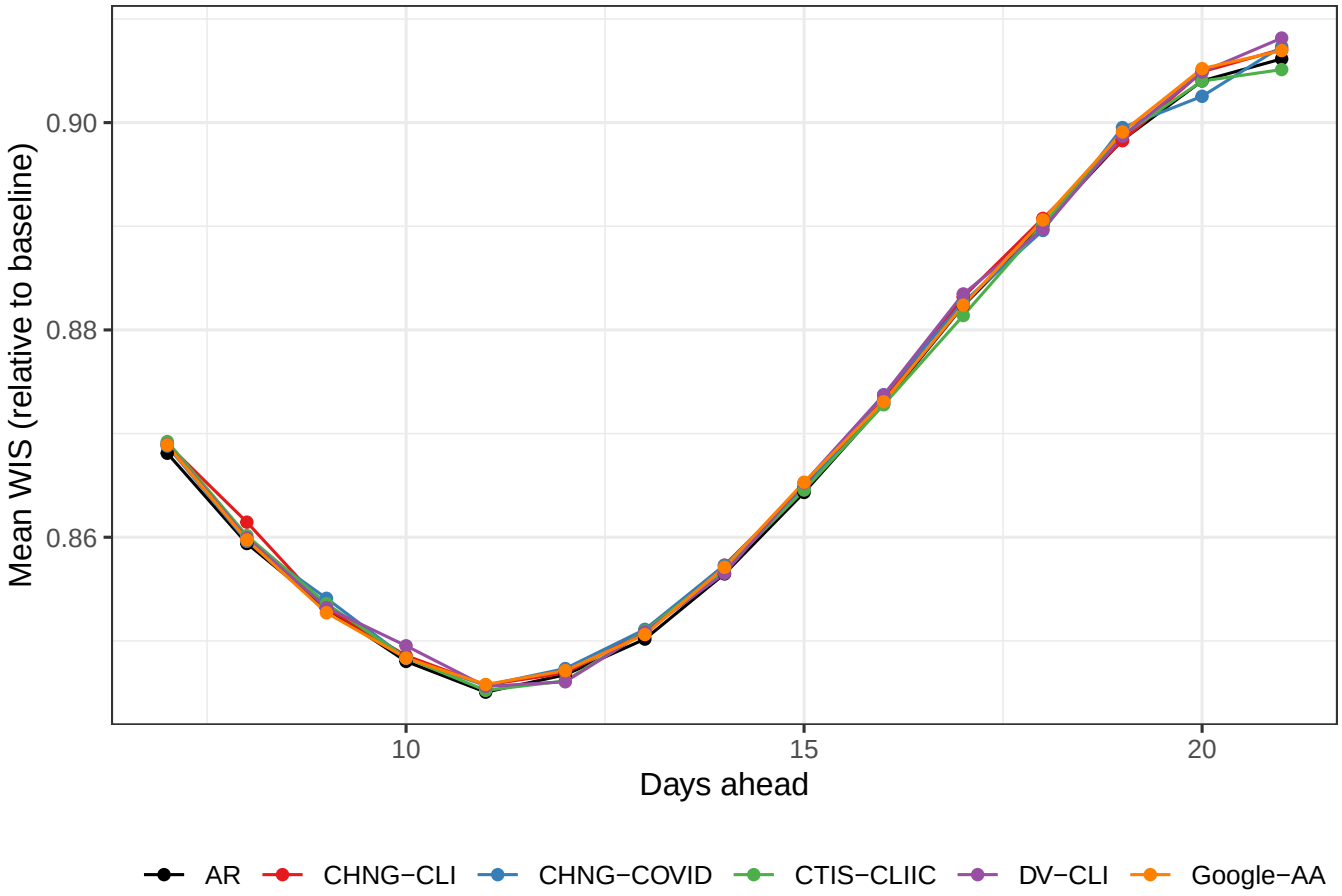

Fig. S7. Forecast performance when indicators are replaced with samples from their empirical distribution. Performance is largely similar to the AR model.

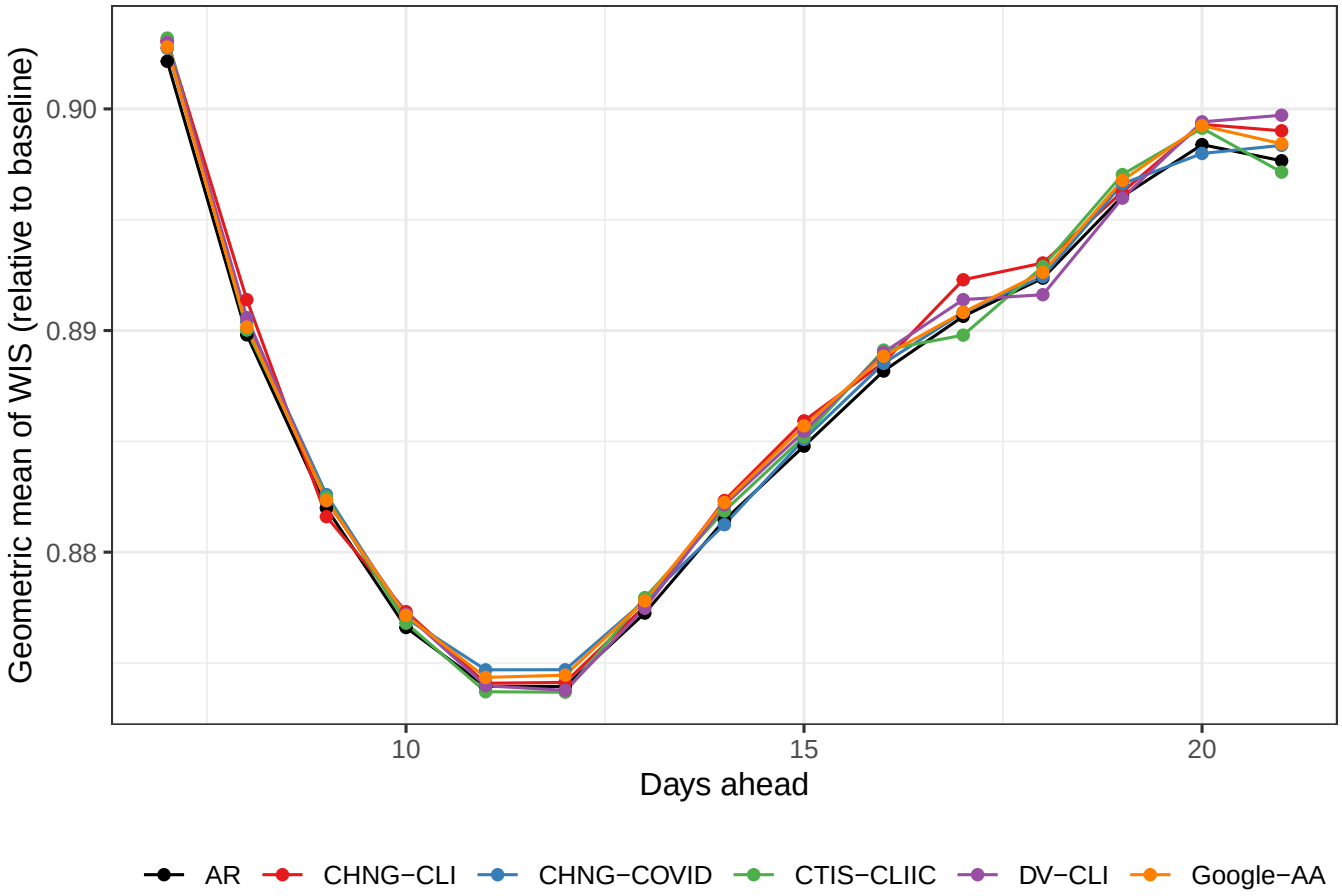

Fig. S8. Forecast performance as measured with the geometric mean when indicators are replaced with samples from their empirical distribution. Performance is largely similar to the AR model.

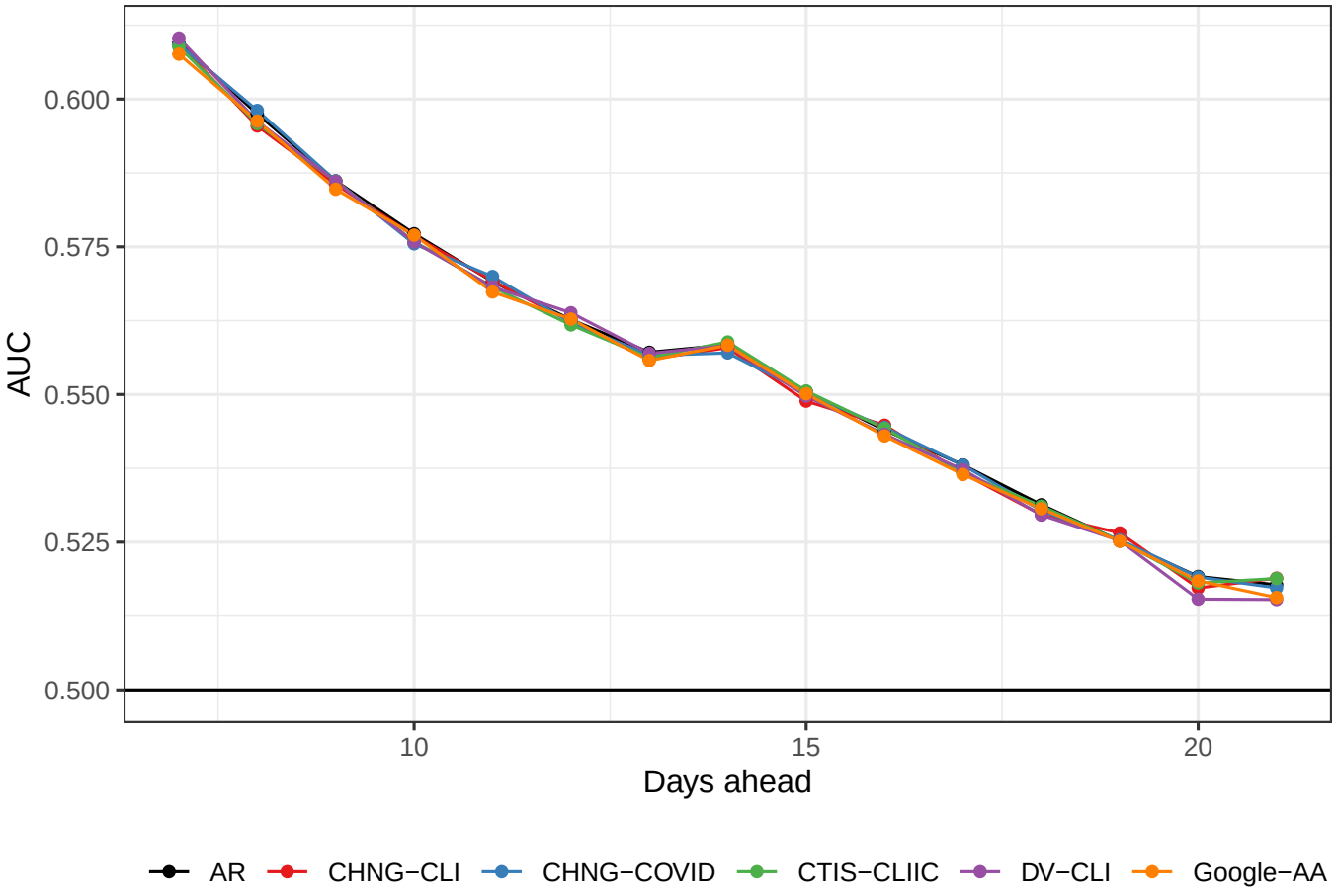

Fig. S9. Hotspot prediction performance when indicators are replaced with samples from their empirical distribution. Performance is largely similar to the AR model.

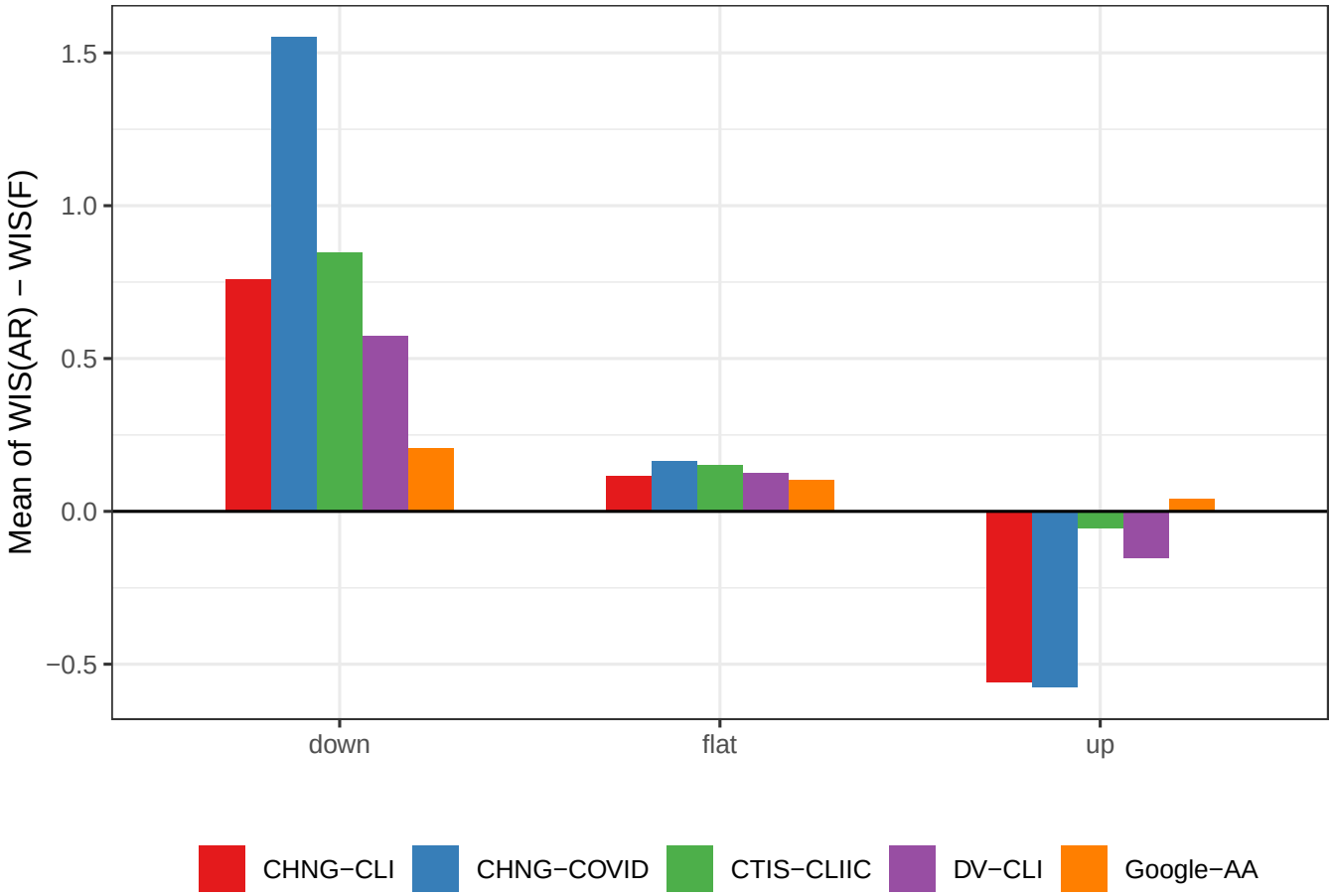

Fig. S10. Average difference between the WIS of the AR model and the WIS of the other forecasters. The indicator-assisted forecasters do best during down and flat periods.

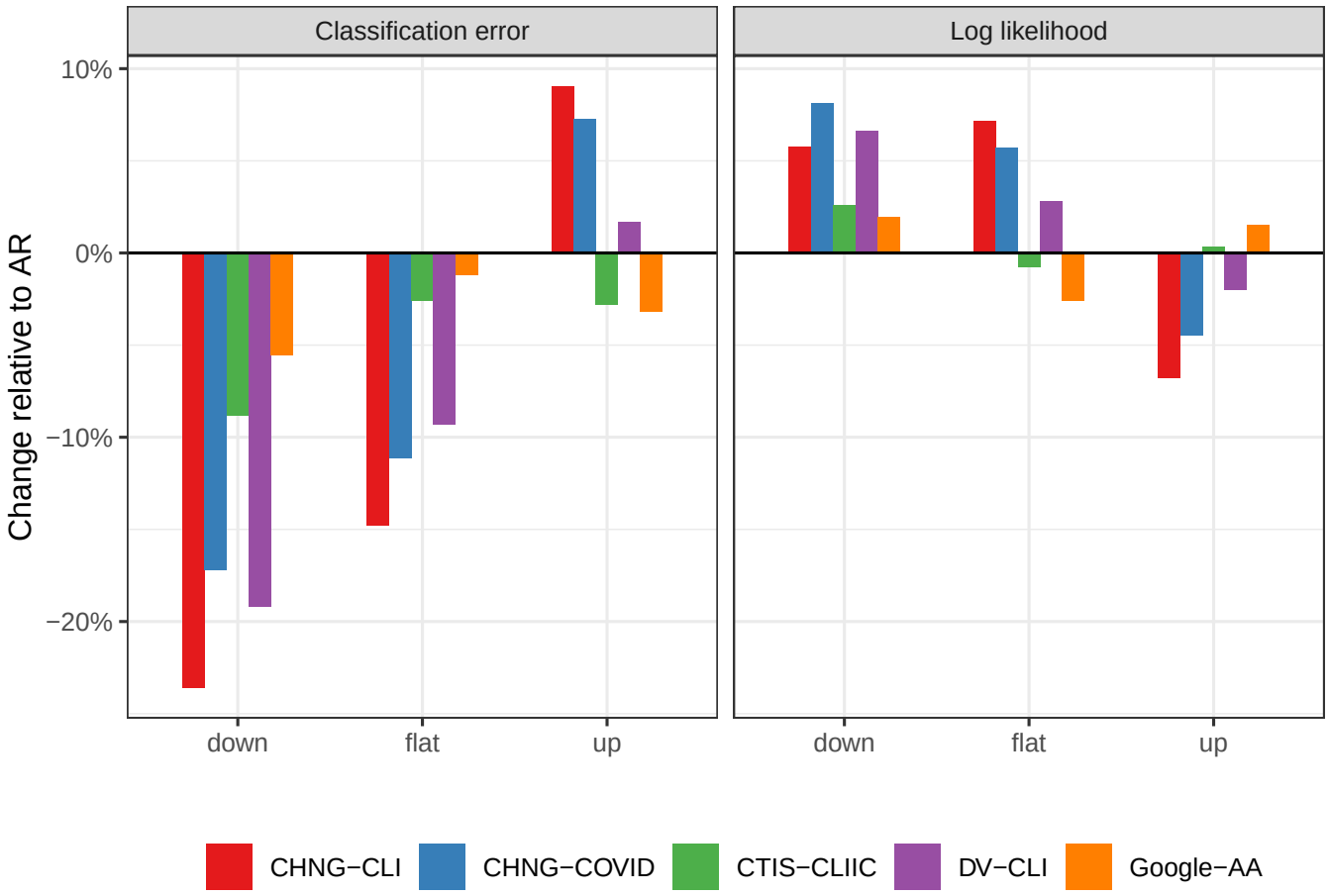

Fig. S11. Classification and loglikelihood separated into periods of upswing, downswing, and flat cases. Like the analysis of the forecasting task in the main paper (see Figure 7), performance is better during down and flat periods.

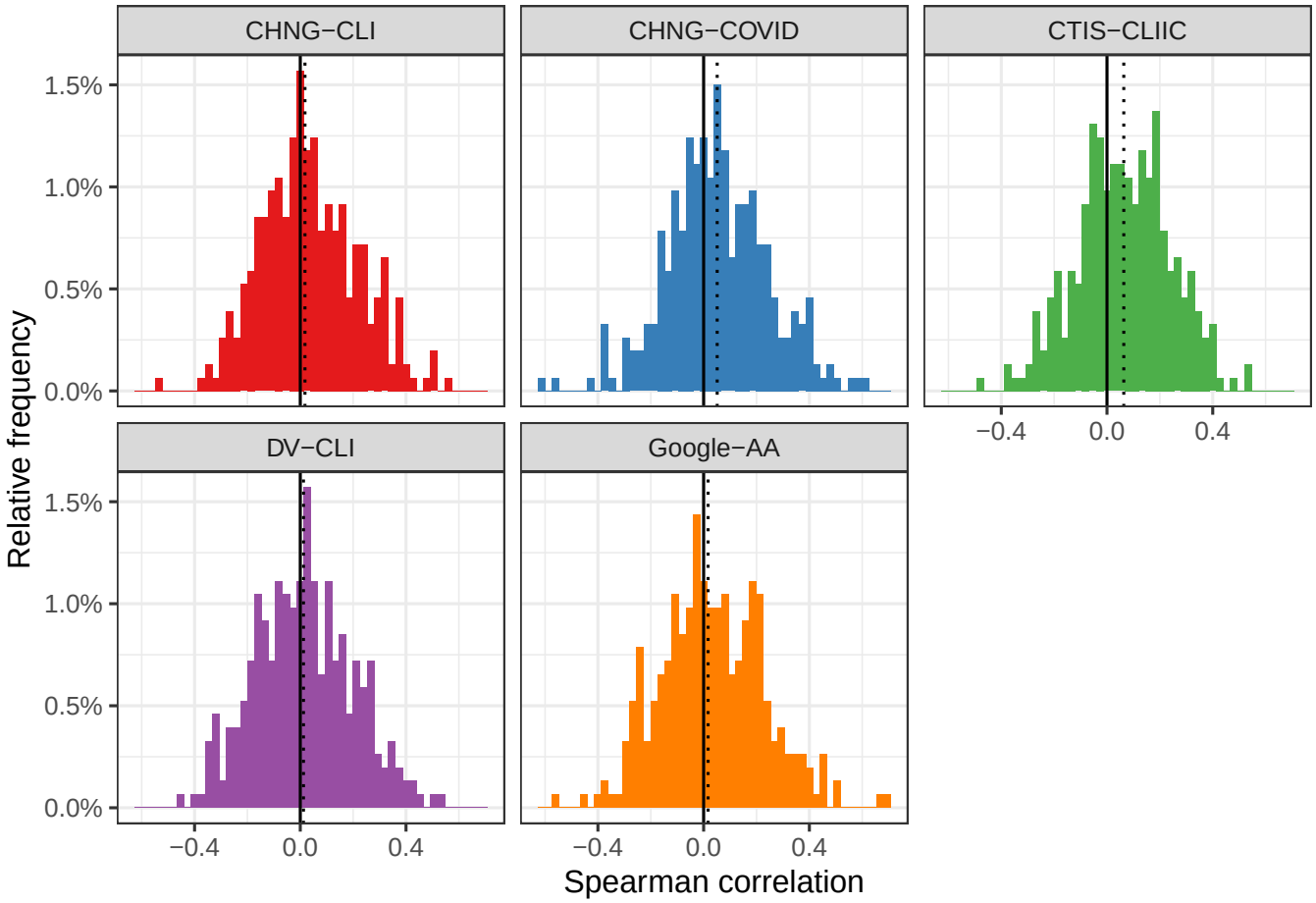

Fig. S12. Histograms of the Spearman correlation between the ratio of AR to AR WIS with the percent change in smoothed case rates relative to 7 days earlier.

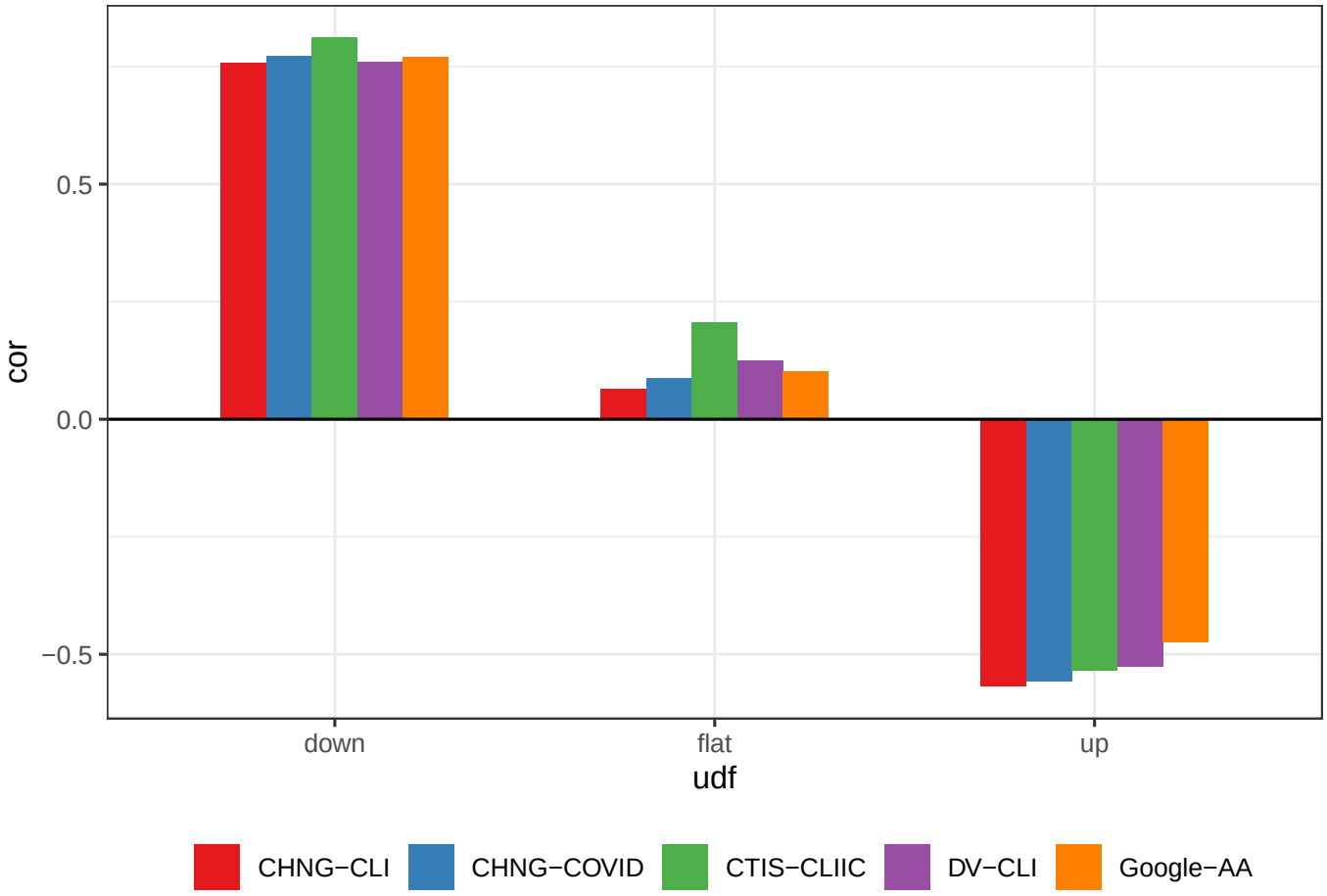

Fig. S13. Correlation of the difference in WIS with the difference in median predictions for the AR model relative to the indicator-assisted forecaster. In down periods, improvements in forecast risk are highlycorrelated with lower median predictions. The opposite is true in up periods. This suggests, as one might expect that improved performance of the indicator-assisted model is attributable to being closer to the truth then the AR model. This conclusion is stronger in down periods then in up periods.

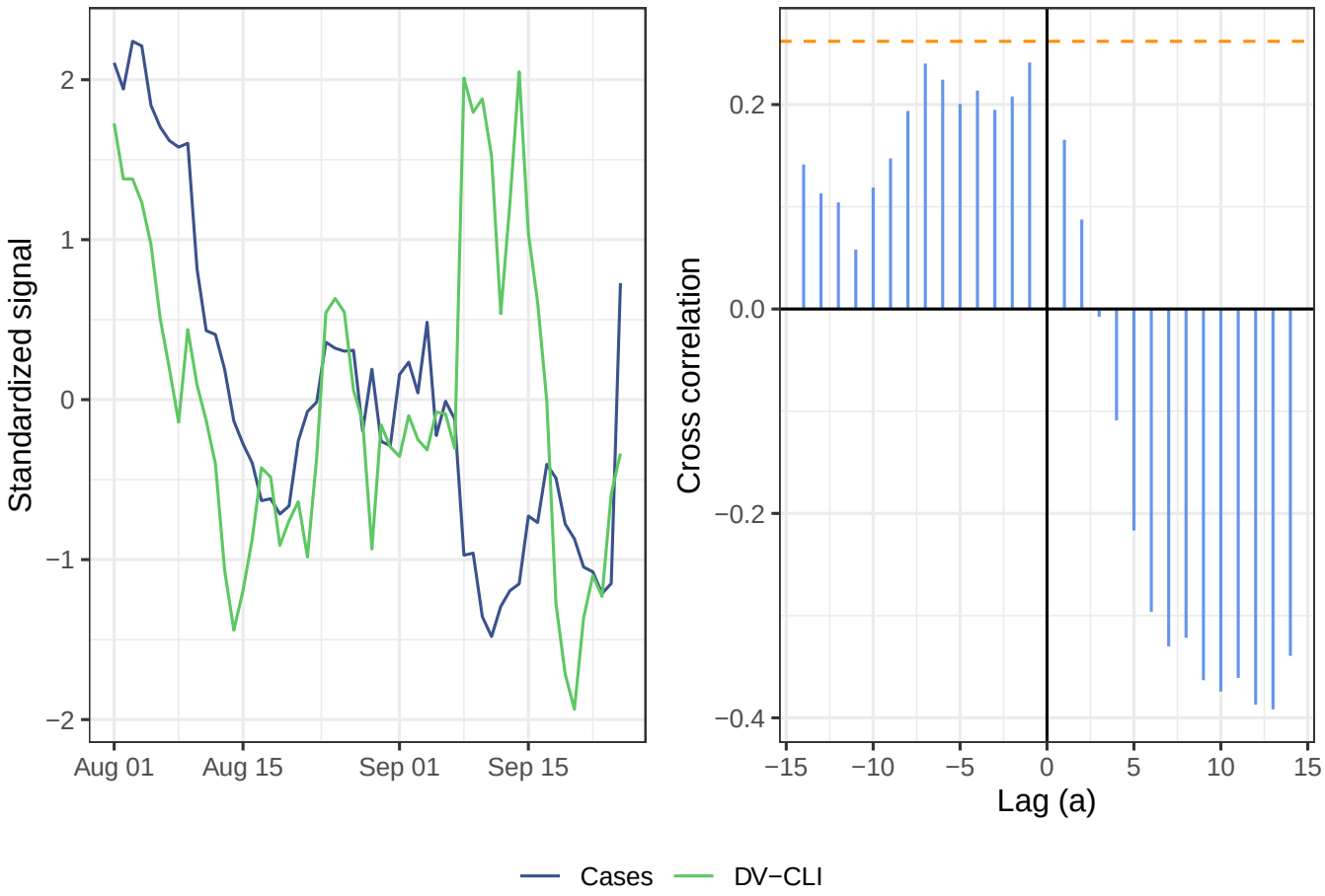

Fig. S14. Illustration of the cross-correlation function between DV-CLI and cases. The left panel shows the standardized signals over the period from August 1 to September 28 (as of September 28, 2020). The right panel shows $\mathrm{CCF}(a)$ for different values of a as vertical blue bars. The orange dashed lines indicate the 95% significance threshold. By our leadingness metric, DV-CLI is neither leading nor lagging cases over this period.

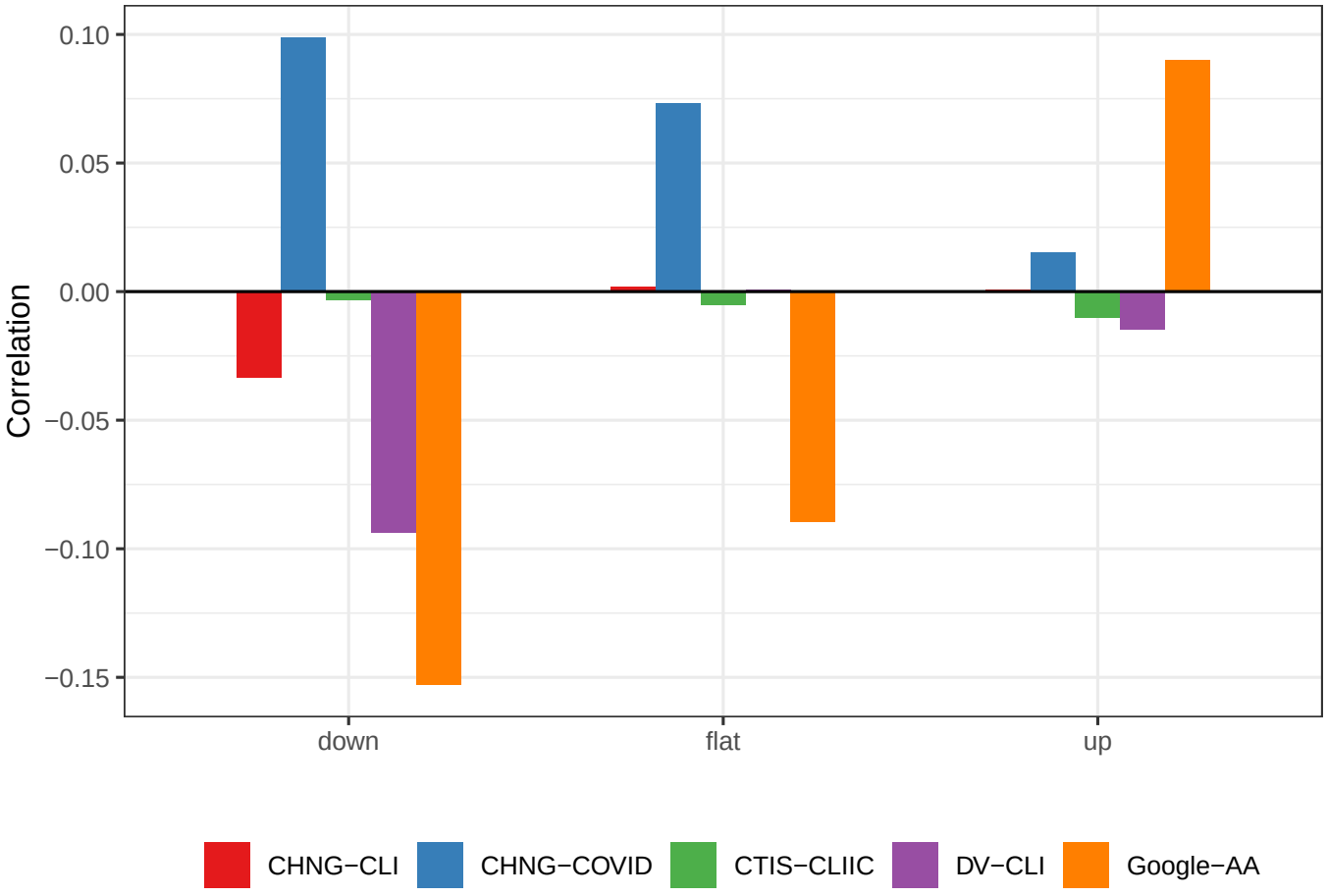

Fig. S15. Correlation of the difference in WIS with the laggingness of the indicator at the target date, stratified by up, down, or flat period. Compare to Figure 5 in the manuscript.

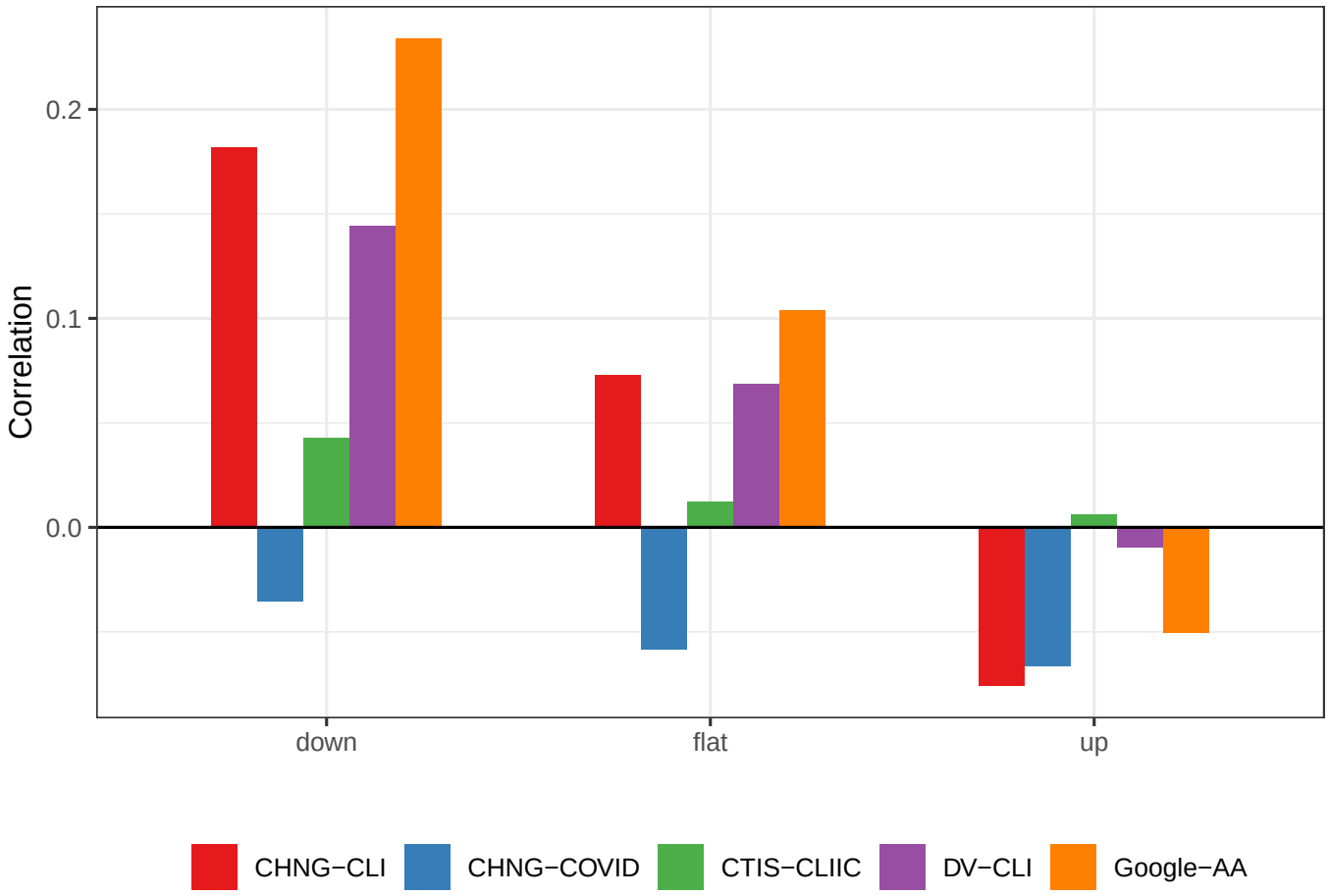

Fig. S16. Correlation of the difference between leadingness and laggingness with the difference in WIS. The relationship is essentially the same as described in the manuscript and shown in Figure 5.

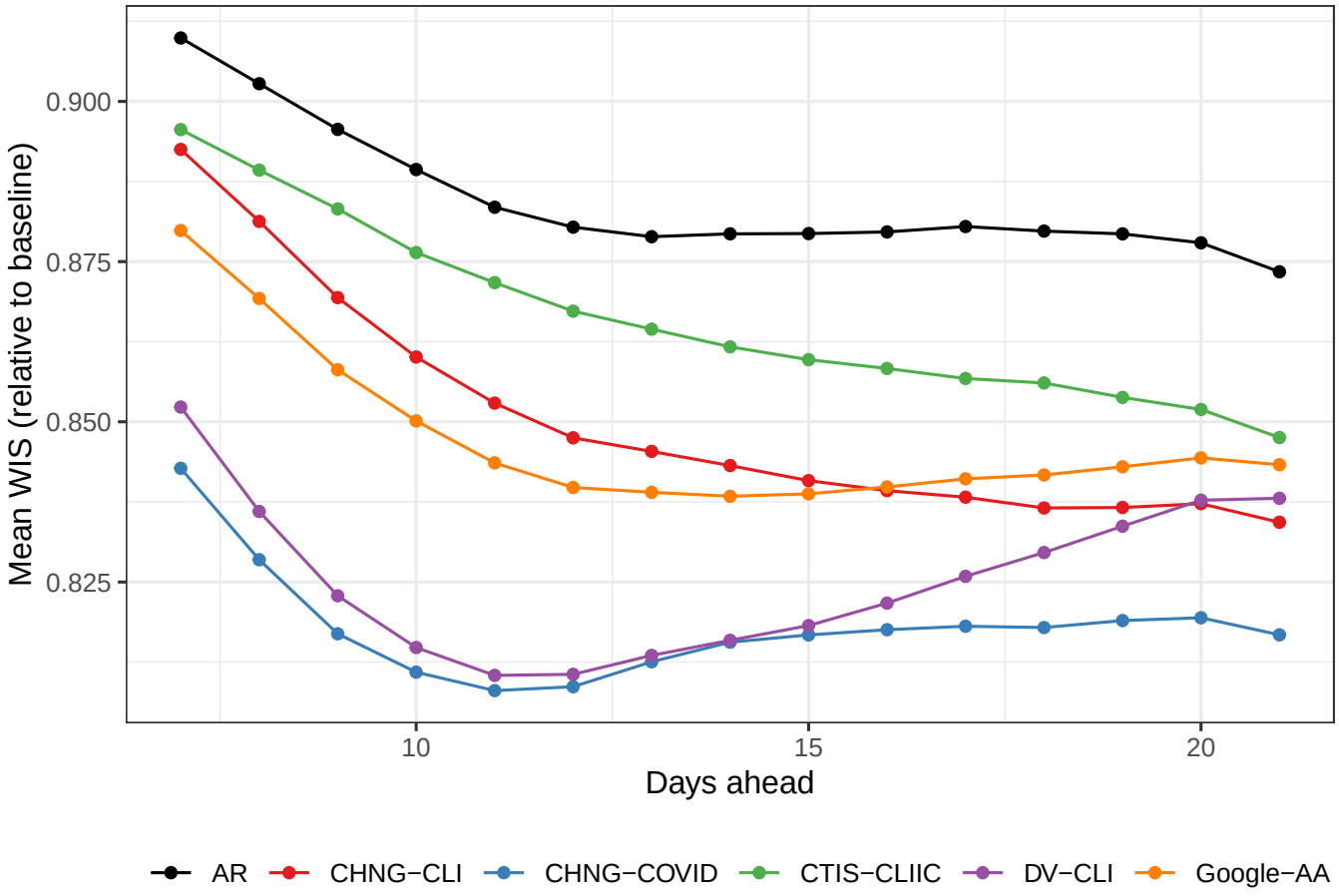

Fig. S17. Forecast performance over all periods. Performance largely improves for all forecasters with the inclusion of data in 2021.

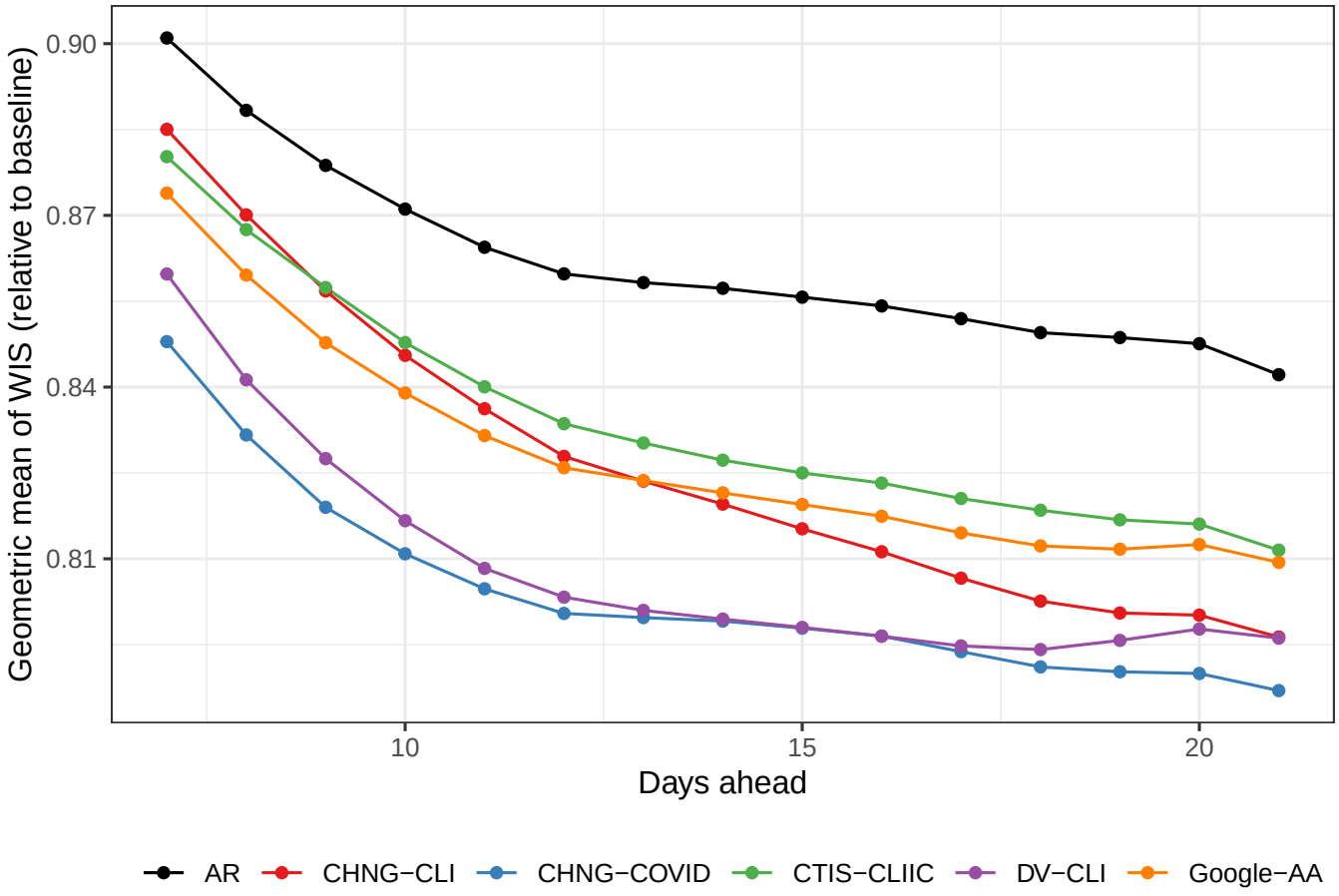

Fig. S18. Forcast performance over all periods aggregaged with the geometric mean. Again, the inclusion of data in 2021 leads to improved performance.

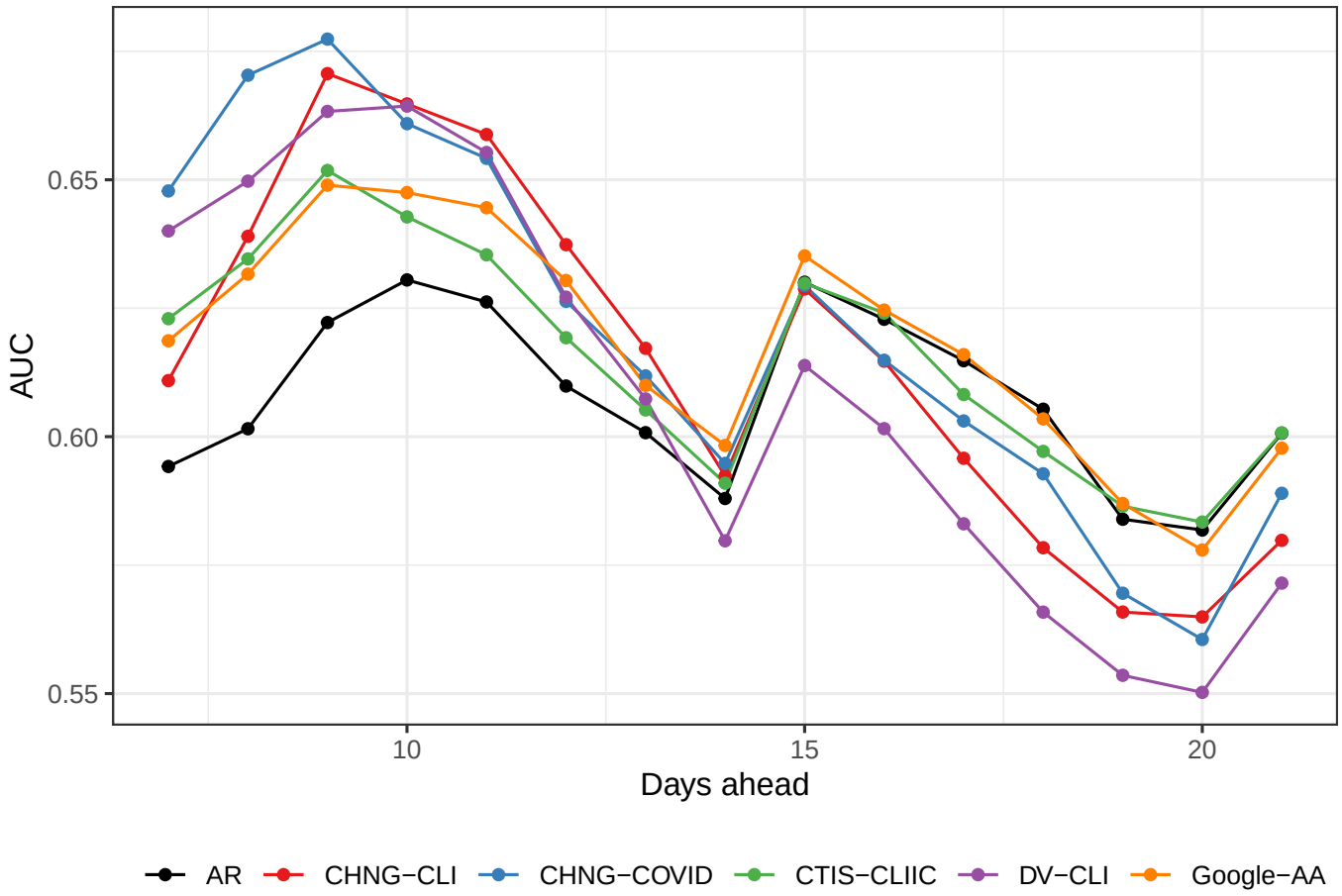

Fig. S19. Area under the curve for hotspot predictions including data in 2021. Performance degrades relative to the period in 2020. However, there are far fewer hotspots during this period as case rates declined in much of the country.

SI Dataset S1 (dataset_one.txt)

Type or paste legend here.

SI Dataset S2 (dataset_two.txt)

Type or paste legend here. Adding longer text to show what happens, to decide on alignment and/or indentations for multi-line or paragraph captions.