Can COVID-19 Indicators Improve Forecasts and Hotspot Prediction?

Author One[a,c,1], **Author Two**[b,1,2], **and Author Three**[a]

[a]Affiliation One; [b]Affiliation Two; [c]Affiliation Three

This manuscript was compiled on June 7, 2021

Please provide an abstract of no more than 250 words in a single paragraph. Abstracts should explain to the general reader the major contributions of the article. References in the abstract must be cited in full within the abstract itself and cited in the text.

Keyword 1 | Keyword 2 | Keyword 3 | ...

Tracking and forecasting indicators from public health reporting streams—such as confirmed cases and deaths in the COVID-19 pandemic—is crucial for understanding disease spread, formulating public policy responses, and anticipating future public health resource needs. In a companion paper, we detail our research group's (Delphi's) efforts in curating and maintaining a database of real-time indicators that track COVID-19 activity and other relevant phenomena. The signals (which we also call "indicators") in this database are accessible through the COVIDcast API (1), with associated R (2) and Python (3) packages for convenient data fetching and processing. In the current paper, we aim to quantify the additional utility provided by a core set of these indicators for two fundamental prediction tasks: probabilistic forecasting of COVID-19 case rates and predicting future hotspots (i.e., where and when in the future there will be a surge in cases).

At the outset, we should be clear that our intent in this paper is *not* to provide an authoritative take on cutting-edge COVID-19 forecasting methods. Instead, our purpose here is to provide a quantitative assessment of the potential of several indicators to add value in forecasting tasks. To measure such a benefit in as simple terms as possible, we work within the framework of a basic autoregressive model (in which the near future is predicted based on a linear combination of the near past), asking whether the inclusion of an indicator improves predictive power over what could be achieved with confirmed COVID-19 case rates alone. While forecasting is a rich area with a plethora of models and techniques, e.g. (**?**), we purposely constrain ourselves to very simple models, avoiding common enhancements such as order selection, correction of outliers or reporting errors in the data, inclusion of regularization or nonlinearities, or use of more complicated nonparametric models. That said, analyses of forecasts submitted to the COVID-19 Forecast Hub (4) by a large community of forecasters have shown that simple, robust models have consistently been among the best-performing (5), including time series models similar to those in this paper.

In our companion paper, we analyze correlations between various indicators and COVID case rates. While correlations and lagged correlations are natural summaries measuring the connection between an indicator and COVID case rates, they fall short of addressing the most practically relevant of questions: Is the information contained in an indicator demonstrably useful for the forecasting tasks we care about? This question about *usefulness* is in fact focused on a much higher standard than simply asking about correlations. To be useful for forecasting, an indicator must provide relevant information that is not otherwise contained in lagged versions of the case rate itself.

For example, Figure 1 shows scaled versions of the COVID-like illness indicator derived from the Delphi-Facebook survey and the COVID-19 case rates (**?**) in Los Angeles. Lagged correlations are suggestive of this indicator having a potential 6-day leading relationship in this region during this period. Indeed, two pronounced surges in the indicator seem to precede two surges in case rates. But does this behavior persist across time and location? Is the relationship stable enough to translate to reliable forecast improvement above a predictive model that uses case rates alone?

Furthermore, adopting this practically-oriented forecasting perspective to indicator evaluation requires one to confront fundamental limitations inherent to public health surveillance systems. In particular, many reporting systems are plagued by latency, with delays in reporting that can sometimes span multiple weeks. Latency manifests itself in different ways in different data sources. Some data sources may only make a signal available one week after the date it describes; other data sources may provide immediate imperfect initial estimates, which are gradually "backfilled" over a period of days, weeks, or months (in which previous estimates are updated as new data trickle in). And few of these latency patterns are stable or predictable, with holidays, extreme weather events, and changes in the reporting process all affecting the dynamics of the delay process. To fully and accurately understand an indicator's usefulness for forecasting requires accounting for the unfortunate reality of data latency.

Retrospective analyses showing the promise of a data source or a forecaster are easy to get wrong, leading to unwittingly over-optimistic assessments. The gold standard for such assessments is of course true prospective evaluation, in which

Significance Statement

Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of speciality. The primary goal of the significance statement is to explain the relevance of the work in broad context to a broad readership. The significance statement appears in the paper itself and is required for all research papers.

Please provide details of author contributions here.

Please declare any competing interests here.

[1]A.O.(Author One) contributed equally to this work with A.T. (Author Two) (remove if not applicable).

[2]To whom correspondence should be addressed. E-mail: author.twoemail.com

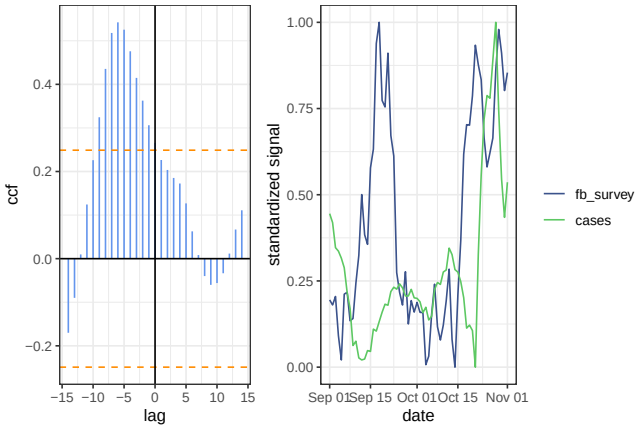

Fig. 1. Right: Normalized case rate and COVID-like illness indicator from the Facebook Survey for the Los Angeles HRR over a 2 month period. Left: cross-correlation plot for the same signals. The survey indicator seems to "lead" the observed case rate by about a week.

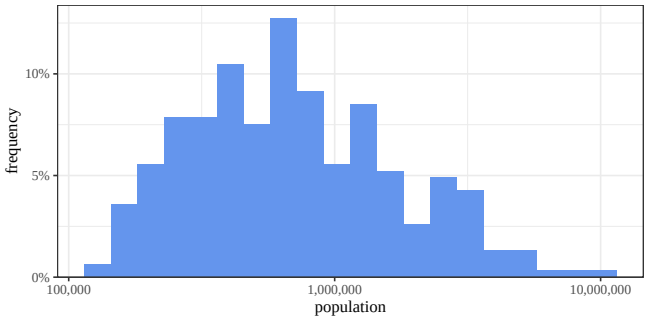

Fig. 2. Histogram of population size for HRR. Note that the x-axis is on the log scale.

all forecasts were made in real-time rather than after the fact in backtesting. To accurately evaluate a forecaster (or in this case an indicator for forecasting) in a retrospective analysis requires great care. In particular, it requires carefully accounting for "what was known when." In our companion piece, we describe the data versioning system developed by our group's API (1). In this present paper, we take this a step further, making extensive use of an R package we have developed, called `evalcast` (), which leverages this data versioning to enable honest backtesting of COVID-19 forecasters.

Examining the importance of additional features for prediction is a core question in inferential statistics, and there are many ways to examine this question under a variety of assumptions. For time series, work in this area can be traced at least to Granger causality (**?**) and related multivariate extensions (**?**) under the condition that the series are linearly related autoregressions among other assumptions. These techniques are especially popular in econometrics and neuroscience (**? ?**).

We choose to take a predictive angle, which is simple but also fairly robust, compared to various other classical inferential approaches. In fact, drawing rigorous inference based on predictions, without (or with lean) assumptions, is an active field of research (**? ? ? ?**). Nonetheless, the predictive approach directly examines the target of interest: do these signals help to produce forecasts? Furthermore, directly comparing forecast errors without strong assumptions through tests of predictive accuracy are common in epidemiology, econometrics, and other fields (**? ?**).

1. Methods

A. Signals and Locations. The indicators we focus on provide information not generally available from standard public health reporting. The [ATTN: COVIDcast Survey via Facebook] reaches roughly 50,000 respondents across the United States daily, and provides information on symptoms, social distancing, and other attitudes and behaviors related to COVID. The particular indicator we examine, which we refer to as `fb`, is the estimated percentage of COVID-like illness within the local community. This indicator is calculated by examining the percent of individuals who report knowing someone else [ATTN: in their community?] who has been diagnosed with COVID. In some ways, this can be seen as a proxy for the strength of social contacts between infected and susceptible populations, a key feature for disease modeling (6) The Google indicator, which we refer to as `gs`, tracks the local intensity of Google searches related to lack-of-smell (anosmia) or lack-of-taste (ageusia), both of which have been connected to COVID19 (7, 8). Two indicators are derived from de-identified insurance claims data from Change Healthcare, a large healthcare technology company that aggregates data from many providers. The first signal, which we refer to as `chng_cov`, is the estimated percentage of outpatient doctor visits with confirmed COVID-19 diagnosis (potentially collected earlier than the related report to the local health authority), while the second, which we refer to as `chng_cli`, is the estimated percentage of outpatient visits with COVID-like illness. Finally, [ATTN: describe `dv` indicator, perhaps best described in conjunction with `chng_cov`]. The hope is that each of these signals serves as a "leading indicator" of future COVID cases, improving the forecastibility of COVID cases and hotspots. For more details on these and other indicators, see the companion paper.

As for geographic resolution, we consider prediction of COVID case rates (and hotspots) aggregated at the level of an individual hospital referral region (HRR). HRRs correspond to groups of counties within the same hospital referral system. The Dartmouth Atlas of Healthcare Policy (**?**), defines these 306 regions based on a number of characteristics. They are contiguous regions containing at least one city where major cardiovascular or neurological precedures are performed. Each HRR has a minimum population of about 120,000, and the majority of hospital services for that population are performed by hospitals within the region. Some HRRs are quite large (Los Angeles has about 10 million), but generally HRRs are much more homogenous than the ∼3100 counties would be but at much higher resolution than states. They would be more relevant for hospitalization forecasting, given their definition. We have chosen to focus on forecasting cases at this resolution because the indicators should provide a bigger boost. Once we observe case information, there should not be much dependence of future hospitalizations on the indicators, but predicting cases should provide more timely warnings of future hospitalization needs as opposed to waiting for the case numbers.

B. Vintage Training Data. In this manuscript, all models are estimated using only "vintage training data". This means

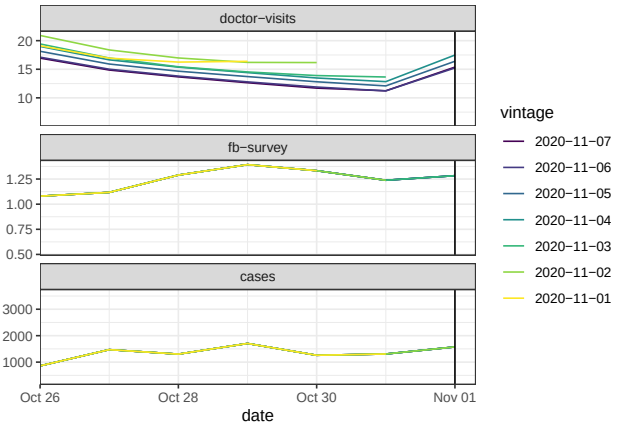

Fig. 3. Changes in availability and revisions for three signals in the Los Angeles HRR. The vertical black line indicates a hypothetical forecast date of November 1 while the colored lines get progressively darker as we collect later and later vintages of data.

that for a given forecast date, say November 1st 2020, we estimate our models using only the data that would have been available to us on November 1st. Essentially, the training data is rewound appropriately to dates in the past. It is likely, for example, that no data would actually exist for the forecast date on the forecast date because there is some latency between the time signals are reported and the dates for which they are reported. Similarly, any future revisions made to the indicators or case numbers would not have been present on the forecast date, and so would not be available to the forecaster. Figure 3 shows three signals for the Los Angeles HRR with a hypothetical forecast date of November 1. On November 1, no signals are available for October 31. The cases data is one day latent while the Facebook survey is 2 days latent. Doctor's visits are 3 days latent while also exhibiting revisions. By training our models in this way, the evaluations are more realistic in the sense that we are accurately mimicking the difficulty faced by real-time epidemic forecasters. The COVIDcast API maintains all versions of the data, thereby allowing these sorts of synthetic forecasting exercises. For more details about data versioning, see the companion paper.

C. Analysis Tasks. On every day between [ATTN: July 1 and December 31], we make two types of forecasts for each of the 306 hospital referral regions in the United States: (1) we forecast the seven-day trailing average of incident cases, or "smoothed incident cases"; and (2) we predict whether the HRR will be a hotspot, a location with a large increase in cases. For both tasks we try to predict the outcome 1, 2, 3, up to [ATTN: 28] days into the future. For both tasks, we estimate a baseline global model across all HRRs using the most recent observation of the target, as well the observations from 1 and 2 weeks earlier. We train a separate model for each forecast date and for each forecast horizon, but we do not assume any HRR-specific effects; that is, a single global model is used to describe all HRRs for each forecast date-horizon pair. This model is re-estimated for every forecast date and each future target resulting in [ATTN: $180 \times 28 = 5040$] estimated baseline models for each task. We then add current and lagged values (analogous to the baseline) for each of the four signals separately. We now provide further details on each of these tasks.

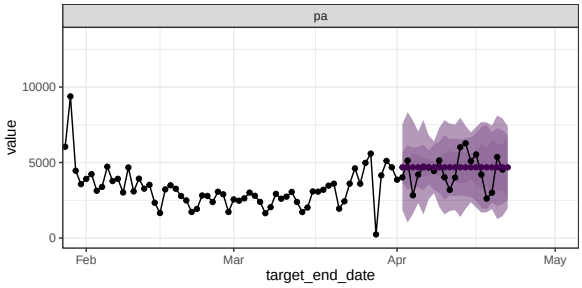

Fig. 4. Forecasts and observed cases for HRR x. [ATTN: Create a figure here for 1 HRR. Show the fan for the quantiles.]

C.1. Task 1: Case Forecasting. To fix notation, let $Y_{\ell,t}$ denote the smoothed COVID-19 case incidence rate per 100,000 people for location (HRR) ℓ and time (day) t. Let $X_{\ell,t}$ denote the one of the indicators described above, for example the Facebook % CLI-in-community signal, for location ℓ and time t. We evaluate five models of the form:

$$Y_{\ell,t+d} \approx \alpha + \sum_{j=0}^{2} \beta_j Y_{\ell,t-7j}$$

$$Y_{\ell,t+d} \approx \alpha + \sum_{j=0}^{2} \beta_j Y_{\ell,t-7j} + \sum_{j=0}^{2} \gamma_j X_{\ell,t-7j},$$

where we also exchange $X_{\ell,t}$ for the signals from [ATTN: CHNG and Google.] Here $d = 1, \ldots, 28$ depending on the target value (number of days ahead we predict).

Informally, the first model, which we will call the "Cases" model, bases its predictions of future case rates on the following three features: current COVID-19 case rates and those 1 and 2 weeks back. The second model, "Cases + Facebook", additionally incorporates the current Facebook signal and the Facebook signal from 1 and 2 weeks in the past. The third model, "Cases + Google", is exactly the same but substitutes the Google signal instead of the Facebook one and similarly for "Cases + Change".

For each model, in order to make a forecast at time t_0 (to predict case rates at time $t_0 + d$), we estimate the unknown parameters using quantile regression (10), training over all locations ℓ, and all times t that are within the most recent 28 days of data available up to and including time t_0. Rather than predicting only the median (or mean) value for each forecast date and each horizon, we estimate seven quantiles, coinciding with the county-level forecast submission guidelines for the COVID-19 Forecast Hub (5, 11). These are {0.025, 0.1, 0.25, 0.5, 0.75, 0.9, 0.975}. This provides a notion of uncertainty in the forecasts. An example of this type of forecast is shown in Figure 4.

C.2. Task 2: Hotspot Prediction. We also examine whether alternative indicators help in predicting hotspots. On a given date, we say an HRR is a hotspot if the average number of observed cases over the past 7 days has increased by at least 25% compared to the preceding week. We remove all data where, on average, fewer than 30 cases were observed over the preceding 7 days. Thus, hotspot status can be represented by a $\{0, 1\}$-valued variable, and hotspot prediction is a binary classification problem.

As in the case forecasting task, in hotspot prediction we train models that use previously observed values of cases and alternative indicators as features to predict a response (in our case, hotspot status). In addition to the differently defined response, there are two other differences between the hotspot prediction and case forecasting tasks. First, in hotspot prediction we always use relative changes in previous observed cases/indicators as features. That is, rather than $Y_{t,\ell}$ we use

$$Y^{\Delta}_{\ell,t} = \frac{Y_{\ell,t} - Y_{\ell,t-7}}{Y_{\ell,t-7}} \quad \text{and} \quad X^{\Delta}_{\ell,t} = \frac{X_{\ell,t} - X_{\ell,t-7}}{X_{\ell,t-7}}$$

Second, because hotspot prediction is a binary classification problem, we estimate the parameters α, β and γ by logistic regression rather than quantile regression.

D. Evaluation Metrics. We evaluate forecasts for each forecast date and horizon using the weighted interval score. Weighted interval score (WIS), a well-known quantile-based approximation of the commonly-used continuous ranked probability score (12). WIS is a proper score, meaning that lower scores imply better forecasters, and can be thought of as a distributional generalization of absolute error. The CDC has embraced this metric in the context of epidemic forecasting and especially for COVID-19 (**?**). Essentially, pairs of quantile forecasts are treated as a central interval. For example, the 0.1 and 0.9 quantiles create an 80% interval that should, if the forecaster is well-calibrated, contain the truth about 80% of the time. A forecaster whose 80% interval contains the truth less frequently is overly confident, while one whose interval contains the truth more frequently (say, by using an infinitely-wide interval) is failing to take a stand on the future outcome.

Here, denoting K predicted intervals along with the median as $\{\alpha_K/2, \ldots, \alpha_2/2, \ \alpha_1/2, \ 0.5, 1-\alpha_1/2, \ 1-\alpha_2/2, \ldots, 1-\alpha_K/2\}$, the associated forecasts as $\{l_K, \ldots, l_2, \ l_1, \ m, \ u_1, \ u_2, \ldots, u_K\}$, and the observed value (after the forecast was made) as y, this metric can be expressed as [ATTN: someone please check this]

$$\text{WIS}_K = \frac{1}{K}|m-y| + \frac{2}{K+1}\sum_{k=1}^{K}\frac{1}{\alpha_k}(u_k - l_k) + \sum_{k=1}^{K}(l_k - y)\mathbf{1}\left(y < l_k\right) + (y - u_k)\mathbf{1}\left(y > u_k\right),$$

where $\mathbf{1}\left(a\right) = 1$ if a is true and 0 otherwise. Essentially, this metric decomposes into 4 parts: (1) a penalty for the median, the point forecast or best guess, being far from the truth; (2) a penalty for the width of each interval $(u_k - l_k)$; (3) a penalty for overprediction, when the observed value falls below l_k and hence underneath the interval; and (4) a corresponding penalty for underprediction. Weighted interval score can also be written in terms of an average of the quantile or "pinball" losses. Setting $\tau_1 = \alpha_K/2, \ \ldots, \ \tau_{K+1} = 0.5, \ \ldots, \ \tau_{2K+1} = 1 - \alpha_K/2$ and similarly rewriting $q_1 = l_K, \ \ldots, \ q_{K+1} = m, \ \ldots, \ q_{2K+1} = u_K$, then

$$\text{WIS}_K = \frac{2}{2K+1}\sum_{k=1}^{2K+1}\left(\mathbf{1}\left(y \le q_k\right) - \tau_k\right)\left(q_k - y\right).$$

While this second expression does not have the convenient decomposition discussed above, the pinball loss is precisely the objective function that quantile regression minimizes. So, mathematically, quantile regression should give parameter estimates that minimize the in-sample weighted interval score.

When comparing the aggregate performance of the baseline quantile autoregression without additional signals with the indicator-assisted quantile autoregression forecasters, we first normalize the individual weighted interval scores by the weighted interval score of a boneheaded model: predicting all future values to be the same as the most recent observation. The dummy model uses the errors such a forecast has made in-sample to form the other quantiles. We normalize by the dummy model to account for the fact that different locations or time periods have different magnitudes of COVID-19 incidence (even after scaling by population). Performing well by this metric means that a particular forecaster is doing well at all locations and time periods. Without the scaling, a forecaster that is good at locations and time periods with relatively high case numbers but poor everywhere else could appear preferable.

For hotspots, we use the area under the curve (AUC) computed by examining the true-positive and false-negative rates across HRRs at each horizon and forecast date. That is, for a particular forecast date and a particular horizon, we ask, how many true hotspots were called hotspots compared to the total number of hotspots (specificity). The false-negative rate (sensitivity) is the ratio of correctly classified non-hotspots to the total number of non-hotspots. A perfect classifier would have both sensitivity and specificity equal to 1. Plotting the sensitivity on the y-axis and 1−specificity on the x-axis where each point is a forecast date produces an ROC curve for each classifier. We compute the area under this curve for each forecast horizon to compare the cases-only model to those adding indicators.

[ATTN: Methods section graveyard, most for Appendix]

- Discuss the "how many days does the indicator buy" interpretation here or below?
- Do we explain why regularization/CV not really helpful for both forecasting and hotspots
- Implementation was via quantgen, cite here?
- some people don't like relative WIS beacuse it's no longer a "proper score". directly address that here?
- Do we talk about log loss?
- Do we explain that we are using infinite-backfilled data?

2. Results

A. Primary Results. Figure 5 displays the primary results of this paper, showing the effect of using indicators on the two forecasting tasks as a function of the horizon of the forecast (i.e., the number of days ahead one is forecasting). For the case rate forecasting task, the left panel shows the ratio of the mean WIS of each forecaster relative to the mean WIS of the baseline. [ATTN: Further description and justification of this metric should be added to the "Evaluation Metrics" section rather than here it seems?] Both means are taken over all HRRs and forecast dates ranging from June 9, 2020 to December 31, 2020 [ATTN: verify]. All curves are well below 1, which means that

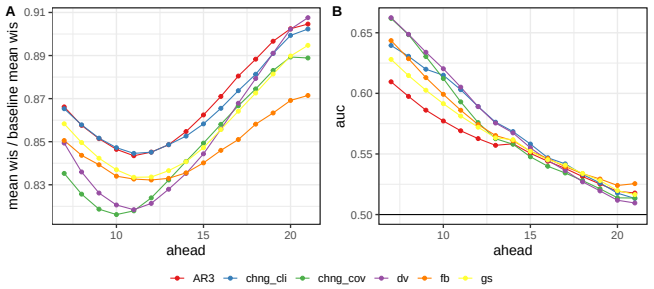

Fig. 5. Forecasting performance as a function of number of days ahead, aggregated across all HRRs and forecast dates. [ATTN: How's this?]

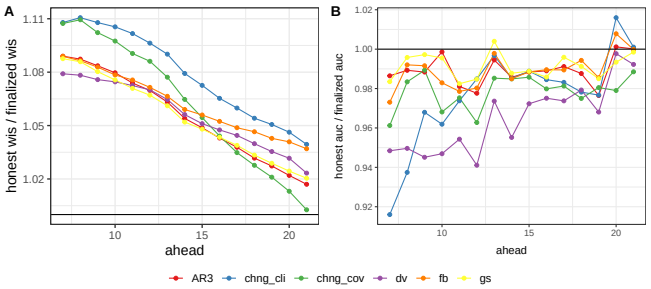

Fig. 6. How misleading is a retrospective analysis that uses the finalized version of the data rather than the version that was in fact available at the time the forecast was made? Plots show the ratio in performance (using mean WIS fore the forecasting task and AUC for the hotspot prediction task). Cheating vs. honest plots.

all methods, including `AR(3)`, outperform the baseline across all considered forecast horizons. The claims-based indicators `dv` and `chng_cov` show the most substantial gains, whereas the `chng_cli` does not show any improvement over `AR(3)`. The `fb` and `gs` indicators also show a non-negligible improvement. Notably, `fb` is the only indicator to offer a clear improvement at far aheads.

For the hotspot prediction task, the right panel of Figure 5 shows the AUC of each method. The monotonic decrease of all curves reflects the fact that hotspot prediction gets harder as the forecasting horizon increases. At the highest considered ahead (nearly one month ahead), the AUCs are close to 0.5, which is no better than random guessing. In the 7-14 day ahead range, the indicators show a noticeable improvement over `AR(3)`. For larger horizons there does not appear to be much difference in methods. The approximate ordering of methods is similar to what is seen in the forecasting task: `chng_cov` and `dv` are the strongest performing and `AR(3)` the weakest. Even for 7-day ahead predictions, the AUC is below 0.7, suggesting that all methods struggle with this classification task. [ATTN: Do people agree that an AUC of 0.7 is not all that great?] That said, `dv` and `chng_cov` attain roughly a 10% increase in AUC over the `AR(3)` classifier.

B. Avoiding misleading retrospective evaluations. Figure 6 quantifies the effect of not properly accounting for the question of "what was known when" in performing retrospective evaluations of forecasters. 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 (falsely) to have better performance. For example, for forecasting case rates 7-days ahead, the WIS of all methods is at least 8% larger than what would have been recorded using the finalized values of the data. This effect diminishes as the forecasting horizon increases, reflecting the fact that these forecasters rely less heavily on recent data than very short-term 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.

The `chng_cli` indicator (along with the other claims-based signals) is the most affected by this distinction, reflecting the latency in claims-based reporting. [ATTN: Actually, `dv` is the least affected for forecasting whereas it is one of the most affected for hotspots.] This supports 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).

Even the `AR(3)` 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.

- For forecasting task: Also, `chng_cli` and `dv` perform very similarly here, which is reassuring because they're measuring the same thing (in principle). Yet they must have very different backfill profiles...
- For hotspots task: dv seems to get a bigger boost than `chng_cov`- interesting dichotomy to forecasting
- But in both forecasting/hotspots `chng_cli` is affected a lot by backfill
- [ATTN: We should be careful about which indicators we were not able to track latency for. In fact, we should probably remove these from the plot since they would be more strongly affected than is shown in the plot.]
- [ATTN: We can also discuss the distinction between indicators that could in principle have low latency (e.g. GS) versus those where latency is inherent (e.g. claims-based indicators). If the goal is assessing the performance of GS "for the next pandemic" then it would make sense to give GS a pass since in theory they could have had it with very low latency (as opposed to claims based signals where backfill may be unavoidable). Of course, in "the next pandemic" there might not be such highly specific symptoms such as A+A as there was in the COVID-19 case.]

C. Additional sensitivity analyses. In addition to whether one properly accounts for signal latency, there are several other choices in evaluating forecasters that can also have a non-trivial impact on results.

For example, in Figure 5, we show the ratio of means. Other choices of aggregation and scaling are also possible. For example, taking the mean of ratios of errors is possible and would give more importance to locations where errors are small (either because the location-time pairs had low case rates or because the forecasting task was itself easier). By contrast, the ratio of the mean of errors may be dominated by location-time pairs where the error is large. Figure **??** of the Supplementary Information shows...

- Note the y-axis: barely an improvement over baseline, and this goes away for AR after 11 days ahead, and for all other models around 13-14 days ahead.

- Using trimmed means here because otherwise the results are extremely unstable (as the denominators can be small, when WIS of baseline is small).

- Similar message to before (non-adjusted), but `dv` and `chng_cov` the best, `fb` and `gs` still good, but now `chng_cli` seems to do much better!

- What's our interpretation here? That somehow `chng_cli` just does particularly bad in tasks where the baseline is also bad?

- We might include a histogram of WIS with a log-transformed x-axis to suggest that the geometric mean may be reasonable here.

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. Figure **??** of the Supplementary Information shows how Figure 5 panel A would change if one replaces the arithmetic mean by the geometric mean.

In Supplementary Information **??**, we investigate the effect of lengthening the time window to include 2021.

- For forecasting: results get better, whether we look at non-adjusted (first plot) or adjusted scores (second plot). Other than that, interpretation is qualitatively similar, except fb now seems to be worse than gs. And `chng_cli` is strong, even in the non-adjusted view.

- For hotspots: Alden will write something explaining why we don't do hotspot detection in 2021.

In Supplementary Information Section **??**, we look at various approaches to accounting for the fact that the `gs` indicator is only available for XX HRRs.

- For forecasting task: clearly gs is not missing at random, and when it's present, it tends to be predictive. Hence high values have a low false positivity rate. Most visible in the adjusted view below, where gs triumphs. Also `chng_cli` gets a lot worse.

- For hotspots: All methods look better, suggesting that the "GS locations" are easier for hotspot prediction. E.g. at 7-days ahead, the AUCs computed based on all locations range from 0.61-0.66; restricted to GS locations, the AUCs range from about 0.65-0.69. (These are all eyeballed, should get actual numbers if we want to say something like this in paper.)

- gs_subset appears to be particularly helped by this subsetting (which makes sense since on those other locations it was 0-imputed)

- Daniel will also try gs_impute as in forecasting

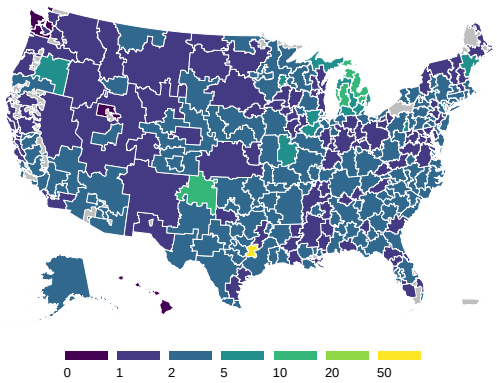

Fig. 7. Weighted interval score for forecasts made on April 1, 2021. Scores are averaged over horizons from 1 to 14 days ahead and are relative to population size.

[ATTN: Additional results section material, some for Discussion]

1. Spatial map of errors across HRRs for both tasks. See Figure 7 for an example.

2. Unlikely we can find it, but we should try anyway. Possibly put in the intro or discussion instead. Find an example of a forecast altered by the presence of an indicator. Plot the two trajectories.

3. Examine a specific HRR over time. What does it mean to predict certain hotspots? How can we understand what is happening?

4. What can we say about statistical significance? NRI, paired t-test?

5. Figures showing hotspots make sense

6. Horizontal line at 0.5 on AUC curve, discribe that this is random guessing.

3. Discussion

We have assessed the performance of [NUMBER] indicators in terms of their usefulness for performing two forecasting tasks: probabilistic forecasting of case rates and hotspot prediction at the HRR-level, 1–[NUMBER] days in advance. In measuring the practical usefulness of an indicator to forecasting, we distinguish between honest versus aspirational assessments. Data latency is an important factor when considering the usefulness of an indicator. Retrospective evaluations that ignore these messy data issues fail to provide the full picture. While one can hope that reporting systems will operate faster and more consistently in "the next pandemic," some reasons for this latency appear unavoidable. The chaos and strain inherent to pandemic times make indicators that can be reported and collected with quickly with minimal effort or manual involvement desirable.

We conclude by observing that the approach we have taken in this paper is somewhat opposite to that of much of the COVID-19 forecasting literature. We have chosen to consider only very simple forecasting models while devoting most of our effort to accounting for as much of the complexity of the data and evaluation as possible. By contrast, many papers

focus on very complicated forecasting approaches but then evaluate them under unrealistic, retrospective conditions.

1. Discussion should contain more detailed lit review ("context"), rather than in the intro, per the instructions.
2. Use of syndromic surveillance for other diseases?
3. Cite any papers that use our signals in forecasting.

[ATTN: Other potential discussion topics]

- When was this data actually available? (As of issue. The argument is: sure we didn't have it, but it would have helped. We should have this next time.)
- Side issues, are there places / times where one signal is especially good? Especially bad?
- What do we do about holidays? How do we understand the performance when the data is trash? If the data is trash, all the more reason to have other data that is perhaps less susceptible to these criticisms.
- Other signals we might want? Mobility? Discuss how SEIR models need contact matrices, relationship with "in community" signals.

ACKNOWLEDGMENTS. Please include your acknowledgments here, set in a single paragraph. Please do not include any acknowledgments in the Supporting Information, or anywhere else in the manuscript.

1 Delphi Research Group, COVIDcast Epidata API (2020).
2 J Bien, et al., *covidcast: R Client for Delphi's COVIDcast Epidata API*, (2020).
3 A Chin, A Reinhart, *COVIDcast Python API client*, (2020).
4 Reich Lab, The COVID-19 Forecast Hub (2020).
5 EY Cramer, et al., Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. medRxiv (2021).
6 J Zhang, et al., Changes in contact patterns shape the dynamics of the covid-19 outbreak in china. *Science* **368**, 1481–1486 (2020).
7 T Klopfenstein, et al., Features of anosmia in covid-19. *Médecine et Maladies Infect.* **50**, 436–439 (2020).
8 LA Vaira, G Salzano, G Deiana, G De Riu, Anosmia and ageusia: common findings in covid-19 patients. *The Laryngoscope* **130**, 1787 (2020).
9 DC Farrow, LC Brooks, A Rumack, RJ Tibshirani, R Rosenfeld, Delphi epidata api (2015).
10 R Koenker, Z Xiao, Quantile autoregression. *J. Am. Stat. Assoc.* **101**, 980–990 (2006).
11 E Cramer, et al., Covid-19 forecast hub: 4 december 2020 snapshot (2020).
12 T Gneiting, AE Raftery, Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).