BMC Global and Public Health

RESEARCH

Open Access

When are predictions useful? A new method for evaluating epidemic forecasts

Maximilian Marshall[1*], Felix Parker[1] and Lauren M. Gardner[1]

Abstract

Background COVID-19 will not be the last pandemic of the twenty-first century. To better prepare for the next one, it is essential that we make honest appraisals of the utility of different responses to COVID. In this paper, we focus specifically on epidemiologic forecasting. Characterizing forecast efficacy over the history of the pandemic is challenging, especially given its significant spatial, temporal, and contextual variability. In this light, we introduce the Weighted Contextual Interval Score (WCIS), a new method for retrospective interval forecast evaluation.

Methods The central tenet of the WCIS is a direct incorporation of contextual utility into the evaluation. This necessitates a specific characterization of forecast efficacy depending on the use case for predictions, accomplished via defining a utility threshold parameter. This idea is generalized to probabilistic interval-form forecasts, which are the preferred prediction format for epidemiological modeling, as an extension of the existing Weighted Interval Score (WIS).

Results We apply the WCIS to two forecasting scenarios: facility-level hospitalizations for a single state, and state-level hospitalizations for the whole of the United States. We observe that an appropriately parameterized application of the WCIS captures both the relative quality and the overall frequency of useful forecasts. Since the WCIS represents the utility of predictions using contextual normalization, it is easily comparable across highly variable pandemic scenarios while remaining intuitively representative of the in-situ quality of individual forecasts.

Conclusions The WCIS provides a pragmatic utility-based characterization of probabilistic predictions. This method is expressly intended to enable practitioners and policymakers who may not have expertise in forecasting but are nevertheless essential partners in epidemic response to use and provide insightful analysis of predictions. We note that the WCIS is intended specifically for retrospective forecast evaluation and should not be used as a minimized penalty in a competitive context as it lacks statistical propriety. Code and data used for our analysis are available at https://github.com/maximilian-marshall/wcis.

Keywords COVID-19, Epidemiology, Public health, Statistics

Background

The advent of the COVID-19 pandemic precipitated a massive public health response, including a significant modeling effort [1, 2]. In the United States, this quickly resulted in the formation of the COVID-19 Forecast Hub, a repository for short-term pandemic predictions [3]. Similar to prior collective forecasting efforts focused on seasonal influenza, dengue, and Ebola, the Forecast Hub solicited predictions from a large and diverse group of modelers, synthesizing their submissions into ensemble

*Correspondence:
Maximilian Marshall
mmarsh29@jhu.edu
[1] Department of Civil and Systems Engineering, Johns Hopkins University, Baltimore, MD, USA

forecasts of COVID-19 cases, deaths, and hospitalizations. These outputs were provided to the United States Centers for Disease Control and Prevention (CDC) for policy making and dissemination to the public [4–9]. In addition to modeling efforts like the Hub at the regional level, COVID prompted a considerable amount of more granular forecasting, such as predictions for individual medical facilities [10, 11]. Despite this abundance of pandemic modeling, translating short-term epidemiological forecasts into applicable, actionable, and insightful decision-making remains a significant challenge [7, 12–17].

Probabilistic predictions are preferred in many disciplines, including the epidemic forecasting community. Unlike single outcome "point" predictions, probabilistic forecasts convey the uncertainty of the underlying model. This is particularly important given the difficulty of correctly interpreting a quickly-evolving pandemic [7, 18]. The extant Weighted Interval Score (WIS), an error metric for quantile forecasts that approximates the Continuous Ranked Probability Score, is the primary method used to evaluate Forecast Hub submissions [19, 20]. As summarized by Bracher et al., "the (Weighted Interval) score can be interpreted heuristically as a measure of distance between the predictive distribution and the true observation, where the units are those of the absolute error" [19]. The WIS is an effective metric for real-time prediction scoring, model comparison, and ensemble forecast creation [20]. However, the WIS is limited in its ability to be used for intuitive forecast utility analysis, in particular because the score is scaled on the order of the prediction data [19]. Retrospective pandemic evaluation involves comparing scenarios of highly different scales. One example of such a comparison would be between regions with large baseline differences in data magnitudes, such as highly vs sparsely populated regions (as in the Forecast Hub). Another situation where scale-related contextualization is essential to consider is the comparison of periods of high vs low epidemic activity (surge vs non-surge). In fact, both of these spatial and temporal scaling challenges are often necessary to consider at the same time (see Additional file 1 for motivating examples of these issues drawn from state-level pandemic scenarios in the United States).

Our work is framed around the two following ideas. First, any meaningful measurement of forecast quality must arise from the context into which predictions are disseminated. In other words, a useful forecast improves real-time knowledge and/or decision-making capabilities. The reverse also holds: a forecast is not useful if it is incapable of providing (or if it provides information detrimental to gaining) better real-time information or improved decision-making. Second, for the purposes of enabling the comparison of forecast performances in disparate scenarios without post-processing, a helpful score should be a relative metric. Taken together, these two concepts form the aim of this study: creating a consistently meaningful probabilistic scoring method with endogenous contextualization. Such a score should normalize forecast performance as a function of the ability of the forecast to be used in the specific environment in which it was made. This way, despite potentially occurring in radically different spatial and temporal scenarios, individual predictions can be meaningfully compared to others. We believe that these attributes are highly important in the context of pandemic preparedness efforts given the need to more strongly connect the modeling and policy-making spheres of the public health community. Decision-makers need to be able to assess whether or not forecasting has the capacity to positively contribute to pandemic response.

To meet this goal, we introduce the WCIS: a score designed to reflect relative forecast quality using a flexible and contextually specific retrospective parameterization of utility. The WCIS was designed as an extension of (not a replacement for) the extant WIS, which functions well for real-time forecast scoring and ensemble generation. In this paper, we detail the technical basis and formulation of the WCIS, and demonstrate using relevant test cases that it is intuitively meaningful, interpretable, and comparable.

Methods

Review of the Weighted Interval Score

The Weighted Contextual Interval Score (WCIS) builds directly from the Weighted Interval Score (WIS). Bracher et al. [19] provide an excellent explanation of the mechanics of the score and its applications in epidemiology, and we endeavor to use the same symbology whenever possible. For brevity, the entire WIS formulation is not reviewed here, but the key elements (that are also important to the WCIS) are necessarily summarized:

$$IS_{\alpha}(F,y) = (u-l) + \frac{2}{\alpha}(l-y)\mathbb{1}\{y<l\} + \frac{2}{\alpha}(y-u)\mathbb{1}\{y>u\} \quad (1)$$

$$WIS_{\alpha_{[0:K]}}(F,y) = \frac{1}{K+\frac{1}{2}}\left(w_0\cdot|y-m| + \sum_{k=1}^{K}\{w_k\cdot IS_{\alpha_k}(F,y)\}\right) \quad (2)$$

- We assume a submission of K interval forecasts drawn from a predicted distribution F, a probabilistic representation of the target variable. Each of the K forecasts represents a $(1-\alpha_k)$ prediction interval (PI). These intervals are delineated by their lower and upper bounds l and u, the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of the predicted distribution, respectively. For example,

a 95% interval would be represented by an α_k of 0.05, its lower and upper bounds defined by the 0.025 and 0.975 quantiles of F.

- A predictive median m (point prediction) is submitted, and the true target value y is known.
- For each interval $k \in \{1, 2, ..., K\}$, an individual Interval Score (IS) is calculated, penalizing both the width/sharpness of the interval: $u - l$, and (if necessary) the amount by which the interval missed the true value: $\frac{2}{\alpha}(l - y)\mathbb{1}\{y < l\} + \frac{2}{\alpha}(y - u)\mathbb{1}\{y > u\}$ [21]. Note that the "miss" component is scaled by the inverse of α, thus narrower prediction intervals are penalized less for missing than are higher confidence submissions.
- The WIS is a weighted average of each of the K Interval Scores and the absolute error of the predictive median, with the weights w_k used for the average corresponding to $\frac{\alpha}{2}$ for each interval.

Contextualizing point forecasts

Although the WCIS (like the WIS) is an interval score, it is framed around a point score that we call the Contextual Relative Error (CRE). The CRE maps the absolute error of a point forecast x to its contextual utility. This is achieved by specifying δ, the utility threshold parameter. (Note that δ is the only parameter in the WCIS formulation that does not already appear in the WIS score.)

$$CRE(x, y, \delta) = min\left\{\frac{|x - y|}{\delta},\ 1\right\} \tag{3}$$

δ is the magnitude of the absolute error above which a forecast loses its utility. The CRE is so named because instead of mapping to the distance between a predicted value and its target like absolute error, it maps to an interval from 0 to 1. A score of 0 indicates a forecast with maximum possible utility (with an absolute error of 0), and a score of 1 indicates a useless forecast (with an absolute error of δ or more). See panel *(a)* of Additional file 1: Fig. S1 for a graphical representation of the CRE. An important feature to note is the "plateau" of the metric when the absolute error exceeds δ. This might seem problematic, given that beyond the δ threshold the absolute error is capable of increasing without any commensurate increase in the CRE. This is, in fact, the desired behavior of the CRE and warrants a slight re-framing of perspective. Selecting δ requires, when applying the CRE (and the WCIS, as it is a generalization of the CRE from point to interval scores), identification of a practical limit for how a forecast is used or interpreted in a particular context or for a particular purpose. For example, in many scenarios we have a finite capacity to respond to an expected outcome. If the "demand" imparted by an incorrect forecast exceeds that capacity, we are unable to alter our response despite an apparent increase in need. Therefore, an incorrect forecast with an absolute error of 2δ wastes exactly as many resources as a incorrect forecast of magnitude δ, where δ precipitates the maximum allocation in response to the forecast. A different way to interpret δ is as an "absorbable error magnitude". The test cases later in the paper frame δ this way, where a decision-maker has limited capacity to recover from plans made according to forecasted outcomes. If the forecast is wrong enough that it precipitates an action that cannot be recovered from, such a forecast has met or exceeded the δ threshold.

Note that δ is both a normalizer and a limit. Thus a forecast with an absolute error greater than δ is not at all useful, and a forecast with an absolute error less than δ is evaluated as a ratio of δ. This gives the CRE (and the WCIS) the ability to provide information about both forecast quality and how frequently forecasts are useful, which, as demonstrated later, is helpful for intuitive analysis and performance visualization.

Contextualizing interval-form forecasts

We begin by introducing the Contextual Interval Score (CIS). The CIS is both a probabilistic extension of the Contextual Relative Error, and a contextualized version of the Interval Score. Like the CRE, it maps a forecast's error to the δ-parameterized utility space, and like the IS, it generates a score for a single-interval forecast. (In fact, the CIS can be equivalently formulated in two different ways, based on either the IS or the CRE. For brevity, we use the IS-based formulation here but, particularly if more intuition about the score is desired, we suggest referencing the explanation of the CRE-based formulation in Additional file 1.)

$$CIS_\alpha(F, y, \delta) = min\left\{\frac{\alpha}{2\delta} IS_\alpha(F, y),\ 1\right\} \tag{4}$$

The WCIS is the simple average of the CIS across all α -intervals and the CRE of the predictive median m:

$$WCIS_{\alpha_{\{0:K\}}}(F, y, \delta) = \frac{1}{K + 1}\left(CRE(m, y, \delta) + \sum_{k=1}^{K} CIS_{\alpha_k}(F, y, \delta)\right) \tag{5}$$

Note that we still retain the descriptor "Weighted" in the WCIS title even though there are no weights directly included in its formulation, whereas each component of the WIS is multiplied by $\frac{\alpha}{2}$. However, in our formulation, the same weights are effectively applied directly to the individual constituent CIS scores. Instead of the "miss" components of the score being multiplied by $\frac{2}{\alpha}$, the "width" term is scaled by $\frac{\alpha}{2}$. Thus when the average is taken to create the WCIS, the scaling effect is the

same as the WIS, but the weights are applied in this way because it preserves the interpretability of the individual single-interval CIS components as described above. Another notable difference is the WCIS uses $K+1$ for the denominator of the average (unlike $K+\frac{1}{2}$ in the WIS) because like the single-interval components, the predictive median component of the score has a maximum penalty of 1. This, and the bound on each CIS term, means the WCIS also takes values only on the interval from 0 to 1. Note the natural equivalence between the WCIS for interval forecasts and the CRE for point forecasts, which mirrors that between the WIS and the absolute error. In both cases, the interval scoring method preserves the behavior and intuitive interpretation of the corresponding point forecast technique. Code that includes the WCIS formulation (and the analysis below) can be found at https://github.com/maximilian-marshall/wcis.

Results

The WCIS is expressly intended to be a flexible scoring method and as such there are many possible and highly variable ways to apply it. We use this Results section to present two demonstrative use cases. Both scenarios evaluate COVID-19 hospitalization forecasts, but each works at a different scale and uses a necessarily different δ formulation. The first scenario applies the WCIS to results from a multi-facility-level forecasting model. We use this first application primarily to develop the intuition for the δ selection process. We show via a direct demonstration how δ can be chosen to represent contextually specific utility as a function of time-varying data, and explore how the choices made during this parameterization map onto the output of the WCIS. Since this section focuses more on the WCIS formulation and less on interpreting the real-world applicability of the predictions, we use forecasts from a model developed in-house. Conversely, the second test case evaluates 4 weeks ahead predictions from the COVID-19 Forecast Hub's ensemble model, examining hospitalization forecasts from May 2021 to May 2022 [3]. This period includes both the Delta and Omicron variant waves and allows for a larger exploration of the utility and communicability of the WCIS. Data for these analyses are sourced from the COVID-19 Reported Patient Impact and Hospital Capacity by Facility dataset for the first section and from the Forecast Hub's repository for the second [3, 22]. Both datasets are publicly available.

Facility-level analysis (first test case)

As introduced above, our first test case evaluates a facility-level hospitalization model. More specifically, the model forecasts daily COVID-19 bed occupancy, for each individual hospital in Maryland, from one to twenty-one days out, from July 2021 to July 2022. Because our δ selection reflects capacity management within the 3-week forecast window, we only use hospitals listed as "short-term" type (this excludes long-term and pediatric facilities) and for relevance only include facilities that had at least ten COVID-19 patients at some point during the time range specified. The particular time range used was chosen because contextualization is vital when comparing and contrasting scenarios with highly different levels of pandemic activity, and July 2021 to July 2022 includes the Omicron wave in Maryland. This scenario and facility selection yields 42 hospitals with an overall capacity range of 30 beds at the smallest facility to 919 beds at the largest facility.

The model used is a Time Series Dense Encoder, using the prior ninety days for each hospital at each time point to predict the following 21 days [23]. For a complete model formulation, see Additional file 1, but in brief, this model type was selected because it is a state-of-the-art general-purpose time series forecaster that is efficient to train and flexible across different covariates, prediction horizons, output types, and loss functions. We note that the purpose of this test case is to explore and explain the formulation and application of the WCIS. Thus, we developed this relatively basic model in order to apply the WCIS to a facility-level scenario, not to refine a specific method for forecasting hospitalizations. The predictions from this section are not necessarily indicative of those performed in real time. Because the data used for training and scoring this model may contain retrospective corrections of errors that were present in the real-time data, it has the potential for higher performance when compared to an equivalent in-situ forecaster.

The δ-parameterization used for this analysis is intended to characterize the capacity of each facility to absorb an incorrect allocation of COVID-19 bed space based on a flawed forecast. We assume that capacity allocations are made at forecast time, under the in-situ assumption that forecasts perfectly reflect future outcomes. Thus, the δ value represents an achievable capacity correction during the time interval separating the making of the forecast and the realization of its true target value. For example, the δ value for a seven-day-ahead forecast for each facility is the amount of COVID-19 beds that each individual hospital can add or take away over a week. Specifically, this δ is determined as follows. The daily capacity change for each hospital is calculated as the mean of all single day, non-zero capacity changes over the entire available time series for each facility. δ for a particular forecast is then set as the product of the forecast horizon and the facility-specific daily change capacity. This means that the further out a forecast is, the larger (and thus more forgiving) the delta value is, based on the idea that the more time a facility has to respond

to a poor allocation of resources, the greater the magnitude of the response can be. Please note that the particular formulation chosen here is not intended to provide an assessment of forecast quality outside the utility scenario posited by the assumptions given above. However, it demonstrates an important capability of utility threshold selection: δ can be a defined as a dynamic function of data that can change in time and space. Since contextually meaningful forecast utility varies significantly over these same dimensions, a broadly applicable and interpretable score must be similarly adaptable.

Using Figs. 1 and 2, we are able to interpret some important aspects of how this selection of δ maps onto the scoring of our facility-level model. First, consider the relationship between the breadth of the confidence intervals and the δ region in Fig. 1, which visualizes a single facility. The larger prediction intervals for the 14-day-ahead forecasts indicate less model certainty than those of the 2-day-ahead predictions and, all else equal, would yield a worse score. However, δ is significantly higher for the 14-day-ahead scenario, given the assumption that facilities have more time to adapt to inaccurate forecasts over longer horizons. This results in generally better performance for the 14-day model. However, there remain in the 14-day scenario several forecasts that still receive a high penalty despite the more forgiving δ parameterization. Note that these instances tend to occur when the forecast median approaches or exceeds the utility threshold. Moving to Fig. 2, we can see that these trends are also visible in the aggregate performance across all 42 facilities. Comparing the WIS to the WCIS over these instances reveals a relatively linear relationship in the more forgiving scenarios, i.e., non-wave with a larger delta. During the wave, when absolute performance was broadly worse (as evidenced by the WIS values), the δ-limit was reached significantly more often. We also draw attention to the clear differences in marginal distributions that are visible in the scatter plot column of Fig. 2. The scaling and limiting action of the WCIS distributes performances significantly more evenly than the WIS (see Additional file 1 for plots with these marginal distributions included).

In general, we are able to observe that given a contextually relevant δ choice, the score is able to simultaneously convey an intuitive sense of both relative quality and the overall frequency of useful forecasts, as shown in the histograms of Fig. 2.

State-level analysis (second test case)

For this test case, we apply the WCIS to real-world predictions drawn from the COVID-19 Forecast Hub's Ensemble model, asking how much contextual utility hospitalization forecasts provided at the state level from May 2021 to May 2022 [3]. (Note that Forecast Hub hospitalization predictions were performed at daily resolution, but for the sake of visualizing a longer-term analysis we aggregate to and evaluate at weekly totals.)

The WCIS always requires a specific interpretation of the use-case for forecasts in the selection of the utility threshold δ. Similar to the facility-level analysis above, we choose to assess hospitalization predictions as a function of potential capacity changes. However, we assume a different decision-making scenario for hospital capacity at the state level than for its facility level counterpart. Due to the disaggregate decision-making apparatus across statewide hospitals and the inherent institutional inertia that must be overcome for larger scale change, we take a more conservative approach to estimating the absorbable error magnitude. Specifically, δ is the 0.9 quantile of the prior deviations in each state's hospital bed capacity over the prediction horizon of the forecast. We assume prior bed capacity deviations are indicative of a state's capacity to make changes, and that it is more difficult to make changes over a shorter timeline. Thus, any deviation over a shorter-term horizon can also occur for longer term horizons, but not the reverse. For example, when examining 1 week ahead predictions, only historical capacity changes over the course of a single week are considered. For 4-week-ahead predictions, capacity changes for 1, 2, 3, and 4 weeks ahead are considered. Finally, the 0.9 quantile is selected as the threshold under the assumption that states are not necessarily able to repeat their largest historical deviations, but can approach them. To be clear, this choice of δ is a heuristic for the amount of resource allocation, staffing changes, and other matters that hospitals might practically accomplish in response to an assumed change in pandemic dynamics. It is intended to demonstrate the WCIS given a reasonable, data-driven parameterization of forecast utility. Namely, a response predicated on a forecast outside the δ-range as defined here would require corrective action of a magnitude that could not be reasonably expected over such a forecast's prediction horizon.

WCIS performance results for 4-week ahead state-level hospitalization predictions are demonstrated in Fig. 3. Since the WCIS was designed primarily as a way to meaningfully evaluate and compare forecasts in disparate contexts, we can easily use it to observe several important aspects of hospitalization forecasting performance. For example, during surges and declines, forecast utility decreases substantially. We can intuit that this is a consistent trend across different locations both by directly observing the large central grid and by examining the lower, spatially averaged array of the figure. In contrast, if we examine the right-side, temporally averaged array, we observe that there is less variability in space than there

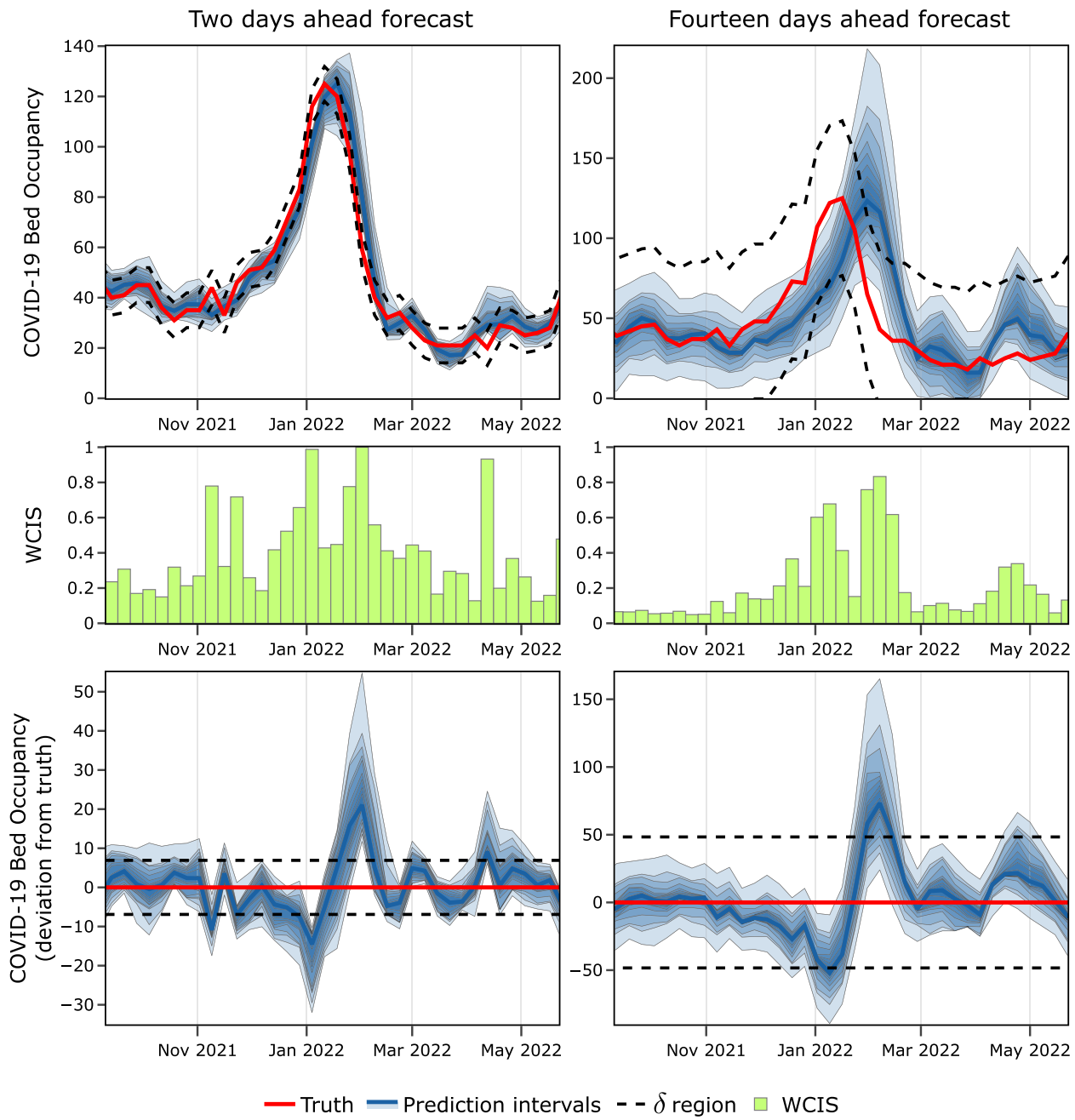

Fig. 1 Illustrated here are facility-level forecasts over two prediction horizons for one hospital: the University of Maryland Medical Center. The top and bottom rows both show the same forecasts, truth data, and δ (utility threshold) region. The top row displays these values normally, whereas the bottom row shows how far each value deviates from the truth. The middle row displays the WCIS, aligned with the data in the other rows. (Note that the facility-level analysis includes more prediction intervals and more dates than are shown in this figure, the extent of both displayed here are reduced for clarity)

is in time. Thus, by making an up-front determination about what constitutes a useful prediction (performing the δ-parameterization), we are capable of making, displaying, and intuitively evaluating forecasts. This allows, given a well-informed choice of δ, for meaningful overall analysis without needing to repeatedly delve into the specific circumstances during which each forecast was made. Without contextual normalization, conveying informative and comparable performance would be much more challenging. This capability, demonstrated by the ease of

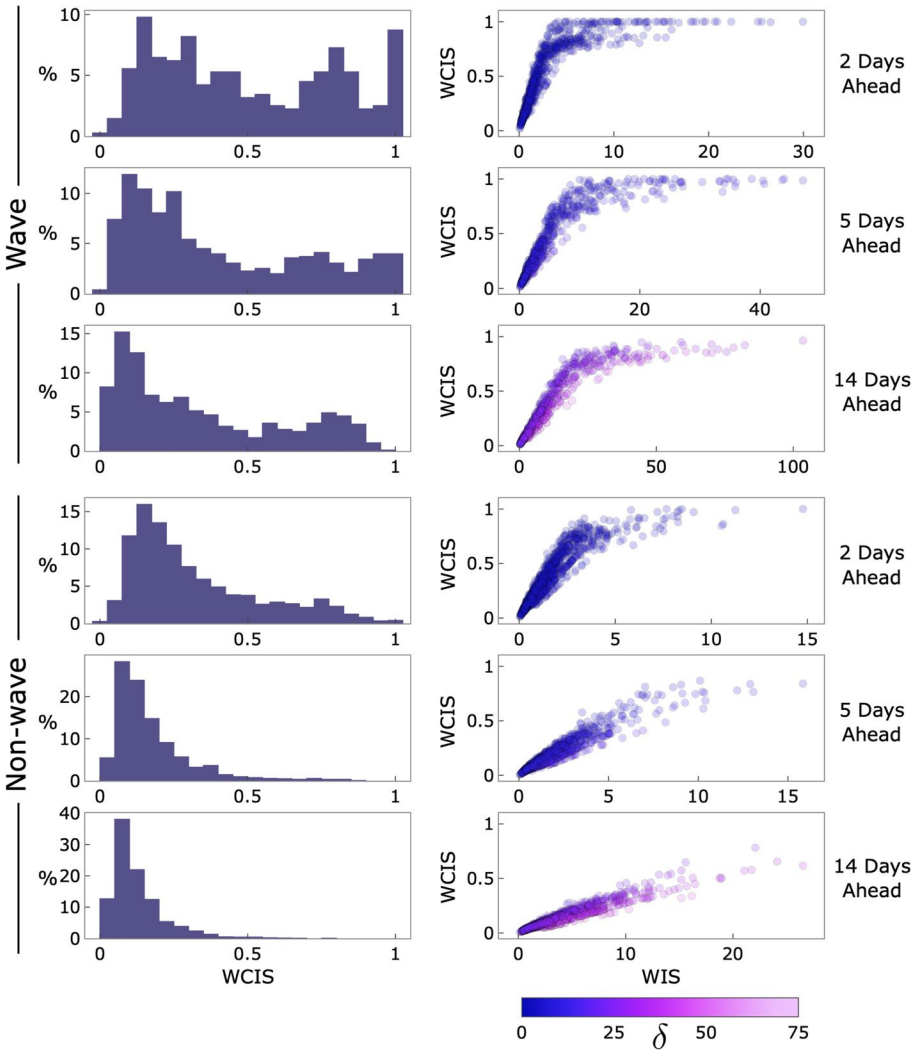

Fig. 2 Results in this figure are generated from all 42 hospitals, for all prediction dates in the facility-level model. The top three rows are from forecasts during the Omicron wave, and the bottom three from before and after the wave. We define the wave as lasting from November 14, 2021, through May 15, 2022, as illustrated in Additional file 1: Fig. S4

interpreting Fig. 3, is the overall aim for our creation of the WCIS. It permits substantive and easily interpretable performance evaluation.

Discussion

Given the devastating impact of COVID-19, and in the face of future pandemic threats, it is incumbent upon the epidemic forecasting community to deploy prediction tools that provide meaningful and actionable utility to those who need them [7, 14]. An important piece of this effort is candid evaluation of forecasting during the COVID-19 pandemic, and the WCIS is designed specifically as a retrospective way to judge whether or not forecasting could have been useful. It is not intended for real-time model ranking and ensemble construction. Instead, the WCIS is meant for broader pandemic preparedness efforts, for taking an honest look at how helpful forecasts could have been and thus potentially could be. This requires understanding the conditions in which forecasts were made. It also requires knowledge about the type of decisions the forecasts would be used to inform. Despite the high spatial and temporal variability of pandemic scenarios, its utility-based normalization scheme enables the WCIS to provide intuitive, meaningful, and comparable characterizations of forecast quality.

The WCIS is framed around our belief that a useful forecast contributes meaningful and/or actionable information given uncertain future outcomes. Determining whether or not forecasts accomplish this necessitates an explicit definition of utility. This brings up an important

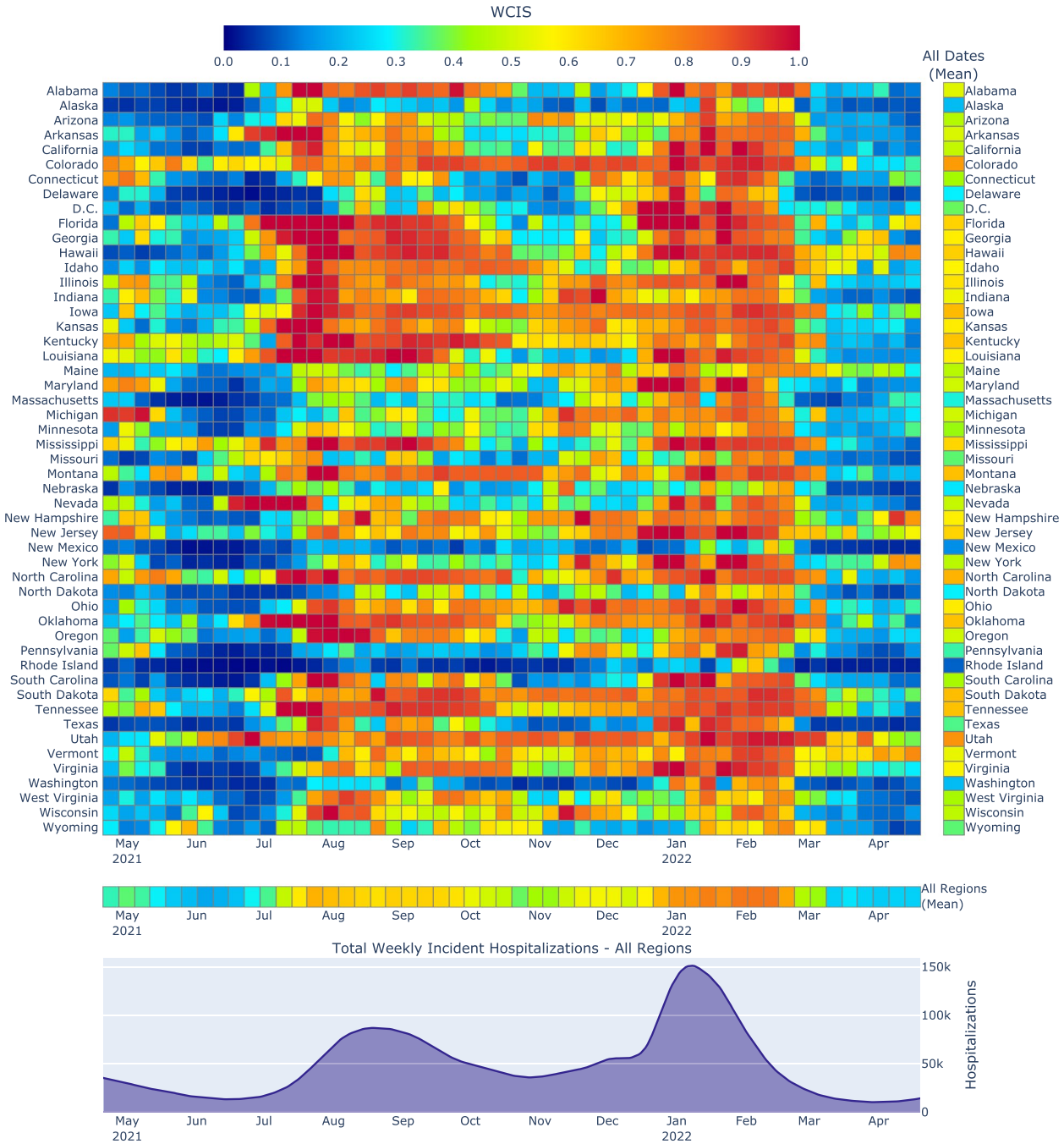

Fig. 3 Heatmap of the WCIS for 4-week-ahead hospitalization forecasts, performed by the Forecast Hub's Ensemble model. The central and largest grid shows the most granular results: region- and time-specific performance. On the right and lower sides of the grid are average performances over time and space, respectively. The shaded line plot at the bottom of the figure is the target hospitalization variable aggregated across all regions. Note that its domain is aligned exactly with those of the time-dependent heatmaps above, to provide insight into the trends of the overall pandemic alongside the more granular information in the heatmaps. (See Additional file 1 for heatmaps over differing prediction horizons)

philosophical difference between the WCIS and other techniques. The WCIS formulation, centered around a user-defined utility threshold δ, arises from our assertion that there will never be a one-size-fits-all solution for assessing and comparing short-term forecast quality. One must always consider prediction context and purpose lest standard metrics tell a misleading story. Additionally, different forecast use-cases yield different judgments

of predictions. The helpfulness of a model that predicts rainfall, for example, will be judged very differently by a user deciding whether or not to bring an umbrella on a walk as compared to a user deciding whether or not to issue regional flood warnings. An incorrect forecast of light rain with a realization of heavy rain is good enough for the first user but may be catastrophic for the second. Forecast purpose is essential to consider. The WCIS ensures this by building a definition of forecast utility directly into the formulation of the score.

The core of the WCIS is the combined normalization and thresholding imposed by the δ parameterization, which incorporates a vital aspect of real-world forecast utility. Namely, past a certain point, changes in a prediction's absolute error do not equate to changes in outcomes predicated on that forecast. Even when one forecast is more accurate than another, if both are beyond the utility horizon then the "better" one is not actually more useful, just arbitrarily closer to the truth. This idea is the basis for the plateaued CRE point scoring function, which in turn is the basis for the WCIS. While a metric that does not always increase the penalty as forecast accuracy diminishes may seem counterintuitive, we believe that for characterizing contextual utility, a score with a limited scope of relevance is actually more intuitive than a score that gets arbitrarily worse (or better) no matter how far away it is from being helpful.

The WCIS builds on the Weighted Interval Score, adding the δ-parameterization to impel users to directly characterize contextual utility. Judging predictions in this way allows for a powerful and effective normalization of the error, making the WCIS easy to interpret and compare across heterogeneous forecasting scenarios. Importantly, this robust efficacy exists *only for each individual definition of utility*. We belabor this point because it is inherent to our overall assertion about forecast interpretability: that a specific use case is necessary to meaningfully evaluate prediction quality. Without a link to how forecasts are used, it is difficult to consistently and meaningfully evaluate them over variable spatial and temporal conditions. Metrics without an explicit connection to forecast utility are in essence arbitrary until they are contextualized. For example, the WIS and the absolute error represent deviation from the truth on the order of the target data. They inherently require analysis in the context of those data in order for forecast quality to be understood. Even scores with temporally and spatially specific normalization can suffer from similar issues. In their evaluation of the COVID-19 Forecast Hub's performance, Cramer et al. use the Relative Weighted Interval Score (RWIS) for comparison and aggregation [20]. The RWIS is the ratio of the WIS of the evaluated model to the WIS of the Forecast Hub's Baseline model, where the Baseline model is a simple forward extension of the most recently observable data at the time of forecasting. In other words, the RWIS normalizes forecast performance relative to the performance of a naive, simple model. While this enables scale-indifferent forecast comparison, it does not ensure that these comparisons are contextually meaningful, since the performance of the Baseline model is highly variable in space and time [20]. We contrast this with an effective WCIS δ-parameterization, which builds contextualization directly into the formulation of the score.

Before concluding, it is necessary to address the intended purpose of the WCIS and its limitations. The WCIS is not a statistically proper score (see empirical demonstration in Additional file 1), which means it should not be used in competitive forecasting contexts. In these situations, such as real-time evaluation of COVID-19 Forecast Hub submissions and ensemble creation, scores that are not statistically proper have the potential to be gamed [20, 21]. The WCIS is not designed for and should not be used for such purposes. Instead, it is intended for retrospective evaluation, when δ is able to be selected in a meaningful way. Importantly, generation of an appropriate δ is another limitation of the WCIS. Without a well-designed and contextually robust threshold, the score loses its power. Finally, the thresholding of the score, despite providing the benefit of collapsing the interpretable range, means it can equate forecasts with dramatically different accuracies. However, and related to the prior limitation, an appropriately selected δ should render this distinction contextually irrelevant.

Determining the future role of pandemic forecasting, as well as identifying areas of forecasting that need improvement, must at some point include the translation of modeling results to policy and decision-makers. The WCIS is expressly intended to function well in this process, allowing for intuitive characterization of forecast utility that can be easily communicated to an audience with less technical expertise. Figure 3 demonstrates this directly. Without effective contextual normalization, generating such a display would be challenging given large differences in error magnitude, likely requiring a transformation (such as log-scaling) that limits interpretability. Instead, the WCIS allows for a direct, clearly defined interpretation of forecast utility to be displayed and compared in a technically meaningful and intuitively understandable way.

Conclusions

We created the WCIS to enable and encourage honest and contextually specific discourse about the utility of short-term epidemic predictions. It incorporates prediction uncertainty, keeps the technical definition of utility as

simple as possible, and generates an intuitively interpretable and comparable numerical output. Our intent is to allow for people without specific technical experience to be able to interact with and evaluate probabilistic forecasting in a meaningful way. As the public health community learns from COVID-19 and prepares for future challenges, explicit analysis of the utility of historical predictions is essential. We hope the WCIS will help with effective and meaningful communication between modelers and practitioners in this effort.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s44263-024-00098-7.

Additional file 1. This document contains the Supplemental Materials for this article. These include sections that provide more detail on and/or motivating examples for the formulation of the score, the impropriety analysis, and the facility-level model formulation. It also includes figures comparing the WIS and the WCIS performance of the facility-level model for different scenarios, and state-level heatmaps of the WCIS for the Forecast Hub Ensemble model for 4 prediction horizons (1, 2, 3, and 4 weeks ahead)

Acknowledgements
Not applicable.

Authors' contributions
M.M. conceived of the study, performed the analysis, and wrote the manuscript. M.M. and F.P. developed the methodology. L.M.G. supervised the project and provided essential methodological guidance. All authors read, edited, and approved the final manuscript.

Funding
This study was supported by the United States National Science Foundation (NSF) under grant no. 2108526.

Data availability
Code and processed data for both the facility- and state-level analyses are accessible from our publicly available GitHub repository [https://github.com/maximilian-marshall/wcis] [24]. Forecast and ground truth data used for our state-level analysis are available from the COVID-19 Forecast Hub repository [https://doi.org/10.1038/s41597-022-01517-w] [3]. The original source for ground truth hospitalization data is the COVID-19 Reported Patient Impact and Hospital Capacity by Facility repository [https://healthdata.gov/d/j4ip-wfsv] [22].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 2 November 2023 Accepted: 19 September 2024
Published online: 03 October 2024

References

1. Horbach SPJM. Pandemic publishing: medical journals strongly speed up their publication process for COVID-19. Quant Sci Stud. 2020;1(3):1056–67. https://doi.org/10.1162/qss_a_00076.
2. Fraser N, Brierley L, Dey G, Polka JK, Pálfy M, Nanni F, et al. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. PLOS Biol. 2021;19(4):e3000959. https://doi.org/10.1371/journal.pbio.3000959.
3. Cramer EY, Huang Y, Wang Y, Ray EL, Cornell M, Bracher J, et al. The United States COVID-19 Forecast Hub dataset. Sci Data. 2022;9(1):462. https://doi.org/10.1038/s41597-022-01517-w.
4. McGowan CJ, Biggerstaff M, Johansson M, Apfeldorf KM, Ben-Nun M, Brooks L, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. Sci Rep. 2019;9(1):683. https://doi.org/10.1038/s41598-018-36361-9.
5. Johansson MA, Apfeldorf KM, Dobson S, Devita J, Buczak AL, Baugher B, et al. An open challenge to advance probabilistic forecasting for dengue epidemics. Proc Natl Acad Sci. 2019;116(48):24268–74. https://doi.org/10.1073/pnas.1909865116.
6. Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, et al. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. Epidemics. 2018;22:13–21. https://doi.org/10.1016/j.epidem.2017.08.002.
7. Reich NG, Ray EL. Collaborative modeling key to improving outbreak response. Proc Natl Acad Sci. 2022;119(14):e2200703119. https://doi.org/10.1073/pnas.2200703119.
8. Ray EL, Brooks LC, Bien J, Biggerstaff M, Bosse NI, Bracher J, et al. Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. Int J Forecast. 2022. https://doi.org/10.1016/j.ijforecast.2022.06.005.
9. Reich NG, McGowan CJ, Yamana TK, Tushar A, Ray EL, Osthus D, et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. PLOS Comput Biol. 2019;15(11):e1007486. https://doi.org/10.1371/journal.pcbi.1007486.
10. Weissman GE, Crane-Droesch A, Chivers C, Luong T, Hanish A, Levy MZ, et al. Locally informed simulation to predict hospital capacity needs during the COVID-19 pandemic. Ann Intern Med. 2020;173(1):21–8. https://doi.org/10.7326/M20-1260.
11. Kociurzynski R, D'Ambrosio A, Papathanassopoulos A, Bürkin F, Hertweck S, Eichel VM, et al. Forecasting local hospital bed demand for COVID-19 using on-request simulations. Sci Rep. 2023;13(1):21321. https://doi.org/10.1038/s41598-023-48601-8.
12. Doms C, Kramer SC, Shaman J. Assessing the use of influenza forecasts and epidemiological modeling in public health decision making in the United States. Sci Rep. 2018;8(1):12406. https://doi.org/10.1038/s41598-018-30378-w.
13. Reich NG, Wang Y, Burns M, Ergas R, Cramer EY, Ray EL. Assessing the utility of COVID-19 case reports as a leading indicator for hospitalization forecasting in the United States. Epidemics. 2023;45:100728. https://doi.org/10.1016/j.epidem.2023.100728.
14. Nixon K, Jindal S, Parker F, Marshall M, Reich NG, Ghobadi K, et al. Real-time COVID-19 forecasting: challenges and opportunities of model performance and translation. Lancet Digit Health. 2022;4(10):e699–701. https://doi.org/10.1016/S2589-7500(22)00167-4.
15. Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. BMC Public Health. 2019;19(1):1659. https://doi.org/10.1186/s12889-019-7966-8.
16. Guerrier C, McDonnell C, Magoc T, Fishe JN, Harle CA. Understanding health care administrators' data and information needs for decision making during the COVID-19 pandemic: a qualitative study at an academic health system. MDM Policy Pract. 2022;7(1):23814683221089844. https://doi.org/10.1177/23814683221089844.
17. Lee TH, Do B, Dantzinger L, Holmes J, Chyba M, Hankins S, et al. Mitigation planning and policies informed by COVID-19 modeling: a framework and case study of the state of Hawaii. Int J Environ Res Public Health. 2022;19(10):6119. https://doi.org/10.3390/ijerph19106119.
18. Nixon K, Jindal S, Parker F, Reich NG, Ghobadi K, Lee EC, et al. An evaluation of prospective COVID-19 modelling studies in the USA: from data to science translation. Lancet Digit Health. 2022;4(10):e738–47. https://doi.org/10.1016/S2589-7500(22)00148-0.

19. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. PLOS Comput Biol. 2021;17(2):e1008618. https://doi.org/10.1371/journal.pcbi.1008618.
20. Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proc Natl Acad Sci. 2022;119(15):e2113561119. https://doi.org/10.1073/pnas.2113561119.
21. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc. 2007;102(477):359–78. https://doi.org/10.1198/016214506000001437.
22. COVID-19 reported patient impact and hospital capacity by facility. United States Department of Health & Human Services; 2020. https://healthdata.gov/d/j4ip-wfsv. Accessed 29 Feb 2024.
23. Das A, Kong W, Leach A, Mathur S, Sen R, Yu R. Long-term forecasting with TiDE: time-eries dense encoder. arXiv [Preprint]. 2023. ArXiv:2304.08424 [cs, stat]. https://doi.org/10.48550/arXiv.2304.08424.
24. Marshall M. Accompanying code for "When are predictions useful? A new method for evaluating epidemic forecasts". Github. 2024. https://github.com/maximilian-marshall/wcis. Accessed 4 Apr 2024.

Publisher's Note