Evaluating infectious disease forecasts with allocation scoring rules

Aaron Gerding, Nicholas G. Reich, Benjamin Rogers, Evan L. Ray

July 11, 2023

Abstract

The COVID-19 pandemic has led to rapid innovation in methods for eliciting and evaluating forecasts of infectious disease burdens, with a primary goal being to help public health workers make informed decisions about how to manage these burdens. However, explicit descriptions or quantifications of the value forecasts add to society through the decisions they support are elusive. Moreover, there has only been limited discussion of how predominant forecast evaluation metrics might indicate the success of policies based in part those forecasts.

Here we pursue one possible tether between multivariate forecasts and policy: the allocation of limited medical resources in response to COVID-19 hospitalizations in various regions so as to minimize expected unmet need. Given probabilistic forecasts of hospitalizations in each region, we formulate an allocation algorithm following techniques developed in operations research. We then score forecasts according to how much unmet need their associated allocations would have allowed. We illustrate this scheme with quantile forecasts of COVID-19 hospitalizations in the US at the state level that are recorded in the COVID-19 Forecast Hub, with the goal of determining the allocation of a hypothetical limited resource across the states. The forecast skill ranking given by this allocation scoring rule can vary substantially from the ranking given by the weighted interval score now used by the CDC, especially during surges in hospitalizations such as in late 2021 as the Omicron wave began. We see this as strong evidence that the allocation scoring rule detects forecast value that is missed by traditional accuracy measures and that the general strategy of designing scoring rules directly linked to policy performance is a promising research direction for epidemic forecast evaluation.

1 Introduction

Infectious disease forecasting models have emerged as important tools in public health decision making. Predictions of disease dynamics have been used to inform decision-making about a wide variety of measures designed to reduce disease spread and/or mitigate the severity of disease outcomes. For example, estimates of expected onset of flu season were used to inform national vaccination strategies [Igboh et al., 2023], and forecasts of ebola dynamics were used to allocate surveillance resources [Meltzer et al., 2014, Rainisch et al., 2015]. In April of 2022, the Centers for Disease Control and Prevention (CDC) announced the launch of the Center for Forecasting and Outbreak Analytics (CFA) to translate disease forecasts into decision-making [for Desease Control and Prevention]. Bertsimas et al. [2021] developed tools to inform decision making from infectious disease forecasts, which have been used to to inform allocation of limited medical supplies such as ventilators, ICU capacity planning, and vaccine distribution strategy. Models developed by Fox et al. [2022] have been used to inform resource and care

site planning, as well as community guidelines for masking, traveling, dining and shopping [of Texas at Austin]. **BWR:[Figure out these news site citations!]**

In decision-making settings where it is possible to quantify the utility or loss associated with a particular action, standard tools of decision theory provide a procedure for developing forecast scoring rules that measure the value of forecasts through the quality of the decisions that they lead to. We give an overview of these procedures in Section 2.1. There is a large history of literature applying these ideas to obtain measures of the value of forecasts that are tied to a decision-making context, primarily in fields such as economics and finance, supply chain management, and meteorology. We review this work only briefly here, and we refer the reader to Yardley and Petropoulos [2021] for a general overview, and to Pesaran and Skouras [2002] and **ELR:[TODO: identify relevant review or book style summary for meteorology]** for discussions focused on applications to economics and meteorology, respectively. In finance, the value of forecasts can often be measured by the profits generated by trading decisions informed by the forecasts, perhaps adjusted for risk levels [e.g., Leitch and Tanner, 1991, Cenesizoglu and Timmermann, 2012]. In applications to supply chain management and meteorology, the value of forecasts has typically been operationalized by considering the costs associated with decisions regarding the amount of inventory to hold or the level of protection against the impacts of extreme weather events to enact Peter Catt (2007), Fotios Petropoulos and colleagues (2019), Nada Sanders and Gregory Graman (2009), T.N. Palmer (2002), Florian Pappenberger and colleagues (2015). For example, in supply chain management these decisions may incur costs related to holding inventory, labor, or providing poor service, while in meteorology we may need to balance the costs of implementing protective measures with the costs of potentially preventable weather damages. In this framework, a forecast has value if it leads to decisions with low total costs. In all of these fields, analyses have consistently found that common measures of statistical forecast accuracy do not necessarily correspond directly to measures of the value of forecasts as an input to decision-making [e.g., Leitch and Tanner, 1991, Cenesizoglu and Timmermann, 2012].

However, we are aware of only a limited body of work that explicitly attempts to measure the value of infectious disease forecasts through their impact on policy, and much of this discussion has proceeded informally. For example, Ioannidis et al. [2022] discuss the possible negative consequences of inaccurate forecasts of infectious disease, but do not attempt to quantify the utility or loss incurred as a result of those forecasts. Separately, there is a thread of literature that does quantify the link between infectious disease modeling and policy making, but this work has been done outside of a forecasting context. As an example, Probert et al. [2016] develop measures of the cost of actions designed to control a hypothetical outbreak of foot-and-mouth disease and use this framework to explore policy recommendations from a variety of simulation-based projection models.

In practice, probabilistic infectious disease forecasts have most often been made for observations that emerge from public health surveillance systems and evaluated with standard, "off-the-shelf" scoring rules. For example, seasonal influenza forecasts in the US and dengue forecasts for Peru and Puerto Rico targeted public health surveillance measures of incidence over time and space, and used log-score and mean absolute errors to evaluate forecast skill [McGowan et al., 2019, Reich et al., 2019, Johansson et al., 2019]. Pandemic COVID-19 forecasts of observed cases, hospitalizations and/or deaths in the US and Europe, as reported by municipal, state, or federal surveillance systems, were evaluated using the weighted interval score (WIS, which is an approximation of the continuous ranked probability score, or CRPS), and prediction interval coverage [Cramer et al., 2022, Fox et al., 2022, Sherratt et al., 2023]. Similarly, CRPS was also used to assess probabilistic forecasts of dengue incidence at the district level in Vietnam [Colón-González et al., 2021]. While some of these scores can be interpreted through the

lens of decision theory, and all of the application-specific papers cited above had authors from public health agencies, none of them (including some for which we were co-authors) make explicit connections between forecast evaluation and how a forecast was used in practice.

The motivating application of this work is one where forecasts of standard public health surveillance metrics (1) are used to help determine the allocation of a limited quantity of medical supplies across multiple regions, and (2) are later evaluated based on their recommended allocations. The analysis presented is "synthetic" in that it does not correspond to an actual analysis that supported decision-making in real-time. However, the framework described in this paper corresponds to real-world decisions that must be made by public health administrators around the globe, and could be adapted in the future for such real-time situations. For example, forecasts for districts in Sierra Leone of bed demand to care for patients with Ebola was the subject of a real-time modeling study in late 2014 and early 2015Camacho et al. [2015]. And, in 2020, a model developed by an academic research group turned predictions of COVID-19 hospitalizations into estimates of ventilator usage and and shortages. This framework was used by the Hartford HealthCare system in Connecticut "to align ventilator supply with projected demand at a time where the [COVID-19] pandemic was on the rise"Bertsimas et al. [2021]. These examples illustrate the potential for forecasts to inform decisions about how to allocate limited supplies such as temporary hospital beds, ventilators, personal protective equipment, or other supplies that are known to be effective at reducing transmission or severity of disease.

As described above, there exists both a rich literature on outbreak forecast evaluation and an emerging body of work on turning forecasts into resource allocation decisions. However, there has been far less attention paid (no papers that we could identify) to exploring whether standard forecast accuracy metrics are sufficient for evaluating forecasts that are used to make allocation decisions. In this work, we address the gap between the ways in which infectious disease forecasts have been used to support public health policy and the ways in which they have traditionally been evaluated. Specifically, we define a new forecast scoring rule –the *allocation score*– that explicitly evaluates forecasts based on how accurate their implied allocations would turn out to be. This stands in contrast to standard evaluation metrics that evaluate the accuracy of a forecast relative to the surveillance data being forecasted.

The remainder of this article is organized as follows. In Section 2, we review the general framework for developing scoring rules for probabilistic forecasts using the tools of decision theory, develop a novel scoring rule that is motivated by the problem of allocating limited medical supplies, and explore the relationship between the proposed allocation score and existing scoring rules such as CRPS. In Section 3 we then illustrate the scoring rule through an application to short-term forecasts of COVID-19 hospital admissions in the US. Section 4 summarizes our contributions and discusses opportunities for further extensions in future work.

1.1 Allocation Scoring

In a nutshell, the allocation score of a multivariate forecast F is the unnecessary realized unmet demand when F is used to allocate so as to minimize expected unmet need — that is, when we distribute a limited amount K of our resource so that total expected unmet need according to F,

$$\sum_i E_{F_i}[\text{unmet need in location } i] \tag{1}$$

is minimized, where the F_i are the marginal distributions of F. By "unnecesary" we mean the unmet need that could have been avoided by an oracle that knows exactly how much need will occur in each

location and divides the amount K so that nothing is wasted in one location while it could be put to use in another. This stochastic allocation problem has an intuitively appealing solution: allocate so that the probabilities of need exceeding allocation in various locations are as close to each other as possible (see methods). This will lead to the allocations provided by F being quantiles of F_i for some *single* probability level τ.

For example, suppose we have a forecast F for need in two locations with $F_1 = \text{Unif}(0, 8)$ and $F_2 = \text{Unif}(4, 8)$. If we have $K = 10$ units of our resource available, the optimal allocation according to F would be 4 units in location 1 and 6 units in location 2. This allocates to both locations at their forecast medians. If, on the other hand, we only have $K = 3$ units available, we will necessarily have to allocate in location 2 at a quantile for $\tau = 0$, and therefore all 3 would go to location 2 since allocating any resources to location 1 would lead to unbalanced probabilities of need exceeding allocation. [1]

Next suppose that we observe needs of 8 and 3 in locations 1 and 2, respectively. With $K = 10$ units of the resource, the oracle would be able to prevent all but 1 of the total 11 units of needs from going unmet. The forecast F's allocation results, at the same time, in $8 - 6 = 2$ units of unmet need in location 1 despite leaving no need unmet in location 2. The allocation score for the forecast when $K = 10$ would therfore be 1 ($= 2$ realized $- 1$ unavoidable) in units of unnecessary unmet need. When $K = 3$, however, the forecast does not "over-allocate" in either location which leads to the same unmet need of $11 - 3 = 8$ as does the oracle's allocations. The allocation score for F when $K = 3$ is therefore 0.

Note the way in which F incurred a positive (i.e., non-optimal) allocation score of 1 for F when $K = 10$. It was not directly directly due to individual misalignments of the marginal forecasts F_i with the observed needs, but rather because the allocations and observed needs were not proportional as vectors. This illustrates a fundamental property of the allocation score. For a given probability level τ, if the quantiles $Q_i(\tau)$ of the marginal forecasts are proportional to the observed needs y_i, then the allocation score is zero for the resource constraint level $K = \sum Q_i(\tau)$. This stands in marked constrast to other common scoring methods for multi-variate forecasts that aggegate univariate scores such as CRPS or WIS for the marginal forecasts where a misalignment in one coordinate is penalized regardless of alignments in other coordinates. **APG:[Should this be illustrated here, or elsewhere?]**

It is therefore often straighforward to constuct forecasts F and $\widetilde{F}$ for a given outcome distribution that switch rankings under the allocation and traditional scores by ensuring that the marginal forecasts of F center sharply around allocations that are proportionally similar to the central tendencies of the outcome distribution but are strongly biased.

While multivariate scoring rules have not seen wide application in infectious disease forecast evalution, it seems that the allocation score would have a similar relationship with them, since while bias penalties in multi-variate scores can be offset by better forecasting of dependence structure, the proportional biases that allocation scoring are insensitive to are not *per se* tolerated by multivariate scoring rules such as the energy score, variogram score, or Dawid-Sebastiani score.

2 Methods

We give a high-level review of a general procedure for developing proper scoring rules that are tailored to a specific decision-making task in section 2.1. In section 2.2 we review how quantile loss can be obtained within this framework in a setting where a decision-maker is required to determine the

[1] Matching probabilities will not always be possible if we allow point masses in the F_i. For example, if F_2 were the point forecast of 8 interpreted as a degenerate probabilistic forecast with mass 1 at 8, the optimal allocation for $K = 10$ would be 2 in location 1 and 8 in location 2. Notice though that both allocations are still technically medians.

quantity of a good to procure while balancing the cost of purchasing an additional unit of the good with loss that may result from under-procurement. We then discuss how the continuous ranked probability score (CRPS) can be obtained as an integral of the quantile loss across values of the cost/loss ratio. These developments mirror the structure of section 2.3. There, we develop a novel *allocation score* that is analogous to the quantile score but is suitable for evaluation of forecasts in the context of decisions about allocation of limited resources across multiple locations when the resource constraint is known. We then describe an *integrated allocation score* that is analogous to the CRPS and is obtained by integrating the allocation score across values of the resource constraint.

2.1 The decision-theoretic setup for forecast evaluation

In this section, we give an overview of the decision-theoretic setup for developing proper scoring rules that measure the value of a forecast as an input to decision making. We keep the discussion here at a somewhat informal level; we refer the reader to [some subset of Brehmer and Gneiting; Grünwald and Dawid; Dawid; Granger and Pesaran 2000; Granger and Machina 2006; Ehm et al. 2016] for more technically precise discussion.

In the framework of decision theory, a decision corresponds to the selection of an action x from some set of possible actions $\mathcal{X}$. For example, x may correspond to the level of investment in a measure designed to mitigate severe disease outcomes such as hospital beds, ventilators, medication, or medical staff, with $\mathcal{X}$ being the set of all possible levels of investment that we might select. The quality of a decision to take a particular action x is measured in relation to an outcome y that is unknown at the time the decision is made. For example, y may correspond to the number of individuals who eventually become sick and would benefit from the mitigation measure, and informally, an action x is successful to the extent that it meets the realized need. In the face of uncertainty, a decision-maker may use a forecast F of the random variable Y to help inform the selection of the action to take. We measure the value of a forecast as an input to this decision-making process by the quality of the decisions that it leads to.

We can formalize the preceding discussion with the following three-step procedure for developing scoring rules for probabilistic forecasts:

1. Specify a *loss function* $s(x, y)$ that measures the loss associated with taking action x when outcome y eventually occurs.

2. Given a probabilistic forecast F, determine the *Bayes act* x^F that minimizes the expected loss under the distribution F.

3. The *scoring rule* for F calculates the score as the loss incurred when the Bayes act was used: $S(F, y) = s(x^F, y)$.

We use the letter s for the loss function to align with the literature on evaluation of forecasts of continuous outcomes, in which context we can often identify the action x with a functional (i.e., a numeric summary such as a mean or a quantile) of the forecast distribution F. In this context, s may be used as a *scoring function*. **ELR:[Consider moving preceding sentences to a footnote or just deleting them?]** This is a general procedure that may be applied in settings where it is possible to specify a quantitative loss function. Subject to certain technical conditions, scoring rules obtained from this procedure are proper (cite cite).

2.2 A review of quantile score, CRPS, and the weighted interval score

We review how the quantile score arises from a particular decision-making problem in section 2.2.1, and how CRPS can be obtained by integrating across values of the parameters of that decision-making problem, as well as the connection to WIS, in section 2.2.2. These results have been thoroughly discussed in the literature [cite cite cite].

2.2.1 Decision-theoretic origins of the quantile score

Suppose that a decision-maker is tasked with determining the quantity x of a protective measure to procure; for example, x might represent the number of hospital beds or amount of medicine to purchase. Additionally, suppose that each unit of this good has cost C so that the total cost of procurement is Cx. The variable y denotes the eventual realized need for this resource, e.g. the number of patients in need of a hospital bed or the amount of medication that is needed. We assume that each unit of unmet need incurs a loss denoted by L, so that if the selected procurement level x is less than the realized need y, a loss of $L(x-y)$ results. At the time that a decision-maker determines the amount x to procure, the demand y is not yet known. We therefore define the random variable Y that represents the as-yet-unknown level of demand. The forecast F specifies a predictive distribution for Y. Here we identify F with its cumulative distribution function (CDF), and F^{-1} denotes the quantile function. With this formalization of the decision-making task, we can proceed to develop a proper scoring rule using the procedure outlined in section 2.1.

Step 1: specify a loss function. Combining the cost of procuring goods at level x with losses due to unmet need, we arrive at the overall loss function

$$s_Q(x, y; C, L) = Cx + L(x-y)_-. \tag{2}$$

Here, $(x-y)_- := \max(-(x-y), 0)$ is 0 if the amount procured, x, is greater than or equal to the realized demand y; otherwise, it is $y - x$, the amount of unmet need.

Step 2: Given a probabilistic forecast F, identify the Bayes act. It can be shown that under the loss function s_Q, the Bayes act is a quantile of the forecast distribution at the probability level $\alpha = 1 - C/L$:

$$x^F = F^{-1}(\alpha). \tag{3}$$

See the supplement for a verification of this result.

Step 3: Define the scoring rule. Following the procedures outlined above, we could score the forecast distribution F with the scoring rule

$$\begin{aligned} S_Q(F, y; C, L) &= s_Q(x^F, y; C, L) = Cx^F + L(x^F - y)_- \\ &= CF^{-1}(\alpha) + L\left(F^{-1}(\alpha) - y\right)_- \end{aligned}$$

We have set up the problem here in terms of the cost and loss parameters C and L, which has the benefit of an intuitive connection to the decision-making context. However, to clarify the connection to the usual notation for the quantile loss, we can divide the loss function s_Q by L to obtain an expression

in terms of only α:

$$\begin{aligned} s_Q(x, y; \alpha) &= s_Q(x, y; C, L)/L \\ &= (C/L)x + (x - y)_- \\ &= (1 - \alpha)x + (x - y)_-. \end{aligned}$$

Because these loss functions are equal up to a constant of proportionality, the Bayes act is the same for both. The associated quantile scoring rule expressed in terms of α is

$$S_Q(F, y; \alpha) = (1 - \alpha)F^{-1}(\alpha) + \left(F^{-1}(\alpha) - y\right)_-. \tag{4}$$

In either formulation, the key observation is that the Bayes act is the quantile of the forecast distribution F at the probability level given by one minus the cost/loss ratio C/L.

2.2.2 CRPS as an integrated quantile score

The scoring rule S_Q of Equation (4) evaluates the forecast distribution F only through its α quantile. While this is faithful to the context of the decision-making problem, it may not be satisfying as a measure of the quality of the full forecast distribution. For this purpose, one option is to integrate the quantile scoring rule across different values of the probability level α, weighting the probability levels according to a specified distribution p. This yields a weighted CRPS:

$$S_{CRPS}(F, y; p) = \int S_Q(F, y; \alpha)p(\alpha)\, d\alpha.$$

This weighted form of CRPS has appeared in the literature before, e.g. see Gneiting and Ranjan [2011]. The usual CRPS results from taking p corresponding to a Uniform$(0, 1)$ distribution, equally weighting all probability levels.

We emphasize that because $\alpha = 1 - C/L$, the distribution $p(\alpha)$ can be interpreted as expressing incomplete knowledge about the cost/loss ratio in the decision-making problem. The equal weighting used by the ordinary CRPS may be appropriate in the absence of any knowledge about the context in which forecasts will be used to support decision-making, but may be inappropriate if more information is known about the cost/loss ratio for a specific decision-making task.

The weighted interval score (WIS) is often used when the full forecast distribution F is not available, as in the U.S. Forecast Hub and similar efforts where forecasts are represented by a collection of prediction intervals. WIS is a discrete approximation to CRPS, and can be obtained by using a distribution p that has point masses at the probability levels corresponding to the endpoints of a finite set of prediction intervals.

2.3 The allocation score

ELR:[If we like the organization of the previous subsection into two subsubsections about the quantile score and the CRPS, we should replicate that here.]

We now develop a scoring rule for probabilistic forecasts that measures the value of a forecast as an input to decision making about how to allocate limited resources to meet demand across multiple locations. As a concrete example, we take the resource to be a good such as ventilators or oxygen supply. An administrator is tasked with determining where to send these resources so as to meet demand among hospital patients in different facilities or states. In contrast to the decision-making

problem in the previous section, the administrator is not able to control the total amount of supply; rather, their task is to determine how to allocate the fixed supply to different locations.

In this decision-making setting, an action $\mathbf{x} = (x_1, \ldots, x_n)$ is a vector specifying the amount that is allocated to each of the n locations. We require that each x_i is non-negative and that the total allocation across all locations does not exceed a constraint K on the total available resources: $\sum_{i=1}^{n} x_i \leq K$. **ELR:[Since we got rid of the C parameter, maybe we should just make this a hard constraint here, $\sum_i x_i = K$?]** The set $\mathcal{X}$ collects all possible allocations that satisfy these constraints. The eventually realized resource demand in each location is denoted by $\mathbf{y} = (y_1, \ldots, y_n)$. Again, these levels of demand are not known at the time of decision making, so we define the random vector $Y = (Y_1, \ldots, Y_n)$ where Y_i represents the as-yet-unknown level of resource demand in location i. Forecasts of demand in each location are collected in $F = (F_1, \ldots, F_n)$. We assume that the forecasts do not allow for the possibility of negative demand, i.e. the support of each F_i is a subset of $\mathbb{R}^+$. As in the previous section, we assume that a loss L is incurred for each unit of unmet need.

We note that a number of generalizations to this loss specification have been formulated in the literature, including an allowance for costs for over-allocation to a particular unit (e.g. if there are storage costs for unused resources), differing losses different units (e.g. if a unit of unmet demand imposes more severe costs in one location than another), and the introduction of a convex function that controls the rate at which costs accrue depending on the scale of need. We consider these and other generalizations in other work.

With this notation in place, we can develop a proper scoring rule following the outline in section 2.1.

Step 1: specify a loss function. The loss associated with a particular allocation is calculated by summing contributions from unmet demand in each location:

$$s_A(\mathbf{x}, \mathbf{y}; L) = \sum_{i=1}^{n} L(x_i - y_i)_-. \tag{5}$$

As in the previous section, $(x_i - y_i)_-$ is 0 if the amount x_i allocated to unit i is greater than or equal to the realized demand y_i in that location; otherwise, it is $y_i - x_i$, the amount of unmet need in that location.

Step 2: Given a probabilistic forecast F, identify the Bayes act. The Bayes act is the allocation that minimizes the expected loss:

$$\mathbf{x}^F = \underset{\mathbf{x} \in \mathbb{R}^N, 0 \leq \mathbf{x}}{\operatorname{argmin}} \; \overline{s}_A^F(\mathbf{x}; L) \text{ subject to } \sum_{i=1}^{N} x_i \leq K, \text{ where} \tag{6}$$

$$\overline{s}_A^F(\mathbf{x}; L) = \mathbb{E}_F[s_A(\mathbf{x}, \mathbf{Y}; L)] = \sum_{i=1}^{N} \mathbb{E}_{F_i}[s_A(x_i, Y_i; L)] \tag{7}$$

It can be shown that with the loss function given in Equation (5), the Bayes act has $x_i^F = F_i^{-1}(1 - \lambda^\star / L)$, where $\lambda^\star$ depends on the problem parameters K and L as well as the forecast distributions and is chosen so as to satisfy the equation

$$\sum_{i=1}^{N} F_i^{-1}(1 - \lambda^\star / L) = K. \tag{8}$$

This partial solution to the allocation problem seems to have first appeared in Hadley and Whitin [1963]; see the supplemental materials for a derivation. Figure *** panel (a) **ELR:[TODO: first figure from notes crps_connection_exp.Rmd]** illustrates the expected loss function $\bar{s}_A^F$ and the allocation given by the Bayes act in a simple example with $n = 2$ locations, $L = 1$, $K = 5$, and forecasts $Y_1 \sim \text{Exp}(1/\sigma_1)$ and $Y_2 \sim \text{Exp}(1/\sigma_2)$, using a scale parameterization of the exponential distribution with $\sigma_1 = 1$ and $\sigma_2 = 5$.

One interpretation of this result is that the Bayes act sets the allocation in each location i to a quantile of the forecast distribution F_i for that location. The quantile is at a probability level $\alpha = 1 - \lambda^\star/L$ that is the same for all locations, and is chosen such that the constraint is satisfied. An alternative interpretation comes from noting that for each location i, $\frac{\partial}{\partial x_i} \bar{s}_A^F(\mathbf{x})|_{\mathbf{x}=\mathbf{x}^F} = \lambda^\star$ (see the supplement for a proof). In words, at the allocation given by the Bayes act, the rate of change of the expected score as a function of the amount allocated to location i is given by $\lambda^\star$. This derivative is the same for all locations, so the optimal allocation divides the available resources across all locations in such a way that according to F, the expected benefit of 1 additional unit of resources is the same in all locations.

APG:[first attempt by APG to rephrase last paragraph: The central mathematical ideas in this construction are that

- **optimization under a *single* constraint requires the rates of change of (a probabilistic forecast's) expected benefit with respect to our decision variables, x_i, to be the same for all locations**
- **these rates of change can be identified with probabilities**
- **and therefore the Bayes act results from using a *shared* probability level, $1-\lambda^\star/L$, to determine allocations as the corresponding quantiles of the location-specific forecast distributions F_i which satisfy the constraint.**

(See the supplement for detailed discussion and derivations of these points.)]

Step 3: Define the scoring rule. We can now define a proper scoring rule for the probabilistic forecast F as

$$S_A(F, y; L, K) = s_A(\mathbf{x}^F, y; L) = \sum_{i=1}^{n} L(F_i^{-1}(1 - \lambda^\star/L) - y_i)_- \tag{9}$$

This score measures the total unmet need across all locations that results from using the Bayes allocation associated with the forecast F when the actual level of need in each location is observed to be y_i.

As with the quantile score S_Q, the allocation score S_A measures the skill of the forecast distributions F based on a single probability level α. By analogy to the method for obtaining CRPS by integrating the quantile score, we develop an *integrated allocation score* (IAS) **ELR:[open to better names]** that integrates the allocation score across values of the problem parameters, weighting by a distribution p:

$$S_{IAS}(F, y) = \int S_A(F, y; L, K) p(L, K)\, dL dK$$

We illustrate the relationship between the IAS and the CRPS in our example with two locations and forecasts given by $\text{Exp}(1/\sigma_i)$ distributions with $\sigma_1 = 1$ and $\sigma_2 = 5$. Note that the quantile functions corresponding to these forecasts are given by $F_i^{-1}(\alpha) = -\sigma_i \log(1 - \alpha)$. For simplicity, we keep $L = 1$ fixed (i.e., p places probability 1 on $L = 1$), and only address varying K. As discussed above, each value

of the constraint K determines a quantile probability level α corresponding to the Bayes act such that $K = F_1^{-1}(\alpha) + F_2^{-1}(\alpha) = -\log(1-\alpha)(\sigma_1 + \sigma_2)$; solving for α, we obtain $\alpha = 1 - \exp[-K/(\sigma_1 + \sigma_2)]$. This link between the constraint level K and the probability level α is the key to the link between the IAS and the CRPS.

We use this link to explore the relationship between IAS and CRPS from two directions. First, suppose the decision-maker has some uncertainty about the value of K, which they express through p. Because α can be regarded as a function of K, this distribution on the resource constraint induces a distribution on quantile levels. Figure *** panel (b) **ELR:[TODO, figures from crps_connection_exp]** illustrates with a Gamma$(500, 0.01)$ distribution for K, and the implied distribution on quantile levels is shown in panel (c). The IAS determined by p corresponds to a weighted CRPS with this induced weighting on quantile levels. However, note that this weighting is specific to this pair of Exponential forecasts; a different pair of forecasts would translate to a different weighting on quantile levels.

Figure *** panel (d) **ELR:[TODO, figures from crps_connection_exp]** illustrates this link by going in the other direction: given a forecast F, we exhibit the distribution on K that would lead to equally weighted CRPS. Now we use the fact that K can be written as a function of α to obtain the distribution on K that corresponds to a Uniform$(0, 1)$ distribution on α. In this example, the implied distribution is $K \sim \text{Exp}(\sigma_1 + \sigma_2)$. We observe that this is a right-skewed distribution that places much of its mass on constraint values near 0, which may not correspond well to actual knowledge about the resource constraints. Again, the distribution on resource constraints that corresponds to unweighted CRPS depends on the forecast distributions.

3 Application

We illustrate with an application to hospital admissions in the U.S., considering the problem of allocation of a limited supply of medical resources to the states.

Case study heading into the Omicron wave. Some more detailed discussion of implications of bad forecasts for specific decision-making purposes – take a "deep dive" into one or two example states like FL.

Look at results over a broader range of time.

4 Discussion

We often conceive of infectious disease forecasts as being useful for decision-making purposes, but it is rare for forecast evaluation to be tied directly to the value of the forecasts for informing those decisions. This work seeks to address that gap.

We have demonstrated that evaluation methods that are tied to decision-making context can yield model rankings that are substantively different from generic measures of forecast skill like WIS.

In practice, there are many users of forecasts with many different decision-making problems. Not all can be easily quantified. Those that can be easily quantified may differ enough that no single score is appropriate for all users. We suggest reporting multiple scores. This may be tricky to operationalize in the setting of a general forecast hub. It matters how you elicit and represent probabilistic forecasts (quantiles? samples? cdfs?).

The allocation score we developed here does not directly account for important considerations such as fairness/equity of allocations.

The allocation score we developed also does not attempt to capture the broader context of decision-making. For example, in practice it may be possible to increase the resource constraint K by shifting funding from other disease mitigation measures.

Forecaster's dilemma: a successful forecast may lead to decisions that change the distribution of the outcome Y. Our framework cannot be used in those settings.

There is much more to do in this general area.

5 References

References

Ledor S Igboh, Katherine Roguski, Perrine Marcenac, Gideon O Emukule, Myrna D Charles, Stefano Tempia, Belinda Herring, Katelijn Vandemaele, Ann Moen, Sonja J Olsen, et al. Timing of seasonal influenza epidemics for 25 countries in africa during 2010–19: a retrospective analysis. *The Lancet Global Health*, 11(5):e729–e739, 2023.

Martin I Meltzer, Charisma Y Atkins, Scott Santibanez, Barbara Knust, Brett W Petersen, Elizabeth D Ervin, Stuart T Nichol, Inger K Damon, and Michael L Washington. Estimating the future number of cases in the ebola epidemic–liberia and sierra leone, 2014–2015. 2014.

Gabriel Rainisch, Manjunath Shankar, Michael Wellman, Toby Merlin, and Martin I Meltzer. Regional spread of ebola virus, west africa, 2014. *Emerging Infectious Diseases*, 21(3):444, 2015.

Centers for Desease Control and Prevention.

Dimitris Bertsimas, Leonard Boussioux, Ryan Cory-Wright, Arthur Delarue, Vassilis Digalakis, Alexandre Jacquillat, Driss Lahlou Kitane, Galit Lukin, Michael Li, Luca Mingardi, Omid Nohadani, Agni Orfanoudaki, Theodore Papalexopoulos, Ivan Paskov, Jean Pauphilet, Omar Skali Lami, Bartolomeo Stellato, Hamza Tazi Bouardi, Kimberly Villalobos Carballo, Holly Wiberg, and Cynthia Zeng. From predictions to prescriptions: A data-driven response to covid-19. *Health Care Management Science*, 24:253–272, 2021.

Spencer J. Fox, Michael Lachmann, Mauricio Tec, Remy Pasco, Spencer Woody, Zhanwei Du, Xutong Wang, Tanvi A. Ingle, Emily Javan, Maytal Dahan, Kelly Gaither, Mark E. Escott, Stephen I. Adler, S. Claiborne Johnston, James G. Scott, and Lauren Ancel Meyers. Real-time pandemic surveillance using hospital admissions and mobility data. *Proceedings of the National Academy of Sciences*, 119 (7):e2111870119, February 2022. doi: 10.1073/pnas.2111870119. URL `https://www.pnas.org/doi/10.1073/pnas.2111870119`. Publisher: Proceedings of the National Academy of Sciences.

University of Texas at Austin.

Elizabeth Yardley and Fotios Petropoulos. Beyond error measures to the utility and cost of the forecasts. *Foresight: The International Journal of Applied Forecasting*, (63):36–45, 2021.

M Hashem Pesaran and Spyros Skouras. Decision-based methods for forecast evaluation. *A companion to economic forecasting*, pages 241–267, 2002.

Gordon Leitch and J Ernest Tanner. Economic forecast evaluation: profits versus the conventional error measures. *The American Economic Review*, 81(3):580–590, 1991.

Tolga Cenesizoglu and Allan Timmermann. Do return prediction models add economic value? *Journal of Banking & Finance*, 36(11):2974–2987, 2012.

John PA Ioannidis, Sally Cripps, and Martin A Tanner. Forecasting for covid-19 has failed. *International journal of forecasting*, 38(2):423–438, 2022.

William J.M. Probert, Katriona Shea, Christopher J. Fonnesbeck, Michael C. Runge, Tim E. Carpenter, Salome Dürr, M. Graeme Garner, Neil Harvey, Mark A. Stevenson, Colleen T. Webb, Marleen Werkman, Michael J. Tildesley, and Matthew J. Ferrari. Decision-making for foot-and-mouth disease control: Objectives matter. *Epidemics*, 15:10–19, 2016. ISSN 1755-4365. doi: https://doi.org/10.1016/j.epidem.2015.11.002. URL `https://www.sciencedirect.com/science/article/pii/S175543651500095X`.

Craig J. McGowan, Matthew Biggerstaff, Michael Johansson, Karyn M. Apfeldorf, Michal Ben-Nun, Logan Brooks, Matteo Convertino, Madhav Erraguntla, David C. Farrow, John Freeze, Saurav Ghosh, Sangwon Hyun, Sasikiran Kandula, Joceline Lega, Yang Liu, Nicholas Michaud, Haruka Morita, Jarad Niemi, Naren Ramakrishnan, Evan L. Ray, Nicholas G. Reich, Pete Riley, Jeffrey Shaman, Ryan Tibshirani, Alessandro Vespignani, Qian Zhang, and Carrie Reed. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Scientific Reports*, 9(1):683, January 2019. ISSN 2045-2322. doi: 10.1038/s41598-018-36361-9. URL `https://www.nature.com/articles/s41598-018-36361-9`.

Nicholas G. Reich, Logan C. Brooks, Spencer J. Fox, Sasikiran Kandula, Craig J. McGowan, Evan Moore, Dave Osthus, Evan L. Ray, Abhinav Tushar, Teresa K. Yamana, Matthew Biggerstaff, Michael A. Johansson, Roni Rosenfeld, and Jeffrey Shaman. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 116(8):3146–3154, February 2019. ISSN 1091-6490. doi: 10.1073/pnas.1812594116.

Michael A. Johansson, Karyn M. Apfeldorf, Scott Dobson, Jason Devita, Anna L. Buczak, Benjamin Baugher, Linda J. Moniz, Thomas Bagley, Steven M. Babin, Erhan Guven, Teresa K. Yamana, Jeffrey Shaman, Terry Moschou, Nick Lothian, Aaron Lane, Grant Osborne, Gao Jiang, Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, Roni Rosenfeld, Justin Lessler, Nicholas G. Reich, Derek A. T. Cummings, Stephen A. Lauer, Sean M. Moore, Hannah E. Clapham, Rachel Lowe, Trevor C. Bailey, Markel García-Díez, Marilia Sá Carvalho, Xavier Rodó, Tridip Sardar, Richard Paul, Evan L. Ray, Krzysztof Sakrejda, Alexandria C. Brown, Xi Meng, Osonde Osoba, Raffaele Vardavas, David Manheim, Melinda Moore, Dhananjai M. Rao, Travis C. Porco, Sarah Ackley, Fengchen Liu, Lee Worden, Matteo Convertino, Yang Liu, Abraham Reddy, Eloy Ortiz, Jorge Rivero, Humberto Brito, Alicia Juarrero, Leah R. Johnson, Robert B. Gramacy, Jeremy M. Cohen, Erin A. Mordecai, Courtney C. Murdock, Jason R. Rohr, Sadie J. Ryan, Anna M. Stewart-Ibarra, Daniel P. Weikel, Antarpreet Jutla, Rakibul Khan, Marissa Poultney, Rita R. Colwell, Brenda Rivera-García, Christopher M. Barker, Jesse E. Bell, Matthew Biggerstaff, David Swerdlow, Luis Mier-Y-Teran-Romero, Brett M. Forshey, Juli Trtanj, Jason Asher, Matt Clay, Harold S. Margolis, Andrew M. Hebbeler, Dylan George, and Jean-Paul Chretien. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 116(48):24268–24274, November 2019. ISSN 1091-6490. doi: 10.1073/pnas.1909865116.

Estee Y. Cramer, Evan L. Ray, Velma K. Lopez, Johannes Bracher, Andrea Brennen, Alvaro J. Castro Rivadeneira, Aaron Gerding, Tilmann Gneiting, Katie H. House, Yuxin Huang, Dasuni Jayawardena, Abdul H. Kanji, Ayush Khandelwal, Khoa Le, Anja Mühlemann, Jarad Niemi, Apurv Shah, Ariane Stark, Yijin Wang, Nutcha Wattanachit, Martha W. Zorn, Youyang Gu, Sansiddh Jain, Nayana Bannur, Ayush Deva, Mihir Kulkarni, Srujana Merugu, Alpan Raval, Siddhant Shingi, Avtansh Tiwari, Jerome White, Neil F. Abernethy, Spencer Woody, Maytal Dahan, Spencer Fox, Kelly Gaither, Michael Lachmann, Lauren Ancel Meyers, James G. Scott, Mauricio Tec, Ajitesh Srivastava, Glover E. George, Jeffrey C. Cegan, Ian D. Dettwiller, William P. England, Matthew W. Farthing, Robert H. Hunter, Brandon Lafferty, Igor Linkov, Michael L. Mayo, Matthew D. Parno, Michael A. Rowland, Benjamin D. Trump, Yanli Zhang-James, Samuel Chen, Stephen V. Faraone, Jonathan Hess, Christopher P. Morley, Asif Salekin, Dongliang Wang, Sabrina M. Corsetti, Thomas M. Baer, Marisa C. Eisenberg, Karl Falb, Yitao Huang, Emily T. Martin, Ella McCauley, Robert L. Myers, Tom Schwarz, Daniel Sheldon, Graham Casey Gibson, Rose Yu, Liyao Gao, Yian Ma, Dongxia Wu, Xifeng Yan, Xiaoyong Jin, Yu-Xiang Wang, YangQuan Chen, Lihong Guo, Yanting Zhao, Quanquan Gu, Jinghui Chen, Lingxiao Wang, Pan Xu, Weitong Zhang, Difan Zou, Hannah Biegel, Joceline Lega, Steve McConnell, V. P. Nagraj, Stephanie L. Guertin, Christopher Hulme-Lowe, Stephen D. Turner, Yunfeng Shi, Xuegang Ban, Robert Walraven, Qi-Jun Hong, Stanley Kong, Axel van de Walle, James A. Turtle, Michal Ben-Nun, Steven Riley, Pete Riley, Ugur Koyluoglu, David DesRoches, Pedro Forli, Bruce Hamory, Christina Kyriakides, Helen Leis, John Milliken, Michael Moloney, James Morgan, Ninad Nirgudkar, Gokce Ozcan, Noah Piwonka, Matt Ravi, Chris Schrader, Elizabeth Shakhnovich, Daniel Siegel, Ryan Spatz, Chris Stiefeling, Barrie Wilkinson, Alexander Wong, Sean Cavany, Guido España, Sean Moore, Rachel Oidtman, Alex Perkins, David Kraus, Andrea Kraus, Zhifeng Gao, Jiang Bian, Wei Cao, Juan Lavista Ferres, Chaozhuo Li, Tie-Yan Liu, Xing Xie, Shun Zhang, Shun Zheng, Alessandro Vespignani, Matteo Chinazzi, Jessica T. Davis, Kunpeng Mu, Ana Pastore y Piontti, Xinyue Xiong, Andrew Zheng, Jackie Baek, Vivek Farias, Andreea Georgescu, Retsef Levi, Deeksha Sinha, Joshua Wilde, Georgia Perakis, Mohammed Amine Bennouna, David Nze-Ndong, Divya Singhvi, Ioannis Spantidakis, Leann Thayaparan, Asterios Tsiourvas, Arnab Sarker, Ali Jadbabaie, Devavrat Shah, Nicolas Della Penna, Leo A. Celi, Saketh Sundar, Russ Wolfinger, Dave Osthus, Lauren Castro, Geoffrey Fairchild, Isaac Michaud, Dean Karlen, Matt Kinsey, Luke C. Mullany, Kaitlin Rainwater-Lovett, Lauren Shin, Katharine Tallaksen, Shelby Wilson, Elizabeth C. Lee, Juan Dent, Kyra H. Grantz, Alison L. Hill, Joshua Kaminsky, Kathryn Kaminsky, Lindsay T. Keegan, Stephen A. Lauer, Joseph C. Lemaitre, Justin Lessler, Hannah R. Meredith, Javier Perez-Saez, Sam Shah, Claire P. Smith, Shaun A. Truelove, Josh Wills, Maximilian Marshall, Lauren Gardner, Kristen Nixon, John C. Burant, Lily Wang, Lei Gao, Zhiling Gu, Myungjin Kim, Xinyi Li, Guannan Wang, Yueying Wang, Shan Yu, Robert C. Reiner, Ryan Barber, Emmanuela Gakidou, Simon I. Hay, Steve Lim, Chris Murray, David Pigott, Heidi L. Gurung, Prasith Baccam, Steven A. Stage, Bradley T. Suchoski, B. Aditya Prakash, Bijaya Adhikari, Jiaming Cui, Alexander Rodríguez, Anika Tabassum, Jiajia Xie, Pinar Keskinocak, John Asplund, Arden Baxter, Buse Eylul Oruc, Nicoleta Serban, Sercan O. Arik, Mike Dusenberry, Arkady Epshteyn, Elli Kanal, Long T. Le, Chun-Liang Li, Tomas Pfister, Dario Sava, Rajarishi Sinha, Thomas Tsai, Nate Yoder, Jinsung Yoon, Leyou Zhang, Sam Abbott, Nikos I. Bosse, Sebastian Funk, Joel Hellewell, Sophie R. Meakin, Katharine Sherratt, Mingyuan Zhou, Rahi Kalantari, Teresa K. Yamana, Sen Pei, Jeffrey Shaman, Michael L. Li, Dimitris Bertsimas, Omar Skali Lami, Saksham Soni, Hamza Tazi Bouardi, Turgay Ayer, Madeline Adee, Jagpreet Chhatwal, Ozden O. Dalgic, Mary A. Ladd, Benjamin P. Linas, Peter Mueller, Jade Xiao, Yuanjia Wang, Qinxia Wang, Shanghong Xie, Donglin

Zeng, Alden Green, Jacob Bien, Logan Brooks, Addison J. Hu, Maria Jahja, Daniel McDonald, Balasubramanian Narasimhan, Collin Politsch, Samyak Rajanala, Aaron Rumack, Noah Simon, Ryan J. Tibshirani, Rob Tibshirani, Valerie Ventura, Larry Wasserman, Eamon B. O'Dea, John M. Drake, Robert Pagano, Quoc T. Tran, Lam Si Tung Ho, Huong Huynh, Jo W. Walker, Rachel B. Slayton, Michael A. Johansson, Matthew Biggerstaff, and Nicholas G. Reich. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119, April 2022. doi: 10.1073/pnas.2113561119. URL `https://www.pnas.org/doi/full/10.1073/pnas.2113561119`. Publisher: Proceedings of the National Academy of Sciences.

Katharine Sherratt, Hugo Gruson, Rok Grah, Helen Johnson, Rene Niehus, Bastian Prasse, Frank Sandmann, Jannik Deuschel, Daniel Wolffram, Sam Abbott, et al. Predictive performance of multi-model ensemble forecasts of covid-19 across european nations. *Elife*, 12:e81916, 2023.

Felipe J. Colón-González, Leonardo Soares Bastos, Barbara Hofmann, Alison Hopkin, Quillon Harpham, Tom Crocker, Rosanna Amato, Iacopo Ferrario, Francesca Moschini, Samuel James, Sajni Malde, Eleanor Ainscoe, Vu Sinh Nam, Dang Quang Tan, Nguyen Duc Khoa, Mark Harrison, Gina Tsarouchi, Darren Lumbroso, Oliver J. Brady, and Rachel Lowe. Probabilistic seasonal dengue forecasting in Vietnam: A modelling study using superensembles. *PLOS Medicine*, 18 (3):e1003542, March 2021. ISSN 1549-1676. doi: 10.1371/journal.pmed.1003542. URL `https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003542`. Publisher: Public Library of Science.

Anton Camacho, Adam Kucharski, Yvonne Aki-Sawyerr, Mark A White, Stefan Flasche, Marc Baguelin, Timothy Pollington, Julia R Carney, Rebecca Glover, Elizabeth Smout, et al. Temporal changes in ebola transmission in sierra leone and implications for control requirements: a real-time modelling study. *PLoS currents*, 7, 2015.

Tilmann Gneiting and Roopesh Ranjan. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422, 2011. doi: 10.1198/jbes.2010.08110. URL `https://doi.org/10.1198/jbes.2010.08110`.

G. Hadley and Thomson M. Whitin. *Analysis of inventory systems.* Prentice-Hall international series in management. Prentice-Hall, 1963.