Evaluating infectious disease forecasts with allocation scoring rules

Authors

May 1, 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

High level points to cover in introduction:

- People are using infectious disease forecasts as an input to decision making
- There are standard ways to evaluate forecasts that are responsive to decision making context, and use of those methods is relatively common in other fields like economics and meteorology
- However, there's not much work in infectious disease that does this
- In practice, infectious disease forecasts have typically been evaluated with ”off the shelf” scoring rules such as the WIS which is an approximation to CRPS, log score, and so on.
- In this work, our goal is to begin to address that gap. We focus on a resource allocation setting.
- There is past work focusing on resource allocation in the operations research literature, but it doesn't take the step of getting to a measure of forecast skill.
- That is, while there is plenty of work on
  - DM under risk, where probabilities are taken as known and expected utility is maximized
  - DM under uncertainty, where probabilities are taken as unknown and robust (i.e. maximin) decisions are sought

> little seems to have been said on how forecasts can be evaluated by how well they convert a problem of DM under uncertainty to a problem of DM under risk.

Infectious disease forecasts 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. Such decision-making applications include the allocation of limited medical supplies such as ventilators [1], implementation of social distancing measures such as stay-at-home policies, planning site selection for vaccine trials [1], and strategies for public health communication campaigns.

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. These methods have been applied with some regularity in fields such as economics and meteorology. [review previous applications of decision-theoretic evaluation to fields like economics and high level description of scoring procedures they use]

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, [3] 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, [4] 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, probablistic infectious disease forecasts have most often been evaluated with standard, off-the-shelf scoring rules such as the log score, continuous ranked probability score (CRPS), or weighted interval score (WIS). [cite some examples] While some of these scores can be interpreted through the lens of decision theory [thinking here of WIS/quantile loss], these connections are not a common focus of infectious disease forecast evaluation.

In this work, we address this 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. We work with a motivating example where forecasts are used to help set the allocation of a limited quantity of medical supplies across multiple regions.

operations research work on constrained allocation

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. We then illustrate the scoring rule through an application to short-term forecasts of COVID-19 hospital admissions in the US in section 3. Section 4 summarizes our contributions and discusses opportunities for further extensions in future work.

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 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 loss, CRPS, and the weighted interval score

We review here how quantile loss arises from a particular decision-making problem, and how CRPS can be obtained by integrating across values of the parameters of that decision-making problem. These results have been thoroughly discussed in the literature [cite cite cite].

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

$$s_Q(x, y; C, L) = Cx + L(x - y)_-. \tag{1}$$

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 $\alpha = 1 - C/L$:

$$x^F = F^{-1}(\alpha). \tag{2}$$

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

$$\begin{aligned} S_Q(F, y; C, L) &= s_Q(x^F, y; C, L) = Cx^F + L(x^F - y)_- &(3)\\ &= CF^{-1}(\alpha) + L\left(F^{-1}(\alpha) - y\right)_- &(4)\end{aligned}$$

This scoring rule 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 loss across different values of the quantile level α, or, equivalently, across different values of the decision problem's parameters C and L. This gives rise to the (weighted) CRPS:

$$S_{CRPS}(F, y) = \int S_Q(F, y; C, L)p(C, L)\, dC dL$$

APG:[Quick note that I don't think this will work without some more identifications. The integral needs to be over probability levels and thresholds and without the some discussion like the one commented out (currently) at line 123 or (better in my mind) one like I'm trying to do with binary choices at the unit-level, I don't think we have a threshold.] The usual CRPS results from taking p to be the density of a distribution such that the induced distribution on the probability level $\alpha = 1 - C/L$ is Uniform$(0, 1)$. The weighted interval score (WIS) can be used when the full forecast distribution F is not available, and amounts to an approximation of CRPS where the distribution p is discrete with point masses at the quantile levels corresponding to the endpoints of a finite set of prediction intervals.

ELR:[maybe better to use different notation for this integral that's less specific about whether p is discrete or continuous?]

2.3 The allocation score

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:

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

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

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 [2]; 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:

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

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 $\text{Gamma}(500, 0.01)$ 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 $\text{Uniform}(0, 1)$ 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

[1] Dimitris Bertsimas et al. "From predictions to prescriptions: A data-driven response to COVID-19". In: *Health Care Management Science* 24 (2021), pp. 253–272.

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

[3] John PA Ioannidis, Sally Cripps, and Martin A Tanner. "Forecasting for COVID-19 has failed". In: *International journal of forecasting* 38.2 (2022), pp. 423–438.

[4] William J.M. Probert et al. "Decision-making for foot-and-mouth disease control: Objectives matter". In: *Epidemics* 15 (2016), pp. 10–19. ISSN: 1755-4365. DOI: https://doi.org/10.1016/j.epidem.2015.11.002. URL: https://www.sciencedirect.com/science/article/pii/S175543651500095X.