Evaluating infectious disease forecasts with allocation scoring rules

2023-04-24

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.

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.

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 (Bertsimas et al. 2021), implementation of social distancing measures such as stay-at-home policies, planning site selection for vaccine trials (Bertsimas et al. 2021), 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 . 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, (Ioannidis, Cripps, and Tanner 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, 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 , 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 . Section summarizes our contributions and discusses opportunities for further extensions in future work.

Methods

We first give a high-level review of a general procedure for developing proper scoring rules that are tailored to a specific decision-making task, and then use that procedure to develop a score that is suitable for evaluation of forecasts in the context of decisions about allocation of limited resources across multiple locations.

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, 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 demand. 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. We use the letter s for this function to align with the literature on forecast evaluation, in which context s may be used as a *scoring function*.
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: $S(F, y) = s(x^F, y)$.

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).

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

I haven't written anything here. Fill it in, or delete this section?

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.

We first specialize our notation to this decision-making setting. We define an action $\mathbf{x} = (x_1, \dots, x_n)$ as 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: $\sum_{i=1}^{n} x_i \leq K$. The set $\mathcal{X}$ collects all possible allocations that satisfy these constraints. The eventually realized resource demand in each location is denoted by $\mathbf{y} = (y_1, \dots, y_n)$. At the time that a decision-maker sets the resource allocation, the demand $\mathbf{y}$ is not yet known. We therefore define the random vector $Y = (Y_1, \dots, Y_n)$ where Y_i represents the as-yet-unknown level of resource demand in location i. The forecast $F = (F_1, \dots, F_n)$ collects forecasts of demand in each location. Here we identify F_i with its cumulative distribution function (CDF), and F_i^{-1} denotes the quantile function. 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 $\mathbb{R}^+$. With this notation in place, we can proceed to develop a proper scoring rule following the outline in the previous section.

Step 1: specify a loss function. The loss associated with a particular allocation is calculated by comparing the amount allocated to each location to the realized resource demand in that location. Specifically, suppose that there is a marginal cost L that accrues for each unit of demand that is not met. We can calculate the allocation loss across all locations as

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

where $(x_i - y_i)_- := \max(-(x_i - y_i), 0)$ 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.

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, but for the present exposition we restrict our attention to the relatively simple loss formulation of Equation (1).

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

$$\mathbf{x}^F = \underset{\mathbf{x}\in\mathbb{R}^N, 0\leq\mathbf{x}}{\operatorname{argmin}} \ \overline{s}_F(\mathbf{x}) \text{ subject to } \sum_{i=1}^{N} x_i \leq K, \text{ where} \tag{2}$$

$$\overline{s}_F(\mathbf{x}) = \mathbb{E}_F s(\mathbf{x}, \mathbf{Y}) = \sum_{i=1}^{N} \mathbb{E}_{F_i}[s(x_i, Y_i)] \tag{3}$$

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

$$\sum_{i=1}^{N} F_i^{-1}(1 - \lambda^\star/L) = K. \tag{4}$$

This partial solution to the allocation problem seems to have first appeared in (Hadley and Whitin 1963); see the supplemental materials for a derivation.

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 $(1 - \lambda^\star/L)$ 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, $\frac{\partial}{\partial x_i}\overline{s}_F(\mathbf{x})|_{\mathbf{x}=\mathbf{x}^F} = \lambda^\star$ (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.

1st 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

$$S(F, y) = s(\mathbf{x}^F, y) = \sum_{i=1}^{n} L(F_i^{-1}(1 - \lambda^\star/L) - y_i)_- \tag{5}$$

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.

ELR: This section would likely benefit from a figure or two to illustrate the ideas...?

Description of what happens if there is uncertainty about the the constraint K. Integration across K results in something like a forecaster-specific weighted CRPS. It might help to introduce notation like $\lambda^\star(K)$ or $S_K(F, y)$ throughout all of the above discussion? Then if we have a prior $p(K)$ on K, we get to something like

$$S(F, y) = \int S_K(F, y)p(K)\, dK$$

as our final score.

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.

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.

References

Bertsimas, Dimitris, Leonard Boussioux, Ryan Cory-Wright, Arthur Delarue, Vassilis Digalakis, Alexandre Jacquillat, Driss Lahlou Kitane, et al. 2021. "From Predictions to Prescriptions: A Data-Driven Response to COVID-19." *Health Care Management Science* 24: 253–72.

Hadley, G., and Thomson M. Whitin. 1963. *Analysis of Inventory Systems.* Prentice-Hall International Series in Management. Prentice-Hall.

Ioannidis, John PA, Sally Cripps, and Martin A Tanner. 2022. "Forecasting for COVID-19 Has Failed." *International Journal of Forecasting* 38 (2): 423–38.

Probert, William J. M., Katriona Shea, Christopher J. Fonnesbeck, Michael C. Runge, Tim E. Carpenter, Salome Dürr, M. Graeme Garner, et al. 2016. "Decision-Making for Foot-and-Mouth Disease Control:

Objectives Matter." *Epidemics* 15: 10–19. https://doi.org/https://doi.org/10.1016/j.epidem.2015.11.002.