SN Computer Science (2021) 2:423
https://doi.org/10.1007/s42979-021-00810-6

ORIGINAL RESEARCH

Decision Intelligence for Nationwide Ventilator Allocation During the COVID-19 Pandemic

Jiajun Xu[1] · **Suvrajeet Sen**[2]

Received: 5 February 2021 / Accepted: 5 August 2021 / Published online: 21 August 2021

Abstract

Many states in the U.S. have faced shortages of medical resources because of the surge in the number of patients suffering from COVID-19. As many projections indicate, the situation will be far worse in coming months. The upcoming challenge is not only due to the exponential growth in cases but also because of inherent uncertainty and lags associated with disease progression. In this paper, we present a collection of models for decision intelligence which provide decision-support for ventilator allocation based on predictions from well-accepted oracles of disease progression. It is clear from our study that without coordination among states, there is a very high risk of ventilator shortages in certain states. However, such shortages can be reduced, provided neighboring states agree to share ventilators as suggested by our models. We show that despite the explosive growth in cases and associated uncertainty in ventilator demand, our simulation results hold the promise of reducing unmet demand, even in the face of significant uncertainty. This paper also provides the first evidence that coordination between neighboring states can lead to significant reduction in ventilator shortages across the U.S.

Keywords COVID-19 · Resource allocation · Stochastic optimization · Decision intelligence · Predictive and Prescriptive modeling

Introduction

The novel Coronavirus pandemic (COVID-19), which appeared in December 2019, has spread to most countries around the world. As of now (Jan. 2021), more than 100 million people have been infected, and more than 2 million of them have died worldwide [6]. The surge in the number of infected patients in several areas is predicted to lead to severe shortages of many essential resources and services, from ventilators to ICU beds, and even qualified health-care professionals. Several states in the U.S., such as New York and New Jersey, have already faced a shortage of ventilators during the first surge in the Spring of 2020. In addition, with predictions of a more severe outbreak in the upcoming Spring [5, 9, 19, 21, 25], it is wise to start planning to mitigate the impact of a worsening pandemic.

Coordination between states might have helped the shortage situation, but there was no effective mechanism to help states meet their ventilator needs in the absence of coordination tools. The tools we refer to here are allocation algorithms that would allow states to share ventilator inventory, thus minimizing shortages across the country. With disease spread being a spatio-temporal process and the resource allocation needs among states leading to combinatorial explosion, it is impossible for human intelligence to provide the type of decision-support which would minimize the number of patients with unmet ventilator needs. What is necessary is a decision intelligence system to augment the human analytical capacity so that collaborative solutions can be obtained and shortages minimized.

In preparation for the next surge, we have designed a system that integrates predictions of state-by-state ventilator demand, together with data on the current location of

This article is part of the topical collection "Computer Aided Methods to Combat COVID-19 Pandemic" guest edited by David Clifton, Matthew Brown, Yuan-Ting Zhang and Tapabrata Chakraborty.

✉ Suvrajeet Sen
suvrajes@usc.edu

[1] Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, USA

[2] Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, USA

all ventilators and cost of transportation to come up with a plan. This plan allows decision makers in all 50 states to work towards a solution that reduces the likelihood of not providing ventilator support for patients throughout the U.S. Unfortunately, the surge appears to be at our doorstep in the U.S. The Institute for Health Metrics and Evaluation (IHME) website for COVID-19[1] reported about 300,000 deaths since the start (until December 17, 2020), and predicted approximately 470,000 deaths by March 1, 2021. As for hospitalizations, the number of cases on November 1, 2020, were approximately 101,000, and by the first week of December 2020, that number (hospitalizations) had more than *doubled* to 205,000 individuals. This snapshot should give the reader pause; the speed with which COVID-19 was spreading disease and death in the U.S. was very alarming.

We hope that by combining modern tools of decision intelligence, together with predictive models of uncertainty, one may slow down this raging pandemic by planning precision deployment of critical resources and hopefully saving lives! As more and more people contract COVID-19 across the U.S., current hospital resources may not meet the demand for critical resources, such as ventilators. The pandemic's pace induces a dynamic demand for ventilators, and an agile nationwide allocation policy is necessary to respond to predictions, especially if and when the next surge does materialize. To allocate resources effectively, the first challenge is to forecast the needs in this fast-changing environment accurately. Due to many unpredictable factors, the forecasts may only be reliable for only a couple of weeks. In addition, overestimating needs in one state might induce shortage in another, while underestimating will automatically lead to insufficient supplies. Without decision intelligence to guide human teams, the response to COVID-19 may continue to look like a "whac-a-mole" game, with the pandemic popping up across the country in its march through the U.S. all winter. In this paper, we wish to provide evidence that coordination between states has the potential to help reduce the number of deaths by giving timely ventilator support as needed on a dynamic basis.

Due to our specific focus on COVID-19, this paper only focuses on ventilators, but the applicability of the approach extends beyond ventilators, or even beyond COVID-19 for that matter. Moreover, due to the potential for broader community interest in this topic, we have organized this paper in a manner that should be accessible to a wider audience. Because of this goal, our recipe for this paper includes a non-mathematical preview in the next two sections dedicated to predictions and optimization for resource allocation. These are then followed by specific modeling approaches for stochastic optimization (SO) in the "Methods" section. We expect that all readers will be able to appreciate that balancing model-fidelity with algorithmic SO bears fruit in the form of a more effective allocation of scarce resources during the COVID-19 pandemic.

Oracle-Driven Demand Prediction

While our optimization models will coordinate allocations based on requirements from all states in the U.S., the predictions of COVID-19 cases will be undertaken on a state-by-state basis. There are several reasons to justify this, most of which have to do with the manner in which the states manage their affairs, especially data for their patients. For this reason, we record the need for ventilator requirements provided by each state, which in turn works with hospitals within the state to predict demand for ventilators over time. One of the challenges with demand prediction for COVID-19 is that there can be a significant lag, as well as the uncertainty associated with disease progression. As a result, allocation methods necessary for this kind of pandemic call for plans which can adapt as the uncertainty unfolds. Moreover, it is crucial to bear in mind that the choice of planning horizons are critical. Indeed, the nature of disease progression appears to suggest a rolling horizon approach, in which plans are put forward weekly for a week-long window, and then these need to be updated as the weeks march on.

One reasonably well-accepted approach for such planning models is a formalism known as "Stochastic Optimization (SO) with Recourse" [3]. Normally, such SO models assume that the distribution associated with future uncertain events is available for use within the decision model. However, the large number of states in the U.S. and the rapidly changing forecasts of the disease suggests that an oracle-driven approach may be the most appropriate. The specific oracle we use is the IHME model [5] mentioned in the first footnote. In other words, we will accept the uncertainty bands created by the IHME model for each state and combine these uncertain forecasts with a two-stage SO model as mentioned above. In the IHME model, the forecasts are based on empirically confirmed COVID-19 population death rate curves, which take the transmission of the virus and the fatality rate into consideration. The data on confirmed COVID-19 deaths by day is collected from different sources, such as local governments, WHO websites with third-party aggregators [6]. A nonlinear mixed-effects model is fitted based on these data sources, where the cumulative death rate at each location is assumed to follow a parametrized Gaussian error function. The advantage of using the death rate is that it is more accurately reported than the infection rate, since the testing capacity is limited, and the patients with severe symptoms are more likely to be tested. With the projected death rates, the demand for hospital resources, such as ICU beds and

[1] https://covid19.healthdata.org/united-states-of-america.

ventilators, can be estimated with an individual-level micro-simulation model.

The plans resulting from such SO models consists of two parts: (a) a first stage plan which is intended to position resources before the uncertain demands of the future are revealed, and (b) a collection of second-stage contingency plans which reflect a "recourse/adjustment" depending on how the future unfolds [3]. While the aforementioned book discusses more general sequential (multi-stage) decision processes as well, the two-stage setup, with weekly stages, appears to be most appropriate for two reasons: (a) there is significant evidence to believe that the incubation period of the virus is about 2 weeks, and most prediction models are also updated on a weekly basis. This suggests a week-by-week decision process, with a "look-ahead" (second) week, which accommodates prediction uncertainty. It is important to recognize that the role of the second stage is mainly to let the first stage decision-making process/model recognize that future decisions will be contingency plans that depend on how the future may unfold. By having these contingency decisions of the second-stage incorporate the "what-if" scenarios, the first stage decisions are able to assess the impact of positioning resources "intelligently", without over-committing to any one future scenario. In the presence of forecast uncertainty, accommodating second-stage contingency plans helps overcome myopic decisions.

In addition to spatio-temporal uncertainty associated with disease progression, the question of coordinating allocation decisions across the entire U.S. is an impossible task for humans because of several factors: (a) the number of states in the U.S. leads to a combinatorial explosion in the uncertainty space, (b) the lack of a coordinated system of nationwide response, and (c) from a methodological perspective, the challenge of coordinating predictive approaches under uncertainty with decision intelligence requires an integration of both predictive and prescriptive approaches. This paper presents a model which accommodates the requirements set forth above, and discusses the computational results demonstrating many of the advantages of using decision intelligence for planning under uncertainty.

Many models have been proposed to predict the spread of the virus and the speed of infection [1, 14, 20, 24]. For our formulation, the prediction oracle we use is the model from IHME. The demand for ventilators released on Dec. 17, 2020, for the U.S. is shown in Fig. 1, where the data before the release date are regarded as the confirmed data, and the demand after Dec. 17, 2020, are based on the predictive models. The upper and lower bounds of the predictions give us the 95% prediction intervals, and the mean is the estimated expected value for ventilator demand. To prepare for the upcoming surge, we use back-testing data during the first surge, from March 25 to June 02, 2020, and the early stage of the upcoming surge, from November 19 to December 16. The ivory areas in Fig. 1 illustrate the back-testing time periods. We will test the performance of alternative decision models using the above data and recommend the winner as the tool for future resource allocation.

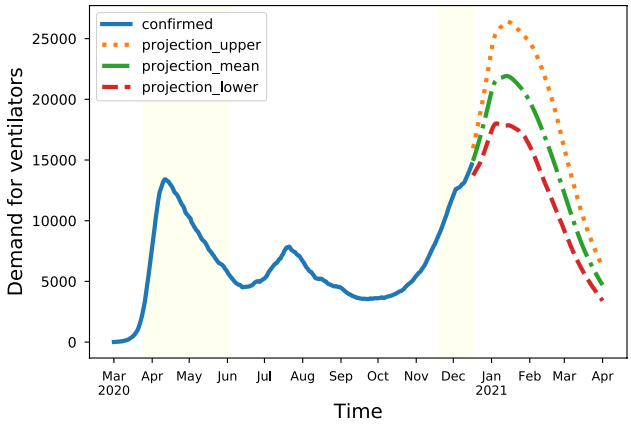


Fig. 1 Demand for ventilators in the U.S., according to data released by [25] on Dec. 17, 2020

Resource Allocation Under Uncertainty

In contrast with parameter optimization required for model-fitting in machine learning, optimization approaches for decision intelligence (also known as decision-support) usually give rise to constrained optimization models that enforce properties that should be satisfied by decisions. For instance, in applications that model resource allocation, such as ventilator allocation, constraints often represent bounds on utilization, e.g., one cannot allocate more ventilators than is available. Accordingly, such prescriptive approaches call for constrained optimization models. Moreover, as the demand for resources changes over time, the requirement constraints also evolve. In this sense, such decision models must track both disease progression as well as resource requirements. It is also essential to recognize that because predictions of disease progression are error-prone, resource utilization decisions lead to stochastic optimization models designed to avoid clairvoyant decisions [22].

The motivation for our models is that with appropriate ventilator allocation, we can satisfy more requests, thus improving the chances of saving more patient-lives. We will also present the situation without any coordination, in which every state is an "island unto itself". In the following, we will begin our study using the expected value from a prediction model commonly used for infectious disease progression [5]. When the error-bars around the prediction are relatively small, point-forecasts may be sufficient for modeling, and a deterministic resource allocation approach may be used to meet expected demand. However, when error-bars

associated with predictions are significant, planning for expected demand may lead to poor (over-confident) decisions, analogous to overfitting in machine learning. Thus, considering the different levels of projected demand, we will adopt a Stochastic Optimization (SO) [3] paradigm that allows us to incorporate forecast uncertainty in the allocation model. Unlike deterministic optimization models formulated with point-forecasts as input parameters, real-world systems often involve uncertainty in demand. The objective of an SO model is to recommend decisions that optimize the expectation or even some measure of risk associated with a function whose value depends on both the decision variables and the random variables. The decision variables will be required to belong to some feasible set of choices, whereas the random variables represent the uncertainty associated with demand predictions.

To curtail the unmet demand for hospital resources (e.g., ventilators) across the United States, we develop a variety of optimization models that include a point-forecasts model, a simple recourse model that considers the variance of the predictions, and a general recourse model. The first one is a deterministic model, and the other two are SO models. Each model reflects a set of assumptions on data and decision processes. Different models lead to alternative allocation of ventilators from their current locations (i.e., all 50 states, and Washington, D.C.) to neighboring states, based on demand projections. Other studies for ventilator allocation have investigated situations covering in-state [18] allocation, or allocation among alternative types of patients [2, 8]. Within the framework of in-state collaboration, studies have also analyzed the risk of shortage with given resources [15]. Hence, our approach will address a macro-allocation problem, where the goal is to satisfy nationwide demands by transferring resources among neighboring states so that most requests may be fulfilled, provided states which ship ventilators to other states have the ability to bring them back when their own needs may be in jeopardy of not being met. In this sense, our model allows a framework that lets each state help their neighbors without hurting the response to their own citizens. We believe that combining these two kinds of allocations will provide better support for patients during the pandemic.

Methods

In our models, we consider an interval of time, say T days, prior to which an allocation is undertaken. The objective of the model is to minimize the unmet demand for resources by optimally distributing them across all states, assuming that only a subset (e.g., neighbors) can share their resources. This is done based on a T day projection of demand across the 50 states and Washington, D.C. Transfer of hospital resources from one location to another is assumed to be completed within a single day. If the resources are transferred on day t, they will be ready for use at the new location on day $t+1$. It is not difficult to see that our models can also accommodate cases, where shipments from state-to-state take longer, or is determined by the pair of states (i, j) under consideration. In this case, the calculation of the unmet demand should consider the resources as being in-transit, and so long as the travel times are known to be deterministic, our model is applicable. Under this assumption, resources which leave the origin i on day t, and require integer travel days of $\tau_{ij} \geq 0$, the resources will arrive at destination j on day $t + \tau_{i,j}$. Due to the fact that most adjacent states are able to ship to each other within a day, we fix $\tau_{i,j} = 1$, although more exact shipment estimates can be easily envisioned.

Before getting into further details regarding the models, let us mention that this section is divided into two subsections, one for deterministic and one for stochastic optimization models. The latter will be subdivided into two further sub-subsections, which will be devoted to SO models under different assumptions. These assumptions lead to different types of models: simple recourse (SR) and general recourse (GR). As the names suggest, the assumptions underlying SR make those models easier to solve, but back-testing suggests that GR model is more effective.

We will begin with notation which span all subsections, and then, proceed to notation which is specific to each particular section (or subsection). We first define the following sets:

- $\mathcal{S} = \{0, 1, ..., 50\}$, the index set of all states and Washington, D.C.
- $\mathcal{S}^+ = \mathcal{S} \cup \{51\}$, where 51 is the index for the Strategic National Stockpile. We assume that once SNS transfers the ventilators to a state, these ventilators belong to the state.
- $\mathcal{S}_j$ = a subset of $\mathcal{S}$, which includes the indexes of the states able to ship resources to location j, $\forall j \in \mathcal{S}^+$.
- $\mathcal{T} = \{1, 2, ..., T\}$, the set of days which the plan covers.

To setup these allocation models, we need to have the data associated with the demand, supply and policy considerations for each state. Our models can also be extended to allocate the multiple resources, if necessary. However, for the current study, we will only consider ventilator allocation.

The supply data includes the number of ventilators that are available for COVID-19 patients in all states and in the Strategic National Stockpile. The policy definitions (or functions) reflect decision rules which may be provided by decision makers (e.g., governors or their surrogates). Such parameters may constrain the maximum number of resources shipped to neighboring states, or set a minimum availability

of resources. The data used for the decision-making models may be summarized as follows:

- a_{jj} = the initial total number of resources that are available at $j \in \mathcal{S}^+$, and belongs to state j.
- a_{ij} = the initial total number of resources that are available at $j \in \mathcal{S}^+$, and belongs to $i \in \mathcal{S}_j$.
- U_{ji} = the maximum number of ventilators belonging to $j \in \mathcal{S}^+$, and can be transferred to $i \in \mathcal{S}$.
- G_j = the minimum number of ventilators belonging to state j which must be retained at location $j \in \mathcal{S}$.
- θ_j = the maximum fraction of a_{jj} used for U_{ji} , that is $U_{ji} = \theta_j a_{jj}$
- ρ_j = the maximum fraction of a_{jj} used for G_j , that is $G_j = \rho_j a_{jj}$

In the above notation, U_{ji}, G_j are guidance from the administration. If the decision makers are conservative and do not wish to share too many resources, they can set a low value for U_{ji} and a high value for G_j. For example, $U_{ji} = 0.2a_{jj}$ means that state j is willing to send at most 20% of its stock to another state. Similarly, $G_j = 0.6a_{jj}$ implies that state j keeps at least 60% of its own ventilators for itself. These two policy parameters can also depend on the predictions, thus creating a non-deterministic policy.

Resource Allocation with Point-Forecasts

We first consider a deterministic decision model which seeks plans according to point-forecasts (i.e., assume that there are no errors in forecasting). This is, of course, unrealistic because of prediction errors are inevitable. However, this unrealistic model will help setup for more realistic SO models which will include the necessary prediction errors. We use the demand data, which includes the daily requirement for ventilators in all states. Let

- d_{jt} = the expected demand for resources at $j \in \mathcal{S}$ on day $t \in \mathcal{T}$.

For a deterministic model (using point forecasts), the decisions we need to make are the number of ventilators shipped from one location to another at $t = 0$. We also define the variables representing the number of available ventilators in all states after the initial allocation, the flow (shipment) variables, and the variables for the unmet demand. The variables are defined as follows:

- x_{ij} = the number of ventilators shipped from location $i \in \mathcal{S}^+$ to location $j \in \mathcal{S}^+$ at time 0. The flow matrix is X, with x_{ij} as the element at (i, j).
- s_{jj} = the total number of ventilators that are available in $j \in \mathcal{S}^+$ after initial allocation, and belongs to j.
- s_{ij} = the total number of ventilators that are available in $j \in \mathcal{S}^+$ after initial allocation, and belongs to $i \in \mathcal{S}_j$.
- Δ_{jt} = unmet demand for ventilators at location $j \in \mathcal{S}$ on day $t \in \mathcal{T}$. The matrix representation is Δ.

The allocation problem aims to minimize the total amount of unmet demand across the country in the periods indexed by $\mathcal{T}$. We assume that once the ventilators are shipped from the SNS to states, they will belong to the state which received them from SNS first. The constraints include flow balance constraints for all locations, policy constraints with G_j and U_{ij}, and the shortfall, which is either zero or the difference between the demand and available ventilators. We further penalize the number of ventilators which are on loan from other states. This penalty can reduce unnecessary sharing. If we choose the penalty coefficient λ between 0 and 1, it is equivalent to the situation that one should not send ventilators to other states unless they can reduce one more unmet demand. The problem can be formulated as

$$\text{Min} \quad \sum_{j\in\mathcal{S},t\in\mathcal{T}} \Delta_{jt} + \lambda\left(\sum_{j\in\mathcal{S},i\in\mathcal{S}_j} s_{ij}\right), \tag{1a}$$

$$\text{s.t.} \quad s_{jj} = a_{jj} - \sum_{i\in\mathcal{S}_j} x_{ji} + x_{51,j}, \forall j \in \mathcal{S}, \tag{1b}$$

$$s_{51,51} = a_{51,51} - \sum_{j\in\mathcal{S}} x_{51,j}, \tag{1c}$$

$$s_{ij} = a_{ij} + x_{ij}, \forall j \in \mathcal{S}, i \in \mathcal{S}_j, \tag{1d}$$

$$\Delta_{jt} \geq \max\left\{0, d_{jt} - s_{jj} - \sum_{i\in\mathcal{S}_j} s_{ij}\right\}, \forall j \in \mathcal{S}, t \in \mathcal{T}, \tag{1e}$$

$$s_{jj} \geq G_j, \forall j \in \mathcal{S}^+, \tag{1f}$$

$$-a_{ij} \leq x_{ij} \leq U_{ij}, \forall i, j \in \mathcal{S}^+. \tag{1g}$$

Note that although one of the inequalities in problem (1) involves a nonlinear function due to the "max" operator, the problem can be solved as a linear program by replacing the "max" operator with two inequalities requiring each term inside the "max" to be less-than-or-equal-to Δ_{jt}. We adopt this formulation style to compress the presentation, especially because some formulations later in the paper have many such conditions. Constraints (1b) indicate the balance of stock in each state, where the right hand side is the initial stock minus the outflows plus the support from SNS.

Constraint (1c) shows the stock in SNS, which equals the initial amount minus the support to states. Constraints (1d) denote shipments of ventilators among states. The unmet demand Δ_{jt} in constraints (1e) is a non-negative value. If the demand is greater than the supply, Δ_{jt} equals their non-negative difference. When the demand is less than the supply, the unmet demand is 0. Equations (1f) and (1g) represent the policy constraints, where the former sets the lower bound for the stock value s_{jj}, and the latter implements the upper bound for the flow value x_{ij}. State i may return the borrowed resources from state j, in which case x_{ij} is negative. In this case we have $-a_{ij} \leq x_{ij}$, as shown in Eq. (1g).

Resource Allocation to Accommodate Forecast Uncertainty

Compared with the prediction model used in the previous subsection, we now consider approaches that accommodate uncertainty. It is important to note that using point-forecasts in an optimization model with data uncertainty gives the optimization algorithm a false sense of data certainty, which in turn, exacerbates the bias due to optimization [3]. Moreover, decisions based on point estimates, as in the previous section, ignore prediction errors, and lead to less generalizable decisions. This concept of generalizability, borrowed from Machine Learning (ML), is explored in greater detail for SO applications in [7].

In this section, we replace point forecasts of the previous subsection with a collection of demand samples, where each sample represents a sample path of potential demands for each day of a week for any given state. In this conceptualization, we model a sample path via a scalar ω which parameterizes demand paths from low to high for an entire period, for a state j. Given this setup, one is able to associate the first percentile of the IHME forecast as the lowest trajectory, and the highest as the 99th percentile. The remaining percentiles can also be simulated by drawing $\omega \in [0, 1]$. Formally then, we define such a random variable for each state as follows.

- $\tilde{\omega}_j \equiv$ a random variable which traces a weekly path representing daily demand for state j (using IHME predictions for the week). We use $(\Omega_j, \mathcal{F}_j, \mathbb{P}_j)$ to denote the probability space consisting of the sample space Ω_j, its measurable collection of subsets $\mathcal{F}_j$, and the measure itself denoted $\mathbb{P}_j$.
- $\tilde{\omega} = [\tilde{\omega}_0, \tilde{\omega}_1, \ldots, \tilde{\omega}_{50}]$. We use $(\Omega, \mathcal{F}, \mathbb{P})$ to denote the probability space, thus $\Omega = \Omega_0 \times \Omega_1 \times \ldots \times \Omega_{50}$. We use ω to denote one sample of $\tilde{\omega}$.
- $d_{jt}(\omega) =$ The demand for ventilators for state $j \in \mathcal{S}$ on day $t \in \mathcal{T}$ under sample/outcome $\omega \in [0, 1]$. Thus, for a given ω, the daily demands for state j is composed of $d_{j1}(\omega), \ldots, d_{jT}(\omega)$, where $T = 7$. Note that the trajectory for a week is clearly correlated according to IHME forecasts. Correlation across states can also be modeled via correlations between ω_k and ω_ℓ $(\ell \neq k)$.

The Simple Recourse (SR) Model

The first stochastic model we consider is the simple recourse model, where we decide to ship ventilators on day $t = 0$, considering all random variables to be independent by state. This assumption of independence among states allows a model which is separable by state. This formulation is called the simple recourse model, since the recourse action will simply calculate linear penalties on the shortage of scarce resources [27]. In the stochastic case, the variables for unmet demands depend on the outcome. We add superscript ω to emphasize this dependency:

- $\Delta^{\omega}_{jt} =$ the unmet demand for ventilators at location $j \in \mathcal{S}$ on day $t \in \mathcal{T}$ under sample $\omega \in \Omega$.

In this case, the problem can be formulated as a two-stage SO model as follows:

$$\text{Min} \quad \lambda \left(\sum_{j \in \mathcal{S}, i \in \mathcal{S}_j} s_{ij} \right) + \mathbb{E}[h(s, \tilde{\omega})], \tag{2a}$$

$$\text{s.t.} \quad (1b) - (1d), (1f) - (1g), \tag{2b}$$

where the expectation is taken with respect to the random variable $\tilde{\omega}$. For any sample $\omega \in \Omega$, we have

$$h(s, \omega) = \text{Min} \sum_{j \in \mathcal{S}, t \in \mathcal{T}} \Delta^{\omega}_{jt}, \tag{3a}$$

$$\text{s.t.} \quad \Delta^{\omega}_{jt} \geq \max\{0, d_{jt}(\omega) - s_{jj} - \sum_{i \in \mathcal{S}_j} s_{ij}\}, \ \forall j \in \mathcal{S}, t \in \mathcal{T}. \tag{3b}$$

In this model, the first stage problem (2) is similar to the deterministic formulation (1), although the objective function is replaced by the expectation (denoted by $\mathbb{E}$) of shortage calculated via the second-stage problem (3). If the predictions for each state are independent of the other states, then a major simplification arises. The objective function evaluation can be subdivided into 51 separable pieces, including 50 states and Washington D.C. In this case, we can define auxiliary variables $u_j = s_{jj} + \sum_{i \in \mathcal{S}_j} s_{ij}$ in Eq. (2). This allows us to separate the unmet demand by state, and then, we have

$$\mathbb{E}[h(s, \tilde{\omega})] = \sum_{j \in \mathcal{S}} \mathbb{E}[h_j(u_j, \tilde{\omega}_j)], \tag{4}$$

where the the expectation on the right hand side is for $\tilde{\omega}_j$. With the sample ω_j, we have

$$h_j(u_j, \omega_j) = \text{Min} \sum_{t \in \mathcal{T}} \Delta_{jt}^{\omega_j}, \tag{5a}$$

$$\text{s.t.} \quad \Delta_{jt}^{\omega_j} \geq \max\{0, d_{jt}(\omega_j) - u_j\}, \ \forall t \in \mathcal{T}. \tag{5b}$$

The function in (5) calculates linear penalties of shortfalls, and for this reason, the SO literature refers to such a model as the "simple recourse" model, which is equivalently known in Machine Learning as a model with weighted ℓ_1 penalty. With this simplification, one is able to evaluate subgradients of h_j for use in the classical Benders/L-Shaped algorithm [3, 26].

General Recourse (GR) Models

Unlike the simple recourse formulation which only considers penalties for deviations from the first-period plan, the general recourse (GR) formulation looks ahead one more period (a week) in order to estimate a reassignment or recourse that may be incurred in the next period. In essence the penalties do not accommodate positioning, and re-positioning of resources in a manner which reflects decisions. The GR model will accommodate such shipments. In fact, the general recourse model is general enough to even allow correlations between ω_ℓ and ω_k, for $\ell \neq k$. However, to avoid confounding our conclusions by choosing a completely different stochastic process with dependence, we continued with the statewise independence assumption from the SR model so that would have fewer uncontrollable outputs of our comparative study. It suffices to say that if IHME produces correlated predictions among states, then, our methodology would be able handle dependent random variables as well.

This second-stage recourse builds in a recognition that as more information becomes available, the allocation of ventilators should be adjusted. Thus depending on the information, there will be a second allocation decision in the GR formulation at the beginning of the second period. We define $\mathcal{T}^+ = \{T+1, ..., 2T-1, 2T\}$ as the time-periods of the second stage. Such a model is also referred to as a "look-ahead" model which tries to foresee adjustments as future contingencies. Accordingly, decisions for the "look-ahead" phase are defined as follows:

- r_{jj}^{ω}= the total number of ventilators that are available in $j \in \mathcal{S}^+$ after initial allocation, and belongs to j after day T under sample $\omega \in \Omega$.
- r_{ij}^{ω} = the total number of ventilators that are available in $j \in \mathcal{S}^+$ after initial allocation, and belongs to $i \in \mathcal{S}_j$ after day T under sample $\omega \in \Omega$
- y_{ij}^{ω} = number of ventilators shipped from location $i \in \mathcal{S}^+$ to location $j \in \mathcal{S}^+$ at time T under sample $\omega \in \Omega$.

For the general recourse model, we allocate ventilators at time $t = 0$, and then adjust the allocation at day T. This model can be formulated as a more general two-stage stochastic optimization problem. The first stage problem is the same as the one in the simple recourse model (i.e., Eq. (2)). Even though the independence assumption is still in effect, we cannot separate one state's model from another, because the allocation linear program creates dependence of the value function across different states. For this reason, we adopt a sampling-based algorithm. Thus, the expectation is replaced by a sub-sampled estimate of the expected value, denoted $\hat{\mathbb{E}}$. Then, the first stage problem is

$$\text{Min} \quad \lambda \left(\sum_{j \in \mathcal{S}, i \in \mathcal{S}_j} s_{ij} \right) + \hat{\mathbb{E}}[h(s, \tilde{\omega})], \tag{6a}$$

$$\text{s.t.} \quad (1b) - (1d), (1f) - (1g). \tag{6b}$$

The second-stage subproblem, which can support alternative formulations, is the contingency model. If we suppose at $t = T$, the total number of ventilators remains the same, then the second-stage problem is as follows:

$$h(s, \omega) = \text{Min} \sum_{j \in \mathcal{S}, t \in \mathcal{T} \cup \mathcal{T}^+} \Delta_{jt}^{\omega} + \lambda \left(\sum_{j \in \mathcal{S}, i \in \mathcal{S}_j} r_{ij}^{\omega} \right), \tag{7a}$$

$$\text{s.t.} \quad \Delta_{jt}^{\omega} \geq \max \left\{ 0, d_{jt}(\omega) - s_{jj} - \sum_{i \in \mathcal{S}_j} s_{ij} \right\}, \ \forall j \in \mathcal{S}, t \in \mathcal{T}, \tag{7b}$$

$$r_{jj}^{\omega} = s_{jj} - \sum_{i \in \mathcal{S}_j} y_{ji}^{\omega} + y_{51,j}^{\omega}, \forall j \in \mathcal{S}, \tag{7c}$$

$$r_{51,51}^{\omega} = s_{51,51} - \sum_{j \in \mathcal{S}} y_{51,j}^{\omega}, \tag{7d}$$

$$r_{ij}^{\omega} = s_{ij} + y_{ij}^{\omega}, \forall j \in \mathcal{S}, i \in \mathcal{S}_j, \tag{7e}$$

$$\Delta_{jt}^{\omega} \geq \max \left\{ 0, d_{jt}(\omega) - r_{jj}^{\omega} - \sum_{i \in \mathcal{S}_j} r_{ij}^{\omega} \right\}, \forall j \in \mathcal{S}, t \in \mathcal{T}^+, \tag{7f}$$

$$r_{jj}^{\omega} \geq G_j, \forall j \in \mathcal{S}, \tag{7g}$$

$$-s_{ij} \le y^{\omega}_{ij} \le U_{ij}, \forall i,j \in \mathcal{S}^{+}. \tag{7h}$$

It is important to distinguish between the "value functions" defined in Eqs. (3) and (7). While they have been defined using the same variables for notational consistency, the former does not use any flows. In contrast, the latter allows the first stage decision to be adjusted based on observations of flows in Eq. (7). The variables r^{ω} represent the stock in the second week, while the variable s is the stock in the first week. The superscript ω shows the dependence on the specific scenario (outcome). Equations (7c)–(7h), which constrain the decision variables in the second week, correspond to the constraints (1b)–(1g), respectively. Moreover, if state $j \in \mathcal{S}$ builds or buys new ventilators that are available at T, we may modify the constraints (7c). If the number of new ventilators is known as n_j, the subproblem is formulated by replacing the constraints (7c) with the following:

$$r^{\omega}_{jj} = s_{jj} - \sum_{i \in \mathcal{S}_j} y^{\omega}_{ji} + y^{\omega}_{51,j} + n_j, \forall j \in \mathcal{S}. \tag{8}$$

Computational Experiments

With the confirmed case data (Mar. 25–Dec. 16, 2020), we observe many states with a shortage of resources. We run back-testing exercises based on out-of-sample historical data. For the following computational experiments, the projected data on patients' need for ventilators are based on predictions made by IHME [5]. We focus on dates around the first demand peak (Mar. 25–June 02, 2020) and the early stage of the next surge (Nov. 19–Dec. 16, 2020). We ensure that for each planning run, we only use data and predictions that were made before the dates for which ventilator allocations are planned. In this exercise, we notice that the demand between June 03 and November 18 is relatively low compared with other weeks used in our study. Due to the lower demand for resources, the supply is able to satisfy requirements for all states. In addition, thus, no shipments of resources are required. The supply remains unchanged during this period. We aim to provide insights to better prepare for the upcoming surge in winter/spring 2021. The need for ventilators in the U.S. is shown in Fig. 1. For our models, we make decisions for sharing resources every week. There are 10 weeks in the first surge, starting from Mar. 25, 2020. Moreover, we make four decisions between November 19 and Dec. 16, 2020. According to [4], Strategic National Stockpile (SNS) has around 12,000 ventilators at the beginning of the pandemic. We collect the number of currently available ventilators in each state and assume that these data were the initial stock on Mar. 25, 2020. Subsequently, ventilator inventory in each week starts from the end state of the previous week. We further assume that the initial allocation on Nov. 19, 2020, is the same as the ending inventory on June 02, 2020; that is, there were no transfers between states during the Fall when there were no significant surges in cases.

We create projections on a rolling-horizon basis, using data generated before the allocation dates. These predicted demands are based on fitting a model using previously confirmed cases. With these projections, we build the decision models to recommend actions, which are then validated against actual demand for the period. The projections are updated weekly, as are the allocation decisions. The available resources in each state are also updated based on these decisions. At time $t (= 0, T, 2T, \ldots)$, we exploit the predicted demand in day $t+1, \ldots, t+T$ (for point-forecasts/SR models), or $t+1, \ldots, t+2T$ (for GR model), to find the optimal allocation decision. We then implement this decision at time t. We should emphasize that in the GR model, only the decision at t will be implemented, although the second- stage decisions are also included in the decision model to avoid "end-of-horizon" effects. When we are at time $t+T$, the decision made on t will be evaluated based on the actual data on day $t+1, \ldots, t+T$. Then, this process will continue for the following week.

For the point-forecasts model, we use the prediction model's expected value as the demand projection. The problems are formulated using Pyomo [10] with CPLEX as the linear programming solver. For the two classes of forecast uncertainty models (i.e., the simple recourse (SR) and the general recourse (GR) model), we restrict each random variable $\tilde{\omega}_j$ to a small set of values, representing the low/medium/high cases. Even with this coarse discretization, the total number of potential multi-dimensional (vector) realizations is as high as $O(3^{51})$(around 10^{24}), assuming that errors around the mean forecast are independent. However, the SR model can be solved without much trouble because of the simplification in Eq. (4). We choose the medium value of demand as the expectation; the midpoint between the predicted upper limit and the medium demand is treated as the high demand; the midpoint between the predicted lower limit and the medium demand is used as the low demand. As in empirical risk minimization, we use an estimated probability of 1/3 for each demand outcome for an individual state. For the SR model, we use the predicted demand in the following week, while for the GR model, we use the prediction model to look ahead one additional week. To find the forecast uncertainty models' solution, we use the codes programmed in C language with CPLEX 12.8 solver. The SR model is solved with the classical Benders/L-Shaped algorithm. As for the general recourse model, we apply a sampling-based algorithm, called Stochastic Decomposition (SD), with compromise decision, where each problem is solved using three replications. We provide a brief overview of SD below.

The SD algorithm is a sequential sampling method that was originally proposed for two-stage stochastic linear programming problems in [12]. It has since been extended by allowing a finite collection of affine functions in [13], and further elaborated in a monograph [11]. As with most sampling-based algorithms, [17] has demonstrated that optimization produces a biased objective function estimate. However, [23] has shown how such bias can be reduced significantly using an extension of "Bagging" which is referred to as a "Compromise Decision". We provide a brief overview of SD in the next couple of paragraphs and point readers to further references below.

The basic idea of the method is to approximate the objective function by the maximum of several piecewise affine functions (subgradients) carried over from past iterations, as well as one generated in the current iteration. By solving a regularized version of the current approximation, one obtains the next decision in the sequence, as well as a collection of Lagrange multipliers during the current iteration. The Lagrange multipliers, which are zero points to those affine pieces which can be purged without sacrificing convergence properties [13]. It can be shown that this process maintains a finite number of affine inequalities throughout the algorithm, and the resulting sequence of iterates provides an optimal solution for the GR version of the SO model with probability one, asymptotically. Nevertheless, the scheme is required to stop in finite time. While the stopping rules of SD are beyond the scope of this presentation, we refer to [16] and [23] for further details on stopping each replication in finite time.

It turns out that the process of producing one decision after running several replications is the key to a variance-reduced decision that can be implemented. This last phase of SD requires one more algorithmic step based on a proximal point iteration that finds an aggregate decision known as the "compromise decision" [23]. This last phase is an extension of the concept of "Bagging" commonly used in Machine Learning, although unlike bagging, the compromise decision also incorporates a stopping rule. It is this stopping rule which *discovers* the sample size based on accuracy requirements imposed on the algorithm via tolerance settings common to numerical optimization. As a result, the user is not required to choose a particular sample size, provided that the data set has a sufficiently large sample. An implementation of the SD algorithm is available through NEOS[2], and its open-source version is available on github[3].

The computational times to obtain the recommended decision for all instances covering 14 weeks of tests is summarized in Table 1. We have not reported the solution times

Table 1 Computational time (in seconds) to find the optimal solution in each period. All computational experiments are conducted on a MacBook Pro with Intel Core i7 processor @2.7 GHz, and 16 GB Memory @2133 MHz

Model	03/25–03/31	04/01–04/07	04/08–04/14	04/15–04/21	04/22–04/28	04/29–05/05	05/06–05/12	05/13–05/19	05/20–05/26	05/27–06/02	–	11/19–11/25	11/26–12/02	12/03–12/09	12/10–12/16
Forecast Uncertainty-SR	0.445	2.099	0.560	1.036	1.144	0.664	0.880	0.313	0.393	0.213	–	0.275	0.251	0.518	0.079
Forecast Uncertainty-GR	6.542	14.023	12.727	23.217	30.091	15.037	25.955	10.057	7.358	3.095	–	5.402	24.761	11.593	2.999

[2] https://neos-server.org/neos/.

[3] https://github.com/USC3DLAB/SD.

Table 2 Total amount of national unmet demand for different models in the period from Mar. 25 to June 02, and from Nov. 19 to Dec. 16, 2020. Each column represents a different percentage of total ventilators available for COVID-19 patients

Model	$P_{\text{covid}} = 50\%$	$P_{\text{covid}} = 60\%$	$P_{\text{covid}} = 70\%$	$P_{\text{covid}} = 80\%$
No-coordination	151,541	124,335	104,330	87,655
Point-Forecasts	46,372	24,176	10,894	6074
Forecast Uncertainty-SR	35,595	8896	372	125
Forecast Uncertainty-GR	28,125	4365	359	127

Table 3 Total unmet demand, in the period from Mar. 25 to Jun. 02, and from Nov. 19 to Dec. 16, 2020, for all models under different policy parameters. P_{covid} is set as 0.6

Model	$\rho_j = 0.6, \theta_j = 0.2$	$\rho_j = 0.5, \theta_j = 0.2$	$\rho_j = 0.5, \theta_j = 0.3$	$\rho_j = 0.4, \theta_j = 0.3$
No-coordination	124,335	124,335	124,335	124,335
Point-Forecasts	23,987	24,176	16,104	15,602
Forecast Uncertainty-SR	9734	8896	5322	4535
Forecast Uncertainty-GR	4468	4365	1396	1170

for the point forecast setup, because it uses an off-the-shelf LP solver. For all SO models, the recommended decision is found in less than a minute, although the general recourse model takes more time than the simple recourse model. However, in the next section, we will show that decisions provided by the GR model perform much better than those provided by SR models.

Model Evaluation for Decision Intelligence

We evaluate the allocation decisions resulting from all three models: one point-forecast (i.e., deterministic) model and two forecast uncertainty models. Each model is tested over a 14-week "rolling horizon" period in which decisions are made 1 week at a time, with the ending state of 1 week providing the starting point of ventilator inventory of the following week. For all models, the connection between states is based on an adjacency matrix[4] which is predefined and essentially captures which states in the database are neighbors of a particular state. The maximum amount of ventilators transferred from state i to state j, and the minimum amount of ventilators that state j want to keep, are policy parameters that can be adjusted by decision makers (e.g., governors). Because ours is a trial experiment, we will explore different setups for these policy parameters.

Since the ventilators will not only be used for COVID-19 patients but also for patients with other conditions (e.g., brain injury, child birth, collapsed lung, coma, etc.), we explore different ratios of ventilators to be used for COVID-19 patients. We let P_{covid} denote the percentage of total available ventilators to be used for the COVID-19 patients. Thus, the initial value of ventilators for the COVID-19 patients in each state is the number of currently available multiplied by P_{covid}. We study situations for which P_{covid} assumes values in the set {50%, 60%, 70%, 80%} of the total available ventilators for the COVID-19 patients. As the pandemic becomes severe, these percentages may change. In our study, we also include a "no-coordination" policy as a baseline. For the "no- coordination" policy, we first distribute the ventilators in SNS to all states proportional to the population. Then, the available resources in each state remain the same at all times. Table 2 shows the total amount of unmet demand in the U.S. in the period from Mar. 25 to June 02, and from Nov. 19 to Dec. 16, with different P_{covid}. Before examining the contents of Table 2, we should remind the reader that these quantities are based on the out-of-sample confirmed demand data with the implementation of decisions at the beginning of each week. As we can see, coordination among states indeed helps the entire nation, and as the supply of resources increases, there is a more significant benefit from coordination. Overall, the forecast uncertainty models outperform the point-forecasts model, since they consider the randomness in the prediction. Besides, the general recourse model, which exploits the projections over a more extended period, has a more significant impact on reducing unmet demand in most cases.

We compared the performance across all models with different policy parameters representing (a) the minimum percentage of ventilators (ρ_j) that state j might want to retain from its own stock, and (b) the maximum percentage of ventilators (θ_j) transferred from state j to its neighbors. Note that we apply a static policy in the experiments. Table 3 shows the total unmet demand for all models under different policy

[4] Although our computations use an adjacency matrix for each state, our models can use any other connection matrix as well. For example, we can connect California with New York or any other state, as required.

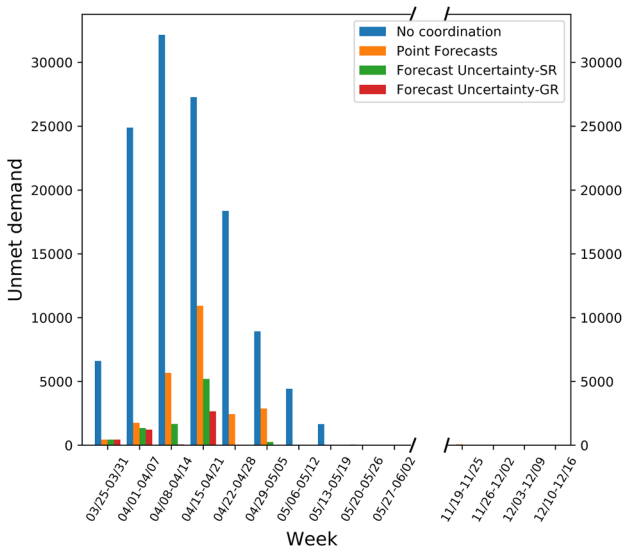


Fig. 2 Amount of unmet demand in different weeks based on the decisions from all models

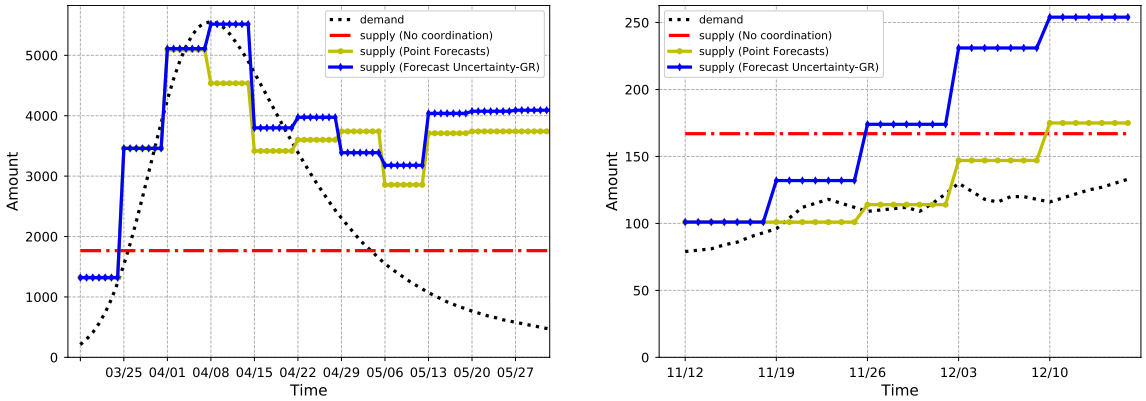
parameters in the covered 14 weeks. We notice from Table 3 that a smaller ρ_j and a larger θ_j will help the nation reduce more unmet demand and exploit the benefits of resource sharing. If a state sends too many extra ventilators to other states, it may face a high risk of not meeting its own demand. However, we can minimize this risk by setting a large penalty parameter λ for lending resources and building a better prediction model.

Figure 2 shows the total amount of unsatisfied demand in each week with $P_{\text{covid}} = 60\%$, $\rho_j = 0.5$ and $\theta_j = 0.2$. The values for the unmet demand are summarized in Table 4. Although the point-forecasts model can substantially reduce the unmet demand, forecast uncertainty models make better decisions under the same policy parameters and quota. The forecast uncertainty models take advantage of the prediction intervals and consider multiple possibilities of future outcomes. Besides, since the general recourse model looks ahead one more week, more information will be considered when making decisions. Since the spread of the disease changes rapidly, the prediction for two successive weeks reflects reality in a reasonable manner. Predictions that are further out in time may be less reliable. Thus, we believe that the general recourse model, which considers the demand in two successive weeks with 1 week being uncertain, is suitable for resource sharing purposes.

In general, an examination of unmet demand in Table 4 reveals that the "no-coordination" strategy of simply shipping out ventilators based on the states' population at the start is not a very smart policy. From the results reported in Table 4, the unmet demands for the "no-coordination" policy were the worst among all policies, because its level of unmet demand is the highest for the period that we tested.

Table 4 Total amount of unmet demand for different models in different weeks of 2020

Model	03/25–03/31	04/01–04/07	04/08–04/14	04/15–04/21	04/22–04/28	04/29–05/05	05/06–05/12	05/13–05/19	05/20–05/26	05/27–06/02	– 11/19–11/25	11/26–12/02	12/03–12/09	12/10–12/16
No-coordination	6603	24,882	32,157	27,269	18,370	8931	4418	1655	50	0	– 0	0	0	0
Point-Forecasts	438	1751	5668	10,927	2433	2880	0	0	0	0	– 70	9	0	0
Forecast Uncertainty-SR	438	1346	1664	5193	0	255	0	0	0	0	– 0	0	0	0
Forecast Uncertainty-GR	438	1219	54	2654	0	0	0	0	0	0	– 0	0	0	0



(a) Supply and demand in NY **(b)** Supply and demand in KA

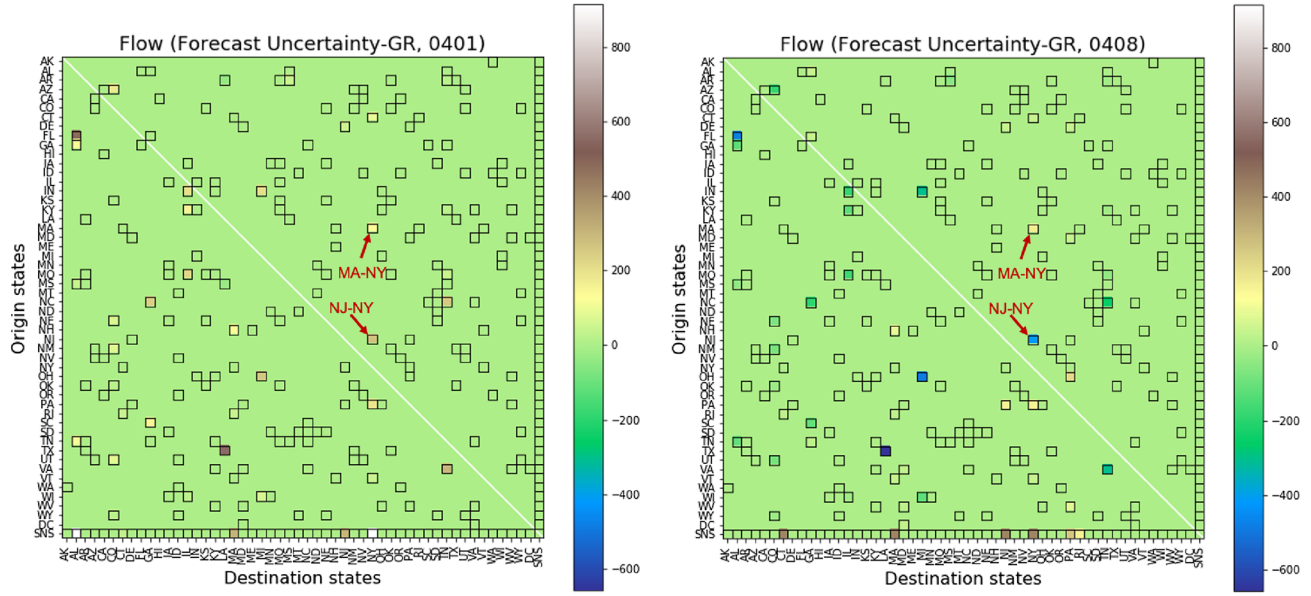
Fig. 3 Supply and demand for the ventilators in New York and Kansas in different time. These figures show the results of the experiment with $P_{\text{covid}} = 60\%, \rho = 0.5, \theta = 0.2$. The states are connected with adjacent states

Of the remaining models, the point-forecasts model is, in general, better than one without coordination. However, because it lacks any foresight regarding uncertainty, it ends up in a poor situation for the start of the second study period in mid-November. As a result, the point forecast model is the only model that fares poorly in the starting phase of the second surge. In fact, the "no-coordination" scheme worked better for the start of the second surge, because the latter did not ship ventilators to other states, which, by our assumptions, had not re-positioned before the start of the second surge. Of course, this situation can be remedied by re-positioning to the initial state during Fall 2020 (September–November 2020).

The GR model's performance is much better than other models, especially during peak-demand. In addition to the "column-by-column" instantaneous advantage of the GR model, there is an accumulated effect as well. To elaborate on this effect, note that during the first 4 weeks of Table 4 the differences between the row for the "Forecast Uncertainty-SR" and that for "Forecast Uncertainty-GR" increases from 0 in week 1, to 127 in week 2, to 1610 in week 3, and then 1539 in week 4, and finally tapers down to 0 in week 5. This pattern is also exhibited for differences when comparing other rows in the same table with the row for "Forecast Uncertainty-GR". These differences emphasize that flexible positioning (due to flows that accommodate demand changes) helps prepare for a surge and potential shortages. Moreover, positioning inventory well makes a continuing difference instead of a model that simply introduces linear penalties, as in the SR model. In other words, accounting for pre-and-re-positioning via the recourse variables r in (7c) supports the kind of decision intelligence which leads to smarter planning.

In periods with lower demand (May 06–Jun. 02, 2020), we can satisfy all the requests, and the unmet demand is reduced to zero. Besides, during the peak of the demand (Apr. 01–21, 2020), the GR model's decisions can reduce the unmet demand up to 95.3% compared with the "no-coordination" case. When the demand is lower, we might not need resource allocation. However, for the other cases, such as the time between April and May, the allocation decisions make a huge contribution to satisfy extra demand for the resources. We observe that the demand from June to October is relatively low. However, starting in November, the number of confirmed cases has rapidly accelerated, and the prediction for the forthcoming spring is much higher than the past surge. Once again, these data suggest that cooperation among states and models such as the GR model may be necessary again for an agile response.

As the three models perform differently, we investigate the decisions made by them. Figure 3 shows the change of supply and demand in New York and Kansas. The demand line shows the confirmed demand across time, while the supply lines indicate the number of available ventilators in each state. We should emphasize that the supply decisions are based on the predicted demand instead of the confirmed demand. In Fig. 3a, based on the GR model, New York acquires more ventilators on Apr. 08, while the point-forecasts model suggests that it should reduce the stock and use fewer ventilators from other states. The main reason is that the demand in the neighboring states, such as New Jersey, increases dramatically. Thus, New York has to return the loaned resources. However, the demand in the following 2 weeks in neighboring states first increase, achieve their peak, then decrease. Since the GR model considers 2-week demand instead of 1 week, it can predict the decrease. Considering that New York has a higher peak value, the rate of demand decrease is slower than others. Thus, it would be more beneficial to send more SNS ventilators to New York instead of other states. Besides, the prediction data



(a) Flow: GR model, Apr. 01 **(b)** Flow: GR model, Apr. 08

Fig. 4 These two figures illustrate the network flow from one state to another based on the decisions from the general recourse model on April 01 and April 08, 2020. The actual connections between the two states are marked with a box. These figures show the results of the experiment with $P_{\text{covid}} = 60\%, \rho_j = 0.5, \theta_j = 0.2$

underestimates the peak demand in New York. Since the GR model considers the projection's variance, more ventilators are transferred from other neighboring states, such as Pennsylvania and Massachusetts. Recall from the results in Table 4, the unmet demand in the entire nation after May 06 is reduced to zero when applying the Point Forecasts and GR models. Although the supply remains quite high in New York in the latter period, compared to the demand (as shown in Fig. 3a), no reallocation is needed. In Fig. 3b, as the demand in Kansas increases starting around Nov. 19, 2020, all the models suggest that the supply should increase as well. However, since the forecast uncertainty models consider various predictions, Kansas obtains more resources from other states.

We also analyze the decisions for "flows", that is, the number of ventilators transferred from one state to another. It helps decision makers understand the reason why the unmet demand decreases when implementing our allocation decisions. Figure 4 provides a visualization of the GR model's "flows" on Apr. 01, and Apr. 08, 2020. The actual connections between two states are marked with a square. A negative value of flow in the figure implies that a state returns ventilators to the origin state. As shown in Fig. 4a, New York (NY) and Alabama (AL) get the most support from the SNS (see the last row). The reason is that, based on the demand prediction data between Apr. 01 and Apr. 07, these two states are the states which suffered most shortages. Moreover, the states neighboring New York, such as New Jersey (NJ) and Massachusetts (MA), send ventilators to meet demand in New York. The arrows, labeled with 'NJ-NY' and 'MA-NY', indicate the number of resources transferred from New Jersey and Massachusetts to New York, respectively. As shown by the 'NJ-NY' arrow in Fig. 4b, on Apr. 08, New York returns the loaned ventilators to New Jersey, since New Jersey is predicted to face increasing demand for the resources. In the meantime, New York gets more supply from Massachusetts (MA), as pointed out by the 'MA-NY' arrow, to meet its surging demand. Such decision intelligence is possible due to the ability of optimization and statistical models to collaborate in a manner which can leverage their strengths in unison.

As the pandemic becomes more severe, more medical resources are ordered from multiple manufacturers. The supply in states as well as in SNS will increase over time. However, the supply may increase linearly, while the demand may have an exponentially increasing rate. Our models provide a system for integrating the demand projections with decisions under uncertainty. They can minimize the loss or unmet demand with limited resources. Besides, our models' rolling horizon scheme can provide the decision starting from any given time and accommodate the case with increasing supply.

Conclusion

The healthcare system in the U.S. encountered extraordinary challenges due to COVID-19, and there are forecasts for variants in the future. The methods of this paper are likely to be useful for managing future pandemics. In this paper, we proposed multiple models, including models with point forecasts (i.e., deterministic) and those which allow forecast uncertainty (i.e., stochastic). Because of the spatio-temporal nature of uncertainty and lags in disease propagation, we propose two stochastic optimization (SO) approaches: one based on the familiar weighted ℓ_1 penalty, also known as "Simple Recourse", and another with a "look ahead" known as "General Recourse". These SO models, especially the latter, provide significant decision intelligence under uncertainty and improve flexibility in responding to the pandemic. Because of their flexibility, the computational results with General Recourse provides evidence that the most reliable plans (with the least number of unmet demands) can relieve the most difficult situations that we have examined in our back-testing exercise. If cases continue to rise, as is being predicted currently, we recommend that the Coronavirus task force prepare to adopt either this or other similar decision intelligence technologies for responding to the pandemic.[5]

Acknowledgements This work was supported by NSF Grant CMMI 1822327, and ONR Grant Number N00014-20-1-2077. We sincerely thank Dr. Julia Higle for introducing this problem to us.

Data/code availability All data, code, and software used in the paper are available online. They are published on the cORe platform.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PloS One. 2020;15(3):e0230405.
2. Beitler JR, Mittel AM, Kallet R, Kacmarek R, Hess D, Branson R, Olson M, Garcia I, Powell B, Wang DS, et al. Ventilator sharing during an acute shortage caused by the COVID-19 pandemic. Am J Respir Crit Care Med. 2020;202(4):600–4.
3. Birge JR, Louveaux F. Introduction to stochastic programming. Berlin: Springer Science & Business Media; 2011.
4. CNN: Administrator for the centers for medicare and medicaid services and member of white house coronavirus task force seema verma interviewed on U.S. preparedness for ongoing coronavirus pandemic; 2020. http://edition.cnn.com/TRANSCRIPTS/2003/13/nday.05.html. Accessed 20 Nov 2020.
5. COVID I, Murray CJ, et al. Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. MedRxiv 2020.
6. CSSE J. Covid-19 data repository by the center for systems science and engineering (csse) at johns hopkins university; 2020. https://github.com/CSSEGISandData/COVID-19. Accessed 2 Nov 2020.
7. Deng Y, Sen S. Predictive stochastic programming. Comput Manag Sci. 2020. https://doi.org/10.1007/s10287-021-00400-0.
8. Emanuel EJ, Persad G, Upshur R, Thome B, Parker M, Glickman A, Zhang C, Boyle C, Smith M, Phillips JP. Fair allocation of scarce medical resources in the time of COVID-19. New England J Med. 2020;382(21):2049–55. https://doi.org/10.1056/NEJMsb2005114.
9. Friedman J, Liu P, Gakidou E, COVID I, Team MC. Predictive performance of international COVID-19 mortality forecasting models. medRxiv 2020. https://doi.org/10.1038/s41467-021-22457-w.
10. Hart WE, Watson JP, Woodruff DL. Pyomo: modeling and solving mathematical programs in python. Math Programm Comput. 2011;3(3):219–60.
11. Higle, J. L., & Sen, S. Stochastic decomposition: a statistical method for large scale stochastic linear programming (Vol. 8). Springer Science & Business Media; 2013.
12. Higle JL, Sen S. Stochastic decomposition: an algorithm for two-stage linear programs with recourse. Math Oper Res. 1991;16(3):650–69.
13. Higle JL, Sen S. Finite master programs in regularized stochastic decomposition. Math Programm. 1994;67:143–68.
14. Hong YR, Lawrence J, Williams D Jr, Mainous Iii A. Population-level interest and telehealth capacity of us hospitals in response to COVID-19: cross-sectional analysis of google search and national hospital survey data. JMIR Public Health Surveill. 2020;6(2):e18961.
15. Huang HC, Araz OM, Morton DP, Johnson GP, Damien P, Clements B, Meyers LA. Stockpiling ventilators for influenza pandemics. Emerg Infect Dis. 2017;23(6):914.
16. Liu J, Sen S. Asymptotic results of stochastic decomposition for two-stage stochastic quadratic programming. SIAM J Optim. 2020;30(1):823–52.
17. Mak WK, Morton D, Wood R. Monte carlo bounding techniques for determining solution quality in stochastic programs. Oper Res Lett. 1999;24:47–56.
18. Piscitello GM, Kapania EM, Miller WD, Rojas JC, Siegler M, Parker WF. Variation in ventilator allocation guidelines by us state during the coronavirus disease 2019 pandemic: a systematic review. JAMA Network Open. 2020;3(6):e2012606–e2012606.
19. Ray EL, Wattanachit N, Niemi J, Kanji AH, House K, Cramer EY, Bracher J, Zheng A, Yamana TK, Xiong X, et al. Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US. MedRXiv 2020. https://doi.org/10.1101/2020.08.19.20177493.
20. Sameni R. Mathematical modeling of epidemic diseases; a case study of the COVID-19 coronavirus; 2020. arXiv preprint arXiv:2003.11371.
21. Santosh K. COVID-19 prediction models and unexploited data. J Med Syst. 2020;44(9):1–4.
22. Sen S, Higle JL. An introductory tutorial on stochastic linear programming models. Interfaces. 1999;29(2):33–61.
23. Sen S, Liu Y. Mitigating uncertainty via compromise decisions in two-stage stochastic linear programming: variance reduction. Oper Res. 2016;64(6):1422–37.
24. Shinde GR, Kalamkar AB, Mahalle PN, Dey N, Chaki J, Hassanien AE. Forecasting models for coronavirus disease (COVID-19): a survey of the state-of-the-art. SN Comput Sci. 2020;1(4):1–15.

[5] https://core.isrd.isi.edu/chaise/record/#1/Core:Instance/RID=WCT4.

25. Team ICF. Modeling COVID-19 scenarios for the united states. Nat Med. 2020. https://doi.org/10.1038/s41591-020-1132-9.
26. Van Slyke RM, Wets R. L-shaped linear programs with applications to optimal control and stochastic programming. SIAM J Appl Math. 1969;17(4):638–63.
27. Wets RJB. Solving stochastic programs with simple recourse. Stochastics. 1983;10(3–4):219–42.