Increasing Healthcare Systems Capacity Resilience: A COVID-19 Study

Felix Parker*, Hamilton Sawczuk, Darius Irani
Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, fparker9@jhu.edu, hsawczu1@jhu.edum dirani2@jhu.edu

Fardin Ganjkhanloo, Farzin Ahmadi, Kimia Ghobadi
Department of Civil and Systems Engineering, The Centre for Systems Science and Engineering, The Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD 21218, fganjkh1@jhu.edu, fahmadi1@jhu.edu, kimia@jhu.edu

At the peak of the pandemic in mid-April, the demand on healthcare resources surpassed capacity in many areas. As the states start to reopen, efficient allocation of resources (e.g., beds, nurses, PPE, ventilators) will be paramount in minimizing the potential impacts since different states will be affected differently in severity and timeline. We present robust mixed-integer linear programming models to minimize resource shortage while encouraging desirable allocation properties such as transfer sparsity, consistency, and locality. We consider models for patient and nurse redistribution based on current capacity and estimated future demand. Our models are validated using COVID-19 surge data in New Jersey and Texas areas.

Key words: COVID-19 Pandemic; Resource Allocation; Hospital Operations; Patient Transfers; Linear Programming, Mixed-Integer Optimization, Robust Optimization

1. Introduction

Shortly after it was first identified in Wuhan, China, COVID-19 has become a global pandemic (Hui et al. 2020). To date, the United States has the most confirmed COVID-19 cases among all countries at more than 13 million, and the world has seen more than 576 thousand deaths to date (Dong et al. 2020). These staggering numbers have created a large imbalance between the demand and the capacity of global healthcare system, notably in regular and Intensive Care Unit (ICU) beds (IHME et al. 2020). The first wave of US COVID-19 pandemic, centered in the northeast region, triggered many local and national interventions including widespread closures and stay-at-home orders. These interventions were mainly designed to relieve some of the pressure from healthcare infrastructure by "flattening the curve" and distributing the demand on the healthcare systems, and while effective, they have come at a high economic cost.

At the same time, the healthcare system has also responded by both reducing its demand through canceling non-urgent procedures and increasing its capacity through creating new COVID-19-suitable beds. Both of these tactics have been effective, however, the former delays medical care for patients and leads to increase demand when hospitals reopen again, both of which may have

[*] Current affiliation: Department of Civil and Systems Engineering, Johns Hopkins University, Baltimore, MD 21218.

negative effects on patients. The latter imposes a large financial cost, detracts from non-COVID-19 capacity, and requires advanced planning to implement. Additionally, these interventions are almost always undertaken on a individual or local level. There have been tremendous efforts in hospitals to optimize the use of resources at the individual level, either through increasing production capacities or extending use of resources like PPE (CDC 2020), however, these methods are often insufficient or unsustainable. While there has been some consortium, it is rare that healthcare responses are implemented at a system level rather than at an individual level. If a cooperation has been made, it has been often in a reactive manner to respond to the crises at hand, such as reported in the local news about the response in Houston to the current wave of COVID-19 in the area (Pereira 2020).

Considering the healthcare system as a whole has several benefits. First, it is known that pooling capacity together often leads to better use of the resources (Vanberkel et al. 2012, Sims et al. 2015, Chod and Rudi 2005). Second, predicting demand and capacity for a collection of hospitals is often more accurate than each individual hospital, which also leads to better use of the current resources and better strategic decision-making for the capacity of each hospital. Finally, the nature of the COVID-19 pandemic and its bias in the severity of symptoms for patients creates a natural opportunity for load-balancing across the system (Onder et al. 2020).

Forecasts show that the pandemic burdens hospitals far beyond their current capacity, especially in Intensive Care Units (IHME et al. 2020). Critically however, different areas experience their peak hospitalizations at different times; while some hospitals are operating near their capacity or running on a shortage of resources, other hospitals may have excess capacity or resources. This calls for a model that can optimally allocate surplus resources between hospitals to minimize overall capacity and resource shortage.

In this paper, we explore how to leverage the capacity of the healthcare systems to better load-balance when the system is under extreme pressure. We focus on the COVID-19 pandemic as the stressor to the system and consider both the first wave in the northeast region (as a retrospective example) and the second wave in the southern region (as a prospective example). While various resources can be considered for load-balancing and optimal allocation, e.g., personal protective equipment, ventilators, beds, and nurses, we concentrate on hospital and ICU beds as primary resources and nurses as complementary resources to the occupied beds.

An important factor contributing to higher demands for hospital resources is inadequate distribution of patients. Patients tend to opt for hospitals and healthcare centers that have a better reputation and more prestige over healthcare centers that can provide quality care for them (Varkevisser et al. 2012, Drevs 2013). During a pandemic, this tendency leads to unbalanced patient loads among hospitals. Therefore, it is crucial to provide guidance to patients allowing them to visit healthcare centers with minimal patient overflow.

We propose a new system that is capable of identifying the excess or shortage of resources among hospitals and optimally redistributing newly admitted patients across a region. Such a problem has been considered previously in other studies for conventional hospital resources but the large amount of data gathered daily through the course of the COVID-19 pandemic calls for new models that can provide insights to the problem of re-distributing patients and resources. Using data on hospital capacities and resources, as well as various forecasts on the number of COVID cases in each region, our model can determine an optimal set of transfers within specific guidelines so that the total number of patients who are properly cared for is maximized. This reallocation decreases strain on healthcare workers and increases the quality of patient care.

To capture the distribution of patient lengths of stay and potential patient transfers between regular and ICU beds, a group model is utilized in which patients are admitted to certain groups, spend time in that group according to some distribution and are then transferred to a new group or discharged. Each patient group utilizes a particular bed type, for example ICU or non-ICU beds. In addition, our formulation allows for nurses to be simultaneously allocated across hospitals and certain secondary solution characteristics to be encouraged or required. The level of value the model places on these secondary characteristics is controlled by a set of hyper-parameters that can be adjusted by the practitioner.

Ultimately, the LP and MILP optimization models (discussed in Sections 3.1–3.3) consider hospitals in the system and aim to find the best bed capacity allocation, or equivalently, the best patients allocation, among hospitals in the same level. Patient allocation is often referred to as patient transfer between hospitals—a fairly common practice that typically moves patients from Primary and Secondary medical centers to Tertiary and Quaternary centers. In our study, we consider patient transfer for COVID-19 inpatients (who are admitted to the hospital) at the point of their entry e.g., transferring COVID-19 patients in Emergency Departments who are identified to be admitted to the hospitals for overnight stay.

The rest of the paper is organized as follows. Related existing literature are discussed in Section 2. Detailed formulations for our proposed system can be found in Section 3. A detailed explanation on the different types and sources of data is provided in Section 4. Results and discussions are explained in Section 5. We conclude in Section 6.

2. Related Work

In this section, we point out at some of the existing literature on the allocation of resources among different entities with the goal of maximizing available supply. We consider previous work on the progression and forecasting of the COVID-19 pandemic and previous efforts to optimize the allocation of resources among hospitals during high demand times like a pandemic. We also highlight some previous studies on methods to relocate patients when some hospitals are operating at capacity levels.

Many works have considered the progression and forecasting of pandemics (Cooper et al. 2006, Tizzoni et al. 2012, Pei et al. 2018) and the related literature on this subject has been growing rapidly since the emergence of the COVID-19 pandemic (Rees et al. 2020, IHME et al. 2020, Halpern and Miller 2020a, Dong et al. 2020). The majority of these works tend to provide an aggregate level of data on how a pandemic spreads while other works provide methods and insights to forecast and hinder the spread of the pandemic. Much of the work on the COVID-19 pandemic has focused on data analytics and forecasting. However, research on the demand for hospital resources during pandemics has also gained attention recently. A number of recent studies have focused on forecasting the spread of the pandemic and providing insights on planning for optimizing medical services during the pandemic (IHME et al. 2020) (Lampariello and Sagratella 2020) (Lewnard et al. 2020). Forecasts show that in the event of a second wave, the hospital demand will readily exceed capacity.

Some studies have considered methods to increase time efficiency of different stages of patient admittance and decrease patient wait times (Judson et al. 2020) (Luscombe and Kozan 2016) while other studies have considered the problem of optimal nurse scheduling in the emergency department (Otegbeye et al. 2015). Judson et al. (2020) provide a self service COVID-19 self-triage and self-scheduling tool designed to increase the time efficiency of the emergency department. Luscombe and Kozan (2016) discuss a dynamic resource allocation system at hospital level in order to improve emergency department efficiency. However, much of the existing literature in scheduling and optimizing patient care times do not consider the high demand settings where some hospitals rapidly hit their capacity or only consider a single hospital.

It has been emphasized in other works that during pandemics, due to unusual circumstances, hospitals cannot accomplish pre-pandemic tasks individually and interplay between hospitals, at least at regional level, becomes crucial (Toner and Waldhorn 2006). This interactive setting is highly constrained and usually includes many objectives. Sun (2011) provided single and multi-objective optimization models for resource and patient allocation between hospitals. In this work, we provide a robust model for patient allocation which is capable of providing alternatives to avoid over-capacity admit rates in locations that experience the peak of the pandemic. While some studies argue that prevention measures are more effective in response to the pandemic rather than increasing the capacities for ICUs and ventilators rapidly (Halpern and Miller 2020b), there is an inevitable need to provide models for allocations of existing and limited resources.

The problem of optimal resource allocation has been extensively studied over the past years and many studies have developed optimization methods using different approaches for this problem in diverse settings (Hegazy 1999, Ren et al. 2018, Kuchuk et al. 2015, Cruz et al. 2004, Tychogiorgos and Leung 2014). Medical resource allocation has also been the focus of some studies in recent years (Li and Xu 2020) (Luscombe and Kozan 2016) and with the emergence of the COVID-19 pandemic, there has been a growing literature regarding the general problem of resource allocations during high demand periods like the COVID-19 pandemic (Emanuel et al. 2020) (White and Lo 2020). While much of the previous work on resource allocation during pandemics is related to hospital level tasks (CDC 2020) (Emanuel et al. 2020) (White and Lo 2020), some studies consider the problem of allocating resources at a larger scale (Mehrotra et al. 2020b) (Arora et al. 2010). Weissman et al. (2020) use Monte-Carlo simulations to estimate the timing of surges in clinical demand and provide different local scenarios for resource allocation decision makers. Mehrotra et al. (2020a) also discuss the idea of converting to virtual practice during the COVID-19 pandemic to reduce the number of hospital visits from suspected COVID-19 cases and the risk of exposure for vulnerable cases.

The existing literature on patient allocation practices is mainly focused on long-term planning or short term disaster management (Sun et al. 2014). Only a handful of previous studies consider the problem of patient allocation during high demand periods like pandemics (Sun et al. 2014) (Lacasa et al. 2020). The closest work in the literature to this paper is the work by Sun et al. (2014). In their work, they provide a multi-objective optimization model for patient allocation during a potential pandemic influenza outbreak. They consider minimizing the total travel distance and maximum distance a patient travels to a hospital. Lacasa et al. (2020) also consider the problem of sharing healthcare resources and relocating COVID-19 patients in need of ventilators or ICUs. They consider geometric feasibility constraints in their algorithm and compare results with ground truth in UK and Spain. In another study, Bai and Zhang (2014) provide an incentive-based approach that lets the patients choose hospitals on their own in order to balance the demand on hospitals.

It is worth noting that all of the above studies consider settings with specific objectives (such as reactive management of patients in dire need of ventilators or incentive-based approaches to direct the public to balance the potential patient load . However, in the uncertain setting of the COVID-19 pandemic, a collective system that is capable of constantly forecasting shortage of resources and providing decision makers with options to manage the high demand by allocation of resources has not been studied. The approach taken in this study provides a model that can prospectively allocate resources to the entities that are expected to be in more limited supplies.

3. Methodology

In this section, we discuss the details of a series of models and frameworks that we developed to solve the load-balance problem in the presence of an extreme stressor. While the primary goal of our proposed methodology is to achieve an optimal load-balance within a system of treatment nodes, a secondary goal is to build a robust and practical capacity-sharing scheme to leverage pooled system resources, increasing the resilience of the system to future stressors. To achieve these goals, we developed a set of LP and MILP load-balancing models and a disaggregation methodology to obtain appropriate data granularity for individual applications.

We assume that the system of treatment nodes is under extreme stress and that individual nodes may be affected by the stressor differently. We also assume that the length of stay (LOS) of each patient is governed by a known probability distribution (Lewnard et al. 2020) for the illness of study, e.g., the LOS distribution of COVID-19 patients. We further refine this model in Section 3.1.2 to allow for different patient groups with distinct length of stay distributions, care requirements, and bed types, e.g., ICU and non-ICU. To allocate patients we consider node capacities and provide optional penalties and constraints to address operational needs such as minimum and maximum transfer quantities, not accepting patients if the receiving hospital is predicted to reach capacity soon, a minimum number of days between a hospital receiving and sending patients, and preferential allocation for less severe nodes that would not experience an overflow with no transfers. We also provide optional penalties and constraints to encourage more sparse transfer relationships between nodes or less transfer quantity. All optional model features are discussed in detail in Section 3.1.3.

We next propose a complementary model where we consider not only primary load-balancing (patients in our study) but also the load-balancing of complementary resources. The nature of the required complementary resources depends on the application domain, and in this study, we focus on nurses (Section 3.2). Other complementary resources including ventilators, personal protective equipment, or other clinical staff could be considered as well, both individually or in groups. We finally consider the range of uncertainties that exist in this problem and robustify the model against them in Section 3.3. There are uncertainties in both the demand and capacity of the system and individual nodes. Therefore, we propose a robust optimization model for the daily number of patients, the available daily number of beds of different types, and supply of nurses. We note that the uncertainty in the LOS of patients is already addressed in the model by the LOS distribution.

3.1. Patient Allocation Models

In this section, we formulate the allocation of newly admitted COVID-19 patients across nodes in the system to achieve a balanced load. The nodes may be healthcare facilities, counties or states depending on the resolution of the problem which we are modelling. We utilize linear models to first derive our main formulation for patient allocation in Section 3.1.1. We then refine this model by adding a feature to distinguish between various groups of patients who may have a different care path (e.g., patients who go to ICU at some point of their care and patients who do not) in Section 3.1.2. We then provide several additional constraints that are optional to the model and can be used to tailor the model to the specific healthcare system of study. The details of our main assumptions, parameters, and variables for all our models are summarized in Table 1 at the end of this section.

3.1.1. Main Patient Allocation Formulation

In this section, we formulate our main model to allocate patients between different healthcare nodes in a given system. We minimize the total patient overflow in all locations over all considered days subject to capacity and operational constraints as follows.

$$\begin{aligned}
\min \quad & \sum_{i\in\mathcal{N}}\sum_{t\in\mathcal{T}} o_{i,t} && \\
\text{subject to} \quad & \sum_{j\in\mathcal{N}} s_{i,j,t} \le p_{i,t} && \forall i\in\mathcal{N}, t\in\mathcal{T} \quad (1)\\
& \alpha_{i,t} + \sum_{j\in\mathcal{N}} s_{i,j,t} - b_i \le o_{i,t} && \forall i\in\mathcal{N}, t\in\mathcal{T} \quad (2)\\
& s_{i,j,t} = 0 && \forall (i,j)\in\overline{E(G)}, t\in\mathcal{T} \quad (3)\\
& s_{i,j,t} \ge 0 && \forall i,j\in\mathcal{N}, t\in\mathcal{T} \quad (4)\\
& o_{i,t} \ge 0 && \forall i\in\mathcal{N}, t\in\mathcal{T} \quad (5)
\end{aligned}$$

$$\text{where}\ \alpha_{i,t} = (p_{i,0} - \sum_{t'=1}^{t} d_{i,t'}) + \sum_{t'=1}^{t}\{[1-\mathcal{L}(t-t')][p_{i,t'} + \sum_{j=1}^{N}(s_{j,i,t'} - s_{i,j,t'})]\} \qquad \forall i\in\mathcal{N}, t\in\mathcal{T} \quad (a)$$

In this formulation, we assume that only newly admitted patients may be transferred between nodes, and this is enforced by Constraint (1), ensuring that the number of patients transferred away does not exceed the number of newly admitted patients $p_{i,t}$ at node i at time t. The overflow for node i at time t is defined in Constraint (2). Since patient transfer between care centers typically requires resources from both centers on the day of the transfer, we count the transferred patients as active at both nodes i and j at time t for the purposes of setting overflow.

Constraint (3) governs which nodes in the system are allowed to send patients to which other nodes (e.g., based on distance or existing relationships between hospitals). These relationships can be captured in a graph G where all healthcare centers are considered as vertices. An edge is added from node i to node j if resources may be transferred accordingly. Constraint (3) disallows transfers from node i to node j if node i is not incident node j in the graph G. More details about the parameters and variables of this formulation are provided in Table 1.

The Expression (a) represents the number of active patients at node i at time t. The first term of this expression captures the number of remaining initial patients at node i at time t. The second term incorporates the cumulative patient length of stay distribution $\mathcal{L}$ to capture the number of admitted and reallocated patients remaining at node i at time t.

One of the primary limitations of this model is that it considers the same care path for all patients and hence, assumes all patients and beds to be interchangeable. In reality however, patients have different care needs and will be assigned to different types of beds accordingly. Additionally, most patients that spend time in an ICU bed also spend time in a regular hospital bed. In the following sections, we introduce more comprehensive models which can tackle these aspects of the problem.

3.1.2. Group Patient LP Formulation

In this section we build on the base LP model discussed in Section 3.1.1 to develop a model capable of handling more features of the problem domain. This extended formulation first introduces a set of patient groups $\mathcal{G}$ and a set of bed types $\mathcal{B}$. It is important to note that we require a function f mapping $\mathcal{G}$ to $\mathcal{B}$, which implies that each patient group is associated with exactly one bed type. We assume that each patient is admitted to a specific patient group. Each patient then undergoes treatment during which they may be transferred between groups and potentially bed types before being discharged. We allow patient transfers between groups along edges in the directed graph G_{group}. In particular, we require that G_{group} be the disjoint union of directed trees whose edges are directed toward their roots, also known as an in-forest. This is to disallow cycles from our patient group transfer scheme and simultaneously require that each patient group has a unique transfer path until they are discharged, i.e., each patient group either discharges patients or transfers to exactly one other patient group . Notice that this does not decrease the generality of the model since any directed forest can be made into an in-forest provided the practitioner is allowed to subdivide vertices with out-degree greater than one.

$$
\begin{array}{lll}
\min & \displaystyle\sum_{\beta\in\mathcal{B}}\sum_{i\in\mathcal{N}}\sum_{t\in\mathcal{T}} o_{\beta,i,t} & \\
\text{subject to} & \displaystyle\sum_{j\in\mathcal{N}} s_{g,i,j,t} \leq p_{g,i,t} & \forall g\in\mathcal{G}, i\in\mathcal{N}, t\in\mathcal{T} \quad (6)\\
& \displaystyle\sum_{g\in img\{\,f^{-1}(\beta)\,\}} (\alpha_{g,i,t} + \sum_{j\in\mathcal{N}} s_{g,i,j,t}) - b_{\beta,i} \leq o_{\beta,i,t} & \forall \beta\in\mathcal{B}, i\in\mathcal{N}, t\in\mathcal{T} \quad (7)\\
& s_{g,i,j,t} = 0 & \forall g\in\mathcal{G}, (i,j)\in\overline{E(G)}, t\in\mathcal{T} \quad (8)\\
& s_{g,i,j,t}, o_{\beta,i,t} \geq 0 & \forall g\in\mathcal{G}, \beta\in\mathcal{B}, i,j\in\mathcal{N}, t\in\mathcal{T} \quad (9)
\end{array}
$$

where

$$\alpha_{g,i,t} = p_{g,i,0} + \sum_{t'=1}^{t} (\chi_{g,i,t'} - \gamma_{g,i,t'}) \qquad \forall g\in\mathcal{G}, i\in\mathcal{N}, t\in\mathcal{T} \tag{b}$$

$$\chi_{g,i,t} = p_{g,i,t} + \sum_{g':g'\sim g} \gamma_{g',i,t} + \sum_{j\in\mathcal{N}} (s_{g,j,i,t} - s_{g,i,j,t}) \qquad \forall g\in\mathcal{G}, i\in\mathcal{N}, t\in\mathcal{T} \tag{c}$$

$$\gamma_{g,i,t} = d_{g,i,t} + \sum_{t'=1}^{t} \ell_g(t-t')\chi_{g,i,t'} \qquad \forall g\in\mathcal{G}, i\in\mathcal{N}, t\in\mathcal{T} \tag{d}$$

Constraints (6) through (9) are generalizations of the constraints (1) through (4) from Section (3.1.1). Expression (b) represents the number of active patients in group g at node i at time t. Expression (c) represents the number of patients entering group g in node i at time t. Finally, expression (d) represents the number of patients exiting group g in node i at time t. The first term of Expression (b) captures initial patients while the second term accounts for the sum of net active patient changes in group g at node i. The first term of Expression (c) captures admitted patients, the second term sums patients leaving other groups that transfer to the current group, and the third term includes net patient transfers, all in group g at node i at time t. The first term of Expression (d) captures initial patients discharged while the second term calculates the number of patients exiting group g at node i at time t.

Notice that although expressions (b) through (d) immediately constitute a closed form expression for the number of active patients, they provide a method for computing one recursively which is guaranteed to terminate by our requirement that G_{group} be a in-forest.

So far our group allocation LP model is capable of minimizing patient overflow across multiple patient groups and bed types, however, there are certain other solution characteristics that the practitioner may wish to encourage. In the following section we provide an optional MILP framework for meeting operational domain needs.

3.1.3. Group Optional Penalties and Constraints

Building on the above group patient LP formulation, we can add additional penalties and constraints to tailor the formulation to desired solution characteristics. In particular, we have added terms to penalize undesirable solutions of the model presented in 3.1.2. Details on the additional variables and parameters are provided in Table 1.

$$\text{penalty} = C_{\text{sent}} \sum_{g\in\mathcal{G}} \sum_{i,j\in\mathcal{N}} \sum_{t\in\mathcal{T}} s_{g,i,j,t} + C_{\text{smooth}} \sum_{i,j\in\mathcal{N}} \sum_{t\in\mathcal{T}} \delta_{i,j,t} + C_{\text{balance}} \sum_{\beta\in\mathcal{B}} \sum_{i\in\mathcal{N}} \sum_{t\in\mathcal{T}} \phi_{\beta,i,t} + C_{\text{setup}} \sum_{i,j\in\mathcal{N}} \rho_{i,j}$$

$$\sum_{g\in\mathcal{G}} (s_{g,i,j,t-1} - s_{g,i,j,t}) \leq \delta_{i,j,t} \qquad \forall i,j\in\mathcal{N}, t\in\mathcal{T}\setminus\{1\} \tag{10}$$

$$-\sum_{g\in\mathcal{G}} (s_{g,i,j,t-1} - s_{g,i,j,t}) \leq \delta_{i,j,t} \qquad \forall i,j\in\mathcal{N}, t\in\mathcal{T}\setminus\{1\} \tag{11}$$

$$\frac{1}{b_{\beta,i}} \sum_{g\in img\{f^{-1}(\beta)\}} \alpha_{g,i,t} - R_{\text{thresh}} \leq \phi_{\beta,i,t} \qquad \forall \beta\in\mathcal{B}, i\in\mathcal{N}, t\in\mathcal{T} \tag{12}$$

$$M \sum_{g\in\mathcal{G}} \sum_{t\in\mathcal{T}} (s_{g,i,j,t} + s_{g,j,i,t}) \geq \rho_{i,j} \qquad \forall i,j\in\mathcal{N}, j>i \tag{13}$$

$$m \sum_{g\in\mathcal{G}} \sum_{t\in\mathcal{T}} (s_{g,i,j,t} + s_{g,j,i,t}) \leq \rho_{i,j} \qquad \forall i,j\in\mathcal{N}, j>i \tag{14}$$

$$M \sum_{g\in\mathcal{G}} \sum_{j\in\mathcal{N}} s_{g,i,j,t} \geq \nu_{1,i,t} \qquad \forall i\in\mathcal{N}, t\in\mathcal{T} \tag{15}$$

$$m \sum_{g\in\mathcal{G}} \sum_{j\in\mathcal{N}} s_{g,i,j,t} \leq \nu_{1,i,t} \qquad \forall i\in\mathcal{N}, t\in\mathcal{T} \tag{16}$$

$$M \sum_{g\in\mathcal{G}} \sum_{j\in\mathcal{N}} s_{g,j,i,t} \geq \nu_{2,i,t} \qquad \forall i\in\mathcal{N}, t\in\mathcal{T} \tag{17}$$

$$m \sum_{g\in\mathcal{G}} \sum_{j\in\mathcal{N}} s_{g,j,i,t} \leq \nu_{2,i,t} \qquad \forall i\in\mathcal{N}, t\in\mathcal{T} \tag{18}$$

$$\nu_{1,i,t} + m \sum_{g\in\mathcal{G}} \sum_{j\in\mathcal{N}} \sum_{t'=t}^{\min\{t+T_{\text{switch}},T\}} s_{g,j,i,t'} \leq 1 \qquad \forall i\in\mathcal{N}, t\in\mathcal{T} \tag{19}$$

$$\nu_{2,i,t} + m \sum_{g\in\mathcal{G}} \sum_{j\in\mathcal{N}} \sum_{t'=t}^{\min\{t+T_{\text{switch}},T\}} s_{g,i,j,t'} \le 1 \qquad \forall i\in\mathcal{N}, t\in\mathcal{T} \tag{20}$$

$$\sum_{g\in img\{f^{-1}(\beta)\}} \alpha_{g,i,t} \le b_{\beta,i} \qquad \forall i\in\mathcal{N}, t\in\mathcal{T}, \alpha'_{\beta,i,t} \le b_{\beta,i} \tag{21}$$

$$\sum_{g\in\mathcal{G}} s_{g,i,j,t} \in \{0\} \cup [S_{\min}, \infty), \delta_{i,j,t} \ge 0 \qquad \forall i,j\in\mathcal{N}, t\in\mathcal{T} \tag{22}$$

$$\phi_{\beta,i,t} \ge 0 \qquad \forall \beta\in\mathcal{B}, i\in\mathcal{N}, t\in\mathcal{T} \tag{23}$$

$$\rho_{i,j}, \nu_{1,i,t}, \nu_{2,i,t} \in \{0,1\} \qquad \forall i,j\in\mathcal{N}, t\in\mathcal{T} \tag{24}$$

where

$$\alpha'_{\chi,i,t} = \sum_{g\in img\{f^{-1}(\beta)\}} [p_{g,i,0} + \sum_{t'=1}^{t} (\chi'_{g,i,t'} - \gamma'_{g,i,t'})] \qquad \forall g\in\mathcal{G}, i\in\mathcal{N}, t\in\mathcal{T} \tag{e}$$

$$\chi'_{g,i,t} = p_{g,i,t} + \sum_{g':g'\sim g} \gamma'_{g',i,t} \qquad \forall g\in\mathcal{G}, i\in\mathcal{N}, t\in\mathcal{T} \tag{f}$$

$$\gamma'_{g,i,t} = d_{g,i,t} + \sum_{t'=1}^{t} \ell_g(t-t')\chi'_{g,i,t'} \qquad \forall g\in\mathcal{G}, i\in\mathcal{N}, t\in\mathcal{T} \tag{g}$$

In the above formulation, the penalty represents a sum of terms that can be added to the group LP formulation from section (3.1.2) along with their corresponding constraints to encourage certain secondary characteristics for the solutions. The first cost coefficient included in the penalty term, C_{sent}, penalizes the objective function for the total number of patients sent by the solution. C_{smooth} penalizes for the absolute value of the difference between patient transfer quantities from node i to node j at time $t-1$ and time t. The corresponding variable $\delta_{i,j,t}$ is determined by constraints (10) and (11). These constraints are motivated by the idea that a more consistent transfer rate from node i to node j may be more operationally feasible. C_{balance} encourages more balanced patient loads by penalizing the objective function when patient load at a node exceeds the balancing threshold ratio, R_{thresh}. Constraint (12) defines the corresponding dummy variable $\phi_{g,i,t}$. Finally, the last term in the penalty function including C_{setup} penalizes the objective function for the first time a transfer takes place between nodes i and j accounting for overhead cost of establishing a transfer relationship. Constraints (13) and (14) define the $\rho_{i,j}$ binary variable accordingly.

Constraints (15) through (20) enforce a minimum gap of T_{switch} days between when a node may send and receive patients. This is done by introducing the binary variables $\nu_{1,i,t}$ and $\nu_{2,i,t}$ and using them to constrain the sent and received patients for the next T_{switch} days. These constraints contribute to smoothness as well, ensuring nodes will be unable to switch quickly between sending patients and receiving them (or do both at once). Then, Constraint (21) ensures that our allocation does not cause any node to experience a patient overflow at a time when they otherwise would not. Next, Constraint (22) contains a non-convex constraint that can be implemented using binary variables which guarantees that if a patient transfer occurs, at least $S_{\min}$ patients are sent. Finally, Constraints (23) and (24) constrain possible variable values to be either non-negative or binary.

Expressions (e) through (g) are analogous to Expressions (b) through (d), representing active patients, entering patients and exiting patients for group, except for these new expressions do not consider patient reallocation and $\alpha'_{\beta,i,t}$ is aggregated by bed type. Therefore Expression (e) gives the number of active patients in bed type β at node i at time t without any patient reallocation.

3.2. Combined Patient and Nurse Allocation Model

We now consider extending the group patient LP formulation from section 3.1.2 to allocate nurses along with patients. In order for a patient to receive proper treatment we require that they both have a bed and adequate nurse care. In this section we first introduce a nurse allocation LP that uses the results of the patient allocation from Section 3.1. This formulation can also be combined with the formulation from Section 3.1 in order to simultaneously allocate both patients and nurses, as we do in 5.1.2. We then include a set of nurse specific optional constraints that address the issue of artificial shortage of static supply for nurses (compared to patients' dynamic demand).

3.2.1. Nurse Allocation Formulation

Much of the modeling infrastructure for patients is reused for the allocation of nurses in the combined model, with the key difference being that nurses are a supply rather than a demand.

$$\min \quad \sum_{i\in\mathcal{N}}\sum_{t\in\mathcal{T}}\theta_{i,t}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{N}}\sigma_{i,j,t} \le \eta_{i,t} \qquad \forall i\in\mathcal{N}, t\in\mathcal{T} \tag{25}$$

$$q_{i,t}-\eta_{i,t} \le \theta_{i,t} \qquad \forall i\in\mathcal{N}, t\in\mathcal{T} \tag{26}$$

$$\sigma_{i,j,t} = 0 \qquad \forall (i,j)\in\overline{E(G)}, t\in\mathcal{T} \tag{27}$$

$$\sigma_{i,j,t} \ge 0 \qquad \forall i,j\in\mathcal{N}, t\in\mathcal{T} \tag{28}$$

$$\theta_{i,j} \ge 0 \qquad \forall i,j\in\mathcal{N} \tag{29}$$

where

$$\eta_{i,t} = n_i + \sum_{j=1}^{N}\sum_{t'=ts(t)}^{t}\left(\sigma_{j,i,t'}-\sigma_{i,j,t'}\right) \qquad \forall i\in\mathcal{N}, t\in\mathcal{T} \tag{h}$$

$$q_{i,t} = \sum_{\beta\in\mathcal{B}}\sum_{g\in img\{\,f^{-1}(\beta)\,\}} Q_\beta\alpha_{g,i,t} \qquad \forall i\in\mathcal{N}, t\in\mathcal{T} \tag{i}$$

Equations (25) through (28) are analogous to equations (1) through (4) in section (3.1). Expression (h) represents the number of active nurses and expression (i) represents the nurse demand which is taken to be a sum weighted by Q_β, which indicates the number of nurses needed per bed of type β, over the number of active patients for all groups at node i at time t. In the following section we provide additional constraints which can be added to the nurse allocation model to make it more capable in capturing the characteristics of the problem.

3.2.2. Nurse Specific Optional Constraints

In this section we provide a set of constraints designed to account for the fact that if a node is experiencing a shortage of nurses, either after or during patient allocation, then that node should not have any nurses deployed to other nodes.

$$n_i \le \eta_{i,t} \qquad \forall i \in \mathcal{N}, t \in T, q_{i,t} \ge n_{i,t} \tag{30}$$

$$m(q_{i,t} - \eta_{i,t}) \le \kappa_{i,t} \qquad \forall i \in \mathcal{N}, t \in \mathcal{T} \tag{31}$$

$$1 + m(q_{i,t} - \eta_{i,t}) \ge \kappa_{i,t} \qquad \forall i \in \mathcal{N}, t \in \mathcal{T} \tag{32}$$

$$\eta_{i,t} \ge n_i \qquad \forall i \in N, t \in T, \kappa_{i,t} = 1 \tag{33}$$

$$\kappa_{i,t} \in \{0, 1\} \qquad \forall i \in \mathcal{N}, t \in \mathcal{T} \tag{34}$$

Equations (31), (33) and (34) which enforce that for any day that a region has a shortage of nurses, it has at least its initial supply of nurses.

3.3. Robust Optimization Model

All of the previous models assumed no uncertainty in problem parameters. In this section, we address two of the primary sources of uncertainty in our models: the number of beds b_i at node i and the number of new patients $p_{i,t}$ at node i at time t.

$$p_{i,t} \in [\hat{p}_{i,t} - \tilde{p}_{i,t}, \hat{p}_{i,t} + \tilde{p}_{i,t}] \qquad \forall i \in \mathcal{N}, t \in \mathcal{T}$$

$$b_i \in [\hat{b}_i - \tilde{b}_i, \hat{b}_i + \tilde{b}_i] \qquad \forall i \in \mathcal{N}$$

In what follows, we build a model based on the methodology presented by Bertsimas and Sim (2004). We utilize an uncertainty budget, $\Gamma = \{\Gamma_1, \Gamma_2\}$ to govern the deviation of our uncertain parameters. Notice that this method is deterministic. In particular, it finds the realizations of our uncertain data that have the worst effect on the optimal objective function value within our uncertainty sets and uncertainty budget. In this paradigm, Γ can be interpreted as the total amount we expect our data realizations to vary from expectation within the uncertainty sets.

The following model is the base patient allocation model extended to our uncertainty set. The additional models presented in Section 3 can be similarly.

$$\min_{s} \quad \max_{\xi} \quad \sum_{i \in \mathcal{N}} \sum_{t \in \mathcal{T}} o_{i,t}$$

$$\text{subject to} \quad \sum_{j \in \mathcal{N}} s_{i,j,t} \le \bar{p}_{i,t} \qquad \forall i \in \mathcal{N}, t \in \mathcal{T} \tag{35}$$

$$\bar{\alpha}_{i,t} + \sum_{j \in \mathcal{N}} s_{i,j,t} - \bar{b}_i \le o_{i,t} \qquad \forall i \in \mathcal{N}, t \in \mathcal{T} \tag{36}$$

$$s_{i,j,t} = 0 \qquad \forall (i,j) \in \overline{E(G)}, t \in \mathcal{T} \tag{37}$$

$$s_{i,j,t} \ge 0 \qquad \forall i,j \in \mathcal{N}, t \in \mathcal{T} \tag{38}$$

$$o_{i,t} \geq 0 \qquad \forall i \in \mathcal{N}, t \in \mathcal{T} \qquad (39)$$

$$||\vec{\xi_i}||_\infty \leq 1, \quad ||\vec{\xi_i}||_1 \leq \Gamma_1 \qquad \forall i \in \mathcal{N} \qquad (40)$$

$$||\vec{\zeta_i}||_\infty \leq 1, \quad ||\vec{\zeta_i}||_1 \leq \Gamma_1 \qquad \forall i \in \mathcal{N} \qquad (41)$$

where

$$\bar{\alpha}_{i,t} = (p_{i,0} - \sum_{t'=1}^{t} d_{i,t'}) + \sum_{t'=1}^{t}\{[1 - \mathcal{L}(t-t')][\bar{p}_{i,t'} + \sum_{j=1}^{N}(s_{j,i,t'} - s_{i,j,t'})]\} \qquad \forall i \in \mathcal{N}, t \in \mathcal{T} \qquad \text{(j)}$$

$$\bar{p}_{i,t} = \hat{p}_{i,t} + \xi_{i,t}\tilde{p}_{i,t} \qquad \forall i \in \mathcal{N}, t \in \mathcal{T} \qquad \text{(k)}$$

$$\bar{b}_i = \hat{b}_i + \zeta_i \tilde{b}_i \qquad \forall i \in \mathcal{N} \qquad \text{(l)}$$

Notice first that Constraints (35) through (39) correspond directly to Constraints (1) through (5) from Section 3.1.1 with the appropriate substitutions for uncertain parameters, denoted by a vertical bar. Similarly, Expression (j) corresponds to (a) from 3.1.1. Then, Constraints (40) and (41) ensure that all realisations of our uncertain parameters are within our uncertainty set and limit the amount of variation in the uncertain parameters by our uncertainty budget Γ.

We solve this model using the JuMPeR library [1] for Robust Optimization which is able to automatically reformulate this model into a certain LP model.

3.4. Forecasting

All the models we have developed require estimates of the number of people who are infected with COVID-19 and require hospitalization. For retrospective studies this data may be available, but for prospective studies we need to forecast the number of hospitalizations for each location considered.

We use a Susceptible, Exposed, Infectious, Removed (SEIR) model developed by COVID-19 Projections [2] to estimate the number of COVID infections given a set of parameters to the model. This model has been cited by the Center for Disease Control (CDC) and performs well in practice.

This requires an estimate of the parameters for each location in the study. The parameters to the model include the initial basic reproduction number (R_0), the date on which people in a region began social distancing, and the initial number of infected individuals, among others. We obtain these estimates using black-box optimization of the simulator with respect to the input parameters. Specifically we use the Distance-weighted Exponential Natural Evolution Strategies Algorithm (DX-NES) proposed by Fukushima et al. (2011) with the score function defined as the ℓ_2-norm of the difference between the number of predicted COVID infections and the true number of confirmed COVID cases. It is important to note that DX-NES is a heuristic algorithm and therefore does not guarantee a global or even local optimum, but given the inherent uncertainty in predicting the future of the COVID pandemic we find this to be acceptable. This process yields a set of parameters which we use to make the final predictions.

[1] https://github.com/IainNZ/JuMPeR.jl

[2] `https://covid19-projections.com/`

Table 1: General information on the models

Patient Modeling Assumptions	• We consider only new patients as potential transfers between nodes. • We assume full knowledge about the number of new patients visiting a hospital each day. • We assume that the only resources limiting patient care are hospital beds, which are fixed in number at each node. • We assume that a fixed proportion of hospital beds are available to COVID-19 patients. • All patients have a length of stay governed by a distribution $\mathcal{L}_g$.
Data	• $p_{g,i,t}$: number of patients admitted to node $i \in \mathcal{N}$ at time $t \in \mathcal{T}$, specified in group $g \in \mathcal{G}$. (zero for gs where $indegree(g) \neq 0$) • $p_{g,i,0}$: number of initial patients in group $g \in \mathcal{G}$ at node $i \in \mathcal{N}$ at time $t = 0$ • $d_{g,i,t}$: number of initial patients who were discharged from group $g \in \mathcal{G}$ at node $i \in \mathcal{N}$ at time $t \in \mathcal{T}$ • $b_{\beta,i}$: number of beds of type $\beta \in \mathcal{B}$ available for COVID-19 patients at node $i \in \mathcal{N}$ • n_i: initial number of nurses at node $i \in \mathcal{N}$ • G: directed graph where $V(G) = \mathcal{N}$ and $(i,j) \in E(G)$ if and only if node i may transfer resources to node j • ℓ_g: distribution over length of stay for patients in group $g \in \mathcal{G}$ • $\mathcal{L}_g$: cumulative distribution over length of stay for patients in group $g \in \mathcal{G}$
Parameters	• R_{thresh}: load ratio at which balancing penalization begins • S_{min}: minimum number of patients that can be included in a transfer • T_{switch}: minimum number of days a node must wait between sending and receiving patients • C_{balance}: cost coefficient for load-balancing penalty $\phi_{i,t}$ • C_{smooth}: cost coefficient for the smoothing penalty $\delta_{i,j,t}$ • C_{sent}: cost coefficient for patient transfers $s_{i,j,t}$ • C_{setup}: cost coefficient for the transfer indicator $\rho_{i,j}$ • G_{group}: a graph where for all $\forall i, j \in \mathcal{G}, i \sim j$ if and only if patients from group $i \in \mathcal{G}$ are transferred to group $j \in \mathcal{G}$ • f: function mapping $\mathcal{G}$ to $\mathcal{B}$ • Q_β: ratio of nurse-days to patient-days for bed type $\beta \in \mathcal{B}$

Nurse Modeling Assumptions	• Nurses may be moved between hospitals. • We have full knowledge of the initial number of nurses in each region, which is constant except for our reallocation. • In addition to bed availability, nurse availability limits patient care. • Some fixed proportion of nurses are available to treat COVID-19 patients.
Variables	• $s_{g,i,j,t}$: number of patients of patient group $g \in \mathcal{G}$ sent from node $i \in \mathcal{N}$ to node $j \in \mathcal{N}$ at time $t \in \mathcal{T}$ • $o_{\beta,i,t}$: dummy variable for patient overflow in bed type $\beta \in \mathcal{B}$ at node $i \in \mathcal{N}$ at time $t \in \mathcal{T}$ • $\delta_{i,j,t}$: dummy variable for the absolute difference in the number of patients sent from node $i \in \mathcal{N}$ to node $j \in \mathcal{N}$ between days $t-1$ and $t \in \mathcal{T} \setminus \{1\}$. • $\phi_{\beta,i,t}$: dummy variable for the amount by which patient load ratio exceeds R_{load} at node $i \in \mathcal{N}$ at time $t \in \mathcal{T}$ • $\rho_{i,j}$: binary dummy variable which is equal to 1 if and only if there is a patient transfer between node $i \in \mathcal{N}$ and node $j \in \mathcal{N}$ • $\nu_{1,i,t}, \nu_{2,i,t}$: binary dummy variables used to enforce the minimum number of days a node must wait between sending and receiving patients • $\sigma_{i,j,t}$: variable for nurses sent from region $i \in \mathcal{N}$ to region $j \in \mathcal{N}$ at time $t \in \mathcal{T}$ • $\theta_{i,t}$: dummy variable for nurse overflow in region $i \in \mathcal{N}$ at time $t \in \mathcal{T}$
Sets	• $\mathcal{N}$: patient treatment nodes, indexed by $n \in \mathcal{N}$ • $\mathcal{T}$: modeling days, indexed by $t \in \mathcal{T} = \{1, 2, 3, ..., T\}$ • $\mathcal{G}$: patients groups, indexed by $g \in \mathcal{G}$ • $\mathcal{B}$: bed types, indexed by $\beta \in \mathcal{B}$

4. Data

Throughout this study, data collection proved to be both challenging and critical. Most data is collected at a local level with minimally standardized reporting methodology. Additionally, different sets of data are available at different resolutions which poses challenges related to aggregation and disaggregation. The data we collected primarily concerned active and admitted patient counts as well as bed and nursing staff availability.

4.1. Data Collection

To aid in this effort, we compiled a thorough collection of data from COVID-19 forecasts to nursing workforce surveys to health and hospital statistics. We provide this collection as an open resource for other researchers to use on our website. We include basic metadata about each data source. These include a short description (application), a longer description, an indication of whether the data is geographical, and the unit resolution. The data can be viewed at: `https://jhu-covid-optimization.github.io/covid-data/`.

4.2. Hospital Beds

The number of beds at each hospital that can be devoted to COVID patients is sourced from Definitive Healthcare USA Hospital Beds dataset located here: `https://coronavirus-resources.esri.com/datasets/1044bb19da8d4dbfb6a96eb1b4ebf629_0`. This dataset provides the locations of every hospital nationwide along with a count of how many specific hospital beds, such as ICU and staffed beds, that hospital contains.

The dataset also contains estimates of typical hospital utilization which we use to estimate the proportion of beds available for COVID patients. Based on this data we assume that 35% of ward beds and 50% of ICU beds at each hospital are made available for COVID-19 patients

4.3. Nurses

The number of registered nurses available to work in each state is estimated from the Area Health Resource Files from the Health Resources and Services Administration. Additionally, the National Sample Survey of Registered Nurses provides more fine-tuned delineations of practicing nurses. From the survey results, the categories of registered nurses can be selected for the model that more accurately represent nurses that could be transferred between facilities. For example, we choose to include Advanced Practice Registered Nurses and Nurse Practitioners, but exclude nurses from nursing home facilities.

We make the assumption that one nurse is needed for every two ICU beds (Lucchini et al. 2020) and every five ward beds. We also assume that 30% of nurse-hours are available to care for COVID patients, and that nurses work 36 hours per week on average. In our group model we require that patients stay for 2 days in a non-ICU bed before ICU admittance (Wunsch et al. 2011) and 5 days after ICU discharge. (Tiruvoipati et al. 2017)

4.4. Geospatial

In our models we ensure that patients are sent along edges of a graph, which enables us to set an upper bound on the amount of time a patient will have to travel. At the state level we query the Google Distance Matrix API [3] to find driving distances between state capitols. At more granular levels we use the latitude and longitude of the midpoint of each location to compute the distance matrix using the Haversine formula. We then threshold this distance matrix to compute the adjacency matrix for the patient transfer graph.

[3] https://developers.google.com/maps/documentation/distance-matrix/start

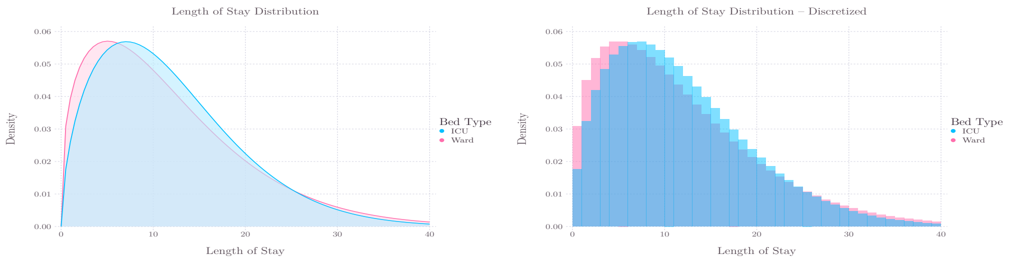

Figure 1 Distribution over the length of stay (LOS) for COVID-19 patients.

4.5. Parameters

The distribution over the length of stay (LOS) in ward and ICU beds for COVID patients is estimated by Lewnard et al. (2020). We use a Weibull($\lambda = 12.88, k = 1.38$) distribution for ward patients and a Weibull($\lambda = 13.32, k = 1.58$) for ICU patients. See Figure 1. We discretize these distributions for use in the model.

4.6. Local Data

Instances of county level data are used to estimate past hospitalizations and forecast future hospitalizations. In what follows, we introduce databases we used to provide such estimates.

4.6.1. New Jersey

We query data on hospitalizations from an API published by the New Jersey Department of Health located here: `https://services7.arcgis.com/Z0rixLlManVefxqY/arcgis/rest/services/PPE_Capacity/FeatureServer/0`. Available fields include number of persons under investigation (PUI) for COVID admitted to each hospital in different bed types and number of available beds. Since we look retrospectively at New Jersey in our analysis we take this data to be ground truth, but assume that any value could diverge from this amount by up to 20%.

4.6.2. Texas

In our analysis we look prospectively at Texas, so we require historical data that we can use to fit the parameters of our forecast model. Currently, the most granular data on COVID available for all of Texas is on the county level, reported here by the Texas Department of State Health Services: `https://dshs.texas.gov/coronavirus/additionaldata/`. This dataset includes confirmed COVID cases and deaths. Hospitalization data is also available from the same source at the Trauma Service Area (TSA) level, which we use to validate our hospitalizations forecast by aggregating from counties to TSAs.

4.6.3. Harris County, Texas

Additional data for Texas is reported by Harris County at the zip-code level here: . Harris County has been among the most impacted counties in the United States during the second wave of the pandemic, and some hospital systems in Harris County have reported transferring patients between hospitals already. We use reported new infections at the zip-code level to fit parameters for the forecast model and predict hospitalizations. We then aggregate from these forecasts to hospitals. This is accomplished by determining the five nearest hospitals to each zip code and distributing the forecasted patients among them, weighted by the number of beds that each hospital has.

5. Results

Our analysis can be divided into two retrospective and prospective results. We apply the models developed in section 3 to the COVID-19 pandemic. Specifically, we first look retrospectively at the first wave of the pandemic, using New Jersey as a case study to demonstrate the efficacy of our methods. We then consider potential flare-ups of the virus in certain states that have re-opened to the date of the study, providing an allocation scheme to reduce future hospital burden.

All models described in Section 3 were implemented in Julia 1.4 using the JuMP library for modeling (Dunning et al. 2017). Models were solved using Gurobi 9.0.1 with default options. All models took less than fifteen minutes to solve on a 2.2 GHz 6-core processor. The code is publically available at: `https://github.com/flixpar/covid-resource-allocation`.

5.1. Retrospective Results Analysis

In this section we look retrospectively at data from the first wave of the COVID-19 pandemic to validate each of our models. First we consider New Jersey as a case study for patient allocation using the basic, patient group, and robust formulations from Section 3.1. We then look at combined patient and nurse allocation between states in the northeastern United States and highlight the nurse allocation results.

5.1.1. Basic Patient Allocation

To validate our methods we apply the various models developed in Section 3 to New Jersey hospitals with a planning horizon from April 6, 2020 to June 1, 2020, covering the peak of the first wave of the COVID-19 pandemic. We chose to target New Jersey specifically because it was among the most impacted states during the first wave, and it reports the number of cases at each hospital which is both uncommon and useful.

The results of the analysis for COVID ICU load for each hospital in New Jersey with and without our proposed patient transfers can be seen in Figure 2, where patient load is computed as the number of COVID PUIs in ICU divided by the number of ICU beds that can be devoted to COVID patients. It can be seen in the figure that without transfers, a number of hospitals have a load greater than 1 at some point in May, which means that they had to create surge capacity in order to provide care for the number of patients they received. However, with our proposed transfers,

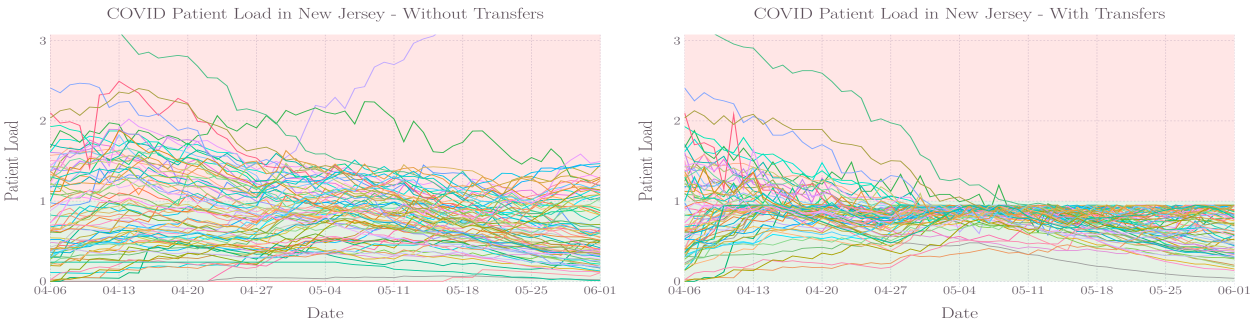

Figure 2 Patient allocation for New Jersey using the base patient allocation model from Section 3.1. Each line represents a hospital. Patient load less than 1 indicates that a hospital has additional capacity for COVID patients, while load greater than 1 indicates that a hospital is over capacity.

all hospitals are able to stay at or below the patient load ratio of 1 throughout May. We used the patient allocation model from 3.1 with one patient group and one bed group representing the ICU and $C_{\text{smooth}} = 10^{-2}, C_{\text{sent}} = 10^{-2}, C_{\text{weight}} = 9$ with all unused parameters set to zero. Final transfers involved sending 4764 patients between 75 participating hospitals, and resulted in a 65.2% reduction in total overflow, from 48602 patient-days to 16856 patient-days.

5.1.2. Combined Patients and Nurses Allocation Model

Nurses availability is crucial when considering transferring patients around. The combined patient and nurse allocation model developed in Section 3.2 takes both bed and nurse availability into account and therefore is able to make optimal transfers of patients and nurses to ensure that all patients are properly cared for. Figure 3 shows the results of this model at a state level for the northeastern US as this is the most granular level for which recent nurse employment data is available. The results show that the model is capable of transferring nurses such that the states have the lowest amount of shortage in the number of nurses when compared with the existing demand. This is, however, dependent on the overall capacity of the system as a whole. As can be seen in Figure 3, the system does not have the enough number of nurses to account for the high demand perion in the state of New York.

5.2. Prospective Analysis Results

In addition to looking retrospectively at what could have been during the first wave of the pandemic, we look prospectively at recent spikes in cases in Texas, providing insights for what steps could be taken to reduce the burden on hospitals in the future. For these analyses we use the same parameters as in the retrospective case.

We first consider all of Texas on the county level. The results of this analysis can be seem in Figure 4. It can be seen from the figure that the model can reduce the excess load on the counties significantly, with exceptions that arise due to the capacity limits of the system. We find a set of patient transfers that results in a 49.6% decrease in total excess patient-days over capacity.

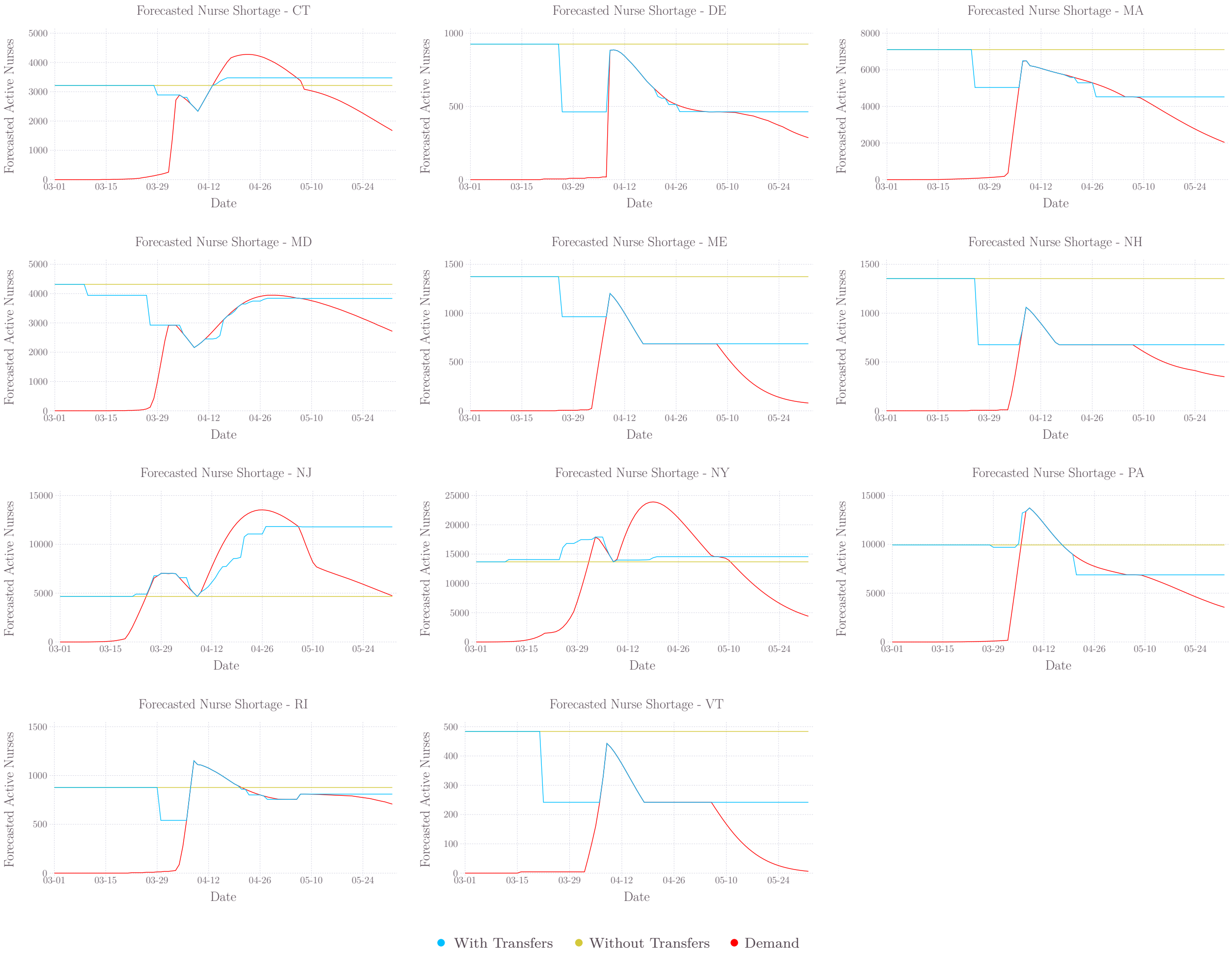

Figure 3 Nurse allocation for the Northeastern US at the state level using the combined allocation model from Section 3.2.

We also consider Harris County at the hospital level. The results of this analysis can be seem in Figure 5. We are able to decrease the total number of excess patient-days over capacity by 44.4%. Note that in both cases, some overflow is unavoidable, as is demonstrated in Figure 6. However, we are able to significantly reduce forecasted overflow while maintaining realistic solutions.

5.3. Website

We have demonstrated the potential of our methods to reduce the burden on hospitals facing an extreme surge in demand. In order to make this a practical tool that hospital systems or governments can use we have provided the models, codes, and data we have used. In addition to this, we have developed a website that people can use to explore our predictions and the potential impact that optimal patient transfers could have. It allows users to modify assumptions and parameters of the model and updates the results in real time.

The website[4] consists of two separate interactive sections; one for patient allocation which provides

[4] `https://covid-hospital-operations.com/`

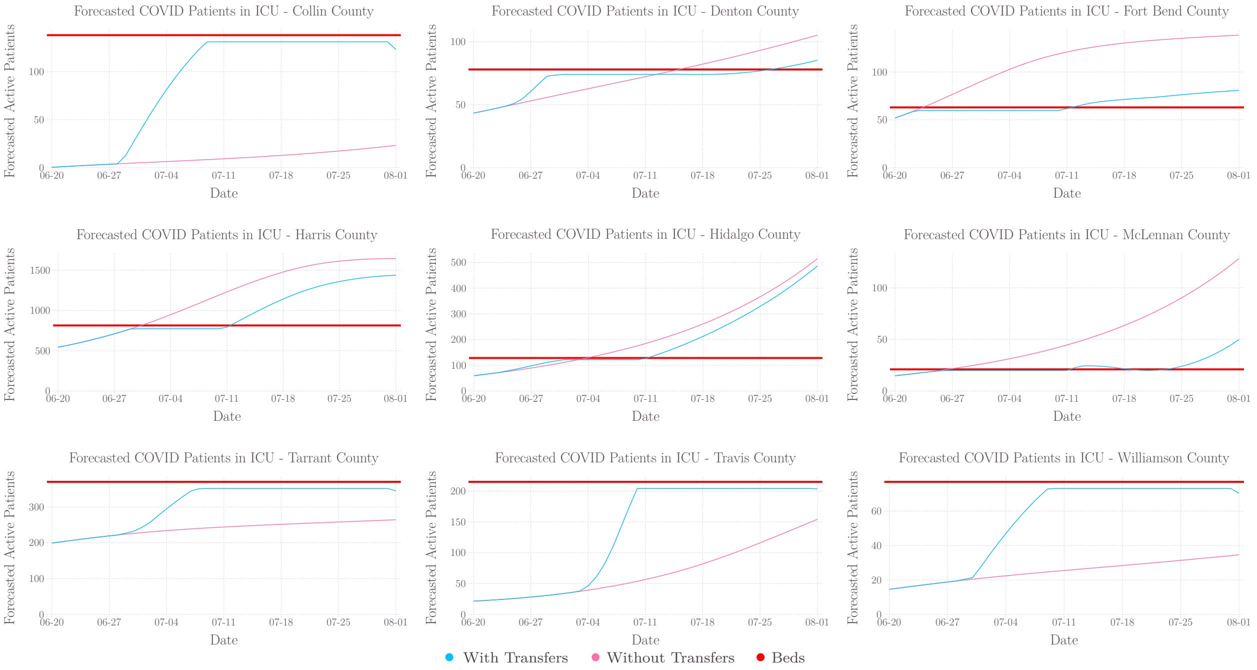

Figure 4 Forecasted COVID patients in the ICU at the county level in Texas, with and without patient transfers. Nine representative counties were selected for display from 254 total counties considered in the model.

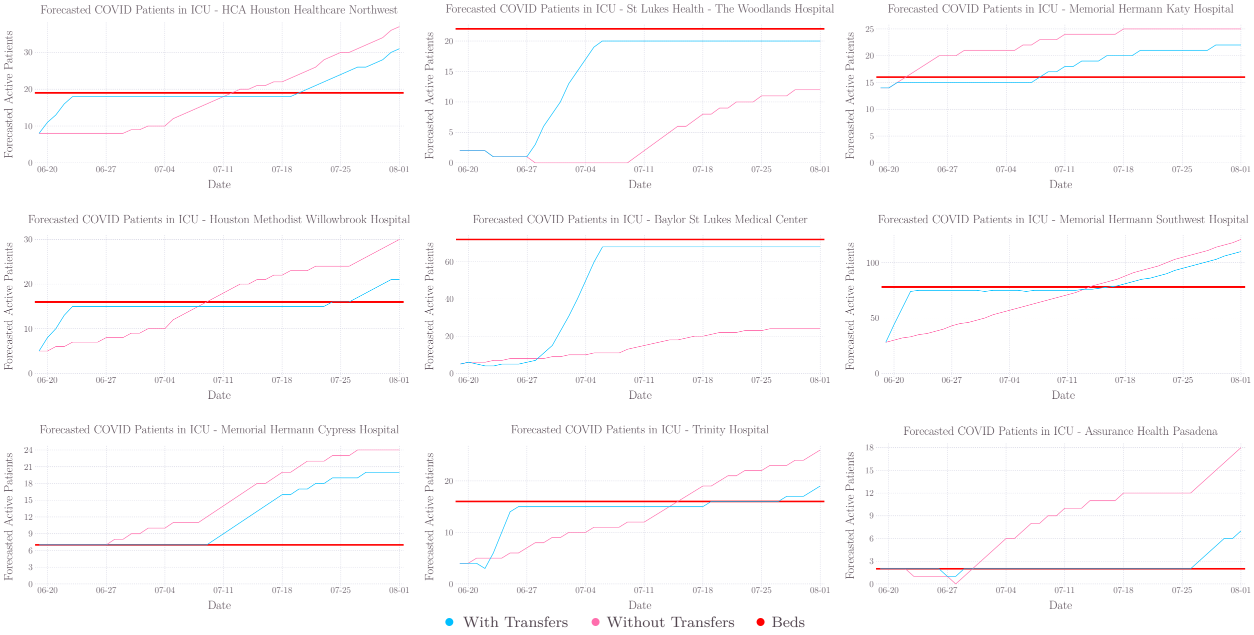

Figure 5 Forecasted COVID patients in the ICU at the hospital level in Harris County, TX (Houston) with and without patient transfers. Nine representative hospials were selected for display from 64 total hospitals considered in the model.

users with the capability to plan for redistributing patients in a time interval with the flexibility of setting several parameters of the problem to the desired level and choosing the preferred forecast model. The other section is designed to solve the nurse allocation model with desired parameter

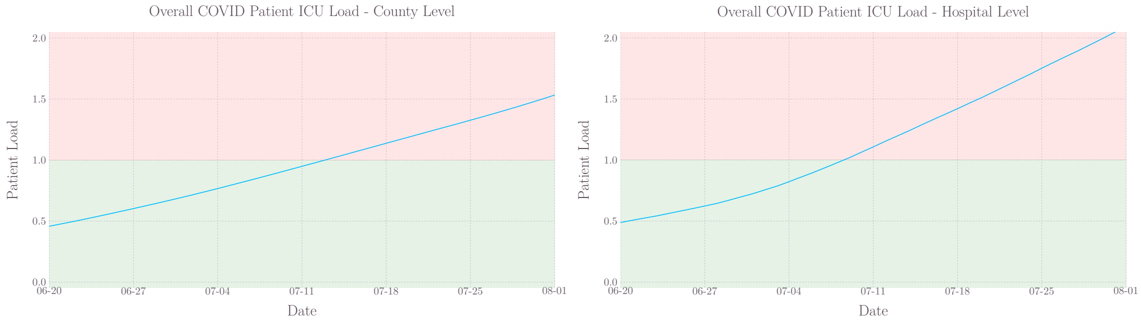

Figure 6 Forecasted total COVID patient ICU load at the county level across Texas (left) and at the hospital level in Harris County (right). Note that both exceed one, so even with perfectly optimal patient re-distribution overflow will be unavoidable.

values. Both tabs provide indicative tables and graphs to further help the procedure of decision making. For the sake of efficiency, they both are based on the base linear programming models for patient allocation and nurse allocation.

6. Conclusions

In this work, we showed that pooling resources in a high demand period such as the COVID-19 pandemic helps hospitals balance their patient loads and provide better care. Results show that by distributing patients among hospitals, it is possible to minimize the adverse effects of resource scarcity on patient care. Additionally, this allocation framework opens the door to systematic collaboration across healthcare entities to prospectively respond to predicted high demand events.

In order to tackle the challenge of patient care with limited capacity and resources during the pandemic we introduced an efficient and flexible patient allocation model. The resources we modeled were beds and nurses, where bed allocation was achieved through redistributing newly admitted patients. We started by introducing base linear programming model for redistribution of the patients from different patient groups with different LOS distributions. We then provide optional MILP constraints to provide more model flexibility in taking into account the policies and preferences of the hospitals at the individual level. We also introduced a linear programming formulation for transferring nurses between healthcare centers which can be deployed in conjunction with our primary patient allocation model for a jointly optimal patient and nurse reallocation scheme. Finally, we made the models robust against a set of uncertainties in the parameters of the problem, namely the number of beds, the number of nurses, and the number of daily admitted patients. In all cases, the models proved to be effective in balancing the loads on the system by relocation.

References

Arora H, Raghu T, Vinze A (2010) Resource allocation for demand surge mitigation during disaster response. *Decision Support Systems* 50(1):304–315.

Bai L, Zhang J (2014) An incentive-based method for hospital capacity management in a pandemic: the assignment approach. *International Journal of Mathematics in Operational Research* 6(4):452–473.

Bertsimas D, Sim M (2004) The price of robustness. *Operations research* 52(1):35–53.

CDC (2020) Strategies for optimizing the supply of facemasks.

Chod J, Rudi N (2005) Resource flexibility with responsive pricing. *Operations Research* 53(3):532–548.

Cooper BS, Pitman RJ, Edmunds WJ, Gay NJ (2006) Delaying the international spread of pandemic influenza. *PLoS Med* 3(6):e212.

Cruz J, Chen G, Li D, Wang X (2004) Particle swarm optimization for resource allocation in uav cooperative control. *AIAA Guid, Ivan ance, Navigation, and Control Conference and Exhibit*, 5250.

Dong E, Du H, Gardner L (2020) An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases* 20(5):533–534.

Drevs F (2013) How patients choose hospitals: Using the stereotypic content model to model trustworthiness, warmth and competence. *Health services management research* 26(2-3):95–101.

Dunning I, Huchette J, Lubin M (2017) Jump: A modeling language for mathematical optimization. *SIAM Review* 59(2):295–320.

Emanuel EJ, Persad G, Upshur R, Thome B, Parker M, Glickman A, Zhang C, Boyle C, Smith M, Phillips JP (2020) Fair allocation of scarce medical resources in the time of covid-19.

Fukushima N, Nagata Y, Kobayashi S, Ono I (2011) Proposal of distance-weighted exponential natural evolution strategies. *2011 IEEE Congress of Evolutionary Computation (CEC)*, 164–171 (IEEE).

Halpern SD, Miller FG (2020a) The Urge to Build More Intensive Care Unit Beds and Ventilators: Intuitive but Errant. *Annals of Internal Medicine* ISSN 0003-4819, URL `http://dx.doi.org/10.7326/M20-2071`, publisher: American College of Physicians.

Halpern SD, Miller FG (2020b) The urge to build more intensive care unit beds and ventilators: Intuitive but errant.

Hegazy T (1999) Optimization of resource allocation and leveling using genetic algorithms. *Journal of construction engineering and management* 125(3):167–175.

Hui DS, Azhar EI, Madani TA, Ntoumi F, Kock R, Dar O, Ippolito G, Mchugh TD, Memish ZA, Drosten C, et al. (2020) The continuing 2019-ncov epidemic threat of novel coronaviruses to global health—the latest 2019 novel coronavirus outbreak in wuhan, china. *International Journal of Infectious Diseases* 91:264–266.

IHME, Murray CJ, et al. (2020) Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. *MedRxiv* .

Judson TJ, Odisho AY, Neinstein AB, Chao J, Williams A, Miller C, Moriarty T, Gleason N, Intinarelli G, Gonzales R (2020) Rapid design and implementation of an integrated patient self-triage and self-scheduling tool for covid-19. *Journal of the American Medical Informatics Association* 27(6):860–866.

Kuchuk G, Nechausov S, Kharchenko V (2015) Two-stage optimization of resource allocation for hybrid cloud data store. *2015 International Conference on Information and Digital Technologies*, 266–271 (IEEE).

Lacasa L, Challen R, Brooks-Pollock E, Danon L (2020) A flexible load sharing system optimising icu demand in the context of covid-19 pandemic. *medRxiv* .

Lampariello L, Sagratella S (2020) Effectively managing diagnostic tests to monitor the covid-19 outbreak in italy. Technical report, Tech. rep. Optimization Online, 2020. url: http://www. optimization-online . . . .

Lewnard JA, Liu VX, Jackson ML, Schmidt MA, Jewell BL, Flores JP, Jentz C, Northrup GR, Mahmud A, Reingold AL, et al. (2020) Incidence, clinical outcomes, and transmission dynamics of severe coronavirus disease 2019 in california and washington: prospective cohort study. *bmj* 369.

Li X, Xu X (2020) A comparative analysis between different resource allocation and operating strategy implementation mechanisms using a system dynamics approach. *International Journal of Production Research* 58(2):367–391.

Lucchini A, Giani M, Elli S, Villa S, Rona R, Foti G (2020) Nursing activities score is increased in covid-19 patients. *Intensive & Critical Care Nursing* .

Luscombe R, Kozan E (2016) Dynamic resource allocation to improve emergency department efficiency in real time. *European Journal of Operational Research* 255(2):593–603.

Mehrotra A, Ray K, Brockmeyer DM, Barnett ML, Bender JA (2020a) Rapidly converting to ”virtual practices”: outpatient care in the era of covid-19. *NEJM catalyst innovations in care delivery* 1(2).

Mehrotra S, Rahimian H, Barah M, Luo F, Schantz K (2020b) A model of supply-chain decisions for resource sharing with an application to ventilator allocation to combat covid-19. *Naval Research Logistics* .

Onder G, Rezza G, Brusaferro S (2020) Case-fatality rate and characteristics of patients dying in relation to covid-19 in italy. *Jama* 323(18):1775–1776.

Otegbeye M, Scriber R, Ducoin D, Glasofer A (2015) Designing a data-driven decision support tool for nurse scheduling in the emergency department: a case study of a southern new jersey emergency department. *Journal of emergency nursing* 41(1):30–35.

Pei S, Kandula S, Yang W, Shaman J (2018) Forecasting the spatial transmission of influenza in the united states. *Proceedings of the National Academy of Sciences* 115(11):2752–2757.

Pereira I (2020) Houston hospitals transferring patients amid covid rise: ’we’re running out of icu beds’. URL `https://abcnews.go.com/US/houston-hospitals-transferring-covid-patients-running-icu-beds/story?id=71552472`.

Rees EM, Nightingale ES, Jafari Y, Waterlow NR, Clifford S, Pearson CA, Jombart T, Procter SR, Knight GM, Group CW, et al. (2020) Covid-19 length of hospital stay: a systematic review and data synthesis .

Ren J, Yu G, Cai Y, He Y (2018) Latency optimization for resource allocation in mobile-edge computation offloading. *IEEE Transactions on Wireless Communications* 17(8):5506–5519.

Sims S, Hewitt G, Harris R (2015) Evidence of collaboration, pooling of resources, learning and role blurring in interprofessional healthcare teams: a realist synthesis. *Journal of Interprofessional Care* 29(1):20–25.

Sun L (2011) Optimization models for patient allocation during a pandemic influenza outbreak. .

Sun L, DePuy GW, Evans GW (2014) Multi-objective optimization models for patient allocation during a pandemic influenza outbreak. *Computers & Operations Research* 51:350–359.

Tiruvoipati R, Botha J, Fletcher J, Gangopadhyay H, Majumdar M, Vij S, Paul E, Pilcher D, Australia, Group NZICSACT (2017) Intensive care discharge delay is associated with increased hospital length of stay: A multicentre prospective observational study. *PloS one* 12(7):e0181827.

Tizzoni M, Bajardi P, Poletto C, Ramasco JJ, Balcan D, Gonçalves B, Perra N, Colizza V, Vespignani A (2012) Real-time numerical forecast of global epidemic spreading: case study of 2009 a/h1n1pdm. *BMC medicine* 10(1):165.

Toner E, Waldhorn R (2006) What hospitals should do to prepare for an influenza pandemic. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* 4(4):397–402.

Tychogiorgos G, Leung KK (2014) Optimization-based resource allocation in communication networks. *Computer Networks* 66:32–45.

Vanberkel PT, Boucherie RJ, Hans EW, Hurink JL, Litvak N (2012) Efficiency evaluation for pooling resources in health care. *OR spectrum* 34(2):371–390.

Varkevisser M, van der Geest SA, Schut FT (2012) Do patients choose hospitals with high quality ratings? empirical evidence from the market for angioplasty in the netherlands. *Journal of Health Economics* 31(2):371–378.

Weissman GE, Crane-Droesch A, Chivers C, Luong T, Hanish A, Levy MZ, Lubken J, Becker M, Draugelis ME, Anesi GL, et al. (2020) Locally informed simulation to predict hospital capacity needs during the covid-19 pandemic. *Annals of internal medicine* .

White DB, Lo B (2020) A framework for rationing ventilators and critical care beds during the covid-19 pandemic. *Jama* 323(18):1773–1774.

Wunsch H, Angus DC, Harrison DA, Linde-Zwirble WT, Rowan KM (2011) Comparison of medical admissions to intensive care units in the united states and united kingdom. *American journal of respiratory and critical care medicine* 183(12):1666–1673.