Received: 21 September 2019 | Accepted: 6 December 2021

DOI: 10.1111/poms.13648

POMS

SPECIAL ISSUE ARTICLE

Operations (management) warp speed: Rapid deployment of hospital-focused predictive/prescriptive analytics for the COVID-19 pandemic

Pengyi Shi[1] | **Jonathan E. Helm**[2] | **Christopher Chen**[2] | **Jeff Lim**[2] | **Rodney P. Parker**[2] | **Troy Tinsley**[3] | **Jacob Cecil**[3]

[1] Krannert School of Management, Purdue University, West Lafayette, Indiana, USA

[2] Kelley School of Business, Indiana University, Bloomington, Indiana, USA

[3] Indiana University Health, Indianapolis, Indiana, USA

Correspondence
Pengyi Shi, Krannert School of Management, Purdue University, West Lafayette, IN 47907, USA.
Email: shi178@purdue.edu

Handling editor: Martin K. Starr

Abstract

At the onset of the COVID-19 pandemic, hospitals were in dire need of data-driven analytics to provide support for critical, expensive, and complex decisions. Yet, the majority of analytics being developed were targeted at state- and national-level policy decisions, with little availability of actionable information to support tactical and operational decision-making and execution at the hospital level. To fill this gap, we developed a multi-method framework leveraging a parsimonious design philosophy that allows for rapid deployment of high-impact predictive and prescriptive analytics in a time-sensitive, dynamic, data-limited environment, such as a novel pandemic. The product of this research is a workload prediction and decision support tool to provide mission-critical, actionable information for individual hospitals. Our framework forecasts time-varying patient workload and demand for critical resources by integrating disease progression models, tailored to data availability during different stages of the pandemic, with a stochastic network model of patient movements among units within individual hospitals. Both components employ adaptive tuning to account for hospital-dependent, time-varying parameters that provide consistently accurate predictions by dynamically learning the impact of latent changes in system dynamics. Our decision support system is designed to be portable and easily implementable across hospital data systems for expeditious expansion and deployment. This work was contextually grounded in close collaboration with IU Health, the largest health system in Indiana, which has 18 hospitals serving over one million residents. Our initial prototype was implemented in April 2020 and has supported managerial decisions, from the operational to the strategic, across multiple functionalities at IU Health.

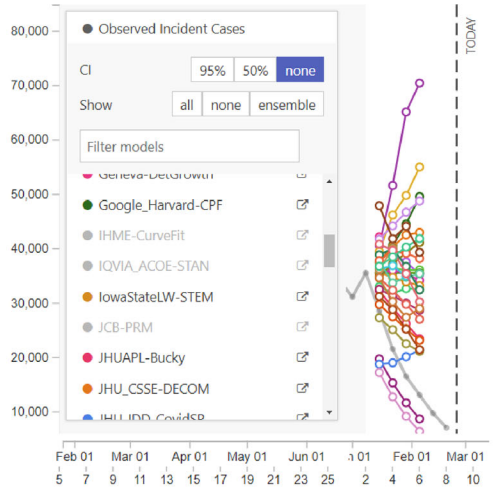
KEYWORDS

epidemiological forecasting, hospital decision support implementation, nurse transshipment, queueing network workload prediction, synthetic control

1 | INTRODUCTION

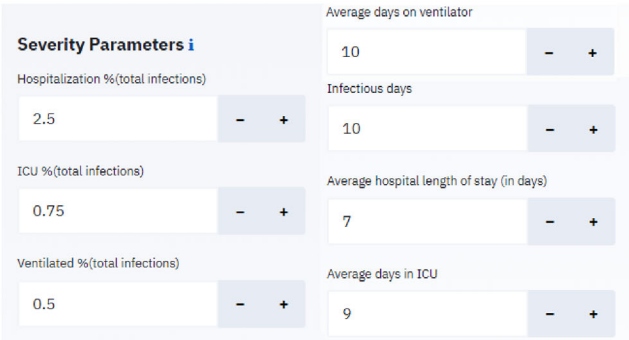
Pandemics place tremendous stress on hospital resources. According to the Office of the Inspector General (Grimm, 2020), hospitals treating COVID-19 (hereafter, COVID) patients in outbreak zones reported severe shortages of specialized hospital beds, nurses, ventilators, PPE, and testing supplies, among others. Effective pandemic response for hospitals relies heavily on the ability to determine (1) the amount of resources needed for COVID patients; (2) how to effectively allocate those resources; and (3) how, when, and how much surge capacity to create. However, the analytics most widely developed and deployed during the COVID pandemic were designed primarily to support large-scale public health agendas. Early in the pandemic, multiple health-care executives in the Indiana Pandemic Information Collaborative expressed concern that the explosion of analytics provided little valuable information to support their

Accepted by Martin K. Starr, after 4 revisions.



(a) Ensemble[a] prediction for Indiana

[a]COVID Forecast Hub: Univ. of Mass



(b) Steady-state parameters for CHIME[a] model

[a]CHIME: Univ. of Pennsylvania

FIGURE 1 Publicly available COVID prediction models [Color figure can be viewed at wileyonlinelibrary.com]

own operational planning. Widely publicized analytics platforms tended to focus on county- or state-level predictions of time-averaged (steady state) bed usage, which was insufficient to support the day-to-day and week-to-week tactical and operational plans being executed by hospitals and hospital systems. This bridges the gap between predictive information and hospital-level execution. Through a close hospital-academia partnership, we developed a suite of analytics to support the rapidly innovated strategic, tactical, and operational actions deployed by hospitals to combat the COVID-19 pandemic.

1.1 | Motivation: The gap between predictions and operational decisions

Many hospitals faced challenges in identifying an analytics driven solution that met their information and decision support needs. The explosion of analytics produced a myriad of widely varying predictions, making it difficult to know which model to use. Further, widely publicized models failed to provide sufficient detail for operational decision-making at the hospital level.

Figure 1a highlights the large variation and generally poor accuracy among popular, publicly available prediction models. Predictions for Indiana for February 1 ranged from 9000 to 75,000 cases, with very few models predicting close to the actual 15,000 (gray line). Figure 1b presents the parameter set used by the well-known CHIME model. CHIME employs a widely used method for translating COVID disease-spread models into hospital resource requirements. This approach predicts COVID demand for hospital resources based on a simple proportional calculation from a prediction of confirmed cases to produce a state-level (county-level at best) average prediction of hospital resource needs. To illustrate CHIME's limitations, consider the Intensive Care Units (ICU) capacity calculation:

$$\text{ICU usage} = \text{daily confirmed cases} \times \text{fraction that need ICU} \times \text{average time in ICU}, \quad (1)$$

which is based on Little's Law and is most appropriate for time-homogeneous, steady-state systems. During a pandemic, these assumptions of Little's Law are violated.

The drawbacks of policy-targeted models for operational decision-making include: (1) failing to account for current state and (transient) time-varying dynamics, (2) failing to capture detailed patient flows in hospitals or distributional information needed for decision support, (3) requiring hospitals to input their own parameters with little guidance on how these inputs should be estimated, and (4) providing state-level predictions that are of little help for individual hospitals.

In contrast to extant models, rather than appending an operational component to a primarily disease-focused model, we develop a multi-stage approach that feeds a disease prediction into a detailed model of hospital operations. This approach required the development of a new theory to overcome the unique challenges of forecasting in the highly dynamic pandemic environment.

Challenges in addressing the gaps in COVID modeling approaches

Although there is a substantial body of literature in patient flow modeling and optimization, most of the research operates under two fundamental assumptions: there is sufficient historical data to accurately estimate patient flows; and patient flow characteristics, such as length-of-stay

distributions, are not changing. In contrast, a pandemic environment has very limited data and exhibits major time-varying disruptions of patient flow primitives. We separate these challenges into four broad categories.

(1) *Limited and censored patient flow data*. At the beginning of the pandemic, March and April 2020 for Indiana, many COVID patients at IU Health had not yet finished their hospital stay, leading to right censoring of the length-of-stay, transitions between units, and mortality statistics. Most patients who were discharged from the hospital early on were those with milder conditions, which causes standard estimation methods to be significantly biased (underestimated).
(2) *Noisy mapping of confirmed cases to individual hospital admissions*. A key feature missing from most COVID analytics models is an estimate of how the number of confirmed cases in a region will translate into an admission at an IU Health hospital and what level of care will be required. The state model in Indiana suggested using market share as a proxy for determining this fraction; however, we found this to be a poor predictor of IU Health admissions. National estimates used in most COVID models for the fraction patients requiring ICU versus medical/surgical (M/S) beds were also inaccurate for IU Health hospitals—the fraction even varied between hospitals. As the pandemic progressed, we also found that bed requirements changed over time with changing demographics of COVID hospitalizations and changes in admission and treatment protocols.
(3) *Evolving patient characteristics*. Our data show that COVID patient characteristics and treatment procedures vary by region of the state and by time. Contributing factors include more elderly patients being admitted in March to April, whereas younger patients constituted the majority of admissions after June, medical professionals becoming more familiar with COVID, and new treatment approaches to streamline the treatment and recovery process. As a result, admissions, unit assignment (e.g., assigning to M/S vs. ICU), length-of-stay, and mortality rates were dynamically evolving. Interestingly, we even found that non-COVID patient characteristics changed during the pandemic, for example, emergency department visits dropped significantly.
(4) *Evolving disease transmission dynamics*. At the onset of the pandemic, testing was extremely limited, and using confirmed cases as the total number of cases led to significant underestimation of actual cases. Further, disease transmission dynamics are driven by complex interactions between public policies (e.g., shelter in place and mask ordinance), human behavior (e.g., policy compliance and COVID fatigue), and re-opening of restaurants and schools, among other factors. These factors are not only difficult to quantify, but are also highly specific to geographical region and culture. This may be a contributing factor to the poor performance of many of the national prediction models when applied to Indiana.

1.2 | Integrated framework and implementation in IU health

Our integrated modeling framework was developed and implemented in close collaboration with the largest health system in the state of Indiana, IU Health; also referred to as IUH (e.g., in Figure 2). When we formed our initial collaboration, most hospitals were using a state-wide prediction tool that failed to provide consistently accurate predictions and actionable information. In response, we jointly developed a suite of analytics based on a data-driven framework designed to bridge the gap between higher level public health predictions and hospital-level pandemic response. Figure 2 illustrates the conceptual model of this framework.

Approach

Our multi-method approach combines a disease prediction model (SIR, Susceptible, Infected, and Recovered) with a stochastic network model of patient flow to predict hospital census in both M/S and ICU units. These two models are joined by a mapping of county-level COVID cases to hospitalizations (workload) at each individual hospital. Novel adaptive tuning methods are employed to determine key parameters in the disease prediction and patient flow models, as well as the mapping of the fraction of COVID cases arriving to a particular hospital.

In creating our framework, we adopt a parsimonious philosophy to capture *first-order effects* with a *minimally complex model* that meets a high-priority hospital need. Two driving principles of this philosophy are eliminate unnecessary model complexity and deliver only actionable information for a need identified by the hospital. Parsimonious design facilitates flexibility, adaptability, and ease of implementation. These features facilitate rapid deployment of high-impact predictive and prescriptive analytics in a time-sensitive, dynamic, data-limited environment such as a pandemic.

Our framework was implemented in mid-April 2020 as a web-based application on IU Health's intranet that was dispatched to their five major service regions, including 18 hospitals serving over one million residents. The tool has the capability of forecasting nurse staffing requirements, ventilator usage, and PPE needs among other resource needs. The tab displayed in Figure 3 shows forecasted demand for ICU and ventilator usage for potential future scenarios of COVID progression. All numbers in the paper are modified to protect sensitive data.

The workload predictions and decision support from our tools have been used by the executives at IU Health and played a critical role in their preparedness for the multiple surges of COVID cases in Indiana. The choice of patient census in the M/S and ICU units as the primary output metric was the result of a joint design effort between IU Health and the academic team; in the design phase, patient census was determined by IU Health leadership as the fundamental input

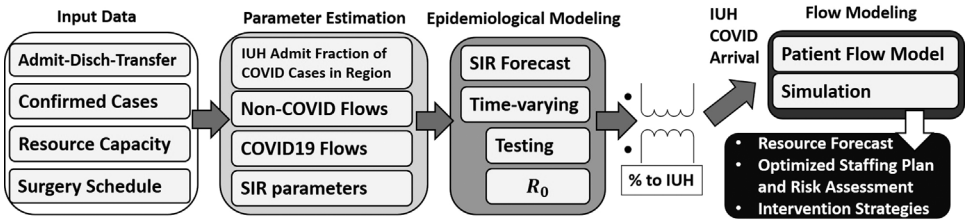


FIGURE 2 Overview of the integrated workload and planning framework and its functionality

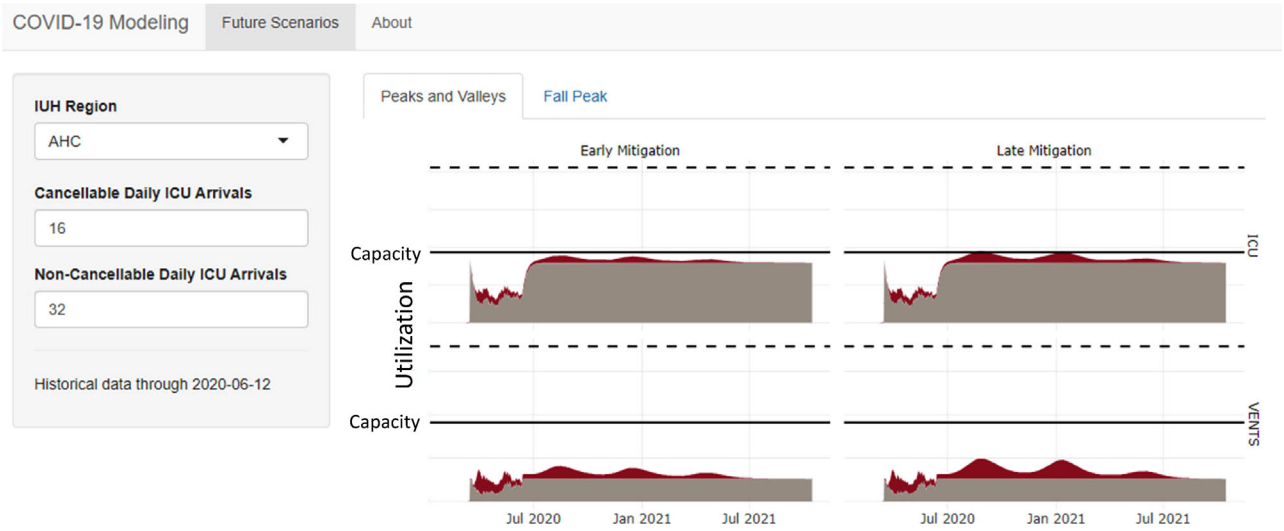


FIGURE 3 Screenshot of one tab of the web-based application for resource forecasting [Color figure can be viewed at wileyonlinelibrary.com]

to their critical operational decisions. Although the research and development was contextually grounded by IU Health's domain knowledge, the product is easily portable to other hospitals and health systems, as evidenced by interest and ongoing conversations with the Indiana Hospital Association as well as other hospital systems in Indiana.

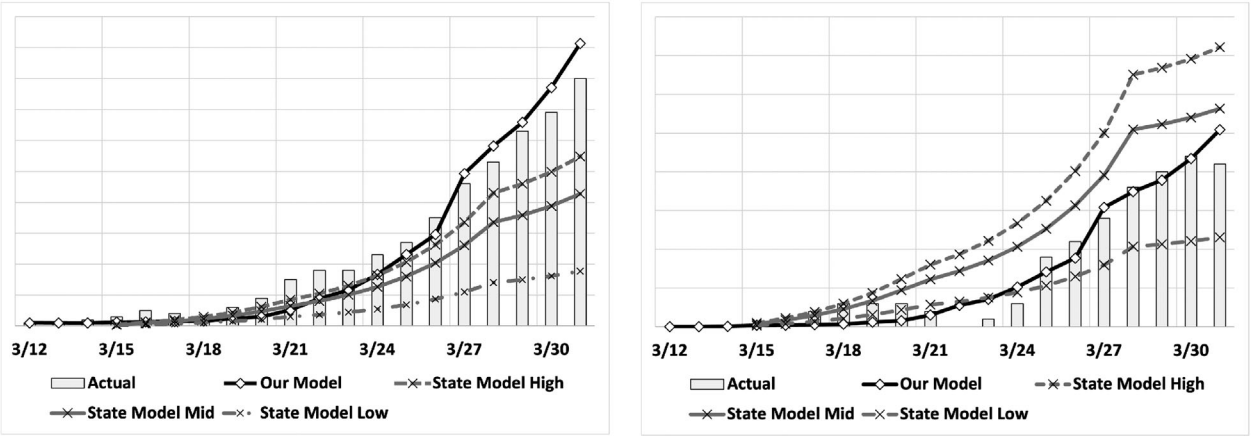
Performance

Prior to implementation, we were asked for a comparative study of our model versus the state model currently being used, shown for Indianapolis in Figure 4. The black solid lines are forecasts from our integrated model. The dark gray lines are the low, medium, and high forecast scenarios for the state model. The light gray bars represent the actual census. The mean absolute percent error (MAPE) for our model was 9.4% and 8.2% for ICU and M/S units. The state model had a MAPE more than four times larger (42.7% and 33.8%) and exhibited an increasingly negative bias, that is, the MAPE for April was >50%. The predicted census numbers shown here are from our integrated model that takes the disease forecast as an input and uses the patient flow model to calculate the census. We implemented a multi-step calibration procedure with adaptive tuning in this integrated model; see details of the calibration in Section 3.5.

1.3 | Contributions

This research contributes to the literature by developing a parsimonious, adaptable, and easily implementable workload tool that is tailored to individual hospitals. Our tool integrates a epidemiology modeling, empirical methods, stochastic network theory, and transshipment optimization, and was implemented in collaboration with IU Health.

- **Disease prediction.** To address severely limited data during the initial phase of the pandemic, we developed an adaptive synthetic control (aSC) method using a weighted portfolio of comparable counties in the United States to predict disease progression in heterogeneous regions. As more data became available, we pivoted to a novel, adaptive SIR compartmental model, enabling short-, medium-, and long-term predictions with consistently high accuracy. Through new adaptive learning methods, this SIR model explicitly accounts for unreported infections and partially observable changes in underlying transmission rates.
- **Workload prediction.** Using a stochastic network model that explicitly incorporates time-varying dynamics, we translate county-level infection predictions into patient census and resource usage at individual hospitals, accounting for movements of COVID patients and non-COVID



(a) ICU census (b) M/S census

FIGURE 4 Comparison between our model, the state model, and the actual census from COVID patients in March. The last five days are used as testing data and the rest as training data. The actual census numbers on the *y*-axis are masked due to nondisclosure requirements. The black solid lines are forecasts from our integrated model; the gray lines are the low, medium, and high forecast scenarios for the state model

patients among different units (ICU and M/S). We leverage an offered-load approximation to provide *distributional* information on the workload process, deriving an appealing analytical form for the workload variance across the network based on prior results for single-station queues. These features bridge the gaps discussed in Section 1.1.

The academic and practical contributions are as follows.

Academic contribution

We develop a *multi-method, multi-layer adaptive method* to integrate the disease and workload models with an adaptively tuned mapping of infections to hospitalizations. This differentiates our work from the patient flow literature in the following ways: (1) we integrate fully developed workload and disease forecasting, (2) we learn model parameters rather than estimating them directly from incomplete or unrepresentative data sources, and (3) we employ adaptive learning as new information becomes available.

Practical contribution

These novel elements combine to create a model that (1) is easily implemented and maintained, (2) produces consistently accurate predictions in environments with severely limited data and changing dynamics that are too complex to directly quantify, (3) can be customized to each individual hospital for actionable information, (4) integrates seamlessly with a suite of decision support optimizations, and (5) is easily portable to other hospitals.

In the rest of the paper, we first discuss the relevant literature in Section 2. We present the patient flow model in Section 3 ahead of the disease prediction model because a hospital-specific workload model filled the largest gap between public health models and models that provided actionable information to individual hospitals. We then present the disease prediction model that feeds the arrival process of the hospital flow model in Section 4. Finally, we illustrate applications, as well as extensions, of our model in Section 5, and we conclude in Section 6.

2 | LITERATURE REVIEW

The COVID pandemic triggered vast quantities of research within the scientific community. The most relevant are works involve disease progression and hospital workload prediction.

2.1 | Disease progression prediction models

COVID spread models can be classified broadly as agent-based and compartmental models. Agent-based models describe a set of rules governing individual behavior and use simulation to examine global trends that emerge from individual movements and interactions with the environment. Although agent-based models that build on compartment models, like Cuevas (2020), Kerr et al. (2021), and Silva et al. (2020), offer tremendous flexibility, they require complex specifications and are computationally intensive. In contrast, our parsimonious approach limits complexity and computational burden by splitting our forecast into a simpler compartmental model for disease forecasting and a stochastic network model for modeling transient patient trajectories. Compartmental models generally involve splitting the population into buckets of SIR but may often be expanded to include other compartments. There are also stochastic compartmental models. References summarizing compartmental

models for infectious diseases include Brauer and Castillo-Chavez (2012) and Diekmann et al. (2013).

Most compartmental models for COVID modify the baseline SIR model by injecting more compartments (e.g., Bertsimas et al., 2021), a time element (e.g., Y.-C. Chen, Lu, et al., 2020), and limited testing capacity (e.g., N. Chen, Hu, et al., 2020), among others. Mamon (2020) includes compartments for Susceptible, Exposed, Infectious, Hospitalized, Critical, Other-recovered, Released, and Dead. Bertsimas et al. (2021) develop DELPHI, a SEIR model extended to include varying states of patient recovery, detection, and quarantine. Although expanding the number of compartments may be appealing, it requires tuning more model parameters with less data. What separates our work is that we use a multi-method approach to employ a tool set designed specifically to capture the dynamics of interest for each component of our forecast: adaptive disease spread models and stochastic queueing networks for patient flow. These simplifications allow greater flexibility to adapt each individual model to account for partially observable and scarce data with time-varying dynamics.

In the severely limited data environment of the early pandemic, we apply the aSC method (C.J. Chen, 2020), which leverages data from counties elsewhere in the United States that have similar characteristics to the focal county to forecast the number of cases. The only other paper using the synthetic controls methodology of Abadie et al. (2010) in a COVID context is Malani et al. (2020), which examines the effect of release from lockdowns. With limited data, we develop an adaptive SIR model with tunable parameters. Other papers using adaptive SIR models include Dos Santos et al. (2021), Shapiro et al. (2021), and Y.-C. Chen, Lu, et al. (2020). We extend these by coupling the SIR to a patient flow model to capture the complexities of additional compartments (e.g., disease severity) while retaining the simplicity of single parameter tuning and simulation to generate prediction intervals to capture forecast uncertainty.

2.2 | Hospital workload models

A central element of our integrated model is the stochastic network model, which captures patient flow among different units in the hospital and predicts unit-level workload (patient census) in Section 3. In the context of workload modeling for COVID patients, several papers incorporate compartments on hospitalization in different units into their disease progression model, for example, Capistran et al. (2020), Garrido et al. (2020), Hill et al. (2020), and Veloz et al. (2020). Bartz-Beielstein et al. (2020) develop a detailed discrete-event simulation model for patient flow and provide resource requirement prediction under various worst-case and best-case scenarios. Compartment models produce average patient demand that is more suitable for systems in the steady state. These models often neglect the detailed patient flow after admission that is key to capturing critical transient dynamics, such as current hospital census, future discharges, and transfers to/from M/S to ICU, whereas higher level models

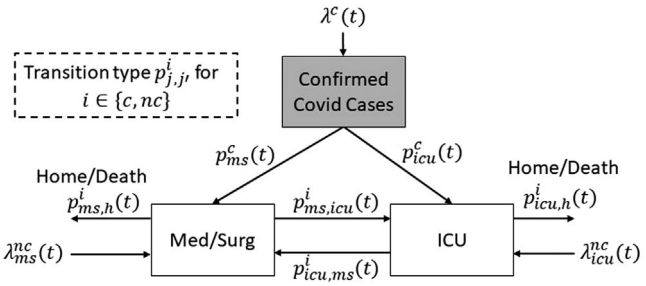


FIGURE 5 Queueing network model

assume a single, static location for each patient. On the other hand, complicated event-based simulation models require many input parameters to be estimated, creating challenges in terms of computation time, parameter identification, and data scarcity. These models can be too cumbersome when there is need for regular updating and rapid implementation.

Structural queueing models for predicting COVID patient demand capture patient flow but often lack the disease spread component. They are also challenged by parameter estimation, which is not a major concern in typical (nonpandemic) patient flow settings. Kaplan (2020) uses a single-station Erlang loss model to evaluate the ICU capacity needed at Yale New Haven Hospital. Bekker et al. (2021) consider a single-station infinite server queue to predict patient census. They first estimate the day-dependent discharge probability from data and then fit a log-normal distribution for LOS. Zhang et al. (2020) consider unit transfers but assume deterministic length-of-stays. They do not consider possible time-varying changes in patient mix or LOS distributions either. Importantly, all these papers initially estimate the patient-flow parameters from data that do not change over time in contrast to the adaptive procedure employed in our methodology. Section 1.1 discusses the importance of accounting for the rapidly changing data environment in estimating these parameters. The most relevant paper is Betcheva et al. (2020). The authors develop an integrated disease forecast and resource planning tool for the East of England region. The paper adopts a similar idea of using the mean-squared error to learn system parameters; however, the authors assume constant system parameters, which differs from our adaptive procedure to estimate time-varying system parameters, which is critical to pandemic modeling and forecasting. It is also unclear from the paper how patient flow dynamics are modeled and whether the model can produce distributional workload information as we have done in Section 3.

3 | STOCHASTIC NETWORK MODEL FOR PATIENT FLOW

Patient flow in each hospital is modeled as a two-station stochastic queueing network, where the stations correspond to ICU units and M/S units in the hospital. Figure 5 illustrates a model with COVID (denoted by c) and non-COVID

(denoted by *nc*) patients. In our implementation, we further separate non-COVID patients into elective and emergency. In the figure, λ represents arrival rates, with $\lambda^c(t)$ being either the predicted (Section 4) or observed confirmed COVID cases in the county, depending on whether t is before or after the last observed data point. p_j^c for $j \in \{MS, ICU\}$ is the adaptively tuned fraction of county-level confirmed cases that enter the target hospital through either the ICU or M/S units, respectively. $p_{j,j'}^i$ is the probability that a patient of type $i \in \{c, nc\}$ will transfer from unit j to $j' \in \{MS, ICU, H, D\}$, where H, D represent discharge to home or death. Each patient type has distinct flow paraters.

In Section 3.1, we describe the motivation behind the choice of the two-station network model. In Section 3.2, we list our assumptions on the arrival, transfer, and service-time processes. In Section 3.3, we specify the offered-load approximation for the workload process. In Section 3.4, we present models for exogenous blocking. In Section 3.5, we introduce our adaptive learning procedure to estimate the flow parameters for the workload process.

3.1 | Two-station network

Patients arrive to either the ICU or M/S units of the hospital and stay for a random time before transferring to another unit or leaving the hospital. This model is sufficiently flexible to capture general networks with multiple stations and arbitrary transition structures. Details for general networks are provided in Appendix B.2 (Supporting Information). Next, we provide justification for our choice to implement this simpler network structure at IU Health as part of our parsimonious design process.

During the development process, limited data rendered the parameterization of a network model capturing a more detailed classification of care units impractical/infeasible. Additionally, many of the major decisions made by IU Health during the pandemic, such as nurse staffing and bed capacity creation, were based on the "level of care" classification, which divides the patient population into *two categories* based on severity of condition: critical and noncritical. These two categories require different types of resources, whereas IU Health has significant flexibility to move resources within a level of care class. As an example, a critical care nurse can generally work in any of the critical care units, but an M/S nurse generally cannot work in critical care units. Early in the pandemic, IU Health took advantage of this flexibility by pooling resources within each level of care to improve response to evolving and uncertain demand. Finally, this two-tier classification facilitated rapid implementation of operational changes by avoiding the challenges of managing the diverse and complex organizational structures and operational procedures across individual units.

Although ICU and M/S generally align with critical and noncritical care, there are some subtleties. We combined ICU and Progressive Care Units (PCU), both of which provide critical care. Mixed-acuity units serve some critical care and some noncritical care patients. We include the critical care portion under the ICU category and the noncritical care portion within the M/S category. These classifications were proposed by IU Health based on their operational structure.

3.2 | Arrival, transfer, and service time processes

For exposition, we focus our presentation on COVID patients; the model architecture extends trivially to non-COVID patients. As such, we omit the superscript $i \in \{c, nc\}$ for patient types. Let $\lambda(t)$ be the deterministic arrival rate function for the number of newly confirmed COVID cases on day t. When instantiating the model on a given day, t', $\lambda(t)$ represents observed cases for previous days, $t' < t$, and predicted cases for future days, $t' \geq t$. New COVID patient arrivals to unit j on day t are modeled by a random variable $\Lambda_j(t)$ with mean $\lambda_j(t) = \lambda(t)p_j(t)$, where $p_j(t)$ represents the time-dependent fraction of confirmed cases that arrive to unit j of a hospital. Arrivals need not follow a Poisson process; non-Poisson arrivals are addressed in Section 3.3.

After admission to a unit, j, a patient remains in the unit for a random amount of time modeled by the random variable $S_j \sim \text{Geo}(\mu_j(t))$, which follows a geometric distribution with time-dependent success probability $\mu_j(t)$. This structure permits the derivation of a closed-form expression for the workload mean and variance and plays an important role in our parsimonious design by facilitating adaptive tuning methods. Tuning this simpler LOS distribution produced better accuracy than attempting to estimate more complicated service-time distributions due to limited data and rapidly evolving patient LOS dynamics over the course of the pandemic.

On discharge from unit j, a patient transitions to their next phase of care, $j' \in \{ICU, MS, H, D\}$ (transfer to ICU, transfer to M/S, discharge to home, death) with probability $p_{j,j'}(t)$. For compactness, we define the set of tuning parameters for this network model as

$$\Theta(t) = \{\mu_j(t), p_j(t), p_{j,j'}(t)\}. \tag{2}$$

During the pandemic, regular tuning was required to capture the evolution of patient condition and care protocols, as well as incorporate the growing volume of new data about COVID-19. This tuning procedure is presented in Section 3.5.

3.3 | An offered-load approximation for workload distribution in a network

In this section, we develop an offered-load approximation for the stochastic workload process $\{X_j(t), t = 0, 1, \ldots\}$ with parameters $\Theta(t)$, where $X_j(t)$ represents the number of patients in unit j on day t. This offered load model is

readily extended to more general networks (Appendix B.2, Supporting Information) and to capture demand for other critical hospital resources such as ventilators and staff, both of which (and others) were included in the model implementation at IU Health.

Let $X_1(t)$ and $X_2(t)$ denote the workloads in the ICU and M/S units, respectively. We approximate $(X_1(t), X_2(t))$ with a multivariate normal distribution, $\mathcal{N}(x(t), v(t))$, with mean $x(t) = (x_1(t), x_2(t))$ and the covariance matrix $v(t)$. The mean for unit j, $x_j(t)$, is calculated from the fluid approximation by solving the dynamic equation

$$x_j(t+1) = x_j(t)\big(1-\mu_j(t)\big) + \lambda_j(t) + p_{j',j}(t)\cdot x_{j'}(t)\mu_{j'}(t), \quad (3)$$

where $j' = ICU$ for $j = MS$ and $j' = MS$ for $j = ICU$. Here, $x_j(t)\mu_j(t)$ corresponds to the expected number of patients leaving unit j on day t, which follows from the geometric LOS. Patients arrive to unit j from two sources: new patient arrivals, with mean $\lambda_j(t)$, and transfers from unit j' to unit j, with mean $x_{j'}(t)\mu_{j'}(t)p_{j',j}(t)$, which is the expected discharges from unit j' multiplied by the probability that a patient discharged from unit j' will transfer to unit j.

To account for variability, we next approximate the hospital workload distribution. We begin by presenting a single-station queueing model of hospital census. We then extend this classical model to the two-station network model implemented at IUH. This analysis reveals an appealing structure for the network variance calculation, where the impact of the network transfer intensity (e.g., $\mu_{j'}(t)p_{j',j}(t)$) appears as an isolated variance term, $\tilde{v}_{2,j}$, added to the formula for the single-station variance, explicitly highlighting the impact of network transitions on workload variance.

Heavy-traffic variance for a single station

Following Whitt and Zhao (2017), we assume the distribution of the arrival process for a single-station $G_t/GI/\infty$ queue is approximately Gaussian any given t, by which we mean that the number of arrivals on interval $[t, t+1]$ is

$$A(t, t+1) \approx \mathcal{N}(\lambda(t), c_a^2\lambda(t)), \quad (4)$$

where c_a^2 is the variability parameter for the index of dispersion of the arrival process ($c_a^2 = 1$ in case of the Poisson arrival process). Assuming the system starts empty in the distant past, Theorem 2.2 of Whitt and Zhao (2017) states that the workload $X(t)$ follows a normal distribution with the same mean as the $M_t/GI/\infty$ queue with arrival rate $\{\lambda(t)\}$, but with variance given by

$$v(t) = \int_0^t \lambda(t-s)V(s)ds, \quad (5)$$

where $V(s) = \bar{G}(s) + (c_a^2 - 1)\bar{G}(s)^2$ and $\bar{G}(s)$ is the complementary cumulative distribution function (CCDF) of the service time random variable. In our discrete-time model with geometric service times and time-dependent success probabilities, the CCDF of the service time for unit j is given by $\bar{G}_j(s) = \prod_{d=1}^{s}(1-\mu_j(d))$ for $s \geq 1$ and $\bar{G}_j(0) = 1$. Without transfers, the workload at unit j at time $t+1$ can be calculated by excluding the last term from (3), with variance being

$$v_j(t+1) = \sum_{s=0}^{t} \lambda(t-s)V_j(s), \quad (6)$$

where $V_j(s) = \bar{G}_j(s) + (c_a^2 - 1)\bar{G}_j(s)^2$ and $V_j(0) = c_{a_j}^2$ is the variability of the arrival process.

Network variance and covariance

We develop a recursive method for calculating the variance and covariance for the workload process at time $t+1$. The workloads on day $t+1$ follow:

$$X_1(t+1) = V_{11}(t) + A_1(t, t+1) + V_{21}(t), \quad (7)$$

$$X_2(t+1) = V_{12}(t) + A_2(t, t+1) + V_{22}(t), \quad (8)$$

where $A_j(t, t+1)$ is the random variable for the arrivals in the interval $[t, t+1)$, with mean $\lambda_j(t)$ and variance $c_{a_j}^2\lambda(t)$. $V_{jj}(t)$ represents patients not discharged from unit j at time t, and $V_{j',j}(t)$ represents patients transferred to unit j from another unit j', for $j = 1, 2$ and $j' = 2, 1$. Finally, let $X_j(t) - V_{jj}(t) - V_{j,j'}(t)$ represent patients discharged home or expired from unit j in $[t, t+1)$. Using these three compartments, not discharged, transferred, and left the system, we can write

$$X_j(t) = V_{jj}(t) + V_{j,j'}(t) + (X_j(t) - V_{jj}(t) - V_{j,j'}(t)). \quad (9)$$

These compartments form a triplet, $(V_{jj}(t), V_{j,j'}(t), X_j(t) - V_{jj}(t) - V_{j,j'}(t))$, that follows a multinomial random variable (r.v.) with the number of trials being the number of patients currently in the unit, $X_j(t)$, and the probability vector being $(1-\mu_j(t), p_{j,j'}\mu_j(t), (1-p_{j,j'})\mu_j(t))$. The expectations of the first two compartments of the triplet are

$$\begin{aligned} E[V_{jj}(t)|X_j(t)] &= (1-\mu_j(t))X_j(t), \\ E[V_{j,j'}(t)|X_j(t)] &= p_{j,j'}\mu_j(t)X_j(t), \end{aligned} \quad (10)$$

the variances are

$$\begin{aligned} \mathrm{Var}[V_{jj}(t)|X_j(t)] &= \mu_j(t)(1-\mu_j(t))X_j(t), \mathrm{Var}[V_{j,j'}(t)|X_j(t)] \\ &= p_{j,j'}\mu_j(t)(1-p_{j,j'}\mu_j(t))X_j(t), \end{aligned} \quad (11)$$

and the covariance between them is

$$\mathrm{Cov}[V_{jj}(t), V_{j,j'}(t)|X_j(t)] = -X_j(t)(1-\mu_j(t))\cdot p_{j,j'}\mu_j(t). \quad (12)$$

Recursive variance calculation

Leveraging the above analysis, we now present our recursive approach to calculate the variance of $X_j(t+1)$. We proceed by analyzing the variance and covariance of each component in (7) and (8) individually. Take (7) as an example. Let $\tilde{v}_{11} = \text{Var}(V_{11}(t))$. The law of total variance implies that

$$\begin{aligned}\tilde{v}_{11} &= \mathbb{E}[\text{Var}(V_{11}(t)|X_1(t), X_2(t))] + \text{Var}[\mathbb{E}(V_{11}(t)|X_1(t), X_2(t))] \\ &= \mathbb{E}[X_1(t)\mu_1(t)(1-\mu_1(t))] + \text{Var}[X_1(t)(1-\mu_1(t))] \\ &= x_1(t)\mu_1(t)(1-\mu_1(t)) + (1-\mu_1(t))^2 v_1(t), \end{aligned} \tag{13}$$

where $v_1(t) = \text{Var}[X_1(t)]$ is the variance of the workload at time t. Similarly, the variance of $V_{21}(t)$, denoted as $\tilde{v}_{21}$, is

$$\begin{aligned}\tilde{v}_{21} &= \mathbb{E}[\text{Var}(V_{21}(t)|X_1(t), X_2(t))] + \text{Var}[\mathbb{E}(V_{21}(t)|X_1(t), X_2(t))] \\ &= \mathbb{E}[X_2(t)p_{21}\mu_2(t)(1-p_{21}\mu_2(t))] + \text{Var}[X_2(t)p_{21}\mu_2(t)] \\ &= x_2(t)p_{21}\mu_2(t)(1-p_{21}\mu_2(t)) + (p_{21}\mu_2(t))^2 v_2(t), \end{aligned} \tag{14}$$

where $v_2(t) = \text{Var}[X_2(t)]$.

Let $\tilde{cv}_{12}$ be the covariance between V_{jj} and $V_{j'j}$. The law of total covariance implies that

$$\begin{aligned}\tilde{cv}_{12} &= \text{Cov}(V_{11}(t), V_{21}(t)) \\ &= \mathbb{E}[\text{Cov}(V_{11}(t), V_{21}(t)|X_1(t), X_2(t))] \qquad (15) \\ &\quad + \text{Cov}[\mathbb{E}(V_{11}(t)|X_1(t), X_2(t)), \mathbb{E}(V_{21}(t)|X_1(t), X_2(t))], \qquad (16)\end{aligned}$$

where (16) equals

$$\begin{aligned}&\text{Cov}[X_1(t)(1-\mu_1(t)), X_2(t)p_{21}\mu_2(t)] \\ &= (1-\mu_1(t))p_{21}\mu_2(t)\cdot cv(t), \end{aligned} \tag{17}$$

and $cv(t) = \text{Cov}(X_1(t), X_2(t))$. The first term, (15), equals 0. To see this, conditioning on $X_1(t) = x_1$, $V_{11}(t) = \sum_{k=1}^{x_1} B_{1k}$ is the sum of outcomes from x_1 independent trials with (marginal) success probability $(1-\mu_1(t))$. Similarly, conditioning on $X_2(t) = x_2$, $V_{21}(t)$ is the sum of outcomes from x_2 independent trials. Given (x_1, x_2), all these trials are independent of one another as the trials are drawn from two different pools of patients in the two units. Hence, $\text{Cov}(V_{11}(t), V_{21}(t)|X_1(t), X_2(t)) = 0$. Assuming the arrival random variable, $A_1(t, t+1)$ is independent of V_{11} and V_{21}, the variance is given by

$$v_1(t+1) = \text{Var}[X_1(t+1)] = \tilde{v}_{11} + c_{a_1}^2\lambda_1(t) + \tilde{v}_{21} + \tilde{cv}_{12}. \tag{18}$$

The variance for unit 2, $v_2(t+1) = \text{Var}[X_2(t+1)]$, can be calculated similarly.

The recursive calculations above require initial values for $v_1(0), v_2(0)$. One method for initializing this recursion is to use the steady-state values of the workload variance from the single-station model by letting $t \to \infty$ in (6).

Covariance calculation

The covariance of the two queue lengths is given by

$$\begin{aligned}cv(t+1) &= \text{Cov}(X_1(t+1), X_2(t+1)) \\ &= -x_1(t)(1-\mu_1(t))p_{12}\mu_1(t) - x_2(t)(1-\mu_2(t))p_{21}\mu_2(t) \\ &\quad + \big((1-\mu_1(t))(1-\mu_2(t)) + p_{21}\mu_2(t)p_{12}\mu_1(t)\big)cv(t) \\ &\quad + (1-\mu_1(t))p_{12}\mu_1(t)v_1(t) + (1-\mu_2(t))p_{21}\mu_2(t)v_2(t). \end{aligned} \tag{19}$$

The details of the derivation are specified in Appendix B.2 (Supporting Information).

Connection between the single-station and network variances

Equation (18) has an elegant structure that explicitly highlights the marginal impact of the network transition matrix on the workload variance through (i) the additional term, $\tilde{v}_{21}$, that can be interpreted as additional variance generated by transfers, and (ii) the covariance term, $\tilde{cv}_{12}$, that explicitly characterizes the correlation between the queue lengths in the two stations. We specify this connection in Appendix B.3 (Supporting Information) by writing the single-station workload variance as a recursive function for direct comparison with the network variance calculation.

Remark 1. The offered-load approximation for the network model is for a given t, that is, approximating the distribution of $(X_1(t), X_2(t))$, not for the joint distribution across different time points or for process-level approximation. When calculating the mean and variance for the workload at t, we leverage the recursive calculation that conditions on starting the approximation at some time $s = 0$, which does not account for the realization for $s < t$. This type of point approximation is sufficient for the workload prediction and associated decision-making in this paper. Developing process approximation involves many-server heavy-traffic analysis such as the one in Whitt and Zhao (2017) and is beyond the scope of this paper.

3.4 | Additional model features

In this section, we discuss how to calculate the total workload by combining the workload generated by different types of patients (e.g., COVID and non-COVID) that have

different flow parameters and how to include exogenous blocking within our framework.

Estimating workload for different types of patients

The offered-load approach introduced in Section 3.3 applies to any set of patient types. Due to the additivity of the offered-load model, the full hospital workload, including an arbitrary number of patient types, can be decomposed into individual and independent offered-load models for each type of patient, where each model has a unique set of flow parameters. The final workload can be obtained by adding the workloads from each individual offered-load model (e.g., the models for different types of patients).

In this paper, we separate the workload for COVID and non-COVID patients, which allows us to tailor the parameter estimation to these disparate patient classes. This is particularly important in a pandemic because flow parameters for COVID patients change rapidly over time, whereas the parameters for non-COVID patients are more stable. For implementation, we further decompose the non-COVID patients into emergency versus elective admissions. This provides flexibility to model elective admissions as a modifiable parameter, which allows hospital managers to explore the effect of canceling/resuming elective surgeries. One could further decompose the workload calculation by medical specialty, though this was not implemented due to the need for model simplicity and rapid implementation.

One practical feature of this decomposition approach is the ability to visualize the forecast at different levels of granularity in terms of types of patients and types of resources. Leveraging this feature, we were able to create tools for management that can display either total ICU and M/S workloads or splice the prediction along patient types, groups of patient types, or even resources required (e.g., ventilators and nurse staff) depending on the decisions to be made. For example, drilling down to focus solely on COVID patient workloads allows hospitals to plan for scarce resources that are commonly utilized by COVID patients but not as highly demanded by non-COVID patients. On the other hand, planning for total bed capacity or nurse staffing requires knowledge of the workload for all patients in ICU and M/S units. Our additive method also allows workloads to be displayed at different levels of granularity for physical resources, for example, individual hospital, a group of flagship hospitals, all the hospitals within a given region, or all the hospitals in the system. Each of these perspectives is important for supporting different types of health system decisions.

Time-varying system with finite capacity

One drawback of the offered-load approach is that the results are based on an infinite-server queueing model, which assumes away blocking and waiting. This underlying assumption can be appropriate when the primary objective is to estimate the *demand for different services* within the hospital. The data obtained from IU Health is also consistent with this demand-driven perspective. In the data, "unit" is determined by the level of care (critical vs. noncritical), which captures true demand for service within the hospital as the level of care is determined strictly by patient condition and not by other factors that can obfuscate the true demand such as physical location. Finally, the primary use cases for our model at IU Health are also consistent with our demand-centered model. Specifically, many of the decisions being supported involved redistributing flexible resources that do not have hard capacity constraints and can be flexed and moved to satisfy surges in demand (e.g., nurse staff and ventilators; see Section 5). As the key concern during a pandemic is the ability to adjust capacity to match surge demand, the offered load model is conceptually a good fit for these types of decision processes.

In contrast to more flexible resources, hospital bed capacity represents a more restrictive constraint. Beds cannot be moved between locations; hence, bed capacity cannot be stretched as easily as modifying nurse to patient ratios in response to demand surges. During standard hospital operations, blocking in hospitals occurs both for exogenous arrivals and internal bed assignments/transfers, the latter being difficult to evaluate analytically. In contrast, during a pandemic, mass pooling of resources and the flexing of physical space in response to a surge in patients serve to minimize the otherwise common occurrence of internal bed blocking. For example, M/S patients can use critical care beds, critical care patients can be flexed to mixed acuity M/S units, and certain units can be repurposed to be more flexible in the types of patients they can care for. In this environment, the primary blocking mechanism is the blocking of external arrivals to the hospital itself.

Due to these simplified blocking dynamics, our offered-load approach can be used to approximate exogenous arrival blocking probabilities and other congestion-related performance metrics. We begin by considering the entire hospital as a $G_t/GI/n$ queue after pooling, with n being the total bed capacity and the service time being a mixture of the service time random variables for the different types of patients. Following Green et al. (2007), we approximate the time-varying system by a stationary finite-server $GI/GI/n$ queue at any given time t, with the arrival rate to the stationary queues being adjusted using the modified offered-load (MOL) approach. That is, the "modified" arrival rate at time t follows

$$\lambda_{\mathrm{MOL}}(t) = \frac{\mathbb{E}[X(t)]}{\mathbb{E}[S]}. \tag{20}$$

Here, $\mathbb{E}[X(t)]$ is the expected workload on day t, which is calculated by adding all the workloads (COVID and non-COVID patients) from the offered-load model in Section 3.3. $\mathbb{E}[S]$ is the expected service time averaged over all patient types. We analyze the corresponding stationary finite-server queue with arrival rate $\lambda_{\mathrm{MOL}}(t)$. If the service time is exponential

or a mixture of exponentials, we can either use numerical methods or heavy-traffic approximations to analyze system performance measures. For example, for an $M/M/n$ queue, when the number of beds is large (e.g., hundreds of beds) and the system runs in the many-server Halfin–Whitt regime, the blocking (delay) probability can be efficiently approximated by

$$\mathbb{P}(X(t) \geq n) \approx [1 + (\beta\Phi(\beta)/\phi(\beta))]^{-1}, \tag{21}$$

where Φ and ϕ are the cumulative distribution function and probability density function of the standard normal distribution, respectively, and β is the "square-root staffing" coefficient such that $n = \mathbb{E}[X(t)] + \beta\sqrt{\mathbb{E}[X(t)]}$. Green et al. (2007) also provide formulas for blocking probabilities in other systems such as those with abandonment. In the context of discrete-time queues with geometric service times, Dai and Shi (2017) and Feng and Shi (2018) provide diffusion approximations that allow for efficient calculation of various performance metrics such as the expected queue length and waiting time. We leave to future research the development of algorithms for approximating internal blocking within the hospital's network of units, for example, solving a multi-dimensional diffusion process to compute blocking and delay performance metrics in these systems.

3.5 | Adaptive learning for COVID patient flow parameters with data scarcity

To address challenges (1)–(3) in Section 1.1, we develop an automated learning procedure for tuning uncertain or censored patient-flow parameters, $\Theta = \{p_j^c, \mu_j^c, p_{j,j'}^c\}$, to achieve the best fit with the workload process for COVID patients. These parameters are learned from historical data (training data), and the accuracy is verified on testing data. Let Θ^* be the parameter set that minimizes the mean-square loss between the observed census and the predicted (mean) census from (3). Letting $\hat{x}_j(t)$ and $x_j(t)|\Theta$ be the observed and predicted census in unit j on day t, respectively, Θ^* is calculated by

$$\Theta^* = \arg\min_{\Theta} \sum_{t=1}^{T} \sum_{j} \left(\hat{x}_j(t) - x_j(t)|\Theta\right)^2, \tag{22}$$

where T is the number of days included in the training data.

Time-dependent parameters

We use nonparametric estimation to model $\Theta(t)$ as a piecewise discrete function, partitioning training data into I time intervals defined by the ordered set $\mathcal{I} = (t_0, t_1, \ldots, t_I)$, where $\Theta(t) = \Theta_i$ for $t \in [t_i, t_{i+1})$. We perform cross-validation using the training data to find the best interval set, $\mathcal{I}$, to avoid overfitting while allowing enough flexibility to capture the underlying changes in the parameters over time.

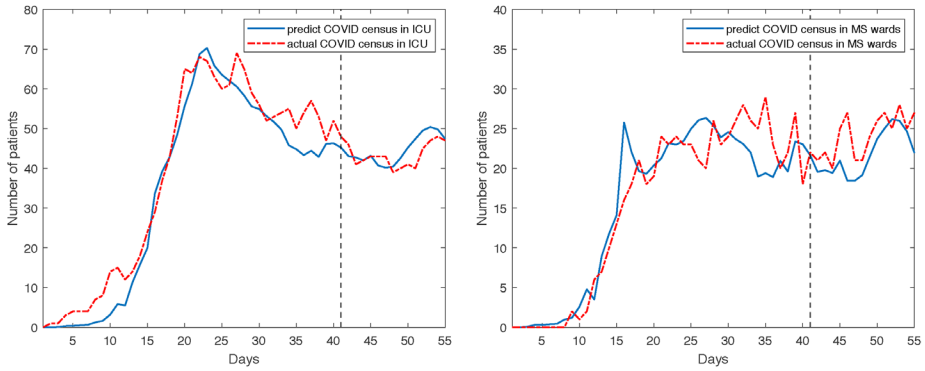
Adaptive learning procedure

During the pandemic, particularly in the early stage when data were scarce and system dynamics were changing rapidly, we reparameterized $\Theta(t)$ on a weekly basis, using incremental daily data dumps automatically generated from IU Health's census database and loaded as a flat file onto a secure virtual machine behind IU Health's firewall. The dataset used for training and testing is updated as new data become available. Parameters are estimated using all historical data up to the current time point. We elected to incorporate all the data rather than a rolling horizon because patient LOS was long relative to the data window. Using an expanding window prioritized stable estimates over change detection and yielded more conservative workload predictions. After each parameter update, we generated the workload prediction for the next 14 days for operational use. We also generated longer term census projections to support strategic decision-making. Figure 6 shows one snapshot from this process comparing the observed and predicted COVID inpatient census at one of IU Health's hospitals from March 12 to May 6, 2020. For this instance, the training set included data up to April 22 (41 days). Using the learned parameters, we generated a prediction over the next 14 days (April 23 to May 6)—the right side of the vertical dashed line in Figure 6. Prediction accuracy is calculated as the MAPE over the testing set. This mid-April prediction can be compared with the first set of predictions generated at the beginning of the pandemic in March (Figure 4) to observe the benefits of adaptive learning. The tuned model is able to capture time-varying fluctuations and adjust LOS and transfer estimates as more patients are discharged from the hospital. In Appendix B (Supporting Information), we show another set of plots for the prediction generated using flow parameters learned from data up to May 19 and tested on data from May 20 to June 2, 2020.

Prediction performance

The MAPE for the ICU and M/S predictions from our integrated framework was 9.4% and 8.2%, respectively, on the testing data for the Indianapolis region during the first month of the pandemic (Figure 4). From April to early May (Figure 6), the MAPE is 6.0% for ICU census and 10.6% for M/S census. The MAPE from our later monitoring/updating process never exceeded 20%.

The predicted census numbers reported here come from the *integrated* model, which takes the disease forecast as an input and uses the calibrated patient flow equations to calculate census. We implement a multi-step calibration procedure with adaptive tuning in this integrated model. First, we use the realized (actual) arrivals to hospitals to tune the length-of-stay



(a) ICU census (b) M/S census

FIGURE 6 Patient census for the largest hospital in our partner health-care network: March 12, 2020 to May 6, 2020. In the plots, census data from the first 41 days (left of the vertical dashed line) are used as the training data, and the remaining 14 days (right of the vertical dashed line) as the testing data [Color figure can be viewed at wileyonlinelibrary.com]

and transfer parameters, that is, $\mu_j(t)$s and $p_{jj'}(t)$s. Second, we calibrate the disease forecast model using the adaptive tuning method specified in Section 4.4. Third, we connect the two components and tune the parameters that represent the fraction of COVID cases that arrive to an IU Health hospital, $p_j(t)$, with respect to the overall prediction performance for ICU and M/S census.

Implications for implementation and sustained use

In the health-care industry, analytics tools that produce inconsistent quality of output over time can quickly lead to mistrust and abandonment. Although the initial accuracy of our model was sufficient to merit undertaking an implementation, it is our model's *consistently* accurate predictions, driven by adaptive learning, that has been key for its continued use in practice.

To emphasize the importance of adaptive learning, we contrast our approach with the static, nonadaptive state model originally employed by hospitals at the outset of the pandemic. The state model performed reasonably well at IU Health until around March 19, 2020 (see Figure 4). However, the pandemic evolved over time in three ways that contributed to inconsistencies in the state model prediction: (1) disease progression of COVID in Indiana ceased resembling the flu-like spread that formed the basis of the state model; (2) as more IU Health-specific data on COVID patients became available, it became clear that COVID flow parameters at IU Health were *not* similar to national estimates on which the state model was built; and (3) as the pandemic progressed, demographics of hospitalized patients along with admission and treatment protocols changed, significantly impacting the patient-flow parameters. Further, the lack of an adaptive component in the state model appeared to have a compounding effect over time causing the prediction error to grow steadily as the model began exhibiting clear prediction bias. The MAPE of the state mode was 42.7% (ICU) and 33.8% (M/S) in March and grew to >50% for both units during April. This is in stark contrast with our adaptive learning methods designed to address problems (1)–(3) in both the patient flow and disease progression models (see Section 4.4.2) to maintain the consistent prediction accuracy highlighted above.

Remark 2. Non-COVID patients' arrival rates and flows are generally more stable than COVID patients. Thus, we follow the existing patient-flow literature to predict non-COVID workloads using admission, discharge, and transfer timestamps to estimate flow parameters. We estimate parameters for emergency and elective patients separately using historical data from the same time period in the last 2 years, accounting for seasonality and day-of-week effects. We model elective admissions as a modifiable parameter to allow IU Health to explore scenarios involving increasing or decreasing the volume of elective admissions.

4 | DISEASE PREDICTION

A critical component of modeling hospital resource needs during a pandemic lies in capturing arrivals of infected patients accurately. In contrast with retrospective epidemiological models that can be fit with sufficient historical data, "real-time" disease prediction in response to a burgeoning pandemic requires methods that can effectively utilize sparse and incomplete data. Further differentiating from other disease forecasting research, our disease forecast acts as one crucial component in the framework for modeling workloads at specific hospitals. That is, prediction of the COVID cases provides the input $\lambda^c(t)$ for the queueing network model presented in Section 3. As such, the metric by which we measure the value of the forecast model must differ from models that are designed to forecast longer term disease trends and/or impacts of public policy decisions. Specifically, our forecasting methods must be sufficiently accurate at predicting the number of new cases on a daily/weekly scale in a small geographic region (county level in our case).

Although there are many models capable of forecasting how a disease might spread in a population with limited historical data by leveraging measures such as mobility data, demographics, and policy levers, we found these models time-consuming to design, difficult to parameterize, and not as accurate in the short-term horizon targeted for operational-level decision making. Although there are many use cases for these types of models, they are not a good fit for our rapid deployment approach. Instead, we adopted a parsimonious design for quick turnaround and incorporated an adaptive learning component to handle sparse data and time-varying dynamics. In our approach, we utilize a blend of epidemiology and data-driven models to develop operational-level forecasts during the initial outbreak while allowing for longer term updating. Long-term forecasts of COVID progression allow for strategic planning and structural changes in health-care systems. Short-term forecasts support operational interventions that can be delivered immediately to improve response to surges in demand.

To position the structure of our forecasting methods within a conceptual framework, we split the phases of the pandemic into two regimes: (1) the *restricted data*, and (2) the *limited data* regimes. We make this distinction specifically because we develop different models for each regime. The data regimes are defined for a target geographical region; for example, the region(s) that the target hospital serves. In the restricted regime, there is little to no data in the target region, such as when the pandemic first presents in a region. We use Indiana in early March as an illustrative example. Indiana's first confirmed case was detected on March 6, 2020. Over the next several weeks, COVID initially spread slowly and testing was limited, providing scarce additional data points for estimating disease spread. For our purposes, we demarcate the boundary separating the restricted and limited data regimes by a threshold (lower-bound) on the number of confirmed cases in the target region, above (below) which the number of data points is sufficient (insufficient) to adequately parameterize a region-specific disease transmission model (e.g., SIR). When the first infection presents in a region, the region enters the restricted data regime. In this regime, we employ our forecasting model that leverages patterns from similar regions that have more data points at the current time period. When the number of confirmed cases surpasses the threshold and the region enters the limited data regime, we switch to our adaptive SIR model.

4.1 | Restricted data regime: Adapted synthetic control

At the onset of the pandemic, data on disease characteristics were limited both because of the novelty as well as bottlenecks in the flow of information. This limited the applicability of traditional epidemiology models, which depend on calibrating parameters tailored to the particular disease and population of interest. In this time frame, we implement the aSC methodology (C.J. Chen, 2020) using counties with earlier outbreaks and similar characteristics to estimate how the disease may spread in the target region.

4.2 | Limited data regime: Adaptive SIR

As more data became available, we transitioned to an adaptive SIR model based on calibrated disease characteristics. The SIR model has the capability of producing longer term forecasts based on structural models of disease transmission dynamics. The integrated learning component of our SIR model reacts to system changes detected from the data using statistical methods rather than trying to anticipate these changes from secondary data sources. This approach has the benefit of being able to infer the true impact of mobility, policy, and human behavior changes (among others) rather than relying on assumptions about the effect of these complex factors on disease spread. The drawback of this approach is a prediction lag between when a system change occurs and when it can be detected. For forecasting COVID cases, we found the prediction lag to be sufficiently short to maintain good accuracy both in predicting cases by region and as an input to our operational workload model.

4.3 | Short-term adaptive disease transmission model: Restricted data regime

For short-term (24–48 h) forecasts of the number of COVID cases, we adopt the aSC methodology of C.J. Chen (2020). The method extends the comparative case study approach based on the Synthetic Controls methodology (SC) (Abadie et al., 2010; Xu, 2017) to estimate the trajectory of COVID spread at the county level. Our aSC method implements a data-driven process for selecting comparable counties across the United States from the subset of counties that had earlier outbreaks of COVID than the county of interest. The disease spread trajectory of the selected comparable counties is then used to generate a forecast. Our method has the advantages of (i) explicitly accounting for observed heterogeneity and implicitly allowing for time-varying unobserved heterogeneity, (ii) generating forecasts at an actionable level for frontline managers, (iii) being interpretable and subject to verification, and (iv) being extendable to forecasting relaxations of policy interventions.

Adapted synthetic control overview

SC formulates a credible counterfactual to the treated unit by taking a weighted combination among the pool of untreated (nonfocal) units that minimizes the differences between the treated and synthetic with respect to both the outcome variable and observed covariates during the pretreatment period. This methodology was first introduced in Abadie and Gardeazabal (2003) and formalized in Abadie et al. (2010). It is often implemented when there is no intuitive method for

identifying a single suitable counterfactual within the pool or when simpler methodologies, such as averaging across all nonfocal units, do not generate credible counterfactual paths from which to estimate the difference-in-differences effect.

aSC uses the pool of U.S. counties that reported COVID cases *before* a focal county in order to construct a synthetic county that is comparable to the focal county along observable dimensions, but exceeded the minimum case threshold at an earlier date. The threshold is chosen ex ante to capture community spread of the disease as opposed to disease being brought in by visitors. The focal and control counties are synchronized based on the respective dates of this threshold. The methodology minimizes the difference between the actual and synthetic county with regard to the number of COVID cases since the synchronizing event, as well as a vector of county characteristics that include population demographics, urban characteristics, and policy interventions, specifically the issuance of a shelter-in-place order.

Implementation and results

We generate synthetic controls for five hospital regions defined by IU Health at the start of the pandemic. The hospital regions are composed of between 1 and 12 counties and represent the catchment areas of individual hospitals within the system. We aggregate county data to form the regions and implement aSC at the regional level. County demographics and characteristics are drawn from the Health Resources and Services Administration's Area Health Resource Files[1] and include total population, population of males, median age, population that is 65 or older, the percentage in poverty, number of 16+ workers who take public transportation, the number of 16+ workers, and population density per square mile. Additionally, we incorporate the institution of shelter-in-place orders. We set the minimum case threshold to 10 confirmed cases.

Across four of the five regions, our forecasts resulted in an average percentage error of 2.2% for forecasts up to 48 h out. However, one region had an error of 72%. In that health region, aSC failed to find a comparable synthetic counterpart as evidenced by the large error in the in-sample fit. This may be due to batch reporting of cases in this region that led to jumps of greater than 50% in a single day. Excluding this abnormal region, this method outperforms benchmarks using simple and exponential moving averages as well as regression approaches that account for heterogeneity and seasonality.

4.4 | Longer term adaptive disease transmission model: Limited data regime

Although short-term forecasts are valuable for managing day-to-day operations, longer term forecasts allow for structural changes that can drive much larger efficiency gains. Changes that modify capacity, such as unit conversions, transshipment of equipment, and redeployment of medical personnel require a longer range forecast (e.g., 1–2 weeks). Given the lack of long-term data due to the novelty of the pandemic, we leverage the epidemiology literature and adapt the SIR modeling framework in making forward forecasts.

The SIR model separates the population into susceptible (S), infected (I), and recovered (R) compartments. Many COVID workload models further divide I into severity buckets; however, such mapping is unnecessary for us as our workload model (Section 3) already accounts for resource demand in a more detailed manner. Although the number of individuals in each compartment is continuously evolving, we discretize it at a daily level. This is consistent not only with case data, which is updated daily, but also with operational decisions (e.g., workforce allocation) that are also most commonly made daily. Individuals move from S to I and then to R, where R captures both recoveries and death. Consistent with other models, we assume recovery removes the individual from the susceptible population.

Adaptive parameter selection

After initial model selection, the next step involves selecting a set of parameters to include in our adaptive learning method. The choice of parameters is particularly important because (1) all the parameters are continuous, which can present computational challenges in the learning step, and (2) attempting to learn too many parameters may compound any identification issues and cause poor prediction performance. Using a combination of intuition and experimentation, we selected two unknown parameters for adaptive learning: the reproductive number, R_0, that determines the rate of disease transmission in a population and the testing fraction r that estimates what proportion of infected individuals are confirmed by a COVID test.

R_0 was selected as a means of summarizing the myriad of complex factors that affect disease transmission within a population. Although there are a number of known external factors that influence disease transmission, such as public policy, mobility, and human behavior, our aim is to predict new cases, and thus we are agnostic to the cause of a change in disease transmission. For parsimonious model development, R_0 provides an attractive alternative to modeling the complexities of human society by focusing on learning the equilibrium outcome of all the factors, summarized with a single learning primitive. As mentioned previously, the concession of this simpler model is that changes in the underlying disease dynamics will be captured with some lag. We compensate for this by developing an adaptive algorithm that estimates R_0 as a piecewise discrete function with increasing weights on newer data to rapidly incorporate underlying changes. Additionally, we perturb the estimates in simulations to formulate confidence intervals. This serves to support operational decision-making through point forecasts with quantifiable prediction error.

Algorithm specification

At a high level, our algorithm calculates the mean percentage error of the predicted number of cases using the SIR model over a set of potential R_0 values. We estimate a distribution for the underlying R_0 for a time interval by taking a weighted mean and standard deviation across the set of R_0 values, where the weights are the inverse of the prediction errors. We then sample from the estimated distribution and simulate disease progression for future dates to yield forecasted point estimates and confidence intervals.

The evolution of COVID transmission under the SIR model is defined by the following set of ordinary differential equations:

$$\frac{dS(t)}{dt} = -\frac{\beta S(t)I(t)}{N}, \tag{23}$$

$$\frac{dI(t)}{dt} = \frac{\beta S(t)I(t)}{N} - \gamma I(t), \tag{24}$$

$$\frac{dR(t)}{dt} = \gamma I(t), \tag{25}$$

where $N = S(t) + I(t) + R(t)$ is the total population, and the parameters β and γ capture the transmission and recovery rates, respectively. R_0 is a function of these two parameters, that is, $R_0 = \beta/\gamma$. Next, we specify the adaptive algorithm to learn the R_0 values from historical data and then present the forecasting algorithm to generate point estimates and confidence intervals based on the learned R_0 values.

Adaptively learning R_0

Let

$$Y(t) = (S(t), I(t), R(t)) \tag{26}$$

be the state of the SIR model at time t. We consider a rolling time interval of length k-days that starts at each $t \in \{t_0, \dots, T\}$ in our training data, where t_0 is the first day that the number of confirmed cases exceeds some threshold (e.g., 100 cases) to allow for sufficient data to fit the SIR model, and T is the last day in the training data. Denote $\mathcal{I}$ as the set of intervals contained in the training data. Within time interval $i \in \mathcal{I}$ that consists of days $\{t_i, \dots, t_i + k\}$, we solve the system of Equations (23)–(25) for a fixed γ and for each $R_0 \in [R_0^{\min}, R_0^{\max}]$. We then set the $R_0^{(i)}$ that gives the minimum relative mean squared error over interval i as the best-fit R_0 for this interval. As each day t will be covered by k overlapping intervals (as we use the rolling intervals with 1 day increments), we take an average of the k best R_0 values from the k intervals to determine the final $R_0(t)$ for day t. We denote the set of these k intervals covering t as $\mathcal{I}_t$. Algorithm 1 details this algorithm.

ALGORITHM 1 Adaptive SIR algorithm

Result: Fitted R_0 for each day $t \in \{t_0, \dots, T\}$
Define $P_0 = \{R_0^1, \dots, R_0^J\}$, a set of R_0's from the range $[R_0^{\min}, R_0^{\max}]$;
Set initial condition $Y(0)$;
for *segment* $i \in \mathcal{I}$ **do**
  for $R_0^j \in P_0$ **do**
    for $t \in \{t_i, \dots, t_i + k\}$ **do**
      Solve (23)-(25);
      Calculate $\hat{\lambda}(t)$ from (27);
    end
    Calculate RMSE_i^j from (28);
  end
  Set $R_0^{(i)} = \arg\min_{R_0^j} \text{RMSE}_i^j$;
  Update $Y(t_{i+1})$ by solving (23)-(25) with $R_0^{(i)}$;
end
Output: $R_0(t) = \frac{1}{k} \sum_{i \in \mathcal{I}_t} R_0^{(i)}$ for each $t \in \{t_0, \dots, T\}$

The number of infected individuals is captured by $I(t)$. Note that not all infected individuals are tested and confirmed due to limitations in testing capacity as well as false-negative results, so $I(t)$ is not the same as the number of reported (confirmed) cases. To account for this, we map the number of infected individuals to the number of predicted reported cases using

$$\hat{\lambda}(t) = r \cdot \Delta I(t), \tag{27}$$

where $\Delta I(t) = S(t-1) - S(t)$ is the change in the noninfected compartment, that is, the number of newly infected individuals at t, and r is the testing fraction. The testing fraction is the joint probability of an individual receiving a test and testing positive when infected.

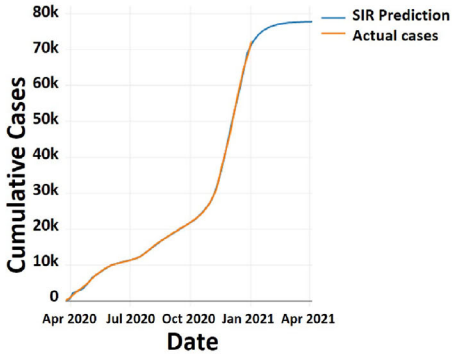
Let $\hat{\Lambda}^j = \{\hat{\lambda}^j(t_i), \dots, \hat{\lambda}^j(t_i + k)\}$ be the predicted cumulative confirmed cases over interval i under reproductive number R_0^j, and let $\Lambda = \{\lambda(t_i), \dots, \lambda(t_i + k)\}$ be the reported observed cases. We define the relative mean squared error over the interval i for R_0^j as

$$\text{RMSE}_i^j = \sum_{t=t_i}^{t_i+k} \left(\frac{\hat{\lambda}^j(t) - \lambda(t))}{\lambda(t)} \right)^2. \tag{28}$$

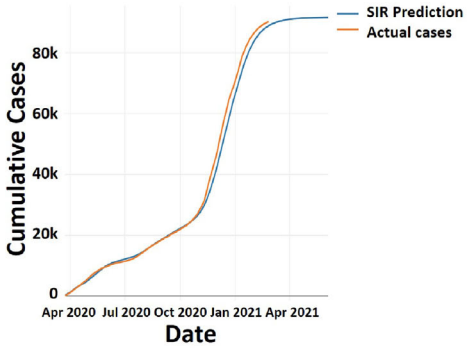
We elect to use the square of the relative errors rather than the absolute difference in order to emphasize limiting the variance of our estimates, which should improve the usefulness of the predictions. We update the parameters and generate forecasts using Algorithm 1.

Forecasting new cases

We leverage the $\{R_0(t) : t = 1, \dots, T\}$ learned from Algorithm 1 to forecast future disease progression for



(a) SIR forecast of cumulative cases: Indianapolis



(b) SIR forecast of cumulative cases: Indianapolis Suburbs

FIGURE 7 Prediction of the census from our application [Color figure can be viewed at wileyonlinelibrary.com]

$t \in [T+1, T+k]$. There are different alternatives to generating the point estimates. One method is to use $\hat{R}_0 = R_0(T)$, that is, the R_0 from the most recent segment, to solve Equations (23)–(25) to generate the predicted $\hat{\lambda}(t)$ for $t \in [T+1, T+k]$. An alternative is to take an average of $R_0(t)$ for $t \in [T-\ell, T]$ for some ℓ days to generate the predicted cases. In our implementation, we use the first method for its simplicity and interpretability.

To generate the confidence intervals, we adopt a Monte Carlo approach. We first estimate the sample standard deviation $\hat{\sigma}$ in R_0 based on the learned values $\{R_0(t)\}$. Then we run M replications using Monte Carlo simulation for future cases. In each replication, we draw a sample R_0 from the $\mathcal{N}(\hat{R}_0, \hat{\sigma}^2)$ distribution. We then simulate COVID spread for the sampled R_0 for the $t \in [T+1, T+k]$. Finally, we use the M simulated paths to generate confidence intervals for the predicted new COVID cases.

Forecasting performance

We implement our SIR prediction algorithm using data from March through June on a 14-day rolling horizon ($k = 14$) with a fixed γ of 0.1 (i.e., 10 day recovery) and a range of $R_0 \in [0.5, 6.0]$ in increments of 0.05. The window size was chosen based on ranges from epidemiology studies on recovery days and other forecasts (Deasy et al., 2020; Fantazzini, 2020). We fixed γ according to the midpoint of CDC guidelines on post-symptom quarantine (Centers for Disease Control, 2020) and selected a wide range of R_0 that covers the minimum and maximum estimates found in the literature. After some experimentation tuning both R_0 and the testing fraction, r, we found that a fixed testing fraction of 15% yielded stable and accurate performance, allowing us to further pare down our learning parameters to include only R_0. Note that 15% implies that true infections are six times the number of reported infections, which falls within estimated ranges (typically 4–10 times), though the literature is sparse (Aizenman et al., 2021; McCulloh et al., 2020). We test our predictions across the same five hospital regions defined by IU Health with updating; that is, we update the SIR compartments as new data become available in the rolling horizon.

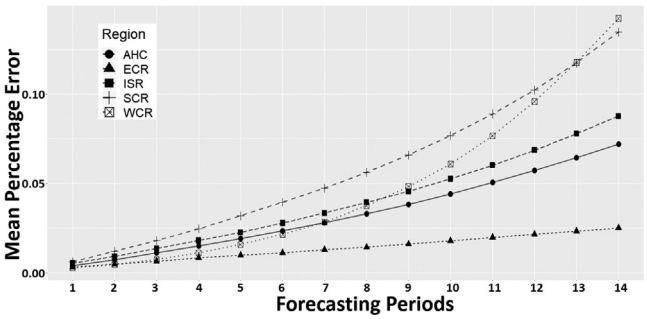


FIGURE 8 SIR forecast mean percentage error by number of predicted days

Using the adaptive SIR model, we find the R_0 that minimizes our in-sample error (shown in Figures 7a and b), and we use it to generate our forecasts. When evaluating our predictions of total cases, we find an out-of-sample MAPE of 3.5% on average with a range of 1.3–5.7% across the regions for all periods. Errors range from 0.4% when forecasting 1 day ahead to 9.0% when looking 14 days out. The breakdown by region is shown in Figure 8, which suggests strong performance across all regions but with notable heterogeneity. By comparison, a static SIR model where R_0 is fixed at the value that minimizes the average error across all data before the first adaptive segment has an average percentage error of 18.8% with a range of 2–38% looking 1–14 days out. Examining specific time frames, the first adaptive segment had an average forecast error of 8.2% while the static forecast error over the same period was 11.8%. The relative outperformance of adaptive SIR increases over the remaining segments.

5 | USE CASES FOR THE WORKLOAD PREDICTION MODEL

The primary product of our multi-method framework is the prediction of the workload distribution for target hospitals. To

provide hospitals with additional information about critical resources needed for responding to the pandemic, we generate secondary outputs, including resource utilization, staffing requirements, ventilators, and personal protective equipment. These resource utilization predictions can be derived directly from the detailed patient demand prediction by estimating a resource usage parameter, α_j^i, for each type of patient (i) and location (j) either directly from the data or using our adaptive tuning methods. As one example from our data (for one hospital during one period of time), 52% of ICU COVID patients ($\alpha^c_{\text{ICU}} = 0.52$) and 2.5% of non-ICU COVID patients ($\alpha^c_{\text{MS}} = 0.25$) required a ventilator. Hence, a point estimate of COVID ventilator needs could be calculated as $x^c_v(t) = \alpha^c_{\text{ICU}} \cdot x^c_{\text{ICU}}(t) + \alpha^c_{\text{MS}} \cdot x^c_{\text{MS}}(t)$, where $x(\cdot)$ is the mean of the workload process $X(\cdot)$ as discussed in Section 3. The distributional estimate of ventilator usage would be the convolution of binomial random variables representing the number of patients of each type/unit pair requiring a ventilator. Let $X^c_{v,j}(t) \sim \text{Bin}(X^c_j(t), \alpha^c_j)$ be the random variable for the number of COVID patients requiring a ventilator in unit $j \in \{ICU, MS\}$ at time t. Then $X^c_v(t) = X^c_{v,ICU}(t) + X^c_{v,MS}(t)$. A similar calculation can be done for ventilators for non-COVID patients, for PPE, etc.

These predictions have been used by IU Health to inform both tactical and operational decisions. To provide additional support for these decisions, we are able to build decision models on top of our prediction, which is facilitated by the parsimonious design. Examples include timing and volume of advanced purchases of nasal canula, PPE, pharmaceuticals, additional diagnostic equipment, CPAP; proactive hiring of temporary agency nurses; optimized staff planning; and flexible resource re/allocation (the rightmost panel in Figure 2); among others. The resource utilization forecast has also been used to support what-if analysis for intervention strategies, surge capacity plans, and logistics planning decisions among others. In this section, we provide details of the decision support systems developed for two main use cases that were requested by IU Health. These use cases demonstrate that our final forecast model is capable of achieving the overarching goal of this research: to provide actionable information to support hospital decision making during a pandemic.

5.1 | Leveraging health system scale for resource transshipment

One of the innovations IU Health implemented during COVID was initiating a program to transfer nurses between hospitals and even between regions, the expansion of which has become a strategic goal for 2021. Implementing such a significant change necessitated data-driven support to justify both the direct cost and change management efforts. We worked closely with IU Health's nursing organization to provide this support by building a nurse transshipment model that delivers both tactical and operational plans. All design specifications were chosen or vetted by IU Health, considering feasibility of implementation and value to the organization, staff, and patients. The complexity of the decision process and careful consideration given by IU Health leadership is reflected in the diverse output metrics present subsequently. Each of the metrics was used during the decision-making process and taken together, the metrics reflect the different perspectives of individual hospitals (source or destination), the overall hospital network, and the nurses.

Structure of nurse transshipment program

The following features provide a high-level overview of the nurse transshipment program. We specify the model formulation in Appendix A.2 (Supporting Information). This optimization takes the (stochastic) workload forecast as an input to generate a 2-week transshipment schedule 1 week in advance. The schedule represents a tentative relocation plan so nurses can have some advance warning if they may need to travel. Nurses receive additional pay plus a relocation bonus when asked to travel. Depending on the distance a nurse travels, there is a minimum number of days they must work at the alternate location, which we call secondment, to avoid an excessive (and possibly dangerous) commute in addition to a typical 12-h shift. When the demand in each hospital/region is realized, IU Health has the option to either cancel a transshipment (call-back), or enact an additional "emergency" transshipment at a premium cost. Both practices are currently being used on a smaller scale within the Indianapolis region, so there is precedent for these types of actions. For comparison, in Appendix A.7 (Supporting Information), we describe a similar ventilator transshipment model. There are two key differences between these transshipment models: (1) when a nurse is transferred, a transfer cost and a salary premium (for the duration of secondment) is paid, whereas only a transfer cost is incurred when a ventilator is borrowed; and (2) a nurse can be transferred only from her "home" location, whereas a ventilator can be borrowed from any location. In this sense, there is no expectation that the ventilators must be returned to their original location.

Transshipment model products

Next, we present some products of the decision model that are used in support of IU Health's data-driven transshipment program. Additional products are described in Appendix A.5 (Supporting Information). The results below are generated by solving the transshipment model using our workload forecast for three different scenarios that are based on different demand patterns observed during the pandemic: unbalanced (e.g., March/April), flat (e.g., May), and increasing (e.g., September). The actual numbers are modified to protect IU Health data, but the underlying results are similar in nature. The model inputs, secondment, additional pay, transfer bonuses, emergency premiums, understaffing costs, and capacities were provided by IU Health based on

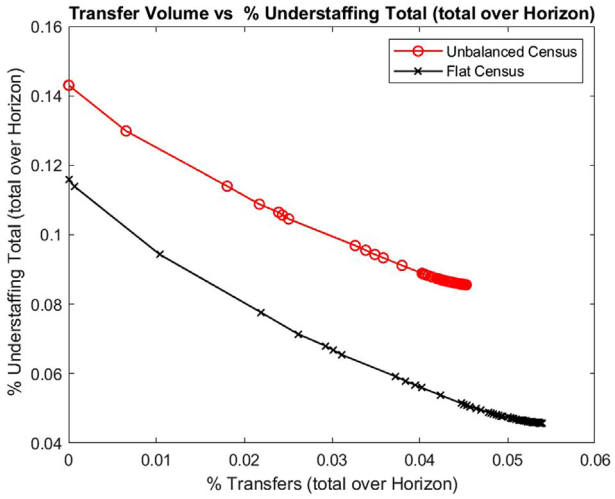


FIGURE 9 Percentage of transfers versus understaffing in the unbalanced and flat census scenarios [Color figure can be viewed at wileyonlinelibrary.com]

surveys soliciting nurse input combined with nurse management strategy sessions. The call-back penalty is set to 0.05 to avoid excessive ex ante transshipments while also capturing the reality that it is desirable to keep a nurse in their "home" region. Two hundred stochastic scenarios are generated from the distributional workload prediction model. All input parameters are summarized in Appendix A.4 (Supporting Information).

Tactical plan

To support IU Health's tactical decision regarding how many transfers they are willing to accept for subsequent improvements in system-wide understaffing, we generate a Pareto curve showing the percentage of transfers (number of transfer shifts divided by total shifts over the 14-day horizon) versus the understaffing percentage (in terms of shifts required vs. shifts worked) across all five regions; see Figure 9. These products allow IU Health to choose a particular point on the curve based on nursing management goals. This tactical selection implies a set of cost parameters that are subsequently used to generate daily operational decisions.

Operational plan

To frame the operational plan based on the chosen tactical plan, we begin with our "bird's eye view" of the impact of the transshipment plan on the five regions in Figure 10 for the unbalanced scenario. The dotted lines and solid lines are the utilization with and without transshipment, respectively, plotted on the left y-axis. For reference, we also plot the expected census (over the 200 scenarios) below the utilization curves, scaled on the right y-axis. These figures provide an easily interpreted, informative visual for IU Health nursing management, who are intimately familiar with the staffing situations in each region. Even at a quick glance, it is clear that Regions 1 and 3—Indianapolis and Indianapolis suburbs—experience significant capacity issues when no transshipments are executed (solid line), while Regions 2 and 4—East Central (Muncie) and South Central (Bloomington)—have some spare capacity. Region 5—West Central (West Lafayette)—progresses from overloaded to underloaded back to overloaded. These unbalanced scenarios are not uncommon given that these five regions serve different communities and span over 150 miles north to south and the entire width of the state. From this figure, the impact of transshipment is clear and provides a strong argument when lobbying for expansion of the transshipment plan, which has since been successfully accomplished. See Appendix A.5 in the Supporting Information for results in other scenarios.

5.2 | Preparation for subsequent waves of the pandemic

Due to COVID, combined with the upcoming flu season, IU Health planned to hire travel nurses to supplement their existing workforce. To identify the maximum number of nurses to cover system-wide need during the highest projected future surge, we designed a regional Newsvendor model (formulated in Appendix A.8, Supporting Information), in line with our rapid deployment approach favoring the simplest valuable solution. The newsvendor takes our forecast as demand and estimates the daily number of nurses needed over the next several months in each region. This model was integrated into IU Health's staffing matrix to project gaps between available and desired staff during different points in the upcoming flu/COVID season from September to February.

As nasal canulas emerged as an effective mechanism for treating COVID, a similar approach was requested to analyze the timing and volume of an advanced purchase of this equipment to meet anticipated need. This was also to avoid waiting and delayed shipments as COVID-relevant medical equipment was in high demand while, at the same time, supply chains had been disrupted causing significant capacity and transportation issues.

6 | CONCLUSION AND FUTURE WORK

In this work, we develop and implement a multi-method model that integrates disease progression and operational workload models in collaboration with the largest health-care network in Indiana for their COVID response and surge planning. We highlight our parsimonious design philosophy in which we trade off model complexity for adaptability, computational efficiency, and speed of development, validation, and implementation. Our parsimonious approach allows us to address key challenges in pandemic modeling: severely limited data, partially observable outcomes of interest, and largely time-varying system dynamics that must be regularly relearned.

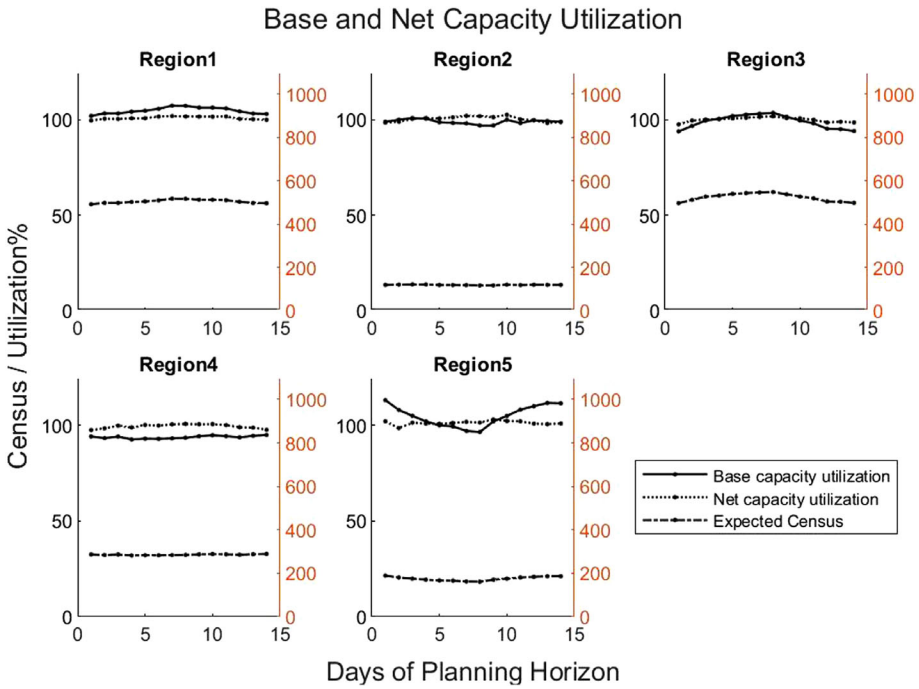


FIGURE 10 System-wide impact of transshipment plan on regional staffing utilization. *Left y-axis:* Capacity utilization. *Right y-axis:* Region census [Color figure can be viewed at wileyonlinelibrary.com]

Via multiple layers of adaptive learning, our tools can adapt to rapidly evolving information during different stages of a pandemic, providing a tailored, consistently accurate prediction for individual hospitals, with a MAPE of less than 10%, while the model used by the state of Indiana exhibited a MAPE of over 30% over the same time frame. Our workload prediction provides the basis for forecasting demand for resources and is detailed enough to support operational decision optimization at the hospital level, in contrast to other existing COVID forecasting models. Facilitated by the simplicity of our prediction model, we were subsequently able to build a suite of decision support optimization tools that have been used to support nurse hiring and nurse transshipment decisions.

In our ongoing collaboration, we are working to provide workload census prediction on an hourly basis for individual units in a hospital, which incorporates a personalized projection of time-dependent resource usage for patients currently in the hospital and predicted elective patient arrivals based on the surgical schedule for future days, among other details. This refined census prediction will be used to support the nurse staffing plan at IU Health, including proactive hiring of contract nurses and allocation of float nurses. According to IU Health, this directly impacts more than 400 nurses in their float pool and indirectly impacts all 9000 nurses in the system.

ORCID

Pengyi Shi https://orcid.org/0000-0003-0905-7858
Jonathan E. Helm https://orcid.org/0000-0001-5577-5530
Rodney P. Parker https://orcid.org/0000-0001-7688-5122

ENDNOTE

[1] https://data.hrsa.gov/data/download

REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505.

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review*, *93*(1), 113–132.

Aizenman, N., Carlsen, A., & Talbot, R. (2021). *Why the pandemic is 10 times worse than you think*. https://www.npr.org/sections/health-shots/2021/02/06/964527835/why-the-pandemic-is-10-times-worse-than-you-think

Bartz-Beielstein, T., Rehbach, F., Mersmann, O. & Bartz, E. (2021). Optimization and Adaptation of a Resource Planning Tool for Hospitals Under Special Consideration of the COVID-19 Pandemic. 2021 IEEE Congress on Evolutionary Computation (CEC), pp. 728-735.

Bekker, R., Koole, G., Broek, M. (2021). Modeling COVID-19 hospital admissions and occupancy in the Netherlands. Preprint, arXiv:2102.11021.

Bertsimas, D., Boussioux, L., Cory-Wright, R., Delarue, A., Digalakis, V., Jacquillat, A., Kitane, D. L., Lukin, G., Li, M., Mingardi, L., Nohadani, O., Orfanoudaki, A., Papalexopoulos, T., Paskov, I., Pauphilet, T., Lami, O., Stellato, B., Bouardi, H., … Zeng, C. (2021). From predictions to prescriptions: A data-driven response to COVID-19. *Health Care Management Science*, *24*, 253–272.

Betcheva, L., Erhun, F., Feylessoufi, A., Gonçalves, P., Jiang, H., Kattuman, P., Pape, T., Pari, A., Scholtes, S., & Tyrrell, C. (2020). Rapid COVID-19 modeling support for regional health systems in England. https://doi.org/10.2139/ssrn.3695258

Brauer, F., & Castillo-Chavez, C. (2012). *Mathematical models in population biology and epidemiology*. Springer.

Capistran, M. A., Capella, A., & Christen, J. A. (2020). Forecasting hospital demand during COVID-19 pandemic outbreaks. Preprint, arXiv:2006.01873.

Centers for Disease Control (2020). *Options to reduce quarantine for contacts of persons with SARS-CoV-2 infection using symptom monitoring and diagnostic testing*. https://www.cdc.gov/coronavirus/2019-ncov/more/scientific-brief-options-to-reduce-quarantine

Chen, C. J. (2020). *Forecasting infection progression: Adapting synthetic controls to predicting COVID-19 spread*. Working Paper.

Chen, N., Hu, M., & Zhang, C. (2020a). Capacitated SIR model with an application to COVID-19. https://ssrn.com/abstract=3692751

Chen, Y.-C., Lu, P.-E., & Chang, C.-S. (2020b). A time-dependent SIR model for COVID-19 with undetectable infected persons. *IEEE Transactions on Network Science and Engineering*, *7*(4), 3279–3294.

Cuevas, E. (2020). An agent-based model to evaluate the COVID-19 transmission risks in facilities. *Computers in Biology and Medicine*, *121*, 103827.

Dai, J., & Shi, P. (2017). A two-time-scale approach to time-varying queues in hospital inpatient flow management. *Operations Research*, *65*(2), 514–536.

Deasy, J., Rocheteau, E., Kohler, K., Stubbs, D. J., Barbiero, P., Liò, P., & Ercole, A. (2020). Forecasting ultra-early intensive care strain from COVID-19 in England. *medRXiv* https://doi.org/10.1101/2020.03.19.20039057

Diekmann, O., Heeserbeek, H., & Britton, T. (2013). *Mathematical tools for understanding infectious disease dynamics*. Princeton University Press.

Dos Santos, I., Almeida, G., & De Moura, F. (2021). Evolution of SARS-CoV-2 in the state of Alagoas-Brazil via an adaptive SIR model. *International Journal of Modern Physics C* (*IJMPC*), *32*(03), 1–6.

Fantazzini, D. (2020). Short-term forecasting of the COVID-19 pandemic using google trends data: Evidence from 158 countries. *Applied Econometrics*, *59*, 33–54.

Feng, J., & Shi, P. (2018). Steady-state diffusion approximations for discrete-time queue in hospital inpatient flow management. *Naval Research Logistics* (*NRL*), *65*(1), 26–65.

Garrido, J. M., Martinez-Rodriguez, D., Rodriguez-Serrano, F., Perez-Villares, J. M., Ferreiro-Marzal, A., del Mar Jimenez-Quintana, M., Grupo de Estudio COVID-19 Granada, & Villanueva, R. J. (2020). Mathematical model optimized for prediction and health care planning for COVID-19. Preprint, arXiv:2012.05804.

Green, L. V., Kolesar, P. J., & Whitt, W. (2007). Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, *16*(1), 13–39.

Grimm, C. A. (2020). *Hospital experiences responding to the COVID-19 pandemic: Results of a national pulse survey March* 23–27, 2020. US Department of Health and Human Services OoIG.

Hill, A., Levy, M., Xie, S., Sheen, J., Shinnick, J., Gheorghe, A., & Rehmann, C. (2020). *Modeling COVID-19 spread vs healthcare capacity*. https://alhill.shinyapps.io/COVID19seir/

Kaplan, E. H. (2020). OM Forum—COVID-19 scratch models to support local decisions. *Manufacturing & Service Operations Management*, *22*(4), 645–655.

Kerr, C. C., Stuart, R. M., Mistry, D., Abeysuriya, R. G., Rosenfeld, K., Hart, G. R., Núñez, R. C., Cohen, J. A., Selvaraj, P., Hagedorn, B., George, L., Jastrzębski, M., Izzo, A., Fowler, G., Palmer, A., Delport, D., Scott, N., Kelly, S., Bennette, C., … Klein, D. (2021). Covasim: An agent-based model of COVID-19 dynamics and interventions. *PLOS Computational Biology*, *17*(7), e1009149.

Malani, A., Soman, S., Asher, S., Novosad, P. M., Imbert, C., Tandel, V., Agarwal, A., Alomar, A., Sarker, A., Shah, D., Shen, D., Gruber, J., Sachdeva, S., Kaiser, D., & Bettencourt, L. (2020). *Adaptive control of COVID-19 outbreaks in India: Local, gradual, and trigger-based exit paths from lockdown*. NBER Working Paper No. w27532.

Mamon, G. A. (2020). Regional analysis of COVID-19 in France from fit of hospital data with different evolutionary models. Preprint, arXiv:2005.06552.

McCulloh, I., Kiernan, K., & Kent, T. (2020). Inferring true COVID-19 infection rates from deaths. *Frontiers in Big Data*, *3*, 37.

Shapiro, M., Karim, F., Muscioni, G., & Augustine, A. (2021). Are we there yet? Adaptive SIR model for continuous estimation of COVID-19 infection rate and reproduction number in the United States. *Journal of Medical Internet Research*, *23*(4), e24389.

Silva, P. C., Batista, P. V., Lima, H. S., Alves, M. A., Guimarães, F. G., & Silva, R. C. (2020). COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals*, *139*, 110088.

Veloz, T., Maldonado, P., Ropert, S., Ravello, C., Mora, S., Barrios, A., Villaseca, T., Valdenegro, C., & Perez-Acle, T. (2020). On the interplay between mobility and hospitalization capacity during the COVID-19 pandemic: The SEIRHUD model. Preprint, arXiv:2006.05357.

Whitt, W., & Zhao, J. (2017). Many-server loss models with non-Poisson time-varying arrivals. *Naval Research Logistics*, *64*(3), 177–202.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, *25*(1), 57–76.

Zhang, T., McFarlane, K., Vallon, J., Yang, L., Xie, J., Blanchet, J., Glynn, P., Staudenmayer, K., Schulman, K., & Scheinker, D. (2020). A model to estimate bed demand for COVID-19 related hospitalization. *medRXiv* https://doi.org/10.1101/2020.03.24.20042762

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Shi, P., Helm, J.E., Chen, C., Lim, J., Parker, R.P., Tinsley, T., & Cecil, J. (2022). Operations (management) warp speed: Rapid deployment of hospital-focused predictive/prescriptive analytics for the COVID-19 pandemic. *Production and Operations Management*, 1–20. https://doi.org/10.1111/poms.13648