This is the published version of the paper.

This version of the publication may differ from the final published version.

POMS

ORIGINAL ARTICLE

ICU capacity expansion under uncertainty in the early stages of a pandemic

Anna Maria Gambaro[1] | Gianluca Fusai[1,2] | ManMohan S. Sodhi[2] | Caterina May[1] | Chiara Morelli[1]

[1]Dipartimento di Studi per l'Economia e l'Impresa, Università del Piemonte Orientale, Novara, Italy

[2]Bayes Business School, City University of London, London, UK

Correspondence
ManMohan S. Sodhi, Bayes Business School, City University of London, London EC1Y 8TZ, UK.
Email: m.sodhi@city.ac.uk

Handling Editor: Sushil Gupta

Abstract

We propose a general modular approach to support decision-makers' response in the early stages of a pandemic with resource expansion, motivated by the shortage of Covid-19-related intensive care units (ICU) capacity in 2020 in Italy. Our approach uses (1) a stochastic extension of an epidemic model for scenarios of projected infections, (2) a capacity load model to translate infections into scenarios of demand for the resources of interest, and (3) an optimization model to allocate this demand to the projected levels of resources based on different values of investment. We demonstrate this approach with the onset of the first and second Covid-19 waves in three Italian regions, using the data available at that time. For epidemic modeling, we used a parsimonious stochastic susceptible-infected-removed model with a robust estimation procedure based on bootstrap resampling, suitable for a noisy and data-limited environment. For capacity loading, we used a Cox queuing model to translate the projected infections into demand for ICU, using stochastic intensity to capture the variability of the patient arrival process. Finally, we used stochastic dynamic optimization to select the best policy (when and how much to expand) to minimize the expected number of patients denied ICU for any level of investment in capacity expansion and obtain an efficient frontier. The frontier allows a trade-off between investment in additional resources and the number of patients denied intensive care. Moreover, in the panic-driven early days of a pandemic, decision-makers can also obtain the time until which they can postpone action, potentially reducing investment costs without increasing the expected number of denied patients.

KEYWORDS
capacity expansion, Covid-19, disaster response, ICU, Italy, pandemic modeling

1 | INTRODUCTION

By the end of 2021, Covid-19 had already taken an estimated 15 million lives as estimated by the WHO, in part due to a shortage of intensive care units (ICUs) and other resources. *We provide a general approach to determine the optimal timing and extent of resource expansion, say, ICUs, in the early stages of a pandemic* with budget constraints and the high uncertainty in the growth of infected people. Furthermore, we

apply it retrospectively for the early stages of the first and second waves of Covid-19 in three regions in Italy that suffered high fatalities.

We frame the problem as that of capacity expansion under demand uncertainty, an established topic in the operations literature (Luss, 1982; Van-Mieghem, 2003). In healthcare, capacity management involves decisions related to the allocation of such critical resources as facilities, equipment, and personnel (Smith-Daniels et al., 1988) that can have life-or-death implications. Moreover, in a pandemic, we must plan for peak demand to avoid waiting lists,

Accepted by Sushil Gupta, after three revisions.

even though the demand surge would eventually subside. Finally, we focus on the *supply-side* response to a pandemic by expanding resource capacity proactively, keeping demand-side responses such as lockdowns, quarantine, and vaccinations that "flatten the epidemic curve" outside the scope of this paper.

The modular approach we propose for resource capacity expansion allows alternative models to be used, depending on the pandemic or the type of resource. The three modules in our approach are (1) a stochastic extension of an epidemic model to create scenarios for the daily number of infections in the coming weeks, (2) a capacity load model to translate the projected infections to demand for the resource of interest, and (3) an optimization model to allocate the demand to the resource at different levels of capacity corresponding to different levels of investment.

We illustrate this approach by applying it retrospectively to the early days of the first and second Covid-19 waves in the three regions in Italy—Lombardia, Piemonte, and Veneto—that have different per capita ICU capacity, taking care to use only the information available at the time. First, we used a parsimonious *susceptible-infected-removed* (SIR) epidemic model, suitable for a noisy and data-limited environment, such as the onset of a pandemic. Second , we used a queuing model to translate the projected infections into demand for ICU. The variance of the number of arrivals in any interval here is larger than the mean in a pandemic, unlike a Poisson process. Finally, we used dynamic programming to allocate the demand to ICU capacity, which changes with different levels of investment. The decision-maker can see the *efficient frontier* of the expected number of patients denied intensive care with investment. We also obtain the optimal time to postpone action, reducing costs without increasing the expected number of denied patients.

In the rest of the paper, Section 2 discusses our contribution to the existing literature. Section 3 describes our modular approach. Sections 4–6.1 detail the three modules in turn, along with the application to the Italian regions for the first two waves of the pandemic. Section 7 concludes.

2 | CONTRIBUTION TO THE LITERATURE

In the literature on capacity expansion under uncertainty, the canonical objective is minimizing the investment while limiting the unmet demand, see Luss (1982) and Van-Mieghem (2003). We invert this perspective by minimizing the unsatisfied demand with a constrained budget. Additionally, we provide the *marginal cost* (shadow price) for admitted patients as an output instead of requiring it as an input to the model. Nor do we need an ex ante penalty for the trade-off between the cost of intervention and the value of a statistical life or a quality life year saved (Li et al., 2023), for example, $10 million in the United States (Kniesner & Viscusi, 2005).

We contribute to the pandemic-focused healthcare operations literature with the economics of capacity expansion,

enabling decision-makers with an approach that is more comprehensive and flexible compared to similar papers. Furthermore, each of the three modules introduces innovations in our retrospective application for Covid-19 in Italy. Important outputs of our approach for decision-makers are (1) the minimum (expected) number of patients denied use of ICU (or other limited resource), given the existing ICU capacity and investment for increasing it; (2) the budget increase needed to accommodate an additional patient, that is, the shadow price; (3) the time until which decision-makers can postpone action, reducing costs without increasing the expected number of denied patients; and (4) the ceiling beyond which further investment is useless.

2.1 | A modular approach to capacity expansion in a pandemic

Our modular approach supports decision-makers in the early stages of a pandemic with planned capacity expansion using (1) a stochastic epidemic model to project scenarios of the infection spread in the early days of a pandemic with high uncertainty, (2) a capacity loading system to translate infections into the need for the resources of interest, and (3) an optimization method to allocate the demand to the resources available at different levels of investment (Figure 1).

The operations literature on Covid-19 focuses either on containing the demand or increasing the supply of needed resources (Table 1). The *demand-focused* stream investigates "flattening the epidemic curve" (Evgeniou et al., 2022; Ferguson et al., 2020; Jain & Rayal, 2022; Perkins & Espana, 2020; Shahmanzari et al., 2022) containment measures such as lockdown or quarantine, given a fixed capacity for the ICU (or other pertinent resource). The *supply-focused* stream, in which our paper is positioned, focuses on capacity expansion of ICU (or other resource) (Alban et al., 2020; Lu et al., 2021; Ouyang et al., 2020; Shi et al., 2022; Wood et al., 2020). Unlike our use of three models—epidemic, demand, and optimization—many researchers have used just two of the modules. Moreover, they have not used stochasticity in their epidemic modeling. Some have used the Poisson process with deterministic intensity in the second module, which is not suitable for the early pandemic stages with data of questionable quality. Other researchers have not used optimization in the third module, limiting themselves to what-if analyses. Table 1 provides an overview; Sections 2.2–2.4 below provide with module-specific details.

Wood et al. (2020) and Gonçalves et al. (2022) straddle both literature in considering both containment measures and ICU (or hospital) capacity. However, they use only what–if analysis rather than optimization in the last stage. Gonçalves et al. (2022) propose a complex compartmental model that captures links between the different quantities (e.g., hospital [or ICU] loading influences the number of infected individuals). However, their model is deterministic and cannot capture the uncertainty in the future demand by simply choosing parameter values randomly chosen from arbitrary ranges. In
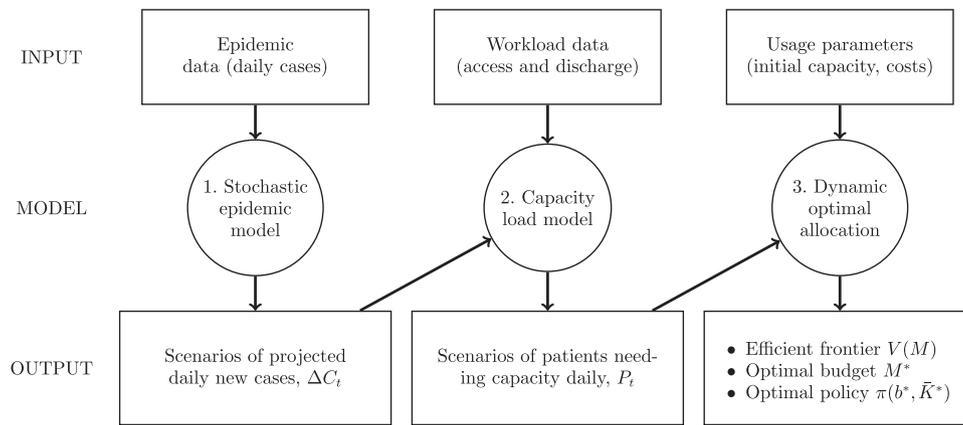
FIGURE 1 Our modular framework comprises three models: (1) The stochastic epidemic model produces infection scenarios, (2) the capacity load model converts these scenarios to the demand for resources, and (3) the optimal allocation model allocates the demand to the resources at different levels of investment.

contrast, for generality of use in future pandemics, we use a parsimonious compartmental epidemic model (SIR) and estimate the parameters for future demand uncertainty and its evolution from the reported infection data to date.

Our work also contributes to the literature specific to each module. We have (a) a robust estimation and forecast procedure for the epidemic model during the early days of a pandemic, (b) the use of a Cox stochastic process rather than, say, a nonhomogeneous Poisson process for the arrival of patients, and (c) a stochastic dynamic programming model for capacity expansion. Below, we discuss each of the three modules.

2.2 | Epidemic modeling

The epidemiology literature offers several models extending the baseline SIR epidemic model for the spread of infections. These are *compartmental* and *agent-based* models. Compartmental models extend the baseline SIR model by adding more "compartments" depicting the status of individuals at any time: susceptible, exposed, infectious, hospitalized, critical, other–recovered, released, or dead (see, for instance, Mamon, 2020). Agent-based models simulate the rule-based interactions of individuals to infer the system's behavior, thus extending compartmental models (Cuevas, 2020; Kerr et al., 2021). Besides agents, compartmental models can be extended with more states. For instance, Bertsimas et al. (2021) present differential equations leads to predictions of hospitalizations and infections, an susceptible-exposed-infectious-removed (SEIR)-based model with 11 states to include varying states of patient recovery, detection, and quarantine. However, such models require estimating many parameters for which data would be difficult to obtain at the start of a pandemic. We use the stochastic version of the SIR compartmental model that has low data requirements at the onset of a pandemic because it has very few parameters.

Other researchers use time-dependent parameters to extend compartmental models specific to a country (Calafiore et al., 2020; Y. Chen et al., 2020; Ferrari et al., 2021). Chatterjee et al. (2020) used a time-dependent infection rate to study the pandemic in India and other countries. A more sophisticated model is *adaptive* SIR (Shapiro et al., 2021; Shi et al., 2022) with a time-dependent reproduction number (and hence the infection rate). The daily value is estimated directly from data without assuming any particular functional form. However, this approach allows only short-term projection for 1 or 2 weeks to manage emergency day-to-day operations. This is because the reproduction number is taken as the last estimated value when projecting forward.

We use a parsimonious SIR model with a parametric time-dependent infection rate. In contrast to the above approaches, we have the flexibility to match past data and the ability to perform robust estimations during the early days of a pandemic when there is very little or noisy data. Like us, Chatterjee et al. (2020) also use a time-dependent infection rate and Shahmanzari et al. (2022) use a stochastic extension of an epidemic model. However, the novelty of our approach lies in (1) the robust estimation method going beyond the point estimate of model parameters by reconstructing the empirical joint distribution of the estimators and (2) creating scenarios of projected daily infections, incorporating the uncertainty in the estimated parameters in the simulated scenarios of infections. In contrast, the recent Covid-19-related literature (Table 1) typically adopts deterministic compartmental epidemiological models, possibly with simulation to vary parameters over an arbitrary range.

2.3 | Capacity demand modeling

The demand-focused literate does not adopt a specific model for ICU (or hospital) demand (Table 1). For instance, Ferguson et al. (2020) consider the number of hospitalized cases requiring critical care simply as a fixed percentage of symptomatic patients. In the supply-focused literature, demand is modeled with *compartmental* or *queuing* models. Z. Chen and Kong (2022) and Gonçalves et al. (2022) propose deterministic compartmental models, which include hospitalized and critically ill patients.

TABLE 1 Covid-19 motivated papers in the OM literature.

Reference	Decision focus	Supply focus	ICU demand model	Stochastic demand	Epidemic model	Stochastic epidemic	Optimization	Real world data	Case study
Ferguson et al. (2020)	Containment	-	-	-	✓	-	-	✓	✓
Perkins and Espana (2020)	Containment	-	-	-	✓	-	✓	✓	✓
Jain and Rayal (2022)	Containment	-	-	-	-	✓	✓	-	-
Evgeniou et al. (2022)	Containment	-	-	-	✓	✓	✓	✓	✓
Shahmanzari et al. (2022)	Containment	-	-	-	✓	✓	✓	✓	-
Alban et al. (2020)	ICU capacity	✓	✓	✓	-	-	-	-	-
Ouyang et al. (2020)	ICU admission/discharge	✓	✓	✓	-	-	✓	-	✓
Z. Chen and Kong (2022)	Hospital capacity and admission policy	✓	✓	-	✓	-	-	✓	✓
Gonçalves et al. (2022)	Containment and hospital/ICU capacity	✓	✓	-	✓	-	-	✓	✓
Lu et al. (2021)	ICU admission policy	-	✓	-	✓	-	-	-	✓
Wood et al. (2020)	Containment and ICU capacity	✓	✓	✓	✓	-	-	✓	✓
Shi et al. (2022)	Resource-transshipment	✓	✓	✓	✓	✓	✓	✓	✓
This paper	Resource (ICU) capacity	✓	✓	✓	✓	✓	✓	✓	✓

Abbreviations: ICU, intensive care units; OM, operations management.

Queuing systems are widely used in the healthcare literature to model emergency unit workloads. Prior to Covid-19, the literature typically used queuing models with a constant arrival rate (McManus et al., 2004; Ridge et al., 1998). More recently, Alban et al. (2020) and Ouyang et al. (2020) also use a constant arrival rate of patients to a fixed ICU capacity. The pandemic also motivated some researchers to make modifications: Lu et al. (2021), Wood et al. (2020), and Shi et al. (2022) assume a deterministic time-dependent arrival rate linked to the average trend of the epidemic. In particular, Lu et al. (2021) simulate ICU demand for Covid-19 patients with a deterministic epidemic SEIR (susceptible-exposed-infected-removed) model and discrete event simulation of Covid-19 patient flow in ICUs. Wood et al. (2020) use a multichannel queuing model to simulate the demand on ICU capacity. Shi et al. (2022) model hospital (and ICU) workload through the Covid pandemic by integrating a deterministic time-dependent SIR epidemic model with a stochastic network model of patient movement among various units within the hospital.

The wider healthcare literature also uses queuing models with an infinite number of servers (hence, no queuing) to capture the future demand for patients needing a constrained resource like intensive care. Heemskerk et al. (2017, 2022) also propose an infinite-server queuing system with a mixed Poisson arrival process in a random environment, a special case of a *stationary Cox process*. Boxma et al. (2019) also analyze an infinite-server queue system where the arrival rate process is a *shot-noise process*.

We contribute by applying a Cox queuing model with a stochastic intensity of arrivals that depends on the *observed* number of newly infected individuals. Doing so has two advantages. First, we can capture the variability of the arrival process (for the resource in question) due to the random nature of the disease spread, particularly in the early stages. Second, a common drawback in using Cox queuing models is that the intensity of the process is not observable, so estimating the model from data is difficult (Rydén, 1996). In our case, the patient arrival rate (for ICU or some other resource) is proportional to the observable number of daily new cases.

2.4 | Optimal allocation of capacity

Not all researchers use an optimization module as we do in our three-pronged approach. Indeed, Ferguson et al. (2020), Alban et al. (2020), Z. Chen and Kong (2022), Gonçalves et al. (2022), Lu et al. (2021), and Wood et al. (2020) do not propose any optimization of capacity or other policy levers. However, Ferguson et al. (2020) perform what–if analyses on nonpharmaceutical interventions (NPIs) on the demand side, while Wood et al. (2020) extend their analyses with different levels of ICU capacity. Lu et al. (2021) study the admission policy of patients concerning their death risk profile, given a fixed level of ICU capacity. Furthermore, Shi et al. (2022) use the projected demand of each hospital in a hospital system for decision models for resource (re)allocation.

In contrast, our work seeks to determine *an optimal policy* for capacity expansion. Moreover, we do so from the public health perspective for an entire region, similar to Jiang and Sodhi (2019) for NHS England, rather than for a ward or a hospital. We use stochastic dynamic programming with selected classes of control policies, similar to Shahmanzari et al. (2022). Like them, we prove the Pareto efficiency of the optimal policies. Moreover, we can demonstrate other analytical results, such as the existence of the optimal policy, the convexity of the efficient frontier, and strong duality. Our work and Shahmanzari et al. (2022) are complementary in that we analyze a supply-focused problem to expand capacity with a given (stochastic) level of demand, while they study a demand-focused one with containment measures to reduce demand with a given ICU capacity. We optimize the ICU capacity, given future epidemic scenarios that could include containment measures.

3 | APPLICATION TO ICU CAPACITY EXPANSION DURING COVID-19

We now apply our approach to a specific pandemic, Covid-19, for ICU capacity as the limited resource in the Italian setting of 2020. Italy had the highest reported per capita death rate globally at the end of 2020, with Lombardia, Piemonte, and Veneto regions being the worst hit (each region being responsible for its public health). One reason for the high death toll in Italy (and other countries) was inadequate ICU capacity.

Expanding the capacity of such a resource is both costly and time critical. ICU costs already constitute a large portion of total hospital costs, accounting for 8–30% of the entire hospital budget (Kilic et al., 2019). The direct daily cost of a single ICU in Germany, Italy, the Netherlands, and the United Kingdom ranges from €1168 to €2025, with labor being the main cost (Tan et al., 2012). The cost was even higher during the pandemic—as much as 20% more for Covid-19 patients—due to comorbidity and the need for more ICU staff per patient.

We obtained region-specific daily data concerning infected, dead, recovered, and hospitalized people published on the official Italian Government site[1] from February 24 to December 31, 2020, for the three regions. Our focus is the onset of the two waves of Covid in 2020, so the decision horizons in our application span (1) March 10 to the end of May 2020 for the first wave and (2) October end to December end for the second. We then carried out the following steps for the *first wave*:

1. *Stochastic epidemic modeling*: First, we estimated the SIR model using the official data on the daily infected people available for the 2 weeks from February 24 to March 10, 2020, for the first wave. Then, we generated 10,000 scenarios of daily new infections for the decision horizon from March 10, 2020, onwards for 90 days to the end of May.

2. *Capacity demand*: Next, we translated infections into demand for ICU capacity. We estimated and simulated a queuing system with a stochastic arrival rate proportional to the epidemic process. We estimated the queuing model using official data of daily admissions in the ICU of Covid-19 patients for the same 2-week period from February 24 to March 10, 2020, relying on clinical estimates of the *average length of stay* (ALOS) of a Covid patient in the ICU. We then simulated the ICU demand from March 10 onwards for 90 days using the demand scenarios from the previous step.

3. *Optimal allocation of capacity*: Finally, we used our proposed optimization procedure to find the optimal capacity expansion policy regarding *when* and *how much* to expand to minimize the expected number of patients denied ICU. The principal optimization problem minimizes the cumulative expected number of patients denied ICU in the remainder of the 90-day decision horizon at any time, given the budget constraint on total investment for this period. The optimal policy provides the optimal number of new beds and its timing by way of the critical ICU load to trigger the capacity expansion.

We followed the same procedure for the *second* wave. By then, Veneto had already acquired more-than-adequate ICU capacity. Therefore, we have omitted the results for the second wave for this region. Next, we describe the application of each successive module of our approach in Sections 4, 5, and 6.1, respectively.

4 | MODULE 1: STOCHASTIC EPIDEMIC MODELING

This section details our stochastic SIR model and the robust estimation and projection procedure for infection scenarios.

4.1 | The epidemic model

The SIR model is a compartmental model with the population divided into *susceptible*, *infected*, and *removed* groups, the last one including both *recovered* and *deceased* people. Recovered people are assumed to become immune *till* the decision horizon.

The stochastic epidemic SIR model. Let S_t and I_t represent the number, respectively, of susceptible and infected at any time $t \geq 0$. In the stochastic extension of the SIR model, susceptible and infected individuals are modeled as a two-dimensional Markov chain process (S_t, I_t); Andersson and Britton (2000) discuss the general stochastic epidemic model. Any increment of the stochastic process is expressed as

$$\begin{cases} \Delta S_t = -\beta_t \dfrac{S_t}{N} I_t \Delta t + \Delta Z_1, \\ \\ \Delta I_t = \beta_t \dfrac{S_t}{N} I_t \Delta t - \gamma \, I_t \Delta t - \Delta Z_1 + \Delta Z_2, \end{cases} \quad (1)$$

where $N = S_t + I_t + R_t$ is the population size which is constant over time $t \geq 0$, given the assumption of the population being closed. The number of *removed* patients can be obtained by difference, that is, $R_t = N - S_t - I_t$. The random variables ΔZ_1 and ΔZ_2 are conditionally centered Poisson increments with zero mean and conditional variances $\beta_t \frac{S_t}{N} I_t \Delta t$ and $\gamma I_t \Delta t$, respectively. If we drop the stochastic terms ΔZ_i from Equations (1) and let $\Delta t \to 0$, the equations reduce to the classical deterministic SIR model (Andersson & Britton, 2000). Finally, $C_t = I_t + R_t$ represents the cumulative total number of cases from the beginning of the epidemic process, while $\Delta C_t = C_{t+\Delta t} - C_t$ is its (daily) variation and $\Delta C_t = -\Delta S_t$.

The rate parameters. The *rate of removal* γ is the sum of the recovery and death rates and is *biological* in that it is virus-specific. The literature assumes γ is constant over time, at least within a single wave of a pandemic, although it does change slowly due to virus mutation and vaccination. In contrast, β_t, the *rate of infection*, varies even within the decision horizon due to containment policies:

$$\beta_t = \begin{cases} \beta_0 & \text{for } 0 \leq t < t_0 \\ (\beta_0 - \beta_F)e^{-\lambda(t-t_0)^2} + \beta_F & \text{for } t \geq t_0, \end{cases}$$

where t_0 is the time when NPIs start to have effects, β_0 is the initial value of the infection rate, λ is a positive parameter that modulates the speed at which the infection rate decreases, and $\beta_F \leq \beta_0$ is the asymptotic long-term value. There are generalizations that allow for negative λ values that correspond to an increased contagion rate because of interventions being terminated or seasonal influences, but these are not considered here.

Basic and effective reproduction numbers. The *reproduction number* determines the progress of the epidemic, and it can be either *basic* or *effective* (Chowell et al., 2007). The *basic reproduction number*, $\mathcal{R}_0$, is the average number of people infected by an infected person when almost all the population is susceptible, that is, at time t_0 $\mathcal{R}_0 = \frac{\beta_0}{\gamma}$. The *effective reproduction number* varies in time with the decrease in the number of susceptible individuals in the population and with the implementation of pandemic containment systems (or the introduction of a vaccine) that reduce the number of daily cases. The effective reproduction number at any time $t \geq t_0$ as

$$\mathcal{R}_t = \frac{S_t}{N} \cdot \frac{\beta_t}{\gamma} = \frac{S_t}{N}\left[(\mathcal{R}_0 - \mathcal{R}_F)e^{-\lambda(t-t_0)^2} + \mathcal{R}_F\right]. \quad (2)$$

Here, $\mathcal{R}_F = (\beta_F/\gamma) \leq \mathcal{R}_0$ is the *asymptotic long-term value of the reproduction number* when we assume $\lambda > 0$. The changing value of $\mathcal{R}_t$ allows us to distinguish the initial exponential growth in infections from their long-run decay to zero as $S_t \to 0$ in time.

4.2 | Model estimation and scenario generation

Epidemic data have two sources of uncertainty, generally ignored in the literature, regarding the new cases each day: (1) the reporting error, such as delays in official registrations of daily cases, and (2) the disease dynamics that are inherently stochastic because recovery time and time to get infected are random variables. We incorporate the uncertainty by using the (joint) probability distribution of the parameter estimators to perform a *Monte Carlo sensitivity analysis* (Gonçalves et al., 2022) of the parameter values, as described below. We can produce longer-term forecasts that are also more robust against reporting errors in the data. Doing so is particularly useful at the beginning of a pandemic, when there is not much data and that too is noisy, in part, due to reporting errors.

Parameter estimation. During the early days of an epidemic, the effects of the depletion of the susceptible population are small, and we can assume exponential growth for most infectious diseases (Anderson & May, 1991). Consequently, the number of new cases per unit of time ΔC_t increases exponentially, and the exponential-growth rate Λ is a function of the basic reproduction number $\mathcal{R}_0$ and the removal rate γ, with $\Lambda(\mathcal{R}_0, \gamma) = (\mathcal{R}_0 - 1)\gamma$. Following Ma (2020), we assume that the number of cases at day t_i, ΔC_{t_i} is independently Poisson distributed with mean $C_0 e^{\Lambda(\mathcal{R}_0,\gamma)t_i}$ to estimate $(\mathcal{R}_0, \gamma)$. We obtain point estimates for the parameters by maximizing the log-likelihood function numerically as follows:

$$(\hat{\mathcal{R}}_0, \hat{\gamma}, \hat{C}_0) = \operatorname*{argmax}_{(\mathcal{R}_0,\gamma,C_0)} \sum_{i=0}^{N-1} -C_0 e^{\Lambda(\mathcal{R}_0,\gamma)t_i}$$
$$+ \Delta C_{t_i} \log C_0 + \Delta C_{t_i} \Lambda(\mathcal{R}_0, \gamma)t_i, \quad (3)$$

where C_0 is the number of infected people at time t_0 and N is the number of days observed.

Interval estimation of the parameters is important as the precision of the estimates. For this reason, we adopt a robust procedure producing an interval estimation of the parameters by reconstructing the (empirical) joint probability distribution of the estimators $(\hat{\mathcal{R}}_0, \hat{\gamma}, \hat{C}_0)$. A natural way to estimate such a distribution is to use a resampling method such as the bootstrap. We do so by generating sets of realizations of the daily cases curve using parametric bootstrap (Chowell et al., 2007; Efron & Tibshirani, 1986). In particular, for day t_i with $i = 1, \ldots, N-1$, each realization of daily cases is sampled from a Poisson distribution with a mean equal to that of the observed data, ΔC_{t_i} (Chowell et al., 2007). Then, the empirical distribution of each parameter, and even their joint distribution, $(\mathcal{R}_0, \gamma)$, is reestimated from each of 10,000 bootstrapped epidemic curves. Confidence intervals are computed from the inverse empirical distribution functions for each parameter.

In estimating R_0 and γ jointly, we improve upon the typical practice for estimating compartmental models (Chowell

et al., 2007; Favier et al., 2020; Shapiro et al., 2021; Shi et al., 2022). Typically, the value of removal rate γ in the SIR model (or the latent period in the SEIR model) is fixed to a value chosen using clinical observations, and the estimation (and bootstrap) procedure focuses only on the reproduction number. However, the estimators of parameters $\mathcal{R}_0$ and γ are strongly dependent, as we confirmed by assessing their joint bootstrap distribution for the three Italian regions. Hence, fixing a particular value for γ can create significant bias in the (unconditional) point estimate of $\mathcal{R}_0$. Indeed, such a choice affects the marginal distribution of the $\mathcal{R}_0$ estimator, something not considered by Chowell et al. (2007).

Forecasting the epidemic curve. We use these bootstrap estimates of (R_0, γ) to simulate the stochastic epidemic model, thus taking into account the uncertainty in the estimated parameters due to the data and that inherent in the epidemic process. We assume a plausible range of values for the asymptotic value $\mathcal{R}_F$ and for the parameter λ of the exponential function, as in Equation (2). The parameter λ modulates the speed at which the reproduction number tends to the asymptotic value $\mathcal{R}_F$. In particular, the quantity $\bar{t} = \sqrt{\log 2 / \lambda}$ is the half-life of the reproduction number, that is, $\mathcal{R}_{\bar{t}} = \frac{1}{2}(\mathcal{R}_0 + \mathcal{R}_F)$. In the early stages of a pandemic, there is only a little data available on the estimation of the reproduction number dynamics $\mathcal{R}_t$. However, we fix a plausible range of values for $\mathcal{R}_F$ and $\bar{t}$ using estimates of $\mathcal{R}_t$ from previous pandemics (Chowell et al., 2007; Cori et al., 2013; Thompson et al., 2019). (We discuss this in more detail in the next two subsections about the two Covid waves in Italy.) Over time, interventions and the decrease in susceptible individuals limit the spread of the virus, eventually reducing the value of the number of reproductions to below one. Hence, we simulate the asymptotic value of the reproduction number $\mathcal{R}_F$ from a uniform distribution over the interval $[LB, 1)$, where $LB > 0$ is a lower bound. Predicting the speed of decline in reproductive numbers is more complex, as it depends on the effectiveness of various interventions. (An extension of our procedure would be to perform a what–if analysis on the intervention effectiveness (Ferguson et al., 2020) and integrate it with our analysis of the ICU demand forecast and capacity optimization.)

Generating future scenarios. We perform the following steps to simulate the future numbers of daily cases ΔC_t: (1) randomly sample values of $(\mathcal{R}_0, \gamma)$ from their bootstrap joint distribution; (2) randomly sample the asymptotic value of the reproduction number $\mathcal{R}_F$ from a uniform distribution in the interval $[LB, 1)$; (3) fix the value of the reproduction number half-life $\bar{t}$; (4) simulate the trajectories of S_t, I_t and R_t via Equation (1) and obtain the total number of cases C_t and daily cases ΔC_t; recall that $\Delta C_t = -\Delta S_t$.

We include both sources of variability in the simulation of future daily cases with this procedure. A Monte Carlo sensitivity analysis captures simultaneous changes in epidemic parameters. While Gonçalves et al. (2022) use a similar procedure, they randomly select all the parameter values from uniform independent distributions over arbi-

TABLE 2 Calibrated parameters values ($\mathcal{R}_0, \gamma$) of the SIR model using data from February 24 to March 10, 2020, with 95% confidence level intervals (CI).

Region	$\mathcal{R}_0$	CI 95%	γ	CI 95%
Lombardia	3.2	[2.8, 3.5]	0.094	[0.091, 0.100]
Piemonte	4.5	[2.7, 5.9]	0.08	[0.07, 0.10]
Veneto	3.5	[2.2, 4.3]	0.09	[0.08, 0.11]

trary ranges. Moreover, they do not consider dependency between the parameter estimators. Instead, we capture the realistic uncertainty in future epidemic scenarios from the reported data.

4.3 | The first Covid-19 wave

We first calibrated the SIR epidemic model using 2020 data only from February 24 to March 10 with maximum likelihood to estimate $\mathcal{R}_0$ and γ. Then, we reconstructed confidence intervals and the empirical joint distribution of the two estimators using the parametric bootstrap described in Section 4.2. Finally, we projected the daily cases for the 90-day decision horizon.

Table 2 presents the estimated parameters of SIR model $\mathcal{R}_0$ and γ and the corresponding 95% confidence intervals. The values of the basic reproduction number $\mathcal{R}_0$ and the recovery rate γ are similar across the three regions and are similar to other preliminary results concerning Covid-19 spread in Italy available at that time (D'Arienzo & Coniglio, 2020). Taking the estimated value of the rate γ, we estimated the average removal time from 9 to 14 days. Figure 2 illustrates the *joint probability distribution* of the parameter estimators and the marginal distribution of the $\mathcal{R}_0$ estimator. Indeed, $\mathcal{R}_0$ and γ show strong dependency. This is why assigning an arbitrary value to γ, as commonly done with SIR modeling, can lead to a strong bias in estimating the value and the marginal distribution of $\mathcal{R}_0$.

We set the start date to be March 10, 2020, when the prime minister (decree of the President of the Council of Ministers) decreed several containment measures. Next, we had to assume a plausible range of values for the asymptotic value $\mathcal{R}_F$ and of the half-life of the reproduction number $\bar{t}$ to forecast the epidemic curve (Section 4.2). Studies estimating the reproduction number $\mathcal{R}_t$ for Covid-19 were not yet available on March 10. Therefore, we consulted the literature on previous pandemics to select a plausible range of values for $\mathcal{R}_F$ and $\bar{t}$. Over time, the number of reproductions reduces to below one. We simulated the asymptotic value $\mathcal{R}_F$ from a uniform distribution over the interval [0.5, 1), thus taking into account the uncertainty in the reproduction number evolution. Furthermore, looking at the literature on previous pandemics (Chowell et al., 2007; Cori et al., 2013; Thompson et al., 2019), we noted that the reproduction number $\mathcal{R}_t$ reaches 1 in around three weeks for cases with similar values of $\mathcal{R}_0$, suggesting a half-life $\bar{t}$ of 7 days. With all the parameters in
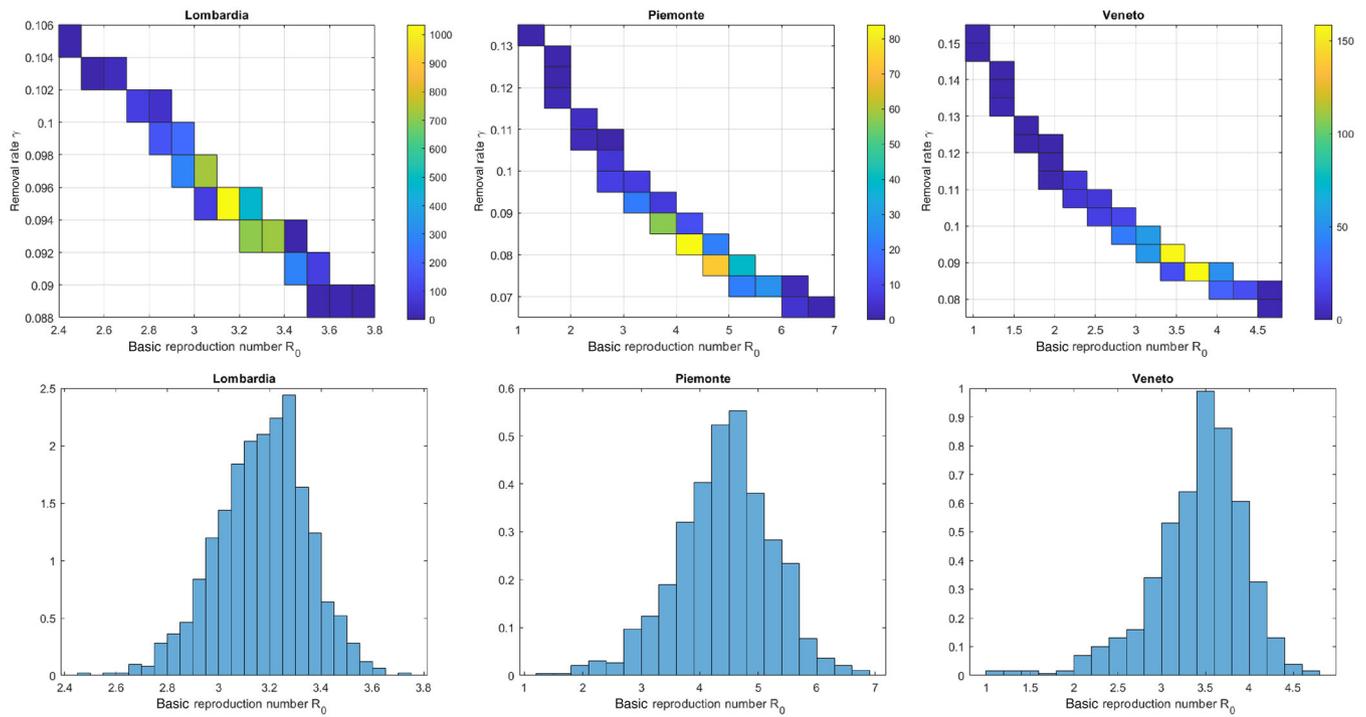
FIGURE 2 Top: The bootstrap joint distribution of the estimator of parameters $(\mathcal{R}_0, \gamma)$. Bottom: The marginal distribution of the estimator of $\mathcal{R}_0$ for the three regions. The distributions are obtained using data from February 24 to March 10, 2020, and 10,000 simulations for the parametric bootstrap. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 The table shows calibrated parameter values $(\mathcal{R}_{t_0}, \gamma)$ of the SIR model using data from October 7 to 30, 2020, with 95% confidence level intervals.

Region	$\mathcal{R}_{t_0}$	CI 95%	γ	CI 95%
Lombardia	2.3	[2.2, 3.1]	0.06	[0.04, 0.07]
Piemonte	2.1	[2.0, 2.3]	0.08	[0.07, 0.10]

Abbreviation: CI, confidence interval.

place, we were able to project daily new-case scenarios for the 90-day decision horizon from March 10, 2020, onwards.

4.4 | The second Covid-19 wave

We recalibrated the SIR parameters for the second wave using data from October 7 to 30, 2020. The 60-day decision horizon takes us to December end. As before, we took the reproduction number $\mathcal{R}_t$ as constant over $0 \le t \le t_0$ and equal to $\mathcal{R}_{t_0}$, with t_0 being October 30, 2020, when the Prime Minister again announced several containment measures across Italy and the reproduction number started to decrease. Our estimation procedure was the same as in the first wave, but the results reflect the differences between the two waves of the pandemic. Table 3 reports the estimated values and confidence intervals of $\mathcal{R}_{t_0}$ and γ for Lombardia and Piemonte, respectively. (We skipped Veneto for the second wave due to overexpansion of ICU capacity in the region.)

The estimated values of the reproduction numbers $\mathcal{R}_{t_0}$ in both Lombardia and Piemonte are smaller than the estimates of the basic reproduction numbers $\mathcal{R}_0$ from the first epidemic wave, possibly due to personal protective equipment and social distancing. Furthermore, for Lombardia, we observe a decrease in the value of the removal rate γ, which implies an increase in the average removal time—from 10 to 25 days—and suggests a possible increase in the average recovery time and thus the ALOS in ICU (or hospital).

We simulated the asymptotic value $\mathcal{R}_F$ again from $uniform[0.5, 1]$. A longer reproduction number half-life, $\bar{t}$ at 14 days was necessary because of the smaller starting value $\mathcal{R}_{t_0}$ and because the announced interventions were reversed to reduce their negative economic impact.

5 | MODULE 2: CAPACITY LOADING

We now detail our $Cox/M/\infty$ queuing model for translating infection scenarios into demand scenarios for ICU. The stochastic arrival intensity at time t is proportional to the daily *reported* new infections ΔC_t. In our parsimonious model, service times have an exponential distribution with the ALOS in ICU as the mean. (We can easily generalize the model to service times with any nonnegative distribution.) We did not limit the number of servers as queuing is not practical here, in line with the review by Worthington et al. (2020).

Creating scenarios for load on ICU. We simulated the queue to forecast its length, that is, the number of patients

in ICU P_t, assuming the daily numbers of arrivals at ICU are independent Poisson random variables, conditional on the daily number of new cases. Given the ith realization of $\{\Delta C_s\}_{s=0}^t$, the number of arrivals in $[t - \Delta t, t]$, $A_t^{(i)}$, is

$$A_t^{(i)} \sim Poisson(a\, \Delta C_t^{(i)}), \qquad (4)$$

where $a > 0$ is the average value of the fraction of the newly infected who needs ICU. Then A_t is a Cox process, with the stochastic intensity proportional to the number of new cases ΔC_t. We took $E_t^{(i)}$, the number of patients leaving the ICU after treatment in $[t - \Delta t, t]$, as a truncated Poisson random variable, conditional on the number of patients in ICU at time $t - \Delta t$, that is, $E_t^{(i)} = \min(\tilde{E}_t^{(i)}, P_{t-\Delta t}^{(i)})$, and $\tilde{E}_t^{(i)} \sim Poisson(\mu\, P_{t-\Delta t}^{(i)})$, where $\mu = 1/ALOS$ is the exit rate. Finally, we updated the number of patients in ICU at time t as

$$P_t^{(i)} = P_{t-\Delta t}^{(i)} + A_t^{(i)} - E_t^{(i)}. \qquad (5)$$

We did not include *uninfected* patients in intensive care in our analysis, but we can accommodate them easily in an extension. In any case, the number of non-Covid-19 patients who needed ICU hospitalization during the first epidemic wave in Italy (March 2020–May2020) was much lower than before the pandemic due to the postponement of all nonurgent surgical interventions, and the reduction of postsurgical intensive care (Naderi et al., 2021; Ridge et al., 1998).

5.1 | Estimating and forecasting the ICU demand model

For each region, we fitted the ICU demand model using historical data of daily cases and ICU patients from February 24 to March 10 for the first wave and from October 7 to 30 for the second wave. We estimated the fraction a of the newly infected Covid patients needing ICU using a maximum likelihood procedure. The daily number of arrivals in ICU, A_t, is distributed as independent Poisson random variables with their mean value proportional to the daily cases, ΔC_t, as in Equation (4). We obtained confidence intervals using parametric bootstrap (Efron & Tibshirani, 1986), similar to our estimation of the SIR model parameters. Additionally, we estimated the ALOS using the available clinical information at the observed date to forecast the future ICU demand of Covid-19 patients using the queuing model.

5.2 | The first Covid-19 wave

The results are given in Table 4. The average fraction of newly infected individuals who need intensive care a in Equation (4), ranged from 4% to 10% (see Table 4). Note that this is the fraction of the *reported* new cases which underestimates the actual ones. Indeed, the estimated ratio between

TABLE 4 Estimated average fraction of new cases that need intensive care, a (Equation 4), using data from February 24 to March 10, 2020.

Region	a (%)	CI 95%
Lombardia	6.0	[5.5, 6.6]
Piemonte	10.2	[7.9, 12.7]
Veneto	4.4	[3.4, 5.4]

Abbreviation: CI, confidence interval.

TABLE 5 Estimated average fraction a (Equation 4) of new cases that need ICU, using data from 7 to 30 October 2020.

Region	a (%)	CI 95%
Lombardia	0.39	[0.35, 0.45]
Piemonte	0.48	[0.40, 0.55]

Abbreviation: CI, confidence interval.

the actual and official number of total cases for Covid-19 typically ranges from 4× to 10× in the literature (McCulloh et al., 2020; Shi et al., 2022), depending on the country and the stage of the pandemic. Thus, our method for estimating a compensates for the underestimation in official data, *ultimately giving the correct projections for the demand for ICU.*

Regarding the ALOS in ICU, no data were available at the beginning of the pandemic on ALOS for Covid-19 patients in ICU. We simply chose ALOS = 12 days, doubling the 6 days in ICU in pre-Covid-19 times in Italy (Agodi et al., 2018). Figure 3 compares the forecasting scenarios with the out-of-sample data of the ICU workload. We see that the actual progression of ICU load in the out-of-sample period for Lombardia and Piemonte falls within the forecast confidence intervals.

For the first wave in Veneto, we performed a sensitivity analysis on the value of the reproduction number half-life $\bar{\imath}$, defined in Section 4.2, that we set equal to 7 days as obtained from the literature. With this choice, simulations of future ICU demand were always below the initial number of available beds in intensive care K_0, therefore making the capacity expansion useless. For this reason, we additionally performed a what–if analysis by considering a more conservative scenario with $\bar{\imath} = 10$ days, the smallest value for which our simulations generated a future demand that exceeded the initial capacity for Veneto.

5.3 | The second Covid-19 wave

We estimated the ICU demand for each region from October 7 to 30, 2020. The results of our analysis reflect the different behavior of the two pandemic waves. For instance, estimated values of the proportion a of Covid patients needing ICU are much smaller than in the first wave (Table 5). Contact tracing was more effective, and even asymptomatic
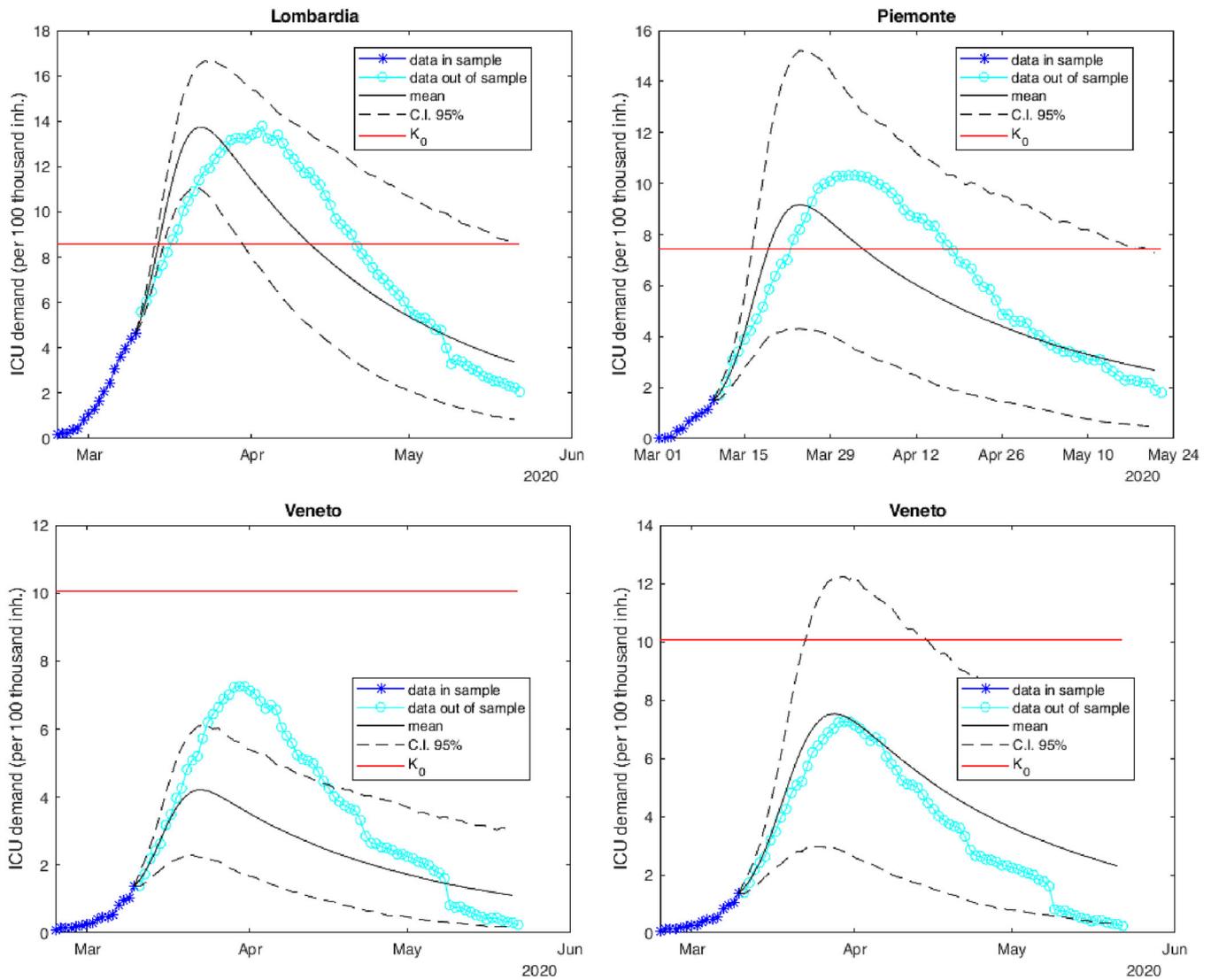
FIGURE 3 Historical ICU workload from February 24 until March 10 (blue) and later (cyan). The solid black line is the mean forecast demand, and the dashed black lines represent the 95% confidence interval around the mean. ICU demand is per 100,000 people for all three regions. ICU ALOS $\frac{1}{\mu} = 12$ days. The half-life of the reproduction number $\bar{t} = 7$ days for (a)–(c) and for sensitivity analysis $\bar{t} = 10$ days for (d). CI, confidence interval; ICU, intensive care unit. [Color figure can be viewed at wileyonlinelibrary.com]

and paucisymptomatic patients were detected. Moreover, improvements in treating Covid-19 reduced the fraction of critically ill patients.

Unlike in the first wave, clinical information was already available on Covid-19 patients by the beginning of October 2020. ALOS for Covid-19 patients in ICU estimated from clinical observations had a wide range of values in the literature of that time. We chose ALOS = 20 days (Lapidus et al., 2020)[2] for the ICU workloads in Figure 4.

6 | MODULE 3: OPTIMAL CAPACITY ALLOCATION

Our ICU capacity management model seeks to help decision-makers achieve the best trade-off between the cost of

expanding ICU capacity and the number of patients denied admission to ICU due to inadequate capacity. We take these two quantities as the decision-makers' ICU capacity policy π, selected from a set of feasible policies Π in Section 6.2.

Let K_t^π be the total number of intensive-care beds available at time t for a capacity (expansion) policy π. Then K_t^π is *endogenous* for the optimization. The stochastic demand process, the number of critically ill patients who need ICU access—or some other critical resource—each day t, P_t, is *exogenous* to the optimization. The couple (K_t^π, P_t) defines the state of the system at each time t. The exogenous part of the state, P_t, evolves according to a known stochastic process as in Section 5 and Equation (5), independent of the policy π. (The independence between endogenous and exogenous components is a classical approximation in dynamic programming literature (Bertsekas, 2007).) At each time t, the
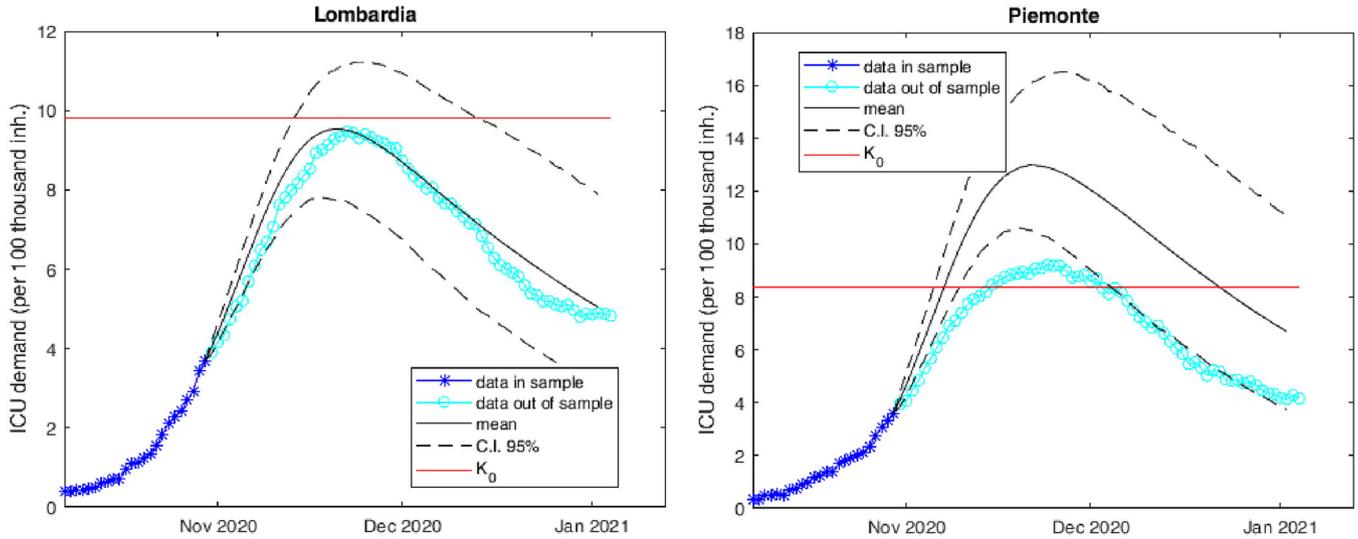
FIGURE 4 Historical ICU workload from October 7 to October 30, 2020 (blue) and later (cyan). The black line is the mean forecast demand after October 30, while the dashed lines are the 95% confidence interval around the mean. ICU demand is per 100,000 inhabitants for both regions. Half-life value of the reproduction number $\bar{t} = 14$ days, ALOS = 20 days for both regions. CI, confidence interval; ICU, intensive care unit. [Color figure can be viewed at wileyonlinelibrary.com]

decision-maker chooses an *action* changing the capacity of the resource ΔK_t^π that depends on the current state (K_t, P_t). The *capacity policy* is the sequence of actions ΔK_t^π over $[0, T]$, that is, $\pi := \{\Delta K_t^\pi\}_{t=0}^T$. The operational state of the system at time $t + 1$, $K_{t+1}^\pi (K_t^\pi, \Delta K_t^\pi)$, depends on the policy π and the state at time t.

The number of patients denied ICU access. The immediate loss l_t signifies the number of patients denied access to the resource (ICU in this context) on day t. If the number of patients P_t exceeds the ICU capacity $K_t^\pi + \Delta K_t^\pi$, then $P_t - K_t^\pi - \Delta K_t^\pi$ patients must be refused ICU access. Otherwise, all infected people needing intensive care would be admitted to ICU at time t, so that

$$l_t(P_t, K_t, \Delta K_t) = \max(P_t - K_t - \Delta K_t, 0). \quad (6)$$

The expected cumulative number of patients, L_T, refused ICU admission at the time $[0, T]$ depends on the policy π through the capacity dynamic $\{K_t^\pi\}_0^T$ and the actions ΔK_t^π:

$$L_T(\pi) = \mathbb{E}\left[\sum_{t=0}^T l_t(P_t, K_t^\pi, \Delta K_t^\pi)\right], \quad (7)$$

where the immediate loss l_t is defined above in Equation (6). If needed, the expected value in the model can be replaced by some other metric, say, the 75th percentile.

The expected cost of capacity expansion. The decision-maker has to consider the costs. Each intensive (or subintensive) therapy unit is subject to setup and maintenance costs. Let the unit setup cost be h_0, the unit maintenance cost be h_1, and the extra cost for a patient due to medicine, hospital staff, etc. be h_2. The total daily cost c_t comprises (1) the setup

cost of additional units of capacity, $h_0 \cdot \max(\Delta K_t^\pi, 0)$; (2) the maintenance cost of all the units at time $h_1 (K_t^\pi + \Delta K_t^\pi)$; and (3) the daily patient-related cost, $h_2(P_t) (K_t^\pi + \Delta K_t^\pi)$. Therefore,

$$c_t(P_t, K_t, \Delta K_t) = h_0 \cdot \max(\Delta K_t, 0) + h_1 \cdot (K_t + \Delta K_t)$$
$$+ h_2(P_t) \cdot (K_t + \Delta K_t). \quad (8)$$

Our model manages healthcare personnel expenditure in a flexible way through the cost for patient $h_2(P_t)$. In fact, the healthcare staff can be hired on a fixed-term basis, or, once the health crisis has been resolved, they can be relocated to other departments. (See details on cost-parameter estimations in Section 6.3.) Given the immediate cost c_t in Equation (8), the *total expected cost* $C_T(\pi)$ for a policy π over $[0, T]$ is given by

$$C_T(\pi) = \mathbb{E}\left[\sum_{t=0}^T c_t(P_t, K_t, \Delta K_t)\right]. \quad (9)$$

6.1 | The capacity expansion optimization

We can now specify the optimization problem with a parameterized investment budget M.

Efficient frontier. We minimize L_T, the expected number of patients denied ICU access, subject to the budget M. Then $V(M)$ is the *efficient frontier* with the optimal value of L_T a function of the budget $M \in [M_0, \infty)$, where M_0 is the minimum total expected cost in Equation (9) without expansion to maintain the initial capacity, K_0. Thus, given some budget

level M, $V(M)$ is the minimum number of patients that we expect to lose. Accordingly, we can solve the optimization problem

$$V(M) := \min_{\pi \in \Pi}\{L_T(\pi) \mid C_T(\pi) < M\}, \qquad (10)$$

with L_T and C_T defined in Equations (7) and (9). The optimization problem in Equation (10) can be equivalently expressed as the recursive Bellman equation (Bertsekas, 2007) for $M \in [M_0, +\infty)$ over $t \in [0, T-1]$ as

$$V_t(M; P_t, K_t) = \min_{\Delta K_t \in \mathcal{A}_t(P_t, K_t)} l_t(P_t, K_t, \Delta K_t)$$
$$+ \mathbb{E}_t[V_{t+1}(M - c_t(P_t, K_t, \Delta K_t); P_{t+1}, K_{t+1})]$$
$$V_T(M; P_T, K_T) = \min_{\Delta K_T \in \mathcal{A}_T(P_T, K_T)} \{l_T(P_T, K_T, \Delta K_T) \mid$$
$$c_T(P_T, K_T, \Delta K_T) \le M\}, \qquad (11)$$

where V_t represents the efficient frontier for each value of resource M and $\mathcal{A}_t$ is the set of feasible actions, given the current state of the system at time t, (P_t, K_t). Next, we make the following assumption to guarantee the existence of the solution to the above problem

Assumption 1. The number of beds K_t is bounded, that is, there exists K^{max} such that $K_t \le K^{max}$ for each t.

Assumption 1 is not a serious limitation in numerical applications. In fact, K^{max} is fixed to a value greater than the maximum number of patients needing ICU across all simulations and all paths, that is $K^{max} > \max_{t \in 1, ..., T} \max_j P_t^{(j)}$, where j is the index of Monte Carlo simulation. Now we can assert

Proposition 1. *Given Assumption 1, there exists an optimal policy for $V(M)$ in Equation (10).*

For proof, apply the Weierstrass theorem to each step $t = 1, ..., T$ in (11), then for each state (P_t, K_t) there exists the optimal quantity $\Delta K_t^*(P_t, K_t)$ and $\pi^* = (\Delta K_1^*, ..., \Delta K_T^*)$. The Weierstrass theorem requires that the objective function $l_t(P_t, K_t, \Delta K_t) + \mathbb{E}_t[V_{t+1}(M - c_t(P_t, K_t, \Delta K_t); P_{t+1}, K_{t+1})]$ is continuous, which is the case here. The theorem also requires that the feasible regions $\mathcal{A}_t(P_t, K_t)$ are closed and bounded. In the most general case, that is, without any restriction on the feasible policy set, the only requirement for an action to be feasible is the total number of resources being nonnegative, that is, $K_t \ge 0$ for all $t \in [0, T]$, so that $\mathcal{A}_t(P_t, K_t) = [-K_t, +\infty)$. If the feasible regions are bounded as per Assumption 1, then $\mathcal{A}_t(P_t, K_t) = [-K_t, K^{max} - K_t]$. □

Next, we present an analytical result on the efficient frontier, which will be useful for the definition of the optimal investment level M^*. For this result, we need

Assumption 2. The set of feasible policy Π is convex.

Assumption 2 is satisfied both in the general case and in the case of irreversible expansion (see Section 6.2). In the general case, Π is the set of all the policies π such that the number of beds K_t^π is positive and lower than K^{max} (Assumption 1) for $t = 1, ..., T$. Then, Π is the set of solutions of the following system of linear inequalities, $-K_0 - \sum_{j=1}^{n-1} \Delta K_j \le \Delta K_n \le K^{max} - K_0 - \sum_{j=1}^{n-1} \Delta K_j$ for $n = 1, ..., T$, and it is convex. In the case of irreversible expansion, the feasible policies are determined by two decision variables $(\bar{K}, b)$, taking values in the convex set $[0, K^{max}] \times [0, 1]$.

Proposition 2. *Given Assumptions 1 and 2, the efficient frontier $V(M)$ in (10) is decreasing in M and convex.*

The objective function $L_T(\pi)$ defined in Equation (7) is convex, being the sum of the immediate losses in Equation (6), which are convex. Moreover, given Assumption 2 and the convexity of the cost function defined in (9), the feasible region $\mathcal{D} := \{\pi \in \Pi \mid C(\pi) \le M\}$ for $M \in [M_0, +\infty)$ is convex. We conclude that if the optimization problem in (10) is convex, then $V(M)$ is convex; see Boyd and Vandenberghe (2004), section 5.6.1. Moreover, as the immediate loss l_t is decreasing in ΔK_t while the immediate cost c_t is increasing, $V(M)$ is decreasing. □

The dual problem. The dual of the optimization problem in Equation (10) is

$$\max_{n>0} \min_{\pi \in \Pi} \mathcal{L}(\pi, n), \qquad (12)$$

where $\mathcal{L}(\pi, n) = L_T(\pi) + n\,(C_T(\pi) - M)$ is the Lagrangian function and n is the Lagrange multiplier. To prove strong duality, we need to assume the smoothness of L_T and C_T functions, that is, differentiable with respect to $(\Delta K_1^\pi, ..., \Delta K_T^\pi)$. This is a reasonable assumption, particularly as both quantities are expected values. The Karush–Kuhn–Tucker conditions are difficult to prove in this case, although we did not see any duality gap in our numerical application. In line with Simon and Blume (1994, Chapter 19) and Boyd and Vandenberghe (2004, Chapter 5), we require the following assumption.

Assumption 3. Assume that $L_T(\pi)$ and $C_T(\pi)$ are smooth enough to be differentiable with respect to $(\Delta K_1^\pi, ..., \Delta K_T^\pi)$ and that the Karush–Kuhn–Tucker conditions are satisfied.

We can now assert

Proposition 3. *Given Assumptions 2 and 3, strong duality holds for the primal problem in (10) and the dual problem in (12).*

Proposition 3 holds given that the primal problem in 10 is convex; for proof, see Boyd and Vandenberghe (2004, Chapter 5).

Shadow price. Solving the dual problem in (12), we obtain the optimal value of the Lagrange multiplier $n(M)$. Under suitable regularity conditions in Assumption 3, the following equality holds:

$$n(M) = -\frac{\partial V(M)}{\partial M}. \tag{13}$$

Given a unit increase in the budget, the shadow price $n(M)$ represents the marginal increase in those given admission, equivalently, the marginal decrease in the number of patients denied access to ICU. From Proposition 2, the function $n(M)$ defined in (13) is always positive and decreasing in the resource level M. So, the larger the budget for investment, the smaller the *additional* number of patients that need to be admitted to ICU.

The optimal investment level M^*. Let M^* be the largest resource level at which $n(M)$ is strictly positive, that is,

$$M^* := \sup\{M \in [M_0, \infty) \mid n(M) > 0\}. \tag{14}$$

M^* is the *threshold resource level* beyond which no further investment is useful for admitting more patients. If M^* is finite, it represents *the optimal investment level*, that is the smallest budget level at which the efficient frontier $V(M)$ reaches its minimum value, that is, $M^* = \min\{M \in [M_0, \infty) \mid M \in \arg\min V(M)\}$. In fact, if M^* is finite and $M \geq M^*$, then $V(M) = V(M^*)$. Allocating an amount larger than M^* will not decrease the expected number of patients denied the critical resource (ICU). The decision-makers can choose to invest an amount smaller than M^* but expect denying access to patients, according to the efficient frontier valuation, that is, if $M < M^*$ then $V(M) > V(M^*)$.

Pareto efficiency of the optimal policy. Our capacity management model trades the number of patients denied access and the total expected cost. We could alternatively express our model as a *multiobjective minimization problem* to minimize both the number of patients denied ICU and the ICU total cost over a set of feasible policies. The optimization problem in Equation (10) is reconstructed, applying the ϵ-constrained method to the multiobjective optimization problem. Then, the optimal policy π^* upon solving (10) is weakly Pareto efficient (Ehrgott, 2005, Section 4.1). Using the ϵ-constrained method, we define a second optimization problem similar to Equation (10), in which we minimize the cost, subject to a limit on the number of patients denied ICU, that is

$$\min_{\pi \in \Pi}\{C_T(\pi) \mid L_T(\pi) \leq N\}, \tag{15}$$

where $N > 0$ represents the tolerated level of expected loss L_T.

Proposition 4. *If there exist two values $(\hat{M}, \hat{N}) \in [M_0, \infty) \times [0, \infty)$, such that a feasible policy $\pi^* \in \Pi$ is solution of problem (10) for $M = \hat{M}$ and problem (15) for $N = \hat{N}$, then, π^* is Pareto efficient and $L_T(\pi^*) = \hat{N}$ and $C_T(\pi^*) = \hat{M}$.*

For proof, see Ehrgott (2005, Section 4.1). In our numerical application, we verified the efficiency of our solutions solving both optimization problems (10) and (15).

Marginal cost of a single patient for ICU. We now consider the dual optimization problem of (15), that is

$$\max_{m>0} \min_{\pi \in \Pi}\{C_T(\pi) + m\,(L_T(\pi) - N)\}, \tag{16}$$

where $N > 0$ represents the tolerated level of expected loss L_T. This optimization problem allows us to quantify the *marginal cost* (shadow price) for an admitted patient, as the optimal value of the Lagrange multiplier $m(N)$, under a suitable regularity condition as we presented in Assumption 3 and Proposition 3.

6.2 | Capacity expansion policy

The two quantities of interest—the cost C_T and the number of patients denied the critical resource (ICU) L_T—depend on the adopted expansion policy π via the actions in changing ICU capacity, ΔK_t^π. In the typical case, as presented in previous sections, a policy $\pi \in \Pi$ is feasible if the capacity at each time t is positive and below the maximum capacity, that is, $0 \leq K_t^\pi \leq K^{max}$.

In an extreme case, the decision-makers could adopt an emergency policy of expanding the resource capacity incrementally every day with updated infection information. However, we avoided such a daily (and possibly reactive) emergency expansion because of the lead times entailed for ICU expansion. Instead, we seek to plan ICU capacity expansion using a medium-term perspective, in this instance, of a few weeks rather than a few hours. Such an expansion is typical in healthcare: structural changes, such as creating new hospital wards, need to be planned and are rarely reversible. Indeed, dismantling the medical equipment of ICUs leads to a minimal recovery of the invested resources. As such, our model anticipates a quantum and irreversible expansion (at least in the near term) of ICU capacity during the decision horizon at the start of a pandemic.

Such an expansion puts restrictions on the set of feasible policies Π. In particular, we choose the set Π, such that a policy π is admissible if the actions ΔK_t^π are defined as

$$\Delta K_t^\pi(\bar{K}, b) = \bar{K}\,\mathbb{1}_{\{t > \tau(b)+d\}},$$
$$(\bar{K}, b) \in \mathbb{R}_+ \times [b_0, 1], \tag{17}$$

where $\mathbb{1}$ is the indicator function, d is the time necessary to make new beds available, b_0 is the initial occupancy percentage of the ICU, and τ is the time at which the preparation for the new beds, including staff training, starts. (Recall that $\bar{K}$ is limited by an upper bound K^{max}.) We take τ to be the first day on which P_t exceeds the threshold value $b\,K_0$ with

$b_0 < b < 1$, that is,

$$\tau(b) = \min\{t > 0 : P_t \geq b \, K_0\}. \tag{18}$$

With K_0 beds initially, $\bar{K}$ is the number of additional units that would become operational at time $\tau + d$. Here, the delay term d captures the time for expanding resource capacity, in this case, for installing the new beds and equipment and recruiting additional trained healthcare staff. The resource (ICU) capacity level at time t is therefore $K_t^\pi(\bar{K}, b) = K_0 + \bar{K} \mathbb{1}_{\{t > \tau(b)+d\}}$.

Percentage occupation as the trigger for capacity expansion. As soon as a predetermined fraction b of the initially available ICU beds (or other critical resource) is occupied, the decision-maker decides whether or not to make an additional $\bar{K}$ new beds available. This approach is consistent with the *color code* policy in Italy when different parts of the regions were assigned different colors depending on the ICU occupancy rate, represented in our model by the parameter b. Therefore, planners have to determine the optimal values of two decision variables when the threshold is exceeded: (1) the threshold value b and (2) how many new units $\bar{K}$ to create. This choice of the expansion policy form simplifies the optimization problems in Equations (12) and (16) for practical application.

6.3 | Application to the two waves of Covid-19

Starting with the 100,000 scenarios for ICU demand described in the previous section on ICU workload, we performed a grid search to determine the *optimal expansion policy*, depending on the decision variables b and $\bar{K}$. The mesh grid for $\bar{K}$ was set to be $[1, 2K_0]$ with a unit step and $[b_0, 1]$ for b with a step size equal to $\frac{1}{K_0}$, where $b_0 = \frac{P_0}{K_0}$ and P_0 and K_0 are the initial number of ICU patients and beds, respectively. For the first phase, the number of pre-Covid-19 critical care beds K_0 was taken for each region from the Italian "Ministero della Salute" 2018 Bulletin.[3] For the second wave, K_0 was equal to the number of beds available at the beginning of October 2020 in each region.

Costs. The cost parameters are the same for all three regions. The fixed cost of setting up each new ICU bed is $h_0 = €80,000$. This cost comprises the bed at €10,000–15,000; a ventilator costing €10,000–25,000, depending on the model; a monitor for at €10,000–20,000; mobile wall units, around €10,000 in total; at least five infusion pumps per bed at €2000 each; and shared equipment (ultrasound, bronchoscope, advanced hemodynamics, and hemofiltration machine) costing at least €10,000/bed and the unit maintenance cost $h_1 = €3000$/year per bed. The additional unit cost for patients, $h_2(P_t)$ is taken as an exponentially decreasing function, ranging from €392,000/year to €200,166/year to incorporate the medications and the remuneration of the resuscitating anesthesiologist and nurses. There is an inten-

sive care doctor every 6–10 patients and a nurse every two or three patients, the proportion dropping as the number of patients increases in critical times like a pandemic.

6.4 | The first Covid-19 wave

Efficient frontier. Figure 5 shows the efficient frontiers obtained for the three regions using the ICU load scenarios developed earlier and obtained by solving Equation (10). The three efficient frontiers are convex and decreasing as proved in Proposition 2. Lombardia and Piemonte are quite different from Veneto in needing a larger ICU capacity expansion. The number of expected patients denied ICU access corresponding to the minimum investment $V(M_0)$ is lower in Veneto than for Lombardia and Piemonte. This substantial difference depends on both the initial conditions and the severity of the pandemic. The number of ICUs per 100,000 inhabitants at the outset of the pandemic was 8.6, 7.4, and 10.1 in the three regions, respectively. Moreover, the maximum expected number of patients needing ICU, that is, $\mathbb{E}[\max_{t \in [0,T]} P_t]$, per 100,000 inhabitants is 14, 9, and 7 for the three regions, respectively.

Timing the decision to expand. Another important aspect of the proposed capacity optimization model is the "timing" of the optimal expansion strategy. The optimal decision time τ represents the time until which the decision maker can postpone action, reducing costs without affecting the expected number of denied patients. In our application, τ is determined by the critical loading level b^* in Equation (18). The critical loading level b^*—and hence the optimal decision time—depends on the initial ICU occupancy level b_0 and the preparation (setup) time d for capacity expansion (adding beds). Therefore, a larger setup time d means a lower critical value b^* to activate the decision-making for expansion. Therefore, as previously mentioned, b^* cannot fall below the initial ICUs occupancy level b_0, as in the case of decision dates of March 10 and October 30 in 2020 in our illustration. In practice, however, the decision-maker can find herself in two opposing scenarios. In the first scenario with b^* much greater than b_0, delays in preparing new beds do not affect the expected number of denied patients; in fact, a greater value of d is compensated by reducing b^* and the efficient frontier does not change with d. In the second scenario, where b^* approaches b_0, the setup delay parameter d does impact the number of patients denied ICU and, thus, the efficient frontier. The effect is more pronounced at high levels of resources, as shown in Figure 6, which reports the efficient frontiers of Lombardia and Piemonte for different values of d. As illustrated in Figure 6, in the case of Piemonte, an increase in the setup time from 5 to 10 days can be managed by reducing the critical loading level b^* without affecting the efficient frontier. However, the impact of the preparation time d on the efficient frontier is more pronounced for Lombardia, characterized by a larger value of the initial level b_0 with respect to Piemonte, that is 54% and 20%, respectively. Moreover, as illustrated in Figure 6, for larger values of d, the expected
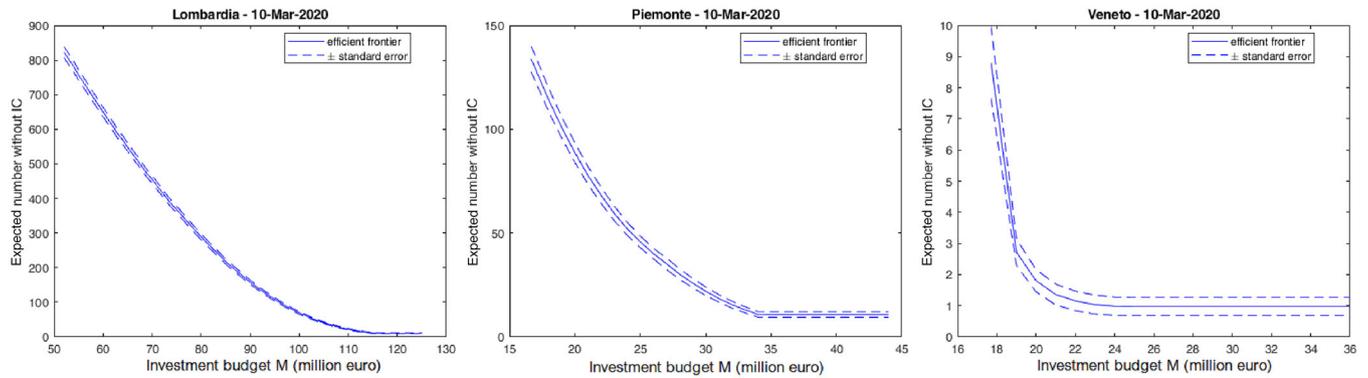
FIGURE 5 Efficient frontier: The minimum achievable expected number of patients denied ICU as a function of the investment budget $V(M)$ as of March 10, 2020, for the three regions. Setup time for additional capacity $d = 5$ days. (For Veneto, $\bar{t} = 10$ days for demonstration only.) IC, intensive care. [Color figure can be viewed at wileyonlinelibrary.com]
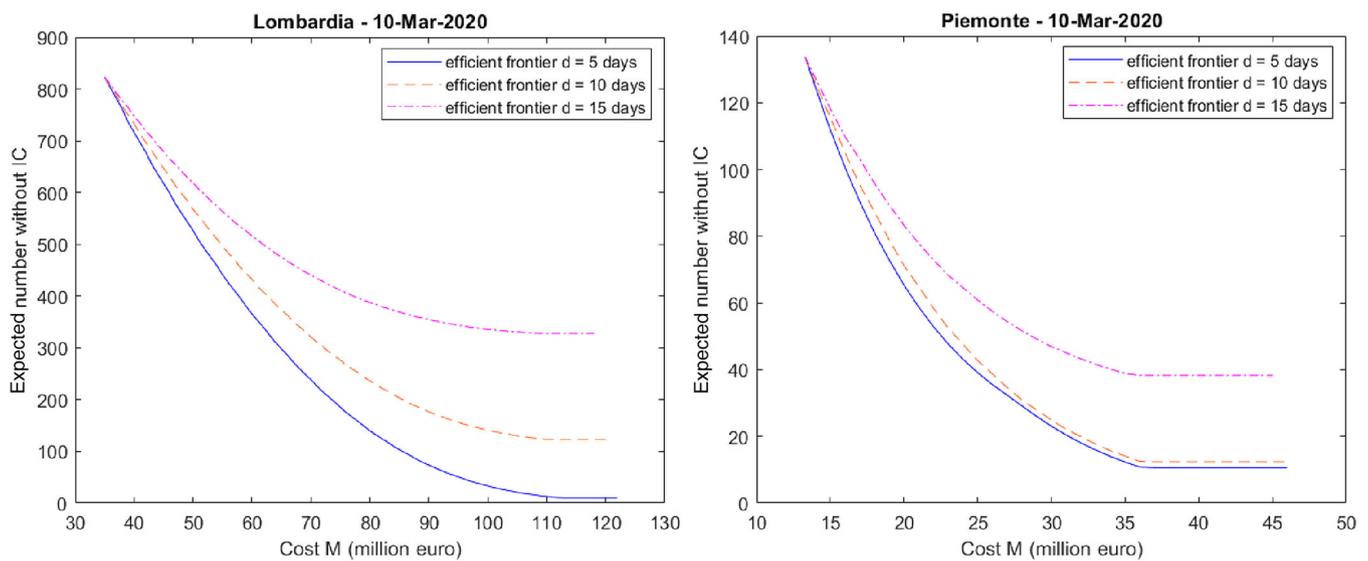

FIGURE 6 Efficient frontier: The minimum achievable expected number of patients denied ICU as a function of the investment budget $V(M)$ as of March 10, 2020, for Lombardia and Piemonte. Setup times for additional capacity is $d = 5$, 10, or 15 days. IC, intensive care. [Color figure can be viewed at wileyonlinelibrary.com]

number of patients denied ICU does not approach zero as the budget increases, that is, $\lim_{M \to \infty} V(M) > 0$ or equivalently $V(M^*) > 0$. This means that there is a threshold delay level d in setup time beyond which not all patients can be admitted, no matter how much resources are invested.

Shadow price of budget increase. Figure 7 reflects the solution of the problem in Equation (12) with the number of additional admitted patients for a given unit increase of the budget, $n(M)$. As expected, the function $n(M)$ is always decreasing in the resource level M. Table 6 shows the optimal investment level M^* as per Equation (14) and the corresponding optimal expansion policy $\pi(b^*, \bar{K}^*)$ for each region.

Contrasting our approach with a deterministic model. Table 6 compares the optimal investment and the optimal expansion policy obtained with our model to that obtained

from a deterministic model with a unique demand scenario equal to the mean value of P_t. Table 6 shows that we underestimate the necessary resources using this deterministic model, highlighting the importance of having a stochastic model. The difference in optimal resource level M^* is 21 million euros for Lombardia, 12 million euros for Piemonte, and 25 million euros for Veneto. The difference in the optimal expansion capacity $\bar{K}^*$ is 1.4 new beds per 100,000 inhabitants for Lombardia, 2.8 beds for Piemonte, and 2.2 beds for Veneto. Also, the timing of the expansion is affected by the choice of the model, this is particularly important for Piemonte, where the critical loading level b^* goes from 30% to 55%, leading to any delay in acting becoming dangerous. So, a stochastic model like ours is valuable compared to a deterministic one.

Sensitivity to a tolerated level of denial of access. Figure 8 shows the marginal cost per admitted patient in ICU $m(N)$,
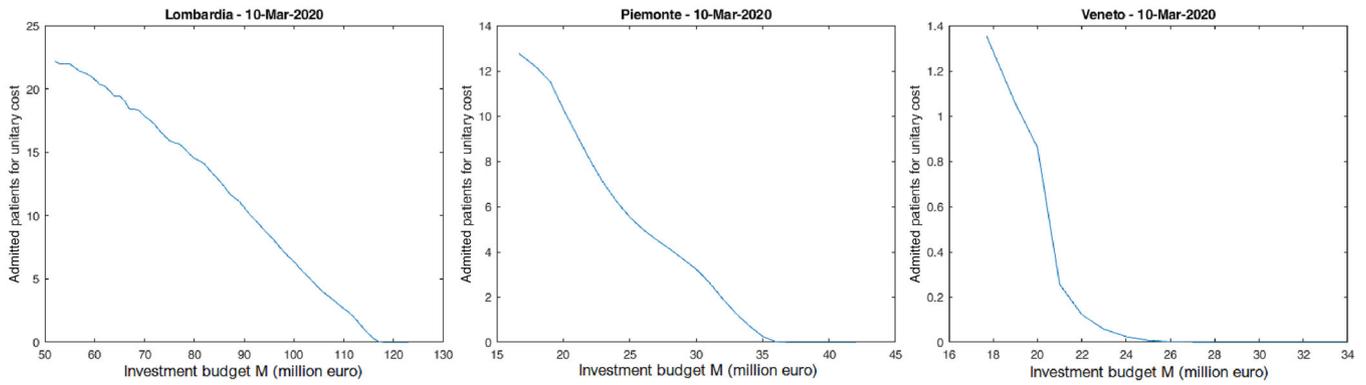
FIGURE 7 Shadow price: the increase in the number of admitted patients given a unit increase of budget $n(M)$ as of March 10, 2020, for the three regions. The capacity setup time $d = 5$ days. [Color figure can be viewed at wileyonlinelibrary.com]
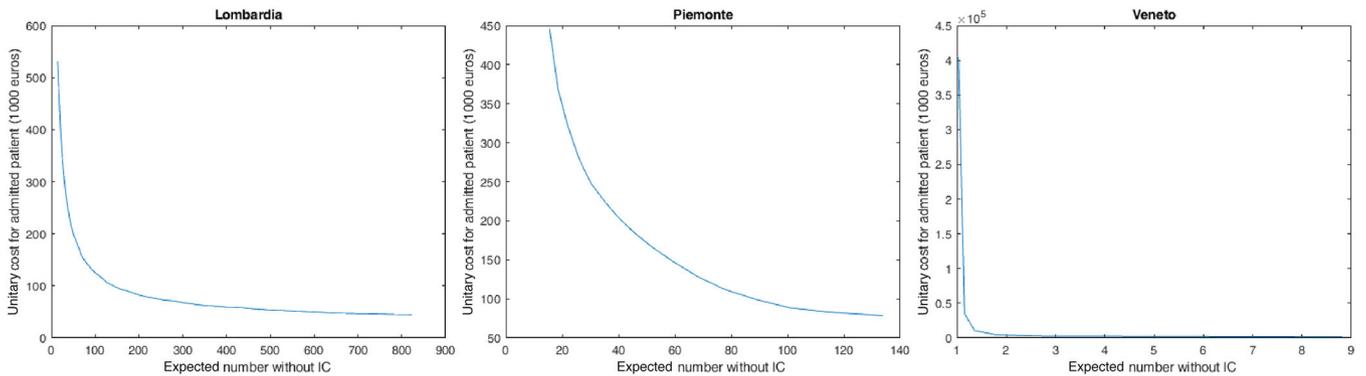

FIGURE 8 The marginal cost for admitted patients as a function of the tolerated loss $m(N)$ (see Equation 16). Preparation time $d = 5$ days. IC, intensive care. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 6 The optimal investment level M^*, the corresponding optimal number of new beds K^*, and the critical loading level b^* from the deterministic model underestimate those needed as per the stochastic model we used. Preparation time is $d = 5$ days. (For Veneto, $\bar{t} = 10$ days for demonstration only.).

Model	Region	M^* (million euro)	b^* (%)	K^* (per 100,000 inhabitants)
Stochastic	Lombardia	118	58	6.5
	Piemonte	36	30	4.5
	Veneto	25	44	2.2
Deterministic	Lombardia	97	54	5.1
	Piemonte	24	55	1.7
	Veneto	0	-	0

as defined in Equation (16). The function $m(N)$ grows indefinitely as the (tolerated) expected number of patients denied ICU N approaches the value $V(M^*)$ (or equivalently the value obtained by $\lim_{M \to \infty} V(M) > 0$). This implies that to push the number of patients denied ICU to its minimum value, the marginal cost is infinite, which is consistent with the number of additional admitted patients for unit investment $n(M) = 0$ for $M > M^*$.

6.5 | The second Covid-19 wave

After the first wave, Veneto expanded the number of ICU beds from 10.1 beds per 100,000 inhabitants to 16.8. Consequently, the region did not need to expand ICU capacity at the start of the second pandemic wave, as we also saw recommended by our models. The expansion carried out by Veneto was much larger than that suggested by our model (6.7 new beds instead of 2.2), incurring in additional and perhaps unnecessary costs. Lombardia increased the number of ICU beds from 8.6 per 100,000 inhabitants in February 2020 by 1.3 beds to 9.9 beds at the beginning of October 2020. During the first pandemic wave, the region reached a maximum of 14 ICU patients per 100,000 inhabitants by using temporary ICUs carved out of operating theatres, inpatient posts, or even military field hospitals. These beds were dismantled by the end of the first wave, leading to a shortfall of ICU capacity again in the second wave. Our model suggested a structural expansion of 6.5 new beds per 100,000 inhabitants, larger than the actual one of 1.3 (Table 6). Similar reasoning can also be applied to Piemonte. Before the second wave, Piemonte expanded the structural ICU capacity from 7.4 by 1.0 beds to 8.4 ICU beds per 100,000 inhabitants. Our model suggested 4.5 new beds per 100,000
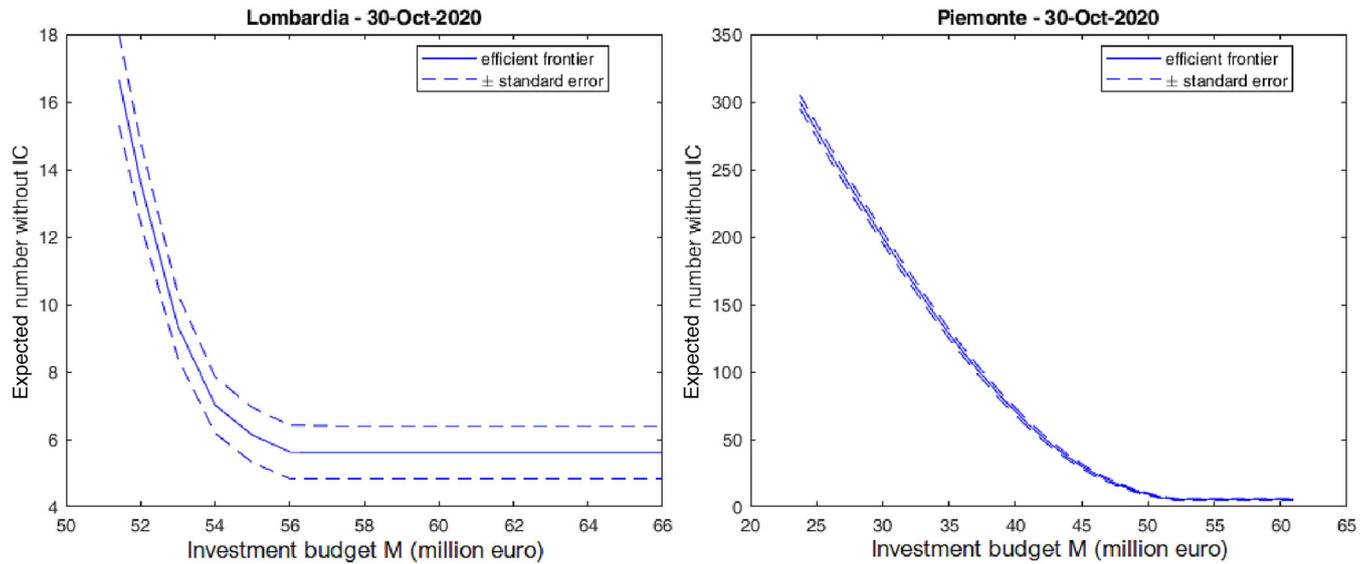
FIGURE 9 Efficient frontier: The minimum achievable expected number of patients denied ICU as a function of the investment budget $V(M)$ as of October 30, 2020, for Lombardia and Piemonte. Preparation time $d = 5$ days. IC, intensive care. [Color figure can be viewed at wileyonlinelibrary.com]
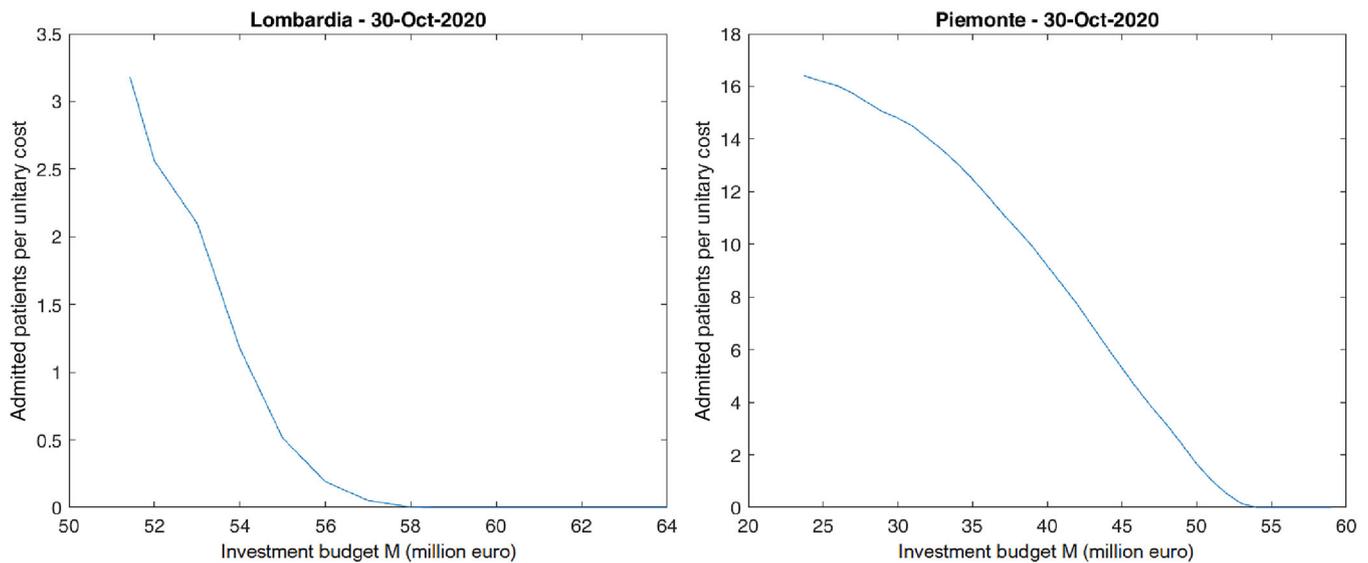

FIGURE 10 Shadow price: the increase in the number of admitted patients given a unit increase of budget $n(M)$ as of October 30, 2020, for Lombardia and Piemonte. Preparation time $d = 5$ days. [Color figure can be viewed at wileyonlinelibrary.com]

inhabitants would have been ideal (Table 6). Figure 9 shows the efficient frontiers from Equation (10) for Lombardia and Piemonte regions, respectively, given the projected demand scenarios explained in the previous section. The two efficient frontiers are convex and decreasing (Proposition 2). The resource level ranges from €52 million–66 million for Lombardia and from €20 million–60 million for Piemonte. The minimum expected cost M_0, necessary to maintain the initial number of intensive care beds without any expansion, depends on the daily numbers of occupied ICU beds through the marginal cost h_2, as in Equation (9). The average value of M_0 is €52 million for Lombardia and €24 million for Piemonte.

TABLE 7 The optimal investment level M^*, the corresponding optimal number of new beds K^* per 100,000 inhabitants in the region, and the critical loading level b^*. Preparation time is $d = 5$ days.

Region	M^* (million euro)	b^*	K^* (per 100,000 inhabitants)
Lombardia	58	80%	0.6
Piemonte	53	59%	6.1

Figure 10 shows the number of admitted patients for a unit increase of budget obtained by solving the dual problem in Equation (12). Table 7 shows the optimal investment level M^*

and the corresponding optimal expansion policy $\pi(b^*, \bar{K}^*)$ in either region.

7 | DISCUSSION

This paper has introduced a general modular approach for capacity expansion under uncertainty in the early stages of a pandemic to meet the expected demand for the resources in the coming weeks. We used the approach—in retrospect—to three regions in Italy for the two waves of the pandemic between February 2020 to January 2021. The application illustrated the inputs and techniques needed and the output obtained for decision-makers. In the Covid-19-related operations management literature, the emphasis is typically on the uncertainty of future demand. However, modeling this uncertainty is not straightforward during the *early* stage of a pandemic. There are a number of challenges at this stage: (1) noisy data from which to estimate parameters, (2) incorporating uncertainty in an optimization framework, and (3) the impact of uncertainty on optimal management decisions. Our work contributes by addressing these challenges. First, we proposed a robust estimation and forecasting procedure of epidemic and demand models, including both the variability due to noisy and limited data (through Monte Carlo sensitivity analysis of the epidemic parameters) and the inherently stochastic dynamics of epidemic and demand processes. Second, we formalized a stochastic dynamic optimization problem that incorporates the uncertainty into the management decision process. Finally, we showed that the uncertainty of future scenarios has a material impact on the optimal results. For instance in the first pandemic wave, for the three Italian regions, we compare the optimal ICU expansion policies obtained with the full model to a deterministic model that considers only the average future demand scenario. We conclude from the large differences that the model error is not negligible and, therefore, having a stochastic extension of the model is valuable.

7.1 | Research implications

Managing health emergencies such as the recent Covid-19 epidemic requires both supply and demand responses. This paper is focused on the supply side, providing an approach to capacity expansion of scarce resources like ICUs in the early stages of a pandemic. Further research may extend this approach to optimally manage the policy levers—the use of lockdown, isolation, quarantine, and vaccination—on the demand side. Note that *our optimal solution is a policy, not a single-number solution*. The expansion is contingent on the filling up of capacity, so if demand-side responses "flatten" the rise in the number of daily cases, the capacity expansion would be delayed as the threshold for expansion would not be met. On the demand side, assessing how delays in imposing restrictions impact the availability and the cost of ICU

capacity is possible by integrating our model with what–if analyses (see, for instance, Ferguson et al., 2020) or with compatible optimization problems (see Shahmanzari et al., 2022). In particular, integrating demand and supply sides in a unique optimization framework present interesting theoretical and computational challenges. Moreover, new variants of Covid-19 emerged even as vaccination drives are underway in many countries, including Italy. There is a need to investigate how to modify the stochastic epidemic model to incorporate vaccination drives with different (and possibly deteriorating) efficacy and new variants. The overall modular approach would still be helpful. Another policy lever is allocating the scarce ICU capacity to needy patients, which requires straightforward prioritization rules for uniform implementation across all hospitals in the region. Extending our approach to that of Lu et al. (2021) can lead to a more comprehensive discussion of the supply side levers, including prioritization of access. Finally, future research could investigate the long-term consequences of the steep decline in ICU admissions of uninfected patients during the pandemic. For instance, researchers could analyze mortality data of oncological or cardiopathy patients.

7.2 | Implications for practice

During a public health emergency response, leaders must make several critical decisions in a rapidly changing environment with limited data. We have shown that even with imperfect data from only the first few days of a pandemic, along with known cost and setup time parameters for expanding resources, our method can provide an efficient frontier of investment needed to limit the denial of beds to patients. Decision-makers can then decide the acceptable trade-off between investment in additional resources and patients denied care. Moreover, our model requires only a small number of variables. Therefore, decision-makers can use our approach to capture the effect of the uncertainty in projecting future demand during the early days of a pandemic. How the high uncertainty in the future spread of infections is modeled is crucial, and our approach provides much more insight than a simplistic what–if analysis on a deterministic model with arbitrary parameter ranges. Additionally, decision-makers would also make decisions at the right time. Our approach captures the nontrivial relationships between the setup time for capacity expansion and the optimal intervention time just as it does between the setup time and the expected number of patients who might be denied access to life-critical resources. Regarding timing, while an early decision on capacity expansion intuitively denies ICU access to fewer patients, our model also shows when delays in expansion can be accepted without affecting the expected number of admitted patients. This is particularly useful as the passage of time provides better information to the decision-maker. Finally, our approach suggests that in future pandemics the chaotic decision-making observed

during the initial days of successive waves of Covid-19 can be replaced by rational decision-making with model-based information that our approach provides.

ORCID

Anna Maria Gambaro https://orcid.org/0000-0002-5129-3929
ManMohan S. Sodhi https://orcid.org/0000-0002-2031-4387

ENDNOTES

[1] PCM-DPC-Covid-19 data (https://github.com/pcm-dpc/Covid-19).

[2] Rees et al. (2020), with their review of eight articles, report ALOS between 5 and 19 days. Lapidus et al. (2020) use a statistical procedure considering patients still in ICU on the date of estimation and with calculations based solely on the patients already discharged. They report ALOS of 23 days (18–30 days, with 95% confidence), much larger than earlier (biased) estimates (Rees et al., 2020).

[3] http://www.dati.salute.gov.it/dati/dettaglioDataset.jsp?menu=dati&idPag=17&fbclid=IwAR23MB90rEn2Xvb49uYo9NrHM2x0PMoAeXBcbUu-sx7BKVQ3R6WOcGIRy8Q

REFERENCES

Alban, A., Chick, S. E., Dongelmans, D., Vlaar, A., & Sent, D. (2020). ICU capacity management during the Covid-19 pandemic using a process simulation. *Intensive Care Medicine*, 46, 1624–1626.

Anderson, R., & May, R. (1991). *Infectious diseases of humans*. Oxford University Press.

Andersson, H., & Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*. Springer-Verlag.

Bertsekas, P. (2007). *Dynamic programming and optimal control* (3rd ed., Vol. 2). Athena Scientific.

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, J., Lami, O. S., Stellato, B., T Bouardi, H., V Carballo, K., … Zeng, C. (2021). From predictions to prescriptions: A data-driven response to Covid-19. *Health Care Management Science*, 24, 253–272.

Boxma, O., Kella, O., & Mandjes, M. (2019). Infinite-server systems with Coxian arrivals. *Queueing Systems*, 92(3), 233–255.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Brusaferro, S., D'Errico, M. M., Montagna, M. T., Pasquarella, C., Tardivo, S., Arrigoni, C., Fabiani, L., Laurenti, P., Mattaliano, A. R., Orsi, G. B., Squeri, R., Torregrossa, M. V., Mura, I., & Collaborators. (2018). Epidemiology of intensive care unit-acquired sepsis in Italy. *Annali di Igiene*, 30(5), 15–21.

Calafiore, G. C., Novara, C., & Possieri, C. (2020). A time-varying SIRD model for the COVID-19 contagion in Italy. *Annual Review in Control*, 50, 361–372.

Chatterjee, S., Sarkar, A., Chatterjee, S., Karmakar, M., & Paul, R. (2020). Studying the progress of Covid-19 outbreak in India using SIRD model. *Indian Journal of Physics*, 95(9), 1941–1957.

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

Chen, Z., & Kong, G. (2022). Hospital admission, facility-based isolation, and social distancing: An SEIR model with constrained medical resources. *Production and Operations Management*. Advanced online publication. https://doi.org/10.1111/poms.13702

Chowell, G., Nishiura, H., & Bettencourt, L. (2007). Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *Journal of the Royal Society, Interface*, 4(12), 155–166.

Cori, A., Ferguson, N. M., Fraser, C., & Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9), 1505–1512.

Cuevas, E. (2020). An agent-based model to evaluate the Covid-19 transmission risks in facilities. *Computers in Biology and Medicine*, 121, 103–827.

D'Arienzo, M., & Coniglio, A. (2020). Assessment of the SARS-COV-2 basic reproduction number, R_0, based on the early phase of Covid-19 outbreak in Italy. *Biosafety and Health*, 2(2), 57–59.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1, 54–75.

Ehrgott, M. (2005). *Multicriteria optimization*. Lecture Notes in Economics and Mathematical Systems, Vol. 491. Springer Science & Business Media.

Evgeniou, T., Fekom, M., Ovchinnikov, A., Porcher, R., Pouchol, C., & Vayatis, N. (2022). Pandemic lockdown, isolation, and exit policies based on machine learning predictions. *Production and Operations Management*. Advanced online publication. https://doi.org/10.1111/poms.13726

Favier, C., Degallier, N., Rosa-Freitas, M., & Tsouris, P. (2020). Early determination of the reproduction number for vector-borne diseases: The case of dengue in Brazil. *Tropical Medicine and International Health*, 11, 332–340.

Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunuba Perez, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L., Van Elsland, S., … Ghani, A. (2020). Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce Covid-19 mortality and healthcare demand. Imperial College London. https://doi.org/10.25561/77482

Ferrari, L., Gerardi, Manzi, G., Micheletti, A., Nicolussi, F., Biganzoli, E., & Salini, S. (2021). Modeling provincial Covid-19 epidemic data using an adjusted time-dependent SIRD model. *International Journal of Environmental Research and Public Health*, 18(12), 6563.

Gonçalves, P., Ferrari, P., Crivelli, L., & Albanese, E. (2022). Model-informed health system reorganization during emergencies. *Production and Operations Management*. Advanced online publication. https://doi.org/10.1111/poms.13710

Heemskerk, M., Mandjes, M., & Mathijsen, B. (2022). Staffing for many-server systems facing non-standard arrival processes. *European Journal of Operational Research*, 296, 900–913.

Heemskerk, M., van Leeuwaarden, J., & Mandjes, M. (2017). Scaling limits for in-finite-server systems in a random environment. *Stochastic Systems*, 7(1), 1–31.

Jain, A., & Rayal, S. (2022). Managing medical equipment capacity with early spread of infection in a region. *Production and Operations Management*. https://doi.org/10.1111/poms.13684

Jiang, H., & Sodhi, M. S. (2019). Analyzing the proposed reconfiguration of accident-and-emergency facilities in England. *Production and Operations Management*, 28(7), 1837–1857.

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

Kilic, M., Yüzkat, N., Soyalp, C., & Gülhas, N. (2019). Cost analysis on intensive care unit costs based on the length of stay. *Turkish Journal of Anaesthesiology and Reanimation*, 47(2), 142–145.

Kniesner, T. J., & Viscusi, W. K. (2005). Value of a statistical life: Relative position vs. relative age. *American Economic Review*, 95(2), 142–146.

Lapidus, N., Zhou, X., Carrat, F., Riou, B., Zhao, Y., & Hejblum, G. (2020). Biased and unbiased estimation of the average length of stay in intensive care units in the Covid-19 pandemic. *Annals of Intensive Care*, 10(1), 135.

Li, M. K., Sodhi, M. S., Tang, C., & Yu, J. (2023). Preparedness with a system integrating inventory, capacity, and capability for future pandemics and other disasters. *Production and Operations Management*, 32, 564–583.

Lu, Y., Guan, Y., Zhong, X., Fishe, J. N., & Hogan, T. (2021). Hospital beds planning and admission control policies for COVID-19 pandemic: A hybrid computer simulation approach. In *17th IEEE International Conference on Automation Science and Engineering (CASE 2021)* (pp. 956–961). IEEE.

Luss, H. (1982). Operations research and capacity expansion problems: A survey. *Operations Research*, 30(5), 907–947.

Ma, J. (2020). Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling*, 5, 129–141.

Mamon, G. A. (2020). Regional analysis of Covid-19 in France from fit of hospital data with different evolutionary models. https://doi.org/10.48550/arXiv.2005.06552

McCulloh, I., Kiernan, K., & Kent, T. (2020). Inferring true COVID-19 infection rates from deaths. *Frontiers in Big Data*, 3, 565589.

McManus, M. L., Long, M. C., Cooper, A., & Litvak, E. (2004). Queuing theory accurately models the need for critical care resources. *Anesthesiology*, 100(5), 1271–1276.

Naderi, B., Roshanaei, V., Begen, M., Aleman, D., & Urbach, D. (2021). Increased surgical capacity without additional resources: Generalized operating room planning and scheduling. *Production and Operations Management*, 30(8), 2608–2635.

Ouyang, H., Argon, N., & Ziya, S. (2020). Allocation of intensive care unit beds in periods of high demand. *Operations Research*, 68(2), 591–608.

Perkins, T., & Espana, G. (2020). Optimal control of the Covid-19 pandemic with non-pharmaceutical interventions. *Bulletin of Mathematical Biology*, 82, 118.

Rees, E. M., Nightingale, E. S., Jafari, Y., Waterlow, N. R., Clifford, S., Pearson, C. A. B., Jombart, T., Procter, S. R., Knight, G. M., & CMMID Working Group. (2020). Covid-19 length of hospital stay: A systematic review and data synthesis. *BMC Medicare*, 18, 270.

Ridge, J. C., Jones, S. K., Nielsen, M., & Shahania, A. K. (1998). Capacity planning for intensive care units. *European Journal of Operational Research*, 105(2), 346–355.

Rydén, T. (1996). An EM algorithm for estimation in Markov-modulated Poisson processes. *Computational Statistics & Data Analysis*, 21, 413–447.

Shahmanzari, M., Tanrisever, F., Eryarsoy, E., & Şensoy, A. (2022). Managing disease containment measures during a pandemic. *Production and Operations Management*. Advanced online publication. https://doi.org/10.1111/poms.13656

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.

Shi, P., Chen, J., Lim, C., Parker, J., Tinsley, R., & Cecil, J. (2022). Operations (management) warp speed: Rapid deployment of hospital-focused predictive/prescriptive analytics for the COVID-19 pandemic. *Production and Operations Management*. Advanced online publication. https://doi.org/10.1111/poms.13648

Simon, C. P., & Blume, L. (1994). *Mathematics for economists*. W. Norton & Co.

Smith-Daniels, V., Schweikhart, S., & Smith-Daniels, D. (1988). Capacity management in health care services: Review and future research. *Decisions Sciences*, 19(4), 889–919.

Tan, S. S., Bakker, J., Hoogendoorn, M. E., Kapila, A., Martin, J., Pezzi, A., Pittoni, G., Spronk, P. E., Welte, R., & Hakkaart-van Roijen, L. (2012). Direct cost analysis of intensive care unit stay in four European countries: Applying a standardized costing methodology. *Value in Health*, 15(1), 81–86.

Thompson, R., Stockwin, J., van Gaalen, R., Polonsky, J. A., Kamvar, Z. N., Demarsh, P. A., Dahlqwist, E., Li, S., Miguel, E., Jombart, T., Lessler, J., Cauchemez, S., & Cori, A. (2019). Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics*, 29, 100356.

Van-Mieghem, J. (2003). Capacity management, investment, and hedging: Review and recent developments. *Manufacturing & Service Operations Management*, 5(4), 269–302.

Wood, R. M., McWilliams, C. J., Thomas, M. J., Bourdeaux, C. P., & Vasilakis, C. (2020). Covid-19 scenario modelling for the mitigation of capacity-dependent deaths in intensive care. *Health Care Management Science*, 23(3), 315–324.

Worthington, D., Utley, M., & Suen, D. (2020). Infinite-server queueing models of demand in healthcare: A review of applications and ideas for further work. *Journal of the Operational Research Society*, 71(8), 1145–1160.