The Hybrid Hospital: Balancing On-Site and Remote Hospitalization

Noa Zychlinski[1], Gal Mendelson[1], Andrew Daw[2]
[1] Faculty of Data and Decision Sciences, Technion – Israel Institute of Technology, Haifa 3200003, Israel
[2] Marshall School of Business, University of Southern California, 3670 Trousdale Pkwy, Los Angeles, CA 90089, USA
noazy@technion.ac.il, galmen@technion.ac.il, dawandre@usc.edu
Dec 7, 2023

Problem definition: We study the dynamics of hybrid hospitals offering on-site and remote hospitalization through telemedicine. These new healthcare models require efficient operational policies to balance costs, efficiency, and patient well-being. Our study addresses two primary operational questions: (i) how to direct patient admission and call-in policies based on individual characteristics and proximity and (ii) how to determine the optimal allocation of medical resources between these two hospitalization options.

Methodology/results: We develop a model that uses Brownian Motions to capture the patient's health evolution during remote/on-site hospitalization. By optimizing call-in policies, we find that remote hospitalization is cost-effective only for moderately distant patients, where the call-in threshold has a non-monotonic relationship with travel time. Additionally, we find that the impact of scarce resources is reflected through simultaneous increase of both remote and on-site costs by the same value, without altering the solution structure under abundant resources. Lastly, we identify a non-monotonic relationship between the total medical resources and the workload allocation, depending on the recovery rates and the hospital proximity.

Managerial implications: Contrary to the widely held view that telemedicine can mitigate rural and non-rural healthcare disparities, our research suggests that on-site care may actually be more cost-effective than remote hospitalization for patients in distant locations, due to increased risks for remote patients who are called in to the hospital. This finding may be of particular concern in light of the growing number of "hospital deserts" amid recent rural hospital closures, as these communities may in fact not be well-served through at-home care. Such insights on cost-effectiveness, proximity, and patient deterioration under remote care can guide healthcare decision-makers and policymakers in shaping future healthcare delivery and design.

Key words: Stochastic modeling, healthcare operations management, telemedicine, resource sharing

1. Introduction

The COVID-19 pandemic has significantly propelled the adoption of virtual services, with telemedicine now playing a prominent role in the realm of healthcare (Bokolo 2020, Kadir 2020). Telemedicine facilitates the remote delivery of clinical services through real-time communication, connecting patients and healthcare providers via video conferencing and remote monitoring (Monaghesh and Hajizadeh 2020). These virtual services offer several advantages, such as cost savings related to travel and reduced exposure to diseases, which may ultimately enhance the efficiency of healthcare delivery (Hur and Chang 2020).

Recent advancements in telemedicine now enable sophisticated remote medical services, including home hospitalization as an alternative to traditional on-site care. Sheba Beyond, a pioneering virtual hospital affiliated with Sheba Medical Center and thus ranked among the world's top medical systems by Newsweek, offers remote examination, monitoring, and online rehabilitation programs. Their goal is to enhance accessibility to top-tier medical expertise for all prospective patients, aligning with the prediction that remote hospitalization will become a widespread offering among major hospital networks. Indeed, virtual hospitals are becoming popular across the world, such as in Australia (Hutchings et al. 2021), China (Francis et al. 2021), and the United States. For example, in the US, this trend is well underway: 186 hospitals participated in the "Acute Hospital Care at Home" program during its inaugural year (Clarke et al. 2021), which permitted Medicare-certified hospitals to deliver inpatient-level care to patients within the comfort of their homes. A recent McKinsey & Company comprehensive report stated that virtual hospitals have the potential to provide significant relief to overburdened healthcare systems. In particular, they project that virtual hospitals could unlock bed capacity, reduce the need to build new hospitals and save hundreds of millions of dollars (Boldt-Christmas et al. 2023). The American Hospital Association (AHA) has similarly promoted the concept through the *Hospital-at-Home* components of their "Value Initiative" public cost-reduction campaign (Americal Hospital Association, 2020).

Often times, such campaigns naturally associate the potential benefits of home hospitalization with rural patients. Frequently referred to as "hospital deserts" due to their significant distance

from healthcare centers, many rural communities face healthcare accessibility challenges worldwide, affecting millions of individuals in large countries like the United States, China, Brazil, and England (Behrman et al. 2021, Jiao et al. 2021, Gong et al. 2021, Noronha et al. 2020, Verhagen et al. 2020). These under-served areas lack proximity to medical facilities, leading to delays in seeking care, limited access to timely interventions, and increased health risks. Transportation hurdles further complicate the problem, as rural residents must contend with limited options and lengthy journeys, often resulting in worsened health conditions by the time they reach a hospital (Kelly et al. 2014). Additionally, as highlighted in a recent report by the Center for Healthcare Quality and Payment Reform (CHQPR), over 600 rural hospitals in the United States, representing more than 30% of the nation's rural healthcare facilities, face the risk of closure (Adams 2023).

This tenuous state of rural healthcare offerings is compounded by growing evidence of a disparity in the health conditions of people in rural areas relative to those who live in non-rural areas (Lewis 2022). For example, in the United States, data from the National Vital Statistics System has shown that in the two decades from 1999 to 2019, although the overall death rate (deaths per 100,000) has declined, a widening gap has emerged between the rates of death in rural and non-rural communities (Curtin and Spencer 2021). Even more troublesome is that this gap is consistent across the 10 leading causes of death, with the widest disparities occurring in the fatality rates for heart disease, cancer, and chronic lower respiratory diseases. Furthermore, these trends are consistent when controlling for demographic factors like age, race, and sex (Cross et al. 2021). Similarly, data shows that the rate of death from the COVID-19 pandemic in non-metropolitan areas has out-paced the same rate in metropolitan ones (Ullrich and Mueller 2023). This heightened deadliness of serious disease in rural communities in the US is coupled with the noted growth of addiction, overdoses, and suicide (so-called "deaths of despair," Case and Deaton 2015, 2017) and increased mortality of unintentional injuries, such as from traffic and firearms (Olaisen et al. 2019).

While telemedicine networks have shown promise in improving healthcare in rural areas (Ishfaq and Raja 2015), and could also potentially serve as a viable alternative to mitigate the impact

of hospital closures, our paper underscores a critical issue: patients residing in remote areas, who ostensibly stand to gain the most from home hospitalization, also face the highest risks when called-in to the hospital. Such calls may result in patients arriving in deteriorated states, ultimately leading to prolonged and more expensive hospitalizations. Hence, hybrid hospitals that offer both on-site and remote hospitalization services present new operational challenges, necessitating the development of innovative models and policies that ensure cost-effectiveness while maintaining the highest standard of patient care.

More specifically, our study focuses on a hybrid hospital setting that incorporates a virtual Emergency Department (ED), which patients can access when they experience illness. In this model, medical professionals conduct remote examinations and consultations with patients. Subsequently, based on their assessment, doctors decide whether to admit the patient for remote hospitalization or advise immediate travel to the hospital for on-site admission. For patients admitted remotely, their physical examination is conducted using telehealth technologies, such as TytoCare®, a digital platform specifically designed for remote physical assessments (Barkai et al. 2022). During these examinations, both data and visual information are recorded and transmitted to the physician. Then, a summary of the visit is provided, which may include orders for blood tests, medication orders and instructions. Remotely admitted patients have two potential outcomes: recovery with subsequent discharge, or, in the event of health deterioration, they are advised to travel to the hospital for on-site admission and the continuation of their treatment. We refer to this as a "call-in" scenario.

Therefore, the first fundamental question we address is how to optimally set the call-in policy so as to minimize the total operational cost. That is, based on each patient's characteristics, decide whether to admit the patient remotely or on-site. In the former case, one also needs to decide at which health condition to call the patient in to the hospital. The marginal improvement cost in each hospitalization option, as well as patient's proximity to the hospital and anticipated further deterioration while traveling—all play important roles in these decisions.

The second question we address in this paper is related to the way the hybrid hospital/ward allocates its resources. In Sheba Beyond, the medical staff of each hybrid ward is divided into two teams, each is responsible either for remote or on-site hospitalized patients, which is what we assume throughout this paper. Therefore, the question is how to allocate these resources to these two groups. This decision goes hand in hand with the call-in policy, since the decision on when to call in patients determines the workload for each group.

To address these two questions, we introduce a model that captures patients' health condition via an acuteness "score" that aggregates clinical measurements for supporting discharge decisions. Such scores are common practice. The Aldrete system, for example, is an acuteness score to determine readiness for discharge post-surgery (Aldrete 1994); other scores were developed for specific diseases such as pneumonia (e.g., Capelastegui et al. 2008), for cardiac patients (the Anderson-Wilkins acuteness score; Anderson et al. 1992), or for patient assessment in SNFs and rehabilitation facilities (e.g., the ADL score; Bowblis and Brunt 2014). We capture the system's dynamics by modeling the individual evolution of the patient's health condition through remote and on-site hospitalization using Brownian Motions (BMs) whose parameters depend on patients' characteristics. That allows us to capture the fact that patient's health score improves, on average, while being treated, yet may nevertheless deteriorate due to the randomness in recovery across patients. The relevant properties in our analysis are hitting time statistics—averages and probabilities—that determine length of stay (LOS) in both hospitalization options and the call-in likelihood due to deterioration at home.

Our work sheds light on the complex operational aspects in managing hybrid hospitals. The questions we address in this paper are ones of *design*. Rather than taking the service content at each location as fixed, we optimize it to meet system-level goals by setting the treatment mix of each patient profile as well as the allocation of resources between the two hospitalization locations.

The following are the main contributions of this paper:

- As a modeling contribution, we study the operations, design, and management of an innovative acute-care system comprising both on-site and remote hospitalization. By capturing the random

dynamics of patients' health scores and this evolution's dependence on the manner of care, we provide a practical framework for determining the optimal treatment blend and call-in policy based on individual patient characteristics and their travel time to the hospital. We explicitly address critical questions that hinge on two essential factors: (i) the disparity in marginal hospitalization costs between on-site and remote hospitalization, and (ii) the feasibility of call-in as opposed to exclusive remote hospitalization.

- For policy-level insights, we find that remote hospitalization will be cost-effective only for patients residing at a moderate distance from the hospital, and only if their marginal hospitalization cost exceed those at the hospital. Our model identifies that the optimal call-in threshold is non-monotonic as a function of patient distance; furthermore, the range of distances for which remote hospitalization is viable will shrink for patients with poorer health scores.

Both intuition and the nascent public policy around home hospitalization suggests that remote hospitalization and telemedicine could offer a geographic panacea for health outcomes, enhancing healthcare access in remote regions and bridge the disparities between rural and non-rural areas. Our research, however, demonstrates that due to the increased risk of deterioration and lengthy travel times to the hospital, it could be preferable for the hospital to direct distantly located patients to on-site care.

In light of the broadly-documented evidence of worse baseline health conditions in rural communities, we discuss how our model's insights caution against the prevailing assumption that remote hospitalization would benefit rural patients.

- To provide insights for managing both on-site and remote hospitalization, we focus on the development of a hybrid hospital model, where a fixed amount of resources, such as medical professionals, must be distributed between the two modes of care. Three distinct behavioral cases, contingent on the ratio of marginal improvement rates, are characterized within the system's workload and feasibility region: they both can exhibit an increasing, decreasing, or unimodal pattern in relation to the call-in threshold. Leveraging these findings, we subsequently derive insights into

optimal resource allocation. We find that the impact of scarce resources is a simultaneous increase of both remote and on-site cost rates by *the same* value, without altering the solution structure and properties from the case where resources are abundant. Notably, we observe that the optimal allocation of resources is non-monotone with the total amount of resources. In some cases, as the total resource pool becomes more limited, a larger proportion may be allocated to one hospitalization option while reducing the allocation to the other. This highlights the dynamic nature of resource allocation in hybrid hospitals.

The rest of the paper is organized as follows. Section 2 includes a brief review of the related literature. In Section 3, we introduce our model and the optimization problem. In Section 4, we include preliminary analyses on the system's workload and feasibility region. The main results of the paper are presented in Section 5. Lastly, in Section 6, we provide some concluding remarks and suggest a few directions for future research. All proofs appear in the appendix.

2. Literature Review

This paper is related to two main lines of literature. The first is applying stochastic modeling to study the operations of health services. The second is related to health progression modeling. We provide here a brief review of the related literature along these two streams.

Stochastic modeling and queueing models have been used to address various healthcare applications to derive operational insights and policies (e.g., Mills et al. 2013, Shi et al. 2016). One of the challenges in managing such complex healthcare systems is how to allocate scarce resources and prioritize patients over these resources (e.g., Sun et al. 2018). While classical models in queuing theory assume that service times are independent random variables with fixed distributions, empirical studies show that there is flexibility in setting transfer/discharge decisions in healthcare; these decisions, in turn, have an effect on patient outcomes (Kc and Terwiesch 2012, Bartel et al. 2020). Here, we build upon prior works that have shown the benefit of modeling in finer detail, such as the controlled queueing models studied in works like Hopp et al. (2007) and Chan et al. (2014).

Protocols for adaptive discharge of individual patients, from a single station, were developed in Shi et al. (2021), where a Markov decision process (MDP) is integrated with data to support discharge decisions from inpatient wards. They suggest an efficient dynamic heuristic that balances personalized readmission-risk prediction and ward congestion. Armony and Yom-Tov (2021) developed discharge rules specifically for hematology patients. For these, a longer hospital stay carries risk (infections) but also the ability to take care of such infections.

We go beyond a single-station analysis to study a new hospital setting—the hybrid hospital, which includes an on-site and remote hospitalization. Our focus includes decisions on patient hospitalization option, call-in thresholds for remote patients, and resource allocation, considering patient characteristics and distance from the hospital.

Different health progression models were developed to address operational questions. Shi et al. (2021), Deo et al. (2013) and Nambiar et al. (2020) explicitly modeled the individual patient progression by using a Markov chain progression model. Grand-Clément et al. (2020) used an MDP to describe the evolution of patients' health condition and derive a proactive transfer policy to a hospital Intensive Care Unit (ICU). Bavafa et al. (2019) modeled patient health dynamics using a Markovian continuous-time framework with three states: "healthy," "intermediate," and "sick." Bavafa et al. (2021) analyzed primary care delivery through e-visits where patients become sick after an office visit, necessitating another visit after a random period—a model with an increasing failure rate, linking longer intervals between visits to a higher sickness likelihood. More recently, Bavafa et al. (2022) introduced a model capturing patients' evolving health condition to study optimal discharge health, impacting readmission probability.

We also use a single aggregated health score to describe patients' health condition. Our model uses Brownian motion dynamics as the underlying mechanism to capture the dynamic evolution of health condition at each location. Being a BM model, it is fully characterized by its mean recovery speed (the drift) and variability (the diffusion coefficient), which can lead to deterioration. Modeling via drifted BMs has been used in sequential decision making and in the modeling of

healthcare decisions (Siegmund 2013, Wang et al. 2010). We use the BM health score progression model to answer macro-level design questions, around which further refinement, such as dynamic control for individual patients, can be done.

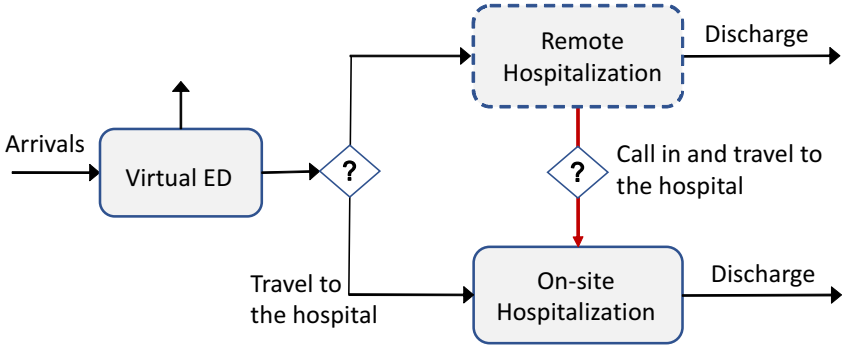
3. Modeling Hybrid Hospitalization and Patient Health Progression

Our modeling perspective in this work will operate on both micro- and macro-levels, capturing both the dynamics of each individual patient's health progression and the operational structure of the hybrid hospitalization network. Let us begin by describing the latter.

3.1. Two-Station Network of Remote and On-Site Hospitalization

Because the decision of whether to hospitalize a patient on-site or remotely must be made as soon as the patient is assessed, our hybrid hospital model begins after a triage through a virtual Emergency Department (ED). The full hospital network is depicted in Figure 1. After assessment at the virtual ED, patients can either be admitted remotely or advised to travel to the hospital for on-site admission. If admitted remotely, they either fully recover and are discharged, or, if their health condition worsens and reaches some predetermined threshold, they are called in to the hospital and complete their hospitalization on-site.

Figure 1 **Illustration of the hybrid hospital service network stations.**



We use the terms "health condition" or "health score" as measures of *clinical acuity*. The higher the score, the worse the health condition is. Patients arrive to the virtual ED stochastically according to a Renewal process with rate λ and an *initial health score* $x \in \mathbb{R}_+$. Upon arrival, a decision must be made as to whether to admit them remotely or on-site (following travel). If admitted

on-site, they remain there until full recovery. If admitted remotely, they stay there as long as their health score does not reach a call-in threshold $x + a$, $a \geq 0$. If a patient's health score reaches 0 before it reaches $x + a$, they are discharged. Otherwise, when their score reaches $x + a$, they travel to the hospital, where they are admitted and stay there until they are healthy. We denote the travel time to the hospital by T. Note that if $a = 0$, the patient is automatically admitted on-site, and if $a > 0$, they are automatically first admitted remotely. Hence, the call-in threshold a parsimoniously captures the health network's primary design decision: when should (or can) a patient be hospitalized remotely?

3.2. Stochastic Dynamics of the Individual Health Score

We model the evolution of patients health conditions through negative-drift BMs, which capture the recovery rates towards improvement during hospitalization as well as the randomness in recovery. Specifically, the health score of a patient during remote hospitalization is given by the process

$$\mathcal{B}^R(t) = x + \sigma_R B^R(t) - \theta_R t,$$

where $B^R(t)$ is a standard BM, $\theta_R > 0$ and $\sigma_R > 0$. Thus, $\mathcal{B}^R(t)$ is a negative-drift BM, starting at the initial score x, with drift $-\theta_R$ and diffusion coefficient σ_R.

While the improvement rate at home being positive implies that home-hospitalized patients tend toward recovery and discharge, randomness allows the health score to increase, meaning that the patient's condition can become more severe. If a remotely hospitalized patient's condition deteriorates too much, they are called in to the hospital and complete the treatment there. Let $a+x$, $a > 0$ denote the call-in threshold for a patient whose initial health score at admission was x. The remote hospitalization LOS is the first time a patient starting from health condition x reaches health condition 0 (healthy) or health condition $x + a$ (called in) and is given by

$$\tau_R(x, a) = \inf\{t \geq 0 : \mathcal{B}^R(t) = 0 \text{ or } \mathcal{B}^R(t) = a + x\}.$$

The call-in likelihood, $\mathbb{P}\left(\mathcal{B}^R(\tau_R(x, a)) = a + x\right)$, is the probability that a patient who starts at health condition x will reach health condition $a+x$ before reaching 0. The expected LOS of remote

hospitalization is $\mathbb{E}[\tau_R(x,a)]$. Both have well known explicit expressions from the solution to the "Gambler's ruin" problem involving a BM. We have

$$p_x(a) := \mathbb{P}\left(\mathcal{B}^R(\tau_R(x,a)) = a + x\right) = \frac{1 - e^{-\rho x}}{e^{\rho a} - e^{-\rho x}},$$

where we defined $\rho := 2\theta_R/\sigma_R^2 > 0$, and

$$\mathbb{E}[\tau_R(x,a)] = \frac{1}{\theta_R}\left(\left(1 - p_x(a)\right)x - p_x(a)a\right).$$

Patients who are called in have to travel to the hospital, and, naturally, their health condition may further degrade while traveling. We assume that their health score deteriorates according to a random variable, $Z(x,a,T)$ on $[0,\infty)$, whose expected value is $T\theta_T$ for $\theta_T > 0$. Therefore, the patient's health score at arrival to the hospital is $x + a + Z(x,a,T)$.

The model dynamics at the hospital are similar to the remote case, but with the difference of the initial starting health score being random, dependent on the patient's condition after the transit. The patient's health score's evolution is determined by

$$\mathcal{B}^H(t) = x + a + Z(x,a,T) + \sigma_R B^H(t) - \theta_H t,$$

where $B^H(t)$ is a standard BM, $\theta_H > 0$ and $\sigma_H > 0$. We assume that the arrival process, B^R, Z and B^H are independent. Define

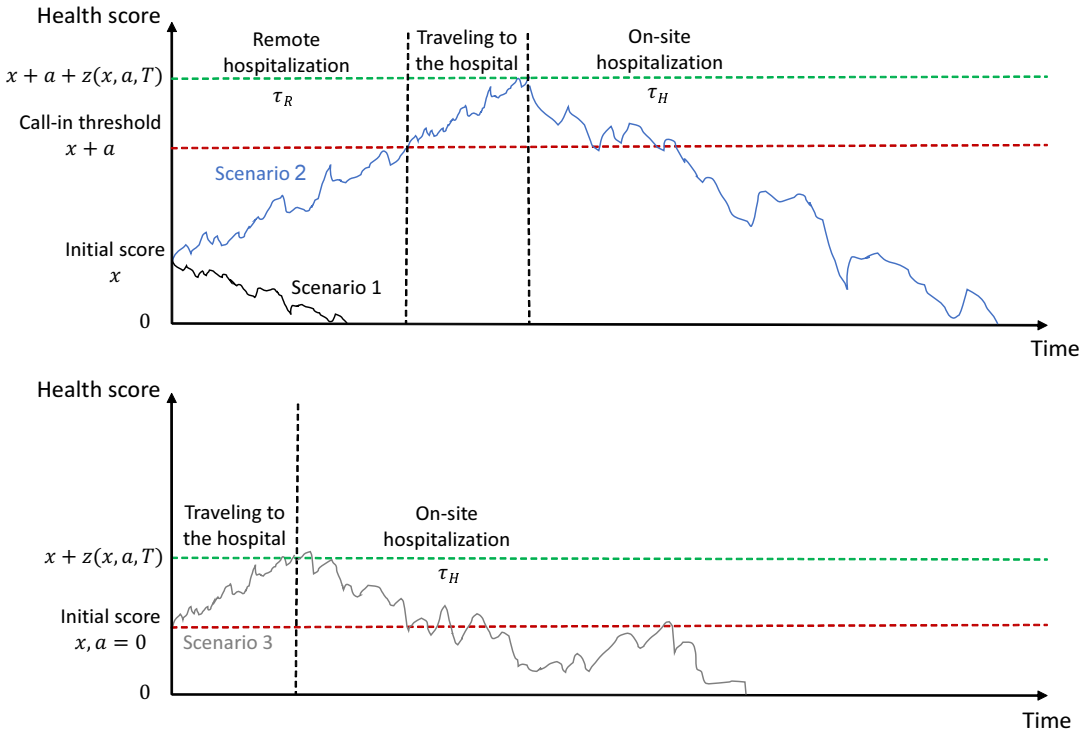
$$\tau_H(x,a,Z) = \inf\{t \geq 0 : \mathcal{B}^H(t) = 0\},$$

to be the patient's LOS at the hospital. Given $Z = Z(x,a,T)$, $\tau_H(x,a,Z)$ is the time it takes a BM with a negative drift $-\theta_H$, starting at $a + x + Z$ to hit zero. The expected LOS at the hospital is therefore

$$\mathbb{E}[\tau_H(x,a,Z)] = \mathbb{E}\left[\mathbb{E}[\tau_H(x,a,Z) \mid Z]\right] = \frac{1}{\theta_H}\mathbb{E}\left[(x + a + Z(a,x,T))\right] = \frac{1}{\theta_H}\left(x + a + T\theta_T\right).$$

Finally, we complete the model by encoding a required clinical constraint, which enforces that the hospital never allows the patient to become too ill while being treated remotely. Let $\bar{S}$ be the

Figure 2 **Three illustrative examples of patient's health score evolution.**



most severe health condition allowed outside the hospital (in expectation). The call-in threshold then must satisfy that $x + a + T\theta_T \leq \bar{S}$. Letting $\bar{A} = \big(0 \vee (\bar{S} - x - T\theta_T)\big)$, this policy constraint implies $a \in [0, \bar{A}] = \mathcal{A}$. Note that when $\bar{S} < x + T\theta_T$, the call-in threshold must be zero.

Figure 2 depicts three sample-path scenarios, all of which commence with a patient's health score at x and involve a transfer time of T. In Scenario 1, the patient is admitted remotely, improves and is discharged once their health score reaches zero. In Scenario 2, the patient is initially admitted remotely but experiences a decline in health. When the patient's health score reaches the predefined call-in threshold of $x + a$, they are called in to the hospital. During the journey to the hospital, the patient's health continues to deteriorate. Upon admission to the hospital, their health score is $x + a + Z(x, a, T)$, and from that point onward, the patient's condition improves. In Scenario 3, the patient is called in to the hospital immediately upon arrival (i.e., $a = 0$). Upon admission and travel to the hospital, the patient's health score is $x + Z(x, a, T)$, and from that point onward, the patient recovers on-site.

3.3. Cost Structure and Optimization

As mentioned, the health network's first-order design decision is captured in the threshold $x+a$. We expound upon that notion in this section. Because the negative drifts inherently capture health conditions that eventually improve, the primary metric by which these decisions are assessed will be the cost of care. Let h_R and h_H denote the holding cost rate for remote and on-site hospitalization, respectively. Similarly, h_T denotes the traveling cost rate. Accordingly, the total long run average cost is:

$$\begin{aligned} V(a) &= \lambda \left(h_R \mathbb{E}\left[\tau_R(x,a)\right] + \left(h_T T + h_H \mathbb{E}\left[\tau_H(x,a,Z)\right] \right) p_x(a) \right) \\ &= \lambda \left(\frac{h_R}{\theta_R} \left((1-p_x(a))\, x - p_x(a) a \right) + p_x(a) \left(h_T T + \frac{h_H}{\theta_H} \left(a + x + \theta_T T \right) \right) \right), \end{aligned} \tag{1}$$

where our goal is to set the optimal call-in threshold $a \in \mathcal{A}$ that minimizes this cost. We find that it is useful to rewrite the value function (1) as

$$V(a) = \lambda \left(\alpha + \beta p_x(a) + \gamma p_x(a) a \right), \tag{2}$$

where the constants α, β and γ are defined as follows:

$$\begin{aligned} \alpha &= h_R x/\theta_R, \\ \beta &= -h_R x/\theta_R + h_T T + h_H \left(x + \theta_T T \right)/\theta_H, \\ \gamma &= -h_R/\theta_R + h_H/\theta_H. \end{aligned}$$

Notice that $\beta = \gamma x + h_T T + h_H \theta_T T/\theta_H = \gamma x + (h_T + h_H \theta_T/\theta_H) T$.

In addition to the shorter expression, each of α, β, and γ offer interpretation to the decision problem. First, γ represents the disparity in marginal costs between on-site and home hospitalization. Then, β is the difference in expected costs of immediate transfer to on-site ($a=0$) and never transferring ($a=\infty$). Hence, β measures the viability of immediate transfer versus exclusively doing remote hospitalization. Lastly, α is the expected cost of never transferring, or simply the expected cost per patient of exclusively doing home hospitalization: $V(0)/\lambda = \alpha + \beta$ and $V(\infty)/\lambda = \alpha$.

Resource constraints. The hospital has to allocate its resources, primarily medical staff, between two groups: the on-site group which treats the on-site patients, and the virtual group, which is responsible for the remotely hospitalized patients. We start by defining the offered workload of each group. The on-site workload is

$$W_H(a) := \frac{\lambda p_x(a)}{\theta_H}\left(a + x + T\theta_T\right),$$

while the remote workload is

$$W_R(a) := \frac{\lambda}{\theta_R}\left(\left(1 - p_x(a)\right)x - p_x(a)a\right).$$

The total workload is therefore,

$$W_T(a) := W_H(a) + W_R(a).$$

Consider a *total* amount of resources, C, that needs to be allocated between the two groups. The corresponding optimization problem is

$$\begin{aligned} \min_{a\in\mathcal{A}} V(a) = \min_{a\in\mathcal{A}} &\lambda\left(\alpha + \beta p_x(a) + \gamma p_x(a)a\right) \\ \text{s.t. } & W_T(a) \le C. \end{aligned} \tag{3}$$

We denote the optimal call-in threshold by a_C^*, to emphasize the dependency of the solution on the total amount of resources. The solution (if it exists) to (3) minimizes the cost $V(a)$, while balancing the on-site and remote workloads, W_H and W_R, so that their sum does not exceed C. In particular, the dependence of W_H and W_R on a dictates which thresholds allow the constraint in (3) to be met and, therefore, encompasses the impact of resource scarcity. We elaborate on this in the Section 4.1. In addition, the existence of a solution to (3) depends on the problem parameters, and, in particular, the values of λ and C. Indeed, if λ is large and C is small, the total workload constraint in (3) might not be satisfied for any threshold $a \in \mathcal{A}$. Section 4.2 is devoted to characterizing the feasibility region in terms of (λ, C) pairs.

4. Preliminary Analyses: Workload and Feasibility

To identify the optimal call-in threshold and the resulting division of work among on-site and home hospitalization, we must first understand how the full operation depends on this level. In this pursuit, this section contains an analysis of the system's workload and a characterization of the problem's feasible region.

4.1. Analyzing the Shape of the Total Workload

We begin by separately characterizing the respective dependence of the on-site and remote hospitalization workloads on a.

LEMMA 1. *$W_H(a)$ is a strictly decreasing function; $W_R(a)$ is a strictly increasing function of a.*

The intuition of Lemma 1 is as follows: with an increase in the value of a, patients, on average, spend more time at home than in the hospital, by design. This translates into a rise in $W_R(a)$ and a decline in $W_H(a)$. The remaining question, tackled in Proposition 1, pertains to the behavior of the sum $W_T(a) = W_H(a) + W_R(a)$ as a function of a. The pivotal factor influencing this behavior is the ratio of relative recovery rates: θ_H/θ_R. Additionally, let $\Delta > 0$ be defined as

$$\Delta = \frac{\rho\theta_T T}{\rho x - 1 + e^{-\rho x}}. \tag{4}$$

Through these two quantities, we can classify the shape of the workload as a function of the call-in threshold.

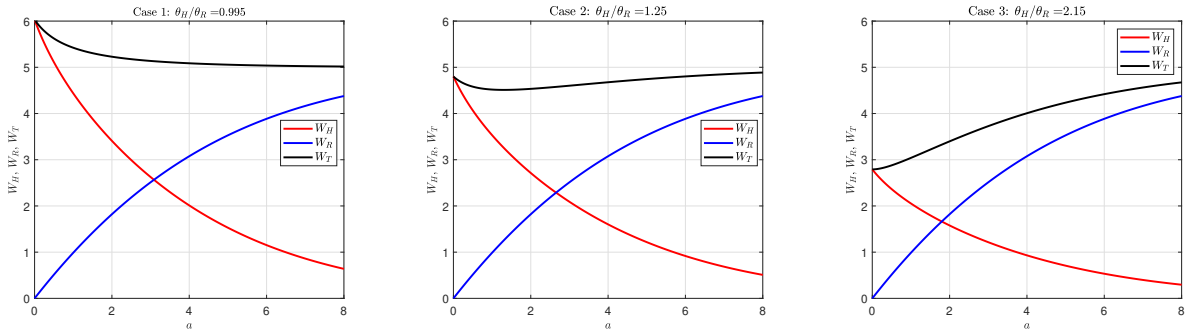
PROPOSITION 1. *The total workload $W_T(a)$ satisfies the following:*

1. ***Case 1:*** *If $\theta_H/\theta_R \leq 1$, then $W_T(a)$ is strictly decreasing.*
2. ***Case 2:*** *If $1 < \theta_H/\theta_R < 1 + \Delta$, then $W_T(a)$ has a unique minimum a_0 in $(0, \infty)$. Moreover, $W_T(a)$ is strictly decreasing in $[0, a_0)$ and strictly increasing in (a_0, ∞).*
3. ***Case 3:*** *If $\theta_H/\theta_R \geq 1 + \Delta$, then $W_T(a)$ is strictly increasing.*

From Proposition 1, we see that the total workload $W_T(a)$ can have three forms. If the average recovery rate at the hospital is slower than at remote hospitalization (Case 1), minimizing the total workload can be achieved by increasing the call-in threshold to its maximum value. On the other

hand, if the recovery rate at the hospital is much faster than under remote hospitalization (Case 3), minimizing the total workload is achieved by setting the call-in threshold to zero. Lastly, in the intermediate range when the on-site recovery rate is only moderately faster than under at home (Case 2), the total workload is unimodal with a unique minimum. Figure 3 illustrates these three cases.

Figure 3 An illustration of the total workload $W_T(a)$.



Considering the problem context, Cases 2 and 3 each seem more realistic than Case 1, and Case 2 is likely the most interesting of all. First, it less likely that the average recovery rate at home would outpace that under the full capabilities available at a hospital. Then, under the same reasoning, it is of the greatest managerial intrigue to consider the setting when the hospital is indeed better on average, but only marginally so.

The insights derived from Proposition 1 will prove valuable in the upcoming section where we analyze the feasibility region. Furthermore, in Section 5.2, these insights will be instrumental as we analyze the capacitated solution.

4.2. Identifying the Feasibility Region

Building on this understanding of the workload, let us now characterize, based on the given problem parameters, the (λ, C) pairs for which there exists an $a \in \mathcal{A}$ satisfying the constraint in the optimization problem (3). The feasibility region of (3) is defined as:

$$\mathcal{C}_{FR} = \{(\lambda, C) \in \mathbb{R}^2_+ : \exists a \in \mathcal{A}, \text{ s.t. } W_T(a) \leq C\}.$$

Let $a_{\min}$ denote the value of $a \in \mathcal{A}$ for which the total workload is minimal, i.e.,

$$a_{\min} = \arg\min_{a \in \mathcal{A}} W_T(a).$$

Note that Proposition 1 guarantees that $a_{\min}$ is unique. However, relative to the a_0 in Proposition 1, $a_{\min}$ is restricted to the range $\mathcal{A} = [0, \bar{A}]$, whereas $a_0 \in \mathbb{R}_+$.

Clearly, $\exists a \in \mathcal{A}$ s.t. $W_T(a) \leq C \iff W_T(a_{\min}) \leq C$. Since $W_T(a_{\min})/\lambda$ does not depend on λ, and $a_{\min}$ minimizes it as well, we are essentially looking for (λ, C) pairs such that $\lambda\,(W_T(a_{\min})/\lambda) \leq C$. Using this and Proposition 1, we obtain the following characterization of the feasibility region.

PROPOSITION 2. *The feasibility region of the optimization problem (3) is given by:*

$$\mathcal{C}_{FR} = \{(\lambda, C) \in \mathbb{R}_+^2 : W_T(a_{\min}) \leq C\},$$

where:

1. ***Case 1:*** *If $\theta_H/\theta_R \leq 1$, then $a_{\min} = \bar{A}$.*
2. ***Case 2:*** *If $1 < \theta_H/\theta_R < 1 + \Delta$, then $a_{\min} = \min\{a_0, \bar{A}\} > 0$, where a_0 is the unique minimum of $W_T(a)$ for $a \in \mathbb{R}_+$ (which does not depend on λ), as in Proposition 1.*
3. ***Case 3:*** *If $\theta_H/\theta_R \geq 1 + \Delta$, then $a_{\min} = 0$.*

Proposition 2 characterizes the feasibility region by considering three cases, mirroring the distinctions established in Proposition 1 regarding the behavior of the total workload – namely, whether it increases, decreases, or exhibits a unimodal pattern. With this understanding of the structure of the arrival rates and capacities for which the resource allocation problem is feasible, let us now analyze the true optimization problem.

5. Minimizing the Cost-of-Care for Hybrid Hospitalization

With the preliminary analyses in hand, we are now prepared to address our focal question of managing the operations of a hybrid hospital. To build insights, we will first study the unconstrained problem, where the resource capacity $C = \infty$, and we will then utilize this solution to analyze the finite C case.

5.1. Identifying the Optimal Call-In Structure with Unlimited Resources

Recalling the simplified notation in Equation (2), our goal is to set the call-in threshold in order to minimize the total expected cost rate:

$$\min_{a\in\mathcal{A}} V(a) = \min_{a\in\mathcal{A}} \lambda\left[\alpha + \beta p_x(a) + \gamma p_x(a) a\right].$$

Proposition 3 characterizes the uncapacitated optimal call-in threshold a^*_∞. In particular, the extreme cases are $a^*_\infty = 0$ and $a^*_\infty = \bar{A} = \bar{S} - T\theta_T - x$. When $a^*_\infty = 0$, remote hospitalization is less cost effective than on-site hospitalization, and thus it is preferable to hospitalize the patient on site. When $a^*_\infty = \bar{S} - T\theta_T - x$, however, remote hospitalization is more cost effective, and thus it is preferable to remotely hospitalize the patient until they reach the worst health condition that can still be treated remotely.

PROPOSITION 3 **(optimal call-in threshold)**. *Let the travel time T and initial condition $x > 0$ be fixed.*

- *If the marginal hospitalization cost is higher at the hospital ($\gamma \geq 0$), then remote hospitalization is always preferable, and the call in threshold is as high as allowable ($a^*_\infty = \bar{A}$).*
- *If the marginal hospitalization cost is smaller at the hospital ($\gamma < 0$):*

  — If immediate transfer to on-site is either not viable ($\beta \geq 0$) or viable but not dominant ($\gamma(1-e^{-\rho x})/\rho < \beta < 0$), then the optimal threshold is given by $a^*_\infty = (\tilde{a} \wedge \bar{A})$, where $\tilde{a} > 0$ is the unique solution to

$$e^{-\rho\tilde{a}} = (1 - \beta\rho/\gamma - \rho\tilde{a})\, e^{\rho x}. \tag{5}$$

which can be expressed by

$$\tilde{a} = \frac{1}{\rho}\left(1 + W\left(-e^{-\rho x + \beta\rho/\gamma - 1}\right)\right) - \frac{\beta}{\gamma}, \tag{6}$$

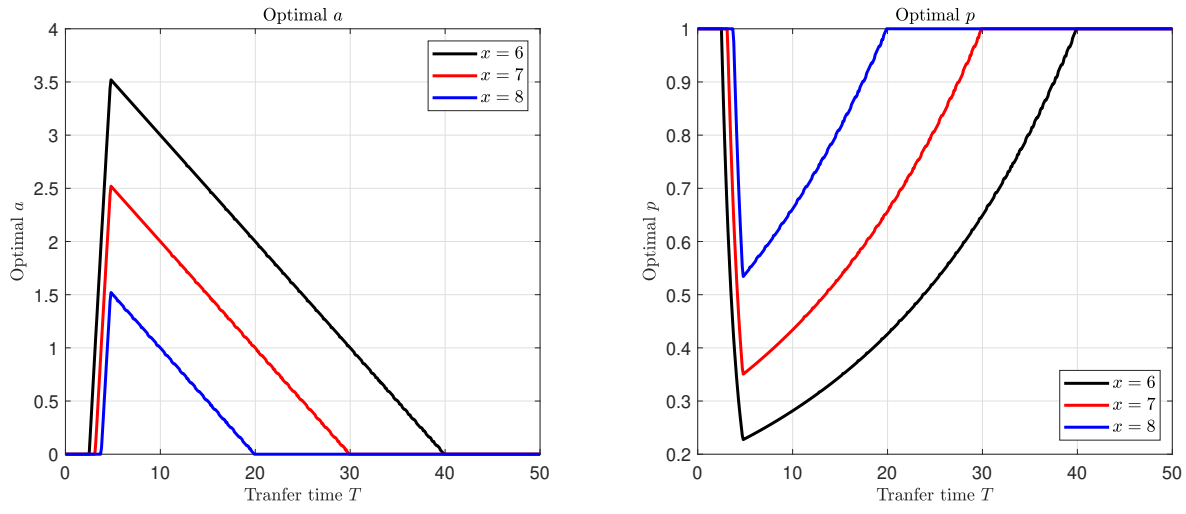
with $W(\cdot)$ as the Lambert-W function (principal branch).

— If immediate transfer is both viable and dominant ($\beta \leq \gamma(1-e^{-\rho x})/\rho$), then $a^*_\infty = 0$; all patients are treated on-site and home hospitalization is not offered.

We find that γ plays an important role in the decision of where to hospitalize patients and in case they were admitted, decide when to call them in to the hospital. Specifically, γ represents the marginal cost difference between on-site and remote hospitalization. If γ is positive, then patients should be admitted to remote hospitalization and stay there as much as possible.

Figure 4 illustrates the optimal call-in threshold and probability for different transfer times to the hospital and different initial conditions. First, we observe that remote hospitalization is not cost-effective for patients in close proximity or those residing at a significant distance from the hospital. In such cases, where $a_\infty = 0$, direct on-site admission is deemed more appropriate. Moreover, the call-in threshold is not monotone in T: it initially rises, reaching a maximum traveling time, $\hat{T}$, that remains consistent across all initial severity levels, before subsequently declining back to zero. The call-in probability has the exact opposite structure: it starts at one, decreases and then increase back. Second, we can see that as the initial condition becomes more severe (x increases), beyond the fact that the call-in threshold decreases, the distance range at which remote hospitalization is cost effective shrinks. Theorem 1 establishes these properties.

Figure 4 **Optimal call-in threshold and call-in probability as a function of travel time for different initial health scores. The parameters for** $[H, R, T]$ **are** $\theta = [0.5, 0.2, 0.1]$, $c = [2.65, 1.4, 2]$, $\gamma = -1.7$, $\lambda = \sigma_R = 1$, $\bar{S} = 10$.



Specifically, to formalize the decision's dependence on distance, let us first clarify how the model parameters depend on T. Recalling the definitions of α, β, and γ for Equation (2), we can rec-

ognize that, among these, only β depends on T. Moreover, if we define $\eta = h_T + h_H\theta_T/\theta_H$ as the marginal cost of travel distance, then β can be simply re-expressed as $\beta = \gamma x + \eta T$. Exploiting this dependence, we formalize the observations from Figure 4 now in Theorem 1.

THEOREM 1. *Let T_{LB} and T_{UB} be defined such that*

$$T_{LB} = -\frac{\gamma}{\eta}\left(x - \frac{1}{\rho}\left(1 - e^{-\rho x}\right)\right) \qquad \text{and} \qquad T_{UB} = \frac{\bar{S} - x}{\theta_T}. \tag{7}$$

Then, if $\gamma \geq 0$, $a^*_\infty > 0$ if and only if $T < T_{UB}$.

Moreover, if $\gamma < 0$, then the $\hat{T}$ which is the unique solution to

$$\bar{S} = \frac{1}{\rho}\left(1 + W\left(-e^{\eta\rho\hat{T}/\gamma - 1}\right) + \left(\theta_T - \frac{\eta}{\gamma}\right)\hat{T}\right), \tag{8}$$

is such that for $T \in (T_{LB}, \hat{T})$,

$$\frac{\partial a^*_\infty}{\partial T} = -\frac{\eta}{\gamma}\frac{1 - W\left(-e^{\rho\eta T/\gamma - 1}\right)}{1 + W\left(-e^{\rho\eta T/\gamma - 1}\right)} > 0, \tag{9}$$

and for $T \in (\hat{T}, T_{UB})$,

$$\frac{\partial a^*_\infty}{\partial T} = -\theta_T < 0, \tag{10}$$

with $a^*_\infty = 0$ for $T \notin (T_{LB}, T_{UB})$.

Theorem 1 reveals that, even when remote hospitalization has lower marginal cost, the cost benefit only applies to patients up to a certain distance from the hospital. While this may seem somewhat paradoxical at first glance, its intuition is clear: the maximum allowable patient severity ($\bar{S}$) and the distance-dependent expected deterioration while in transit ($\theta_T T > 0$) together imply that there is a "shorter leash" to risk on home hospitalization for patients who live far from the facility. This observation is further complicated by the recognition that the range of cases in which home hospitalization is viable narrows as the initial severity increases: T_{LB} increases in x, T_{UB} decreases, and the rates that a^*_∞ changes with T are exactly parallel across x and thus are unaffected. Recalling the growing recognition of more dire health states in rural areas (Lewis

2022), we see that Theorem 1 identifies a problematic combination. That is, if greater distance and worsened initial condition each restrict the feasibility of remote hospitalization, then this mode of care may not benefit the exact populations for which it seems intended.

Let us emphasize that this conundrum is not a consequence of scarce resources – thus far, our results have assumed an unlimited amount of resources. Furthermore, as we show in the next section, the capacitated solution retains the same properties to the uncapacitated counterpart. Consequently, the diminished effectiveness of remote hospitalization with distance and severity persists also in the presence of resource scarcity; in fact, in what may be the most realistic parameter settings, this reduction is exacerbated even further.

5.2. Identifying the Optimal Call-In Structure with Limited Resources

We now go back to our original capacitated problem in (3). The goal is two-fold. First, we wish set the call-in policy under finite amount of resources. Second, we wish to allocate the total amount of resources between the two hospitalization modes: on-site and remote.

To begin, Theorem 2 characterizes the solution of the capacitated problem (3).

THEOREM 2. *Assume that $W_T(a_{\min}) \le C$ (i.e. the feasibility region is not empty). Then, problem (3) has a unique solution $a_C^* \in \mathcal{A}$, such that:*

- *If $W_T(a_{\min}) = C$, then $a_C^* = a_{\min}$.*
- *If $W_T(a_{\min}) < C$, then:*
  - *If $W_T(a_\infty^*) \le C$, then $a_C^* = a_\infty^*$,*
  - *If $W_T(a_\infty^*) > C$, then $a_{\min} \neq a_\infty^*$ and a_C^* is the unique value of $a \in \mathcal{A}$ strictly between $a_{\min}$ and a_∞^* such that $W_T(a) = C$.*

Note that depending on the parameters, both $a_{\min} > a_\infty^*$ and $a_{\min} < a_\infty^*$ are possible. In either case, when $W_T(a_\infty^*) > C$, and $W_T(a_{\min}) < C$, the call-in threshold a_C^* is strictly between them and satisfies (uniquely) $W_T(a_C^*) = C$.

To interpret this structure and make its solution explicit, let us now establish an equivalence between the capacitated and uncapacitated solutions. To emphasize the dependence of the function V on holding costs, we denote it as $V(h_R, h_H, a)$. Recall the uncapacitated minimization problem:

$$\min_{a\in\mathcal{A}} V(h_R, h_H, a), \tag{11}$$

which per Proposition 3, has a unique solution $a^*_\infty \in \mathcal{A}$. Recall also the capacitated minimization problem:

$$\begin{aligned} &\min_{a\in\mathcal{A}} V(h_R, h_H, a) \\ &\text{s.t. } W_T(a) \le C, \end{aligned} \tag{12}$$

which per Theorem 2, assuming that $W_T(a_{\min}) \le C$, has a unique solution $a^*_C \in \mathcal{A}$.

Define

$$\Gamma = \begin{cases} -\frac{V'(h_R,h_H,a^*_C)}{W'_T(a^*_C)}, & \text{if } W_T(a_{\min}) < C \text{ and } W_T(a^*_\infty) > C \\ 0, & \text{otherwise.} \end{cases}$$

Note that $\Gamma \ge 0$, since in the case where $W_T(a_{\min}) < C$ and $W_T(a^*_\infty) > C$, $V'(h_R, h_H, a^*_C)$ and $W'_T(a^*_C)$ must be non zero and with opposite signs (see the proofs of Lemmas 1 and 2). Now, consider a similar uncapacitated optimization problem with Γ-modified costs:

$$\min_{a\in\mathcal{A}} V(h_R + \Gamma, h_H + \Gamma, a). \tag{13}$$

Proposition 4 establishes the equivalence between the solutions of (12) and (13). This equivalence implies that all properties of the uncapacitated problem apply to the capacitated problem. Notably, the solution structure, characterized by (modified) α, β, and γ as outlined in Proposition 3, and the influence of traveling distance on the optimal call-in policy, as indicated in Theorem 1, remain consistent.

PROPOSITION 4. *Assume that $W_T(a_{\min}) < C$ (i.e. the feasibility region of (12) contains more than one point). Then, the problem (13) has a unique solution in $\mathcal{A}$ which equals a^*_C.*

The parameter Γ captures the effect of scarce resources. In essence, the capacity constraint effect is reflected through the simultaneous increase of both remote and on-site costs by Γ. Consequently, the call-in threshold will experience an increase or decrease contingent upon the initial cost rates and recovery rates associated with each hospitalization option. Specifically, the revised parameter $\gamma(\Gamma)$ would be

$$\gamma(\Gamma) = \frac{-(h_R+\Gamma)}{\theta_R} + \frac{h_H+\Gamma}{\theta_H} = \gamma + \Gamma\left(\frac{1}{\theta_H} - \frac{1}{\theta_R}\right).$$

Since $\Gamma \geq 0$, the value of $\gamma(\Gamma)$ may increase/decrease depending on the relation between θ_H and θ_R. Per Proposition 3, the value of $\gamma(\Gamma)$, and in particular its sign, determines the optimal call-in threshold and whether, if at all, patients should be sent to remote hospitalization. If, for example, $\theta_H > \theta_R$, then $\gamma(\Gamma) < \gamma$. When $\gamma > 0$ and $\gamma(\Gamma) < 0$, patients who under ample resources would remain in remote hospitalization until their health score reaches $\bar{S}$, would be called in at a lower threshold under a finite number of resources, or even directly admitted on-site. Furthermore, recalling Theorem 1, we can notice that if $\theta_H > \theta_R$ and $\gamma < 0$, then the fact that $\gamma(\Gamma) < 0$ implies that an even smaller range of distances will be suitable for remote hospitalization, and this range again shrinks with the initial severity x. On the other hand, if $\theta_H < \theta_R$, then $\gamma(\Gamma) > \gamma$. When $\gamma < 0$ and $\gamma(\Gamma) > 0$, patients who under ample resources would be admitted on-site, or be admitted remotely with a call-in threshold that is smaller than $\bar{S}$, would under a finite number of resources, be called in at $\bar{S}$.

Figures 5 and 6 illustrate the optimal capacitated solution for different resource levels and travel times. Figure 5 corresponds to Case 3 in Proposition 1; in the top plots ($T = 2$): $2.5 = \theta_H/\theta_R > \Delta + 1 = 2.14$: When there is ample resources ($C \geq 4$), $a^*_\infty \approx 4$. As resources becomes scarce, the workload $W_T(a)$ decreases to satisfy the capacity constraint. Since in this case $W_T(a)$ is strictly increasing, the call-in threshold a^*_C decreases up until $C \approx 2.5$ – the boundary of the feasibility region. The bottom plots are for $T = 8$. The feasibility region, which is smaller since T is larger, ends at $C \approx 3.3$. In other words, more resources are needed when patients are distant.

The right plots show the optimal resource allocation W_H and W_R. We see that when resources are scarce, most of them (80% when $T = 2$) are allocated to the hospital; as the total amount

increases, fewer resources are allocated to the hospital, while more are allocated to remote hospitalization. When there are ample resources, most of them (82% when $T = 2$) are allocated to remote hospitalization.

Figure 5 **Optimal capacitated solution. The parameters for (H,R,T) are $\theta = (0.5, 0.2, 0.1)$, $h = (2.65, 1.4, 2)$, $x = 1$, $\bar{S} = 15$, $\lambda = 1$, $\sigma_R = 1$.**

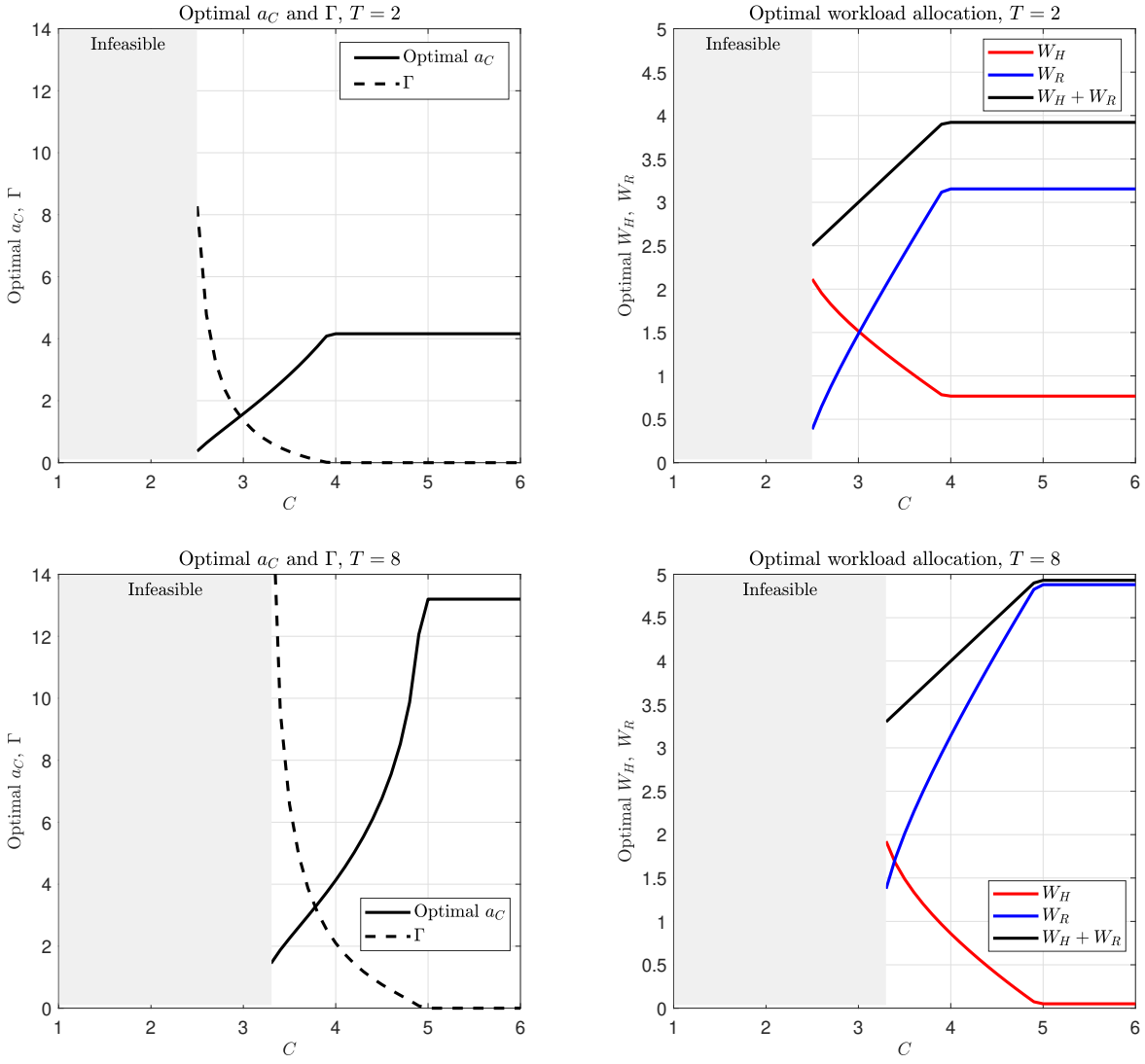
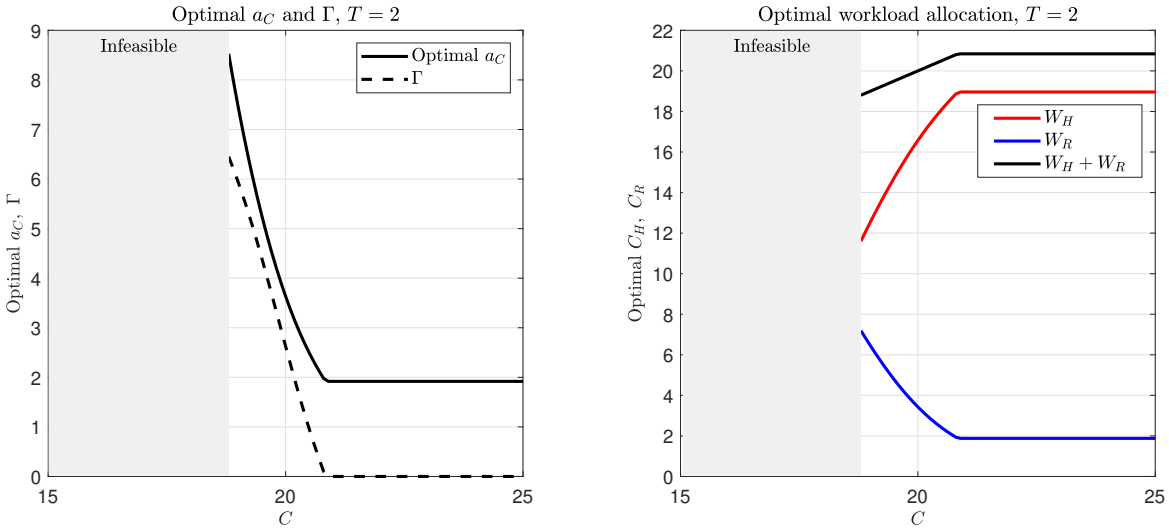


Figure 6 corresponds to Case 1 in Proposition 1, where $0.83 = \theta_H/\theta_R < 1$. In this case, $W_T(a)$ is strictly increasing. Therefore, when there is ample amount of resources, $a^*_\infty \approx 2$; as resources become scarce, a^*_C increases. The right plot shows that, as opposed to the cases in Figure 5, as the

total amount of resources increase, fewer are allocated to remote hospitalization, while more are allocated to the hospital.

Figure 6 **Optimal capacitated solution. The parameters for (H,R,T) are $\theta = (0.05, 0.06, 0.1)$, $h = (2.65, 5.1, 2)$, $x = 1$, $\bar{S} = 10$, $\lambda = 1$, $\sigma_R = 1$.**



6. Conclusions and Direction for Future Research

The hybrid hospital model constitutes a service network design problem. The decision of whether to admit a patient remotely or on-site entails the efficient allocation of resources across the two hospitalization modes. To address this, we adopt a modeling approach that captures the dynamic progression of individual health conditions within the network and during travel. System design optimization, in this context, revolves around establishing the call-in threshold that minimizes the total operational costs, consequently influencing the optimal resource allocation.

Managerially, our results both offer guidance on how hospitals should allocate resources between on-site and remote care, and identify a potential cautionary tale, in that distant patients may not actually be best served by remote hospitalization. Qualitatively, these insights may also be relevant in other parts of public life, e.g., if online education is considered as an alternative for a rural school with declining enrollment. Much like in the case of remote hospitalization, these challenges may

be further heightened by inequities of internet access for those who live in rural areas (Lai and Widmar 2021). Much like we have discussed for the pitfalls of relying on remote hospitalization to serve rural communities, we believe our results also offer a word of caution towards potential over-reliance on online education, in which recent data from the pandemic has revealed stark and concerning disparities in rates of learning relative to just before Covid-19 began (Halloran et al. 2021, Goldhaber et al. 2022).

Whether in the focal hospitalization application or in other relevant areas, one possible limitation of our model is the underlying assumption that all patients eventually recover. Indeed, the negative drifts of the Brownian motions naturally imply that every patient's severity will hit 0 in finite time almost surely. This presents a natural opportunity to generalize and model mortality or another form of negative outcome. While the assumption of guaranteed recovery may be conservative, we believe that this actually emphasizes both the importance of careful hybrid hospital design and the fragility of the relationship between the remote format and patience distance. That is, viewing our results with the eventual recovery assumption in mind, we see that even when the worst that can happen is *added cost*, remote hospitalization is still only viable for a limited range of patient distances (which may be further limited by the initial health severities), even when the operations are designed optimally as we describe.

There are several additional future research directions. The baseline model can be expanded to include additional features such as starting the hospitalization on-site and possibly completing it remotely. The model can be expanded to capture more elaborate process protocols/networks. Specifically, in our model, once patients are called in to the hospital, they stay there until full recovery. In some cases, however, once patients are stabilized at the hospital, they may be sent home to complete their hospitalization remotely. Furthermore, our solution serves as a foundational framework for the development of dynamic control policies. The call-in threshold establishes a reference point that can be further refined through real-time performance enhancements. Ultimately, the optimal control strategy would introduce state-dependent call-in decisions, which adapt based on perturbations from the initial optimal baseline decisions.

References

Adams K (2023) More than 30% of rural hospitals are at risk of closure. *Center for Healthcare Quality and Payment Reform (CHQPR)* https://medcitynews.com/2023/07/rural-hospital-insurance-finance/.

Aldrete J (1994) Discharge criteria. *Baillière's Clinical Anaesthesiology* 8(4):763–773.

Americal Hospital Association, 2020 (2023) https://www.aha.org/hospitalathome.

Anderson S, Wilkins M, Weaver W, Selvester R, Wagner G (1992) Electrocardiographic phasing of acute myocardial infarction. *Journal of Electrocardiology* 25:3–5.

Armony M, Yom-Tov G (2021) Hospitalization versus home care: Balancing mortality and infection risks for hematology patients. *Working paper* .

Barkai G, Amir H, Dulberg O, Itelman E, Gez G, Carmon T, Merhav L, Zigler S, Atamne A, Pinhasov O (2022) "staying at home": A pivotal trial of telemedicine-based internal medicine hospitalization at a nursing home. *Digital Health* 8:1–7.

Bartel A, Chan C, Kim SH (2020) Should hospitals keep their patients longer? The role of inpatient care in reducing post-discharge mortality. *Management Science* 66(6):2326–2346.

Bavafa H, Örmeci L, Savin S, Virudachalam V (2022) Surgical case-mix and discharge decisions: Does within-hospital coordination matter? *Operations Research* 70(2):990–1007.

Bavafa H, Savin S, Terwiesch C (2019) Managing patient panels with non-physician providers. *Production and Operations Management* 28(6):1577–1593.

Bavafa H, Savin S, Terwiesch C (2021) Customizing primary care delivery using E-visits. *Production and Operations Management* 30(11):4306–4327.

Behrman P, Fitzgibbon M, Dulin A, Wang M, Baskin M (2021) Society of behavioral medicine statement on COVID-19 and rural health. *Translational Behavioral Medicine* 11(2):625–630.

Bokolo A (2020) Use of telemedicine and virtual care for remote treatment in response to COVID-19 pandemic. *Journal of Medical Systems* 44(7):1–9.

Boldt-Christmas O, Kannourakis R, M M, D U (2023) Virtual hospitals could offer respite to overwhelmed health systems. *McKinsey & Company* https://www.mckinsey.com/industries/healthcare/our-insights/virtual-hospitals-could-offer-respite-to-overwhelmed-health-systems.

Bowblis J, Brunt C (2014) Medicare skilled nursing facility reimbursement and upcoding. *Health economics* 23(7):821–840.

Capelastegui A, España P, Bilbao A, Martinez-Vazquez M, Gorordo I, Oribe M, Urrutia I, Quintana J (2008) Pneumonia: Criteria for patient instability on hospital discharge. *Chest* 134(3):595–600.

Case A, Deaton A (2015) Rising morbidity and mortality in midlife among white non-hispanic americans in the 21st century. *Proceedings of the National Academy of Sciences* 112(49):15078–15083.

Case A, Deaton A (2017) Mortality and morbidity in the 21st century. *Brookings papers on economic activity* 2017:397.

Chan C, Yom-Tov G, Escobar G (2014) When to use speedup: An examination of service systems with returns. *Operations Research* 62(2):462–482.

Clarke D, Newsam J, Olson D, Adams D, Wolfe A, Fleisher L (2021) Acute hospital care at home: the cms waiver experience. *NEJM Catalyst Innovations in Care Delivery* 2(6).

Cross SH, Califf RM, Warraich HJ (2021) Rural-urban disparity in mortality in the us from 1999 to 2019. *JAMA* 325(22):2312–2314.

Curtin S, Spencer M (2021) Trends in death rates in urban and rural areas: United states, 1999-2019. *NCHS Data Brief* (417):1–8.

Deo S, Iravani S, Jiang T, Smilowitz K, Samuelson S (2013) Improving health outcomes through better capacity allocation in a community-based chronic care model. *Operations Research* 61(6):1277–1294.

Francis N, Stuart B, Knight M, Vancheeswaran R, Oliver C, Willcox M, Barlow A, Moore M (2021) Predictors of clinical deterioration in patients with suspected covid-19 managed in a 'virtual hospital'setting: A cohort study. *BMJ Open* 11(3):e045356.

Goldhaber D, Kane TJ, McEachin A, Morton E, Patterson T, Staiger DO (2022) The consequences of remote and hybrid instruction during the pandemic. Technical report, National Bureau of Economic Research.

Gong S, Gao Y, Zhang F, Mu L, Kang C, Liu Y (2021) Evaluating healthcare resource inequality in Beijing, China based on an improved spatial accessibility measurement. *Transactions in GIS* 25(3):1504–1521.

Grand-Clément J, Chan C, Goyal V, Escobar G (2020) Robust policies for proactive ICU transfers. *Operations Research, forthcoming* .

Halloran C, Jack R, Okun JC, Oster E (2021) Pandemic schooling mode and student test scores: Evidence from us states. Technical report, National Bureau of Economic Research.

Hopp W, Iravani S, Yuen G (2007) Operations systems with discretionary task completion. *Management Science* 53(1):61–77.

Hur J, Chang M (2020) Usefulness of an online preliminary questionnaire under the COVID-19 pandemic. *Journal of Medical Systems* 44:1–2.

Hutchings O, Dearing C, Jagers D, Shaw M, Raffan F, Jones A, Taggart R, Sinclair T, Anderson T, Ritchie A (2021) Virtual health care for community management of patients with covid-19 in australia: observational cohort study. *Journal of Medical Internet Research* 23(3):e21064.

Ishfaq R, Raja U (2015) Bridging the healthcare access divide: A strategic planning model for rural telemedicine network. *Decision Sciences* 46(4):755–790.

Jiao J, Degen N, Azimian A (2021) Identifying hospital deserts in Texas before and during the COVID-19 outbreak. *Available at SSRN 3800281* .

Kadir M (2020) Role of telemedicine in healthcare during COVID-19 pandemic in developing countries. *Telehealth and Medicine Today* .

Kc D, Terwiesch C (2012) An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* 14(1):50–65.

Kelly J, Dwyer J, Willis E, Pekarsky B (2014) Travelling to the city for hospital care: Access factors in country a boriginal patient journeys. *Australian Journal of Rural Health* 22(3):109–113.

Lai J, Widmar NO (2021) Revisiting the digital divide in the covid-19 era. *Applied economic perspectives and policy* 43(1):458–464.

Lewis T (2022) People in rural areas die at higher rates than those in urban areas. *Scientific American* URL https://www.scientificamerican.com/article/people-in-rural-areas-die-at-higher-rates-than-those-in-urban-areas/.

Mills A, Argon N, Ziya S (2013) Resource-based patient prioritization in mass-casualty incidents. *Manufacturing & Service Operations Management* 15(3):361–377.

Monaghesh E, Hajizadeh A (2020) The role of telehealth during COVID-19 outbreak: A systematic review based on current evidence. *BMC Public Health* 20(1):1–9.

Nambiar S, Mayorga M, Capan M (2020) Resource allocation strategies under dynamically changing health conditions. *Wrking paper* .

Noronha K, Guedes G, Turra C, Andrade M, Botega L, Nogueira D, Calazans J, Carvalho L, Servo L, Ferreira M (2020) The COVID-19 pandemic in Brazil: Analysis of supply and demand of hospital and ICU beds and mechanical ventilators under different scenarios. *Cadernos de Saúde Pública* 36.

Olaisen RH, Rossen LM, Warner M, Anderson RN (2019) Unintentional injury death rates in rural and urban areas: United states, 1999–2017. *NCHS Data Brief* (343):1–8.

Shi P, Chou M, Dai J, Ding D, Sim J (2016) Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science* 62(1):1–28.

Shi P, Helm J, Deglise-Hawkinson J, Pan J (2021) Timing it right: Balancing inpatient congestion vs. readmission risk at discharge. *Operations Research* 69(6):1842–1865.

Siegmund D (2013) *Sequential Analysis: Tests and Confidence intervals* (Springer Science & Business Media).

Sun Z, Argon NT, Ziya S (2018) Patient triage and prioritization under austere conditions. *Management Science* 64.

Ullrich F, Mueller K (2023) Covid-19 cases and deaths, metropolitan and nonmetropolitan counties over time (update). *Policy File* 2020.

Verhagen M, Brazel D, Dowd J, Kashnitsky I, Mills M (2020) Mapping hospital demand: Demographics, spatial variation, and the risk of "hospital deserts" during COVID-19 in England and wales. *OSF Preprints* .

Wang X, Debo L, Scheller-Wolf A, Smith S (2010) Design and analysis of diagnostic service centers. *Management Science* 56(11):1873–1890.

Appendix A: Proofs

Proof of Lemma 1. Recall that $p_x(a)$ is the hitting probability given by

$$p_x(a) := \mathbb{P}\left(\mathcal{B}^R(\tau_R(x,a)) = a + x\right) = \frac{1 - e^{-\rho x}}{e^{\rho a} - e^{-\rho x}},$$

and therefore, its first derivative with respect to a is

$$p'_x(a) = -\frac{\rho e^{\rho a}}{e^{\rho a} - e^{-\rho x}} p_x(a) < 0.$$

Our goal is to prove that $W'_H(a) < 0$ and $W'_R(a) > 0$. We begin with W_H. Recall that:

$$W_H(a) = \frac{\lambda p_x(a)}{\theta_H}\left(a + x + T\theta_T\right).$$

Therefore and since $p'_x(a) \neq 0$,

$$W'_H(a) = \frac{\lambda}{\theta_H}\left(p'_x(a)(a + x + \theta_T T) + p_x(a)\right) = \frac{\lambda}{\theta_H} p'_x(a)\left(a + x + \theta_T T + \frac{p_x(a)}{p'_x(a)}\right). \tag{14}$$

Now,

$$\frac{p_x(a)}{p'_x(a)} = -\frac{e^{\rho a} - e^{-\rho x}}{\rho e^{\rho a}} = -\frac{1}{\rho}(1 - e^{-\rho(a+x)}) > -(a + x), \tag{15}$$

where the inequality is because $1 - e^{-x} < x$ for $x > 0$. Therefore,

$$a + x + \theta_T T + \frac{p_x(a)}{p'_x(a)} > \theta_T T \quad \Rightarrow \quad a + x + \frac{p_x(a)}{p'_x(a)} > 0. \tag{16}$$

Multiplying both sides by $\frac{\lambda}{\theta_H} p'_x(a)$, the result follows since $p'_x(a) < 0$. We turn to $W_R(a)$. We have:

$$W_R(a) = \frac{\lambda}{\theta_R}\left((1 - p_x(a))\, x - p_x(a) a\right) = \frac{\lambda}{\theta_R} x - \frac{\lambda}{\theta_R} p_x(a)(a + x),$$

and therefore (and again, because $p'_x(a) \neq 0$),

$$W'_R(a) = -\frac{\lambda}{\theta_R}\left(p'_x(a)(a + x) + p_x(a)\right) = -\frac{\lambda}{\theta_R} p'_x(a)\left(a + x + \frac{p_x(a)}{p'_x(a)}\right) > 0, \tag{17}$$

where the inequality is from (16) and since $p'_x(a) < 0$. Q.E.D.

Proof of Proposition 1. Recall that $W_T(a) = W_H(a) + W_R(a)$ and that we wish to characterize the dependence of W_T on a. For ease of notation, denote $r = \theta_H/\theta_R$. We have:

$$\begin{aligned} W_T'(a) = W_H'(a) + W_R'(a) &\overset{(14),(17)}{=} \frac{\lambda}{\theta_H} p_x'(a) \left(a + x + \theta_T T + \frac{p_x(a)}{p_x'(a)} \right) - \frac{\lambda}{\theta_R} p_x'(a) \left(a + x + \frac{p_x(a)}{p_x'(a)} \right) \\ &\overset{(4)}{=} \frac{\lambda p_x'(a)}{\theta_H} \left((1-r) \left(a + x + \frac{p_x(a)}{p_x'(a)} \right) + \theta_T T \right) \\ &\overset{(15)}{=} \frac{\lambda |p_x'(a)|}{\rho \theta_H} \left((1-r)(1 - \rho(a+x) - e^{-\rho(a+x)}) - \theta_T T \rho \right), \end{aligned}$$

where the last inequality is because $p_x'(a) < 0$.

We start with Case 1, where $\theta_H/\theta_R \le 1 \iff 1 - r \ge 0$. Using the inequality $1 - e^{-x} \le x$, we get that $1 - \rho(a+x) - e^{-\rho(a+x)} \le 0$. Thus, if $1 - r \ge 0$, then $W_T'(a) < 0$.

We turn to Case 3. Note that when $r > 1$, $1 - r = -|1 - r|$. In this case,

$$W_T'(a) \overset{\text{if } r>1}{=} \frac{\lambda |p_x'(a)| |1-r|}{\rho} \left(-1 + \rho(a+x) + e^{-\rho(a+x)} - \frac{\theta_T T \rho}{|1-r|} \right).$$

Denoting $h(a) := -1 + \rho(a+x) + e^{-\rho(a+x)}$, we have:

$$\begin{aligned} &h(0) = \rho x - 1 + e^{-\rho x} > 0 \qquad \text{because } e^{-x} > 1 - x \text{ for } x > 0, \\ &h'(a) = \rho - \rho e^{-\rho(a+x)} = \rho(1 - e^{-\rho(a+x)}) > 0, \\ &h(a) \xrightarrow{a \to \infty} \infty. \end{aligned} \tag{18}$$

Thus, $W_T'(a)$ can be negative, if and only if its value at $a = 0$ is negative. Clearly, this *does not* happen if $h(0) \ge \frac{\theta_T T \rho}{|1-r|}$, or, written differently, if

$$|1 - r| \ge \theta_T T \rho \left(\rho x - 1 + e^{-\rho x} \right)^{-1} \overset{\text{if } r>1}{\iff} r \ge 1 + \Delta,$$

where Δ is defined in (4).

Lastly, we turn to Case 2 where $1 < r < 1 + \Delta$. This implies that $h(0) < \frac{\theta_T T \rho}{|1-r|}$. In this case, by (18), it is clear that $W_T'(0) < 0$, and that there exists $a_0 > 0$ such that $W_T'(a) < 0$ for $a \in [0, a_0)$, $W_T'(a_0) = 0$, and $W_T'(a) > 0$ for $a \in (a_0, \infty)$. This concludes the proof. Q.E.D.

Proof of Proposition 2. Per section 4.2, the feasibility region of the optimization problem (3) is

$$\mathcal{C}_{FR} = \{(\lambda, C) \in \mathbb{R}_+^2 : W_T(a_{\min}) \le C\}.$$

All that is left is to characterize $a_{\min}$. By Proposition 1, if $\theta_H/\theta_R \le 1$, then $W_T(a)$ is strictly increasing in a. Thus, in this case, $a_{\min} = 0$ which proves the first item. Again by Proposition 1, if $\theta_H/\theta_R \ge 1 + \Delta$, then $W_T(a)$ is strictly decreasing in a. Thus, in this case, $a_{\min} = \bar{S} - x - T\theta_T$, which proves the second item.

Finally, in the case where $1 < \theta_H/\theta_R < 1+\Delta$, by Proposition 1, there exists $a_0 > 0$ such that $W_T'(a) < 0$ for $a \in [0, a_0)$, $W_T'(a_0) = 0$, and $W_T'(a) > 0$ for $a \in (a_0, \infty)$. If $a_0 < \bar{S} - x - T\theta_T$, then $a_{\min} = a_0$. Otherwise, $\bar{S} - x - T\theta_T \in (0, a_0]$, and since $W_T(a)$ is strictly decreasing in this interval, we have $a_{\min} = \bar{S} - x - T\theta_T$ which completes the proof of the third item and the proposition. Q.E.D.

Proof of Proposition 3. We first provide technical characterizations of the objective function $V(a)$.

LEMMA 2. *The value function $V(a)$ satisfies the following:*

If $\gamma \geq 0$, then $V(a)$ is strictly decreasing in $a \in \mathbb{R}_+$; else ($\gamma < 0$),

- *If $\beta \geq 0$, or $\gamma(1-e^{-\rho x})/\rho < \beta < 0$, then $V(a)$ is unimodal with a unique minimum over $a \in \mathbb{R}_+$.*
- *If $\beta \leq \gamma(1-e^{-\rho x})/\rho$, then $V(a)$ is strictly increasing in $a \in \mathbb{R}_+$.*

From Lemma 2, $\gamma \geq 0$ yields that $V(a)$ is strictly decreasing, and thus the largest allowable threshold is optimal: $a^* = \bar{A}$. Then, if $\gamma < 0$ and $\beta \geq \gamma(1-e^{-\rho x})/\rho$, then Lemma 2 provides that there is a unique optimal solution. Moreover, because the definition of $\tilde{a}$ in Equation (5) is precisely the first order condition in Equation (21) within the proof of Lemma 2, we can see that $\tilde{a}$ is the unique maximizer of $V(a)$. If $\tilde{a}$ is within the maximum allowable threshold size, then it is optimal for the unlimited capacity problem, but if $\tilde{a} > \bar{A}$, then we can see that $V'(a) > 0$ for all $a \in [0, \bar{A}]$, meaning $\bar{A}$ is optimal. Hence, $a^* = (\tilde{a} \wedge \bar{A})$ is the optimal threshold. Finally, for the remaining case and again by Lemma 2, if $\beta \leq \gamma(1-e^{-\rho x})/\rho$, then $V(a)$ is always increasing, and thus the optimal threshold is as low as possible, directing all patients immediately to on-site hospitalization: $a^* = 0$.

To complete the proof, let us verify that $\tilde{a}$ given by the solution in Equation (6) is indeed positive and obtained by the principal branch of the Lambert-W function. Note that the existence of a unique, positive solution to first order condition in Equation (21) is already guaranteed for $\gamma < 0$ and $\beta > \gamma(1-e^{-\rho x})/\rho$ through the preceding linear-and-exponential-function arguments; the focus now is simply on proving the correctness of Equation (6). Rearranging (21) and multiplying both sides by $e^{\beta\rho/\gamma - 1}$, we have that $\tilde{a}$ will be the a that solves

$$-e^{-\rho x + \beta\rho/\gamma - 1} = (\rho a + \beta\rho/\gamma - 1)\, e^{\rho a + \beta\rho/\gamma - 1}.$$

Before further manipulating this equation, let us inspect the terms in the exponent on the left-hand side. If $\beta \geq 0$, it is clear that $-\rho x + \beta\rho/\gamma - 1 < 0$, so let us focus on $\gamma(1-e^{-\rho x})/\rho < \beta < 0$. Dividing by $\gamma/\rho <$

0, we have that $0 < \beta\rho/\gamma < 1 - e^{-\rho x}$, and, furthermore, by adding $-1 - \rho x$ to each side, we have that $-\rho x + \beta\rho/\gamma - 1 < -\rho x - e^{-\rho x} < 0$. Hence, for all $\beta > \gamma(1 - e^{-\rho x})/\rho$, $-e^{-\rho x + \beta\rho/\gamma - 1} \in (-1/e, 0)$.

For the identity $W(ze^z) = z$ to hold on the principal branch of the Lambert-W, we must have $z \geq -1$. Hence, as a final step, let us show that $\rho\tilde{a} + \beta\rho/\gamma - 1 \geq -1$. If $\beta \leq 0$, this is immediately true by the fact that $\rho > 0$, $\tilde{a} > 0$, and $\gamma < 0$, so let us focus on the $\beta > 0$ case. If $\tilde{a} > -\beta/\gamma$, then we can apply the Lambert-W principal branch identity, and Equation (6) will follow immediately. To see that this is indeed true, we return to the linear-and-exponential-function arguments. Notice that, at $a = -\beta/\gamma > 0$, the left-hand side of Equation (21) is $e^{\rho\beta/\gamma} < 1$, whereas the right-hand side simplifies to $e^{\rho x} > 1$. Therefore, the linear function has not yet crossed the exponential function, implying $\tilde{a} > -\beta/\gamma$. Q.E.D.

Proof of Lemma 2. To begin, let us obtain a first order condition for $V(a)$. The derivative of the cost function with respect to a is

$$V'(a) = \beta p'_x(a) + \gamma a p'_x(a) + \gamma p_x(a).$$

Since $p'_x(a) < 0$, the cost derivative simplifies to

$$\begin{aligned} V'(a) &= p'_x(a)\left(\beta + \gamma a + \gamma\frac{p_x(a)}{p'_x(a)}\right) \overset{(15)}{=} p'_x(a)(\beta + \gamma a - \gamma\frac{1}{\rho}(1 - e^{-\rho(a+x)})) \\ &= \frac{|p'_x(a)|}{\rho}\left(\gamma(1 - e^{-\rho(x+a)}) - \beta\rho - \gamma\rho a\right). \end{aligned} \tag{19}$$

Since $\rho > 0$, and given that $x > 0$, $|p'_x(a)|$ is strictly positive for all $a \geq 0$, the sign of $dV/da = 0$ matches the sign of $\gamma(1 - e^{-\rho(x+a)}) - \beta\rho - \gamma\rho a$. We can see that the a-derivative of this expression is

$$\left(\gamma(1 - e^{-\rho(x+a)}) - \rho\beta - \gamma\rho a\right)' = -\gamma\rho\left(1 - e^{-\rho(x+a)}\right), \tag{20}$$

and thus we can recognize that whether or not $V'(a)$ will be 0 for some $a \in \mathbb{R}_+$ purely depends on the sign of γ and the initial sign of $V'(a)$ at $a = 0$. Note that this does *not* necessarily imply convexity or concavity: $V''(a)$ need not match γ in sign. Hence, $V'(a)$ may fluctuate between increases and decreases across values of $a \in \mathbb{R}_+$, but it will cross 0 at most once on this range.

This leads us to consider when $V'(a) = 0$. Rearranging $\gamma(1 - e^{-\rho(x+a)}) - \rho\beta - \gamma\rho a$, we find the following first order condition: a is a candidate optimal threshold solution, if and only if

$$e^{-\rho a} = (1 - \rho\beta/\gamma - \rho a)\, e^{\rho x}. \tag{21}$$

Now, let us notice that, as functions of a, the left-hand side of Equation (21) is a decaying exponential (exponential with negative rate $-\rho < 0$) and right-hand side is simply a linear function with slope $-\rho e^{\rho x} < -\rho$.

Hence, the right-hand side function will intersect the left-hand side function at most once on $a \in \mathbb{R}_+$. To evaluate where this occurs, let us proceed case wise.

Beginning with $\gamma > 0$, we can see that, by definition, this implies that $\beta > 0$ also. Furthermore, the definitions of β and γ also reveal that

$$\frac{\beta}{\gamma} = x + \frac{1}{\gamma}\left(h_T T + \frac{h_H \theta_T T}{\theta_H}\right) > x,$$

and thus we have that

$$\left(1 - \frac{\rho\beta}{\gamma}\right) e^{\rho x} < (1 - \rho x) e^{\rho x} \leq 1.$$

Therefore, the left-hand side of Equation (21) at $a = 0$ is strictly greater than the right-hand side of (21) at $a = 0$, implying that, respectively, this exponential function is always above the negative slope line, and thus there is no solution to the first order condition in this setting. By applying these arguments to Equation (19) and recalling that $\gamma > 0$ in this case, Equation (20) then shows that $V'(a) < 0$ for all $a \in \mathbb{R}_+$. For $\gamma = 0$, we can quickly recognize from Equation (19) that, again, $V'(a) < 0$ for all a.

Let us now suppose that $\gamma < 0$. Through Equation (20), we have that, once $V'(a) > 0$, it will remain positive for all increasing values of a. So, we now partition the $\gamma < 0$ case into sub-cases evaluating the initial sign of $V'(a)$ at $a = 0$. Here, we see that, at $a = 0$, $\gamma(1 - e^{-\rho(x+a)}) - \rho\beta - \gamma\rho a = \gamma(1 - e^{-\rho x}) - \rho\beta$. In sub-case that $\beta \geq 0$, or, equivalently, $-\left(h_T T + h_H \theta_T T/\theta_H\right)/x \leq \gamma < 0$, we find that

$$\gamma(1 - e^{-\rho x}) - \rho\beta < 0,$$

and thus $V(a)$ is decreasing at $a = 0$. This also implies that the right-hand side of Equation (21) starts above the exponential in the left-hand side of (21), ensuring that there will be a unique solution to the first order condition on $\mathbb{R}_+$. Similarly, if $\beta < 0$ but $\gamma(1 - e^{-\rho x}) < \rho\beta$ still holds, then the same arguments apply.

Finally, if $\beta \leq \gamma(1 - e^{-\rho x})/\rho$ with $\gamma < 0$, then $V'(a) \geq 0$ at $a = 0$, and, by Equation (20), it will remain so for all $a \in \mathbb{R}_+$. Q.E.D.

Proof of Theorem 1. We begin by proving the first statement: $a^*_\infty > 0$ if and only if $T_{LB} < T < T_{UB}$ under the case that $\gamma \geq 0$. If $\gamma \geq 0$, then by Proposition 3, $a^*_\infty = \bar{A} = \bar{S} - x - T\theta_T$. Hence, it is immediately true that $a^*_\infty > 0$ if and only if $T < T_{UB}$. Since $T_{LB} \leq 0$ by consequence of $\gamma \geq 0$, we complete the proof in this setting.

If $\gamma < 0$, Proposition 3 provides that $a^*_\infty = (\tilde{a} \wedge \bar{A})$ if $\beta > \gamma(1-e^{-\rho x})/\rho$, where $\tilde{a} > 0$ is given by Equation (6). By the preceding arguments, notice that if and only if $T \geq T_{UB}$, then $\bar{A} = 0$. Now, we can further observe that among the streamlined model coefficients, α, β, γ, only β depends on T. Specifically, with the additionally defined η, we have that $\beta = \gamma x + \eta T$. Hence, the condition for $\tilde{a} > 0$ can be re-expressed to

$$\gamma x + \eta T > \gamma(1 - e^{-\rho x})/\rho,$$

and this immediately simplifies to $T > T_{LB}$. Hence, we have that $\tilde{a} > 0$ if and only if $T > T_{LB}$ and that $\bar{A} > 0$ if and only if $T < T_{UB}$, which proves that $a^*_\infty > 0$ if and only if $T \in (T_{LB}, T_{UB})$. In particular, $a^*_\infty = (\tilde{a} \wedge \bar{A}) > 0$ on this interval. Moreover, let us observe that the argument of the Lambert-W function in the expression for $\tilde{a}$ in (6) simplifies to

$$-e^{-\rho x + \frac{\beta\rho}{\gamma} - 1} = -e^{-\rho x + \frac{\rho}{\gamma}(\eta T + \gamma x) - 1} = -e^{\frac{\eta\rho}{\gamma}T - 1}.$$

Likewise, Equation (6) itself simplifies to

$$x + \tilde{a} = \frac{1}{\rho}\left(1 + W\left(-e^{\frac{\eta\rho}{\gamma}T - 1}\right)\right) - \frac{\eta}{\gamma}T. \tag{22}$$

Considering each of the two components of $(\tilde{a} \wedge \bar{A})$ individually, let us observe how they each depend on T. Starting with $\tilde{a}$, by Equation (22), we can see that

$$\frac{\partial \tilde{a}}{\partial T} = \frac{1}{\rho}\frac{\partial}{\partial T} W\left(-e^{\rho\eta T/\gamma - 1}\right) - \frac{\eta}{\gamma}.$$

Using the fact that $\mathrm{d}W(z)/\mathrm{d}z = W(z)/(z(1+W(z)))$ for $z \in (-1/e, 0)$, this simplifies to

$$\frac{\partial \tilde{a}}{\partial T} = -\frac{\eta}{\gamma}\frac{1 - W\left(-e^{\rho\eta T/\gamma - 1}\right)}{1 + W\left(-e^{\rho\eta T/\gamma - 1}\right)}.$$

Because $\gamma < 0$ and because the principal branch Lambert-W function is greater than -1 for all arguments greater than $-1/e$, we have that $\partial\tilde{a}/\partial T > 0$ for all values of T. Turning to the second component within the minimum, we can quickly observe from the definition of $\bar{A}$ that

$$\frac{\partial \bar{A}}{\partial T} = -\theta_T.$$

Thus, the dependence of a^*_∞ on T is clear: starting from T_{LB}, a^*_∞ increases according to $\tilde{a}$ until $\tilde{a}$ intersects $\bar{A}$, and then decreases from this point until reaching T_{UB}. Hence, we can find that this change point is given by the unique T at which $\tilde{a} = \bar{A}$. Setting the two quantities equal to one another, we have

$$\frac{1}{\rho}\left(1 + W\left(-e^{\frac{\eta\rho}{\gamma}T - 1}\right)\right) - \frac{\eta}{\gamma}T - x = \bar{S} - x - T\theta_T,$$

and this simplifies to the definition of $\hat{T}$ in Equation (8). Q.E.D.

Proof of Theorem 2. Recall the following notation and previously proven results

1. $a^*_\infty := \arg\min_{a\in\mathcal{A}} V(a)$
2. $a_{\min} := \arg\min_{a\in\mathcal{A}} W_T(a)$
3. Both a^*_∞ and $a_{\min}$ are unique.
4. Based on the analysis in the proof of Proposition 1, depending on the problem parameters, there are 3 possible ways $W_T(a)$ behaves as a function of a.
   (a) $W_T(a)$ is strictly increasing, then $a_{\min} = 0$. Importantly and in particular, $W_T(a)$ is strictly increasing to the right of $a_{\min}$.
   (b) $W_T(a)$ is strictly decreasing, then $a_{\min} = a_{max}$. Importantly and in particular, $W_T(a)$ is strictly decreasing to the left of $a_{\min}$. Meaning, as we decrease a, starting from $a_{\min}$, the value of $W_T(a)$ increases.
   (c) $W_T(a)$ has a unique minimum in $(0, a_{max})$, it strictly decreases before it and strictly increases after.
5. The conclusion from the item above is that if we pick any $\tilde{a} \in \mathcal{A}$ which satisfies $\tilde{a} \neq a_{\min}$ (but both $\tilde{a} > a_{\min}$ and $\tilde{a} < a_{\min}$ are possible), then if we move from $\tilde{a}$ to $a_{\min}$, the value of $W_T(a)$ **strictly decreases**.
6. Based on the analysis in the proof of Proposition 3, depending on the problem parameters, there are 3 possible ways $V(a)$ behaves as a function of a.
   (a) $V(a)$ is strictly increasing.
   (b) $V(a)$ is strictly decreasing
   (c) $V(a)$ decreases, then has a unique minimum in $\mathbb{R}_+$, then strictly increases.
7. From the last item, we can conclude that if we move from a^* to any other $\tilde{a} \in \mathcal{A}$, $V(a)$ **strictly increases**. We can also deduce that $V'(a)$ can be zero at most once, and that if it does, then this point is a minimum.

First, if $W_T(a_{\min}) = C$, since $a_{\min}$ is unique, $a_{\min}$ is the only feasible value for a in $\mathcal{A}$, and therefore it is the unique solution, i.e., $a^*_C = a_{\min}$. Next, assume that $W_T(a_{\min}) < C$. If $W_T(a^*_\infty) \leq C$, then a^*_∞ is feasible and uniquely minimizes $V(a)$ in $\mathcal{A}$. Thus it is the unique solution, i.e., $a^*_C = a^*_\infty$.

We are left with the case where $W_T(a_{\min}) < C$ and $W_T(a^*_\infty) > C$. In particular, we must have $a_{\min} \neq a^*_\infty$. By the properties listed above, when we start at a^*_∞ and go towards $a_{\min}$, $W_T(a)$ must strictly decrease and

$V(a)$ must strictly increase. Since $W_T(a)$ is continuous, there must be a value for a, call it $\tilde{a}$, strictly between a^*_∞ and $a_{\min}$ for which $W_T(\tilde{a}) = C$, which also means $\tilde{a}$ is feasible. Additionally, any other value for a before we reach $\tilde{a}$ must have $W_T(a) > C$ and hence is not feasible. Any value of a after $\tilde{a}$ must have a larger value for $V(a)$, which we are trying to minimize. We can conclude that there exists a unique solution a^*_C for the optimization problem and it is given by the unique solution to the equation $W_T(a) = C$. Q.E.D.

Proof of Proposition 4. Throughout this proof we assume that $W_T(a_{\min}) < C$. First, if $W_T(a^*_\infty) \leq C$, then $\Gamma = 0$, and problems (11) and (13) are identical and their solution is a^*_∞. Theorem 2 assures us that in this case, the solution to (12) satisfies that $a^*_C = a^*_\infty$, which proves the desired result.

We turn to the case where $W_T(a^*_\infty) > C$. In this case, $\Gamma > 0$ and Theorem 2 assures us that $a_{\min} \neq a^*_\infty$ and that a^*_C is the unique value of $a \in \mathcal{A}$ strictly between $a_{\min}$ and a^*_∞ such that $W_T(a) = C$. In particular, a^*_C must be an internal point in $\mathcal{A}$ and $W'_T(a^*_C) \neq 0$.

Next, we leverage a structural property inherent in $V(a)$. Recall that

$$V(h_R, h_H, a) = h_R W_R(a) + p_x(a) h_T T + h_H W_H(a),$$

and, therefore,

$$\begin{aligned} V(h_R, h_H, a) + \Gamma W_T(a) &= h_R W_R(a) + p_x(a) h_T T + h_H W_H(a) + \Gamma W_R(a) + \Gamma W_H(a) \\ &= (h_R + \Gamma) W_R(a) + p_x(a) h_T T + (h_H + \Gamma) W_H(a) = V(h_R + \Gamma, h_H + \Gamma, a), \end{aligned}$$

Namely,

$$V(h_R + \Gamma, h_H + \Gamma, a) = V(h_R, h_H, a) + \Gamma W_T(a). \tag{23}$$

Taking the derivative of the right-hand side with respect to a and using the definition of Γ, we obtain:

$$(V(h_R, h_H, a) + \Gamma W_T(a))' = V'(h_R, h_H, a) - \frac{V'(h_R, h_H, a^*_C)}{W'_T(a^*_C)} W'_T(a).$$

Clearly, this derivative equals zero for $a = a^*_C$. By (23), this also means that the derivative of the left hand-side is zero for $a = a^*_C$. However, from the analysis in the proof of Proposition 3, we know that $V'(h_R, h_H, a)$ (for *any* $h_R, h_H > 0$) can be zero at most once in $\mathbb{R}_+$. Moreover, if $V'(h_R, h_H, \tilde{a}) = 0$ for $\tilde{a} \in (0, \bar{A})$, then $\tilde{a}$ is a unique global minimum of $V(h_R, h_H, a)$ in $\mathcal{A}$. Therefore, a^*_C is the unique solution to (12), which concludes the proof. Q.E.D.