Optimal Call-In Policies under Travel-Induced Risk: Application to Hybrid 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

Hybrid hospitals combine on-site hospitalization with remote care via telemedicine, requiring new operational policies to balance costs, efficiency, and patient well-being across both care modalities. We address two key questions: (i) how to assign patients to remote or on-site care based on individual characteristics and proximity, and (ii) how to optimally allocate shared medical resources between care modes and patient types.

We develop a stochastic model using Brownian Motion to capture the randomness in recovery and travel-related risk during remote and on-site care. While the optimal call-in threshold is shaped by a cost-minimization objective, its behavior – specifically, its non-monotonicity in travel time and the narrowing of the effective distance range for more severe cases – is also driven by clinical constraints on allowable delays before hospital admission. These constraints, motivated by medical guidelines, limit the threshold and lead to cases where remote hospitalization becomes infeasible for very distant or severely ill patients. Under resource constraints, the optimal solution mirrors a simultaneous increase in remote and on-site costs relative to the abundant-resource case. For multiple patient types, we characterize how optimal thresholds shift with resource availability.

Our findings indicate that distant patients may at times be better served by on-site care. This outcome arises not purely from economic trade-offs, but from the interplay between clinical constraints (e.g., safe limits on call-in delays), operational considerations, and treatment costs. These insights can help inform healthcare decision-makers and policymakers in designing hybrid care systems.

Key words: State-dependent service, individual-level modeling, healthcare resource sharing, telemedicine

1. Introduction

The COVID-19 pandemic has significantly accelerated the adoption of telemedicine, which now plays a major role in healthcare delivery (Bokolo 2020, Kadir 2020). Telemedicine enables real-time remote clinical services, connecting patients and providers via video conferencing and monitoring (Monaghesh and Hajizadeh 2020), with benefits such as travel cost savings, reduced disease exposure, and improved healthcare efficiency (Hur and Chang 2020).

Recent advances have extended telemedicine to remote hospitalization, offering home-based alternatives to traditional care (Zychlinski et al. 2024). Institutions like Sheba Beyond, affiliated with Sheba Medical Center, deliver remote examinations, monitoring, and rehabilitation, aiming

to improve access to top-tier medical care. Similar programs are emerging worldwide, including in Australia (Hutchings et al. 2021), China (Francis et al. 2021), and the U.S. In the U.S., 186 hospitals joined Medicare's "Acute Hospital Care at Home" program in its first year (Clarke et al. 2021), allowing inpatient-level care at home. McKinsey & Company projects that virtual hospitals could free bed capacity, reduce hospital construction needs, and save hundreds of millions of dollars (Boldt-Christmas et al. 2023). The American Hospital Association similarly promotes this model via its "Value Initiative."

Naturally, these programs often focus on rural patients, who face significant barriers to care due to distance. Millions in the U.S., China, Brazil, and England live in "hospital deserts" (Behrman et al. 2021, Jiao et al. 2021, Gong et al. 2021, Noronha et al. 2020, Verhagen et al. 2020), where access delays and limited transportation options lead to worsened health outcomes (Kelly et al. 2014). A 2024 report by the Center for Healthcare Quality and Payment Reform warns that over 100 U.S. rural hospitals have closed in the past decade, with over 700 more at risk, including 360 facing immediate threat.

This strained rural healthcare landscape is compounded by growing health disparities between rural and non-rural populations (Lewis 2022). In the U.S., despite overall mortality declines from 1999 to 2019, rural communities have seen widening death gaps across leading causes such as heart disease, cancer, and respiratory illness (Curtin and Spencer 2021, Cross et al. 2021). Rural areas have also faced higher COVID-19 death rates (Ullrich and Mueller 2023), rising "deaths of despair" from addiction, overdoses, and suicides (Case and Deaton 2015, 2017), and increased mortality from injuries (Olaisen et al. 2019).

While telemedicine offers promise for improving rural healthcare access (Ishfaq and Raja 2015) and could mitigate hospital closures, this paper highlights a critical operational issue: the patients who stand to gain most from home hospitalization also face elevated risk when called in for hospital care. Distant patients who must travel greater distances may arrive in worse condition, resulting in longer and more costly hospital stays. Hybrid hospitals that combine on-site and remote care thus create new operational challenges, requiring models and policies that balance cost-effectiveness with clinical and operational considerations. The often-overlooked impact of patient travel motivates this paper's first-order, static-planning analysis of hybrid hospitalization.

More specifically, our study focuses on a hybrid hospital model incorporating a virtual Emergency Department (ED), where patients undergo remote examinations and consultations upon experiencing illness. Based on these assessments, physicians either admit patients to remote hospitalization or refer them for immediate on-site admission. For remotely admitted patients, physical examinations are performed using telehealth technologies such as TytoCare®, which enable remote collection of clinical data, visuals, and diagnostic information (Zychlinski et al. 2024). Physicians

review these inputs, provide visit summaries, and may issue lab orders, prescriptions, or treatment instructions. Remote patients either recover and are discharged or, if their condition worsens, are called in for on-site hospitalization, a scenario we refer to as "call-in."

A key feature of our model is the explicit incorporation of deterioration risk during patient transport, a phenomenon supported by well-documented clinical evidence. Patients transferred to the hospital after receiving remote care typically represent a vulnerable subset whose condition has already worsened, making them particularly susceptible to further deterioration during transport. Several studies across various clinical domains reinforce this concern. For example, Parmentier-Decrucq et al. (2013) report that 45% of critically ill patients experience adverse events during intrahospital transport, with 26% classified as serious, including hypotension and hypoxia. Similarly, Jia et al. (2016) find that 79.8% of transports involve at least one adverse event, with 33.2% being severe, such as cardiac arrest or severe hypotension. The physiological challenges associated with transport—including limited monitoring, movement-induced instability, and physical and psychological stress—are well recognized (Fanara et al. 2010, Droogh et al. 2015, Murata et al. 2022). In the specific context of hospital-at-home programs, Levine et al. (2020) and Leff et al. (2005) emphasize that patients who require escalation to in-hospital care following remote monitoring are typically those who have already experienced clinical deterioration, further highlighting their vulnerability during transfer.

The first question we address is how to optimally set the call-in policy to minimize total operational cost. Based on each patient's characteristics, the hybrid hospital must decide whether to admit the patient remotely or on-site. For remote admissions, it must also determine at which health condition the patient should be called in. These decisions depend on marginal treatment costs, patient proximity, and the risk of deterioration during travel.

The second question concerns resource allocation across remote and on-site care. In our motivating example from Sheba Beyond, medical staff are split into two teams responsible for either remote or on-site patients—a structure we adopt in our model. Thus, resource allocation interacts directly with the call-in policy, as call-in decisions affect the workload of each team.

To address both questions, we develop a model centered on the stochastic progression of each patient's health, represented by an acuteness score aggregating clinical measures to support discharge decisions. Such scores are widely used in practice, including the Aldrete score for post-surgical discharge (Aldrete 1994), pneumonia severity indices (Capelastegui et al. 2008), the Anderson-Wilkins score for cardiac patients (Anderson et al. 1992), and the ADL score for SNFs and rehabilitation (Bowblis and Brunt 2014).

We capture the system's dynamics by modeling the evolution of each patient's health condition during remote and on-site hospitalization using drifted Brownian motions (BMs), with parameters

that depend on patient-specific characteristics. This level of modeling detail allows us to reflect the fact that, while patients generally improve on average during treatment, random fluctuations in the recovery process may still lead to temporary or permanent deterioration. The relevant quantities in our analysis are hitting-time statistics—both expectations and probabilities—that determine the length of stay (LOS) in each care modality and parsimoniously capture the risk that a patient under remote care may need to be called in to the hospital due to clinical deterioration.

Modeling each patient's health evolution as a drifted Brownian motion provides both flexibility and analytical tractability in linking service design decisions to clinical and operational outcomes. While we do not intend this model to serve as a literal depiction of how a patient's severity evolves in practice, it offers a parsimonious abstraction that captures the essential interplay between recovery, deterioration, and travel-related risk, enabling transparent analysis without sacrificing key qualitative behaviors. This model enables the derivation of practically meaningful and implementable policies. In contrast, deriving informative solutions under more general stochastic processes would be significantly more complex and potentially impractical.

An additional advantage of the BM modeling approach lies in its empirical accessibility: its parameters can be estimated directly from historical data using relatively simple procedures. In particular, the hitting times of a drifted Brownian motion follow an inverse Gaussian distribution, a distribution that has been widely used to model LOS in healthcare systems. This well-established statistical foundation further supports the relevance and applicability of our modeling framework (see Remark 2 and Appendix D for further details).

Our work sheds light on the complex operational challenges in managing hybrid hospitals. We focus on the *design* of the health network, optimizing both the treatment mix for each patient profile and the allocation of shared resources across remote and on-site care. We address these questions through a static-planning model of the hybrid system.

The main contributions of this paper are as follows:

Optimal design of hybrid hospitalization: We study the operations, design, and management of a hybrid acute-care system combining on-site and remote hospitalization. By modeling patients' health evolution with stochastic dynamics that depend on care type, we derive optimal treatment mixes and call-in policies based on patient characteristics and travel distance. We explicitly analyze how marginal cost differences and patient proximity jointly determine when remote hospitalization is preferable.

First-principles modeling at the patient level: We model patient health trajectories directly, rather than adopting a queueing-first approach. Brownian motion offers a natural framework to capture how travel time interacts with clinical severity. This structure yields tractable, flexible models directly linked to key managerial decisions—call-in thresholds and resource allocation—without requiring restrictive queueing assumptions.

Consequences of patient travel and the risk of distance: While home hospitalization is often viewed as a means to improve access for rural patients, our model reveals that distant patients may at times be better served by on-site care. This outcome stems from a clinically motivated constraint that limits the maximum allowable delay before patient transfer, which becomes increasingly restrictive as travel time grows. As a result, the range of distances for which remote hospitalization is feasible contracts with worsening patient condition, even if remote care is otherwise less costly. These findings highlight how operational and clinical constraints, beyond economic trade-offs, can shape care decisions in hybrid hospital systems.

Allocation strategies in hybrid healthcare networks: We extend the model to settings with limited resources, such as medical staff, allocated between care modes. Depending on the ratio of improvement rates, we identify three structural cases governing the system workload. Scarcity shifts both on-site and remote costs upward by the same amount while preserving the solution structure from the unconstrained case. Importantly, optimal resource allocation can be non-monotonic: as total resources decline, a larger fraction may sometimes be allocated to one care mode over the other.

Management across heterogeneous populations: When multiple patient types compete for limited resources, optimal call-in thresholds and workload allocations shift in complex ways as resources change. Call-in policies must adjust to resource availability to maintain cost-effective care. By first analyzing the single-type case, we derive structural insights that enable full characterization of the multi-type optimization despite its non-convexity.

The rest of the paper is organized as follows. Section 2 reviews related literature. Section 3 presents the stochastic model and optimization framework. Section 4 develops the core results for a homogeneous patient population, and Section 5 extends the analysis to multiple patient types competing for limited resources. Section 6 concludes and outlines directions for future research. All proofs are provided in the appendix.

2. Literature Review

This paper is related to two main lines of literature. The first applies queueing systems and stochastic processes 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 models and queueing systems have been used to address many different healthcare applications and derive associated 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 queueing theory assume that service times are independent random variables with fixed, if not

identical, 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 and LOS (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 suggested an efficient dynamic heuristic that balances personalized readmission-risk prediction and ward congestion. Perhaps most similar to our setting is Armony and Yom-Tov (2021), which 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. Relative to Shi et al. (2021), Armony and Yom-Tov (2021), and the prior literature, we go beyond a single-station analysis to study a new setting: the hybrid hospital, which includes both on-site and remote hospitalization. The direction in which our patients may move is also a relevant contrast; for instance, Armony and Yom-Tov (2021) focuses on the flow of patients from the on-site hospital to home, whereas we are concerned with the cases in which a patient's deterioration in health status (or the risk thereof) necessitates that they travel from home to be treated on-site.

Our focus includes decisions on each patient's hospitalization option, call-in thresholds for remote patients, and resource allocation, considering patient characteristics and distance from the hospital. Furthermore, to the best of our knowledge, the present paper is the first to study how patient travel time (or distance) impacts *both* the severity of the patient's health condition and the performance of a telemedicine system. Our results show that travel is indeed an important operational factor for hybrid health networks. The stochastic model we use to capture the evolution of patients' health over time (in both hospitalization options and during travel) enables us to derive structural solutions and insights. These insights also address the question of whether and how telemedicine-based hospitalization can mitigate rural and non-rural healthcare disparities.

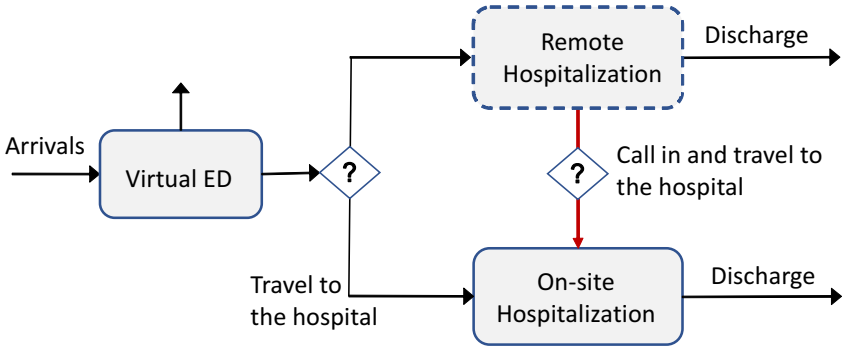
Our work also contributes to the literature on health progression modeling, which has primarily focused on discrete state models. For instance, Shi et al. (2021), Deo et al. (2013), and Nambiar et al. (2020) explicitly modeled the individual patient progression by using a Markov chain model. Grand-Clément 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, through 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. In this work, 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 an individual patient's health condition at each location. Being a BM model, it is fully characterized by its mean recovery speed (the drift) and variability of actual recovery (the diffusion coefficient). Modeling via drifted BMs has been used in sequential decision making and in the modeling of healthcare decisions (Siegmund 2013, Wang et al. 2010), but, to the best of our knowledge, it has not yet been used to model the progression of the patient's health condition. Nevertheless, we will show how this model reproduces some LOS distributional assumptions commonly made in the literature, and this connection shows how our model parameters can be driven by data. 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 Progressions

Since the decision to hospitalize on-site or remotely is made at patient assessment, our hybrid hospital model begins after triage via a virtual Emergency Department (ED), as depicted in Figure 1. Following assessment, patients are either admitted remotely or referred for immediate on-site care. Remotely admitted patients either recover and are discharged or, if their condition deteriorates to a threshold (set by the hospital and potentially patient-specific), they are called in to complete treatment on-site.

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



We use the terms "severity," "health condition," or "health score" interchangeably in reference to a measure of *clinical acuity*. The higher the score, the worse the health condition is. This health score will be the state by which the hybrid health model makes decisions at the patient-level. Let us illustrate this now at a sample path level for the patient severity.

Suppose that a patient arrives to the virtual ED for triage and routing with 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, we will assume that the hybrid hospital employs a policy such that the patient will remain on-site until they fully recover, meaning their health score reaches 0. If admitted remotely, the patient will likewise stay in at-home hospitalizations until their health score first reaches one of two thresholds: 0, which again denotes recovery, and $x + a(x)$ for some $a(x) \geq 0$, which is the "call-in" threshold. That is, if an at-home patient's health score reaches 0 before it reaches $x + a(x)$, they are discharged from the hybrid hospital's care. Otherwise, when their score reaches $x + a(x)$, they travel to the hospital, where they are admitted and stay there until they are healthy. Note that the call-in threshold depends on the initial health score. Nevertheless, to keep the notation concise, we will refer to $a(x)$ simply as a in what follows.

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 patient-level design decision: under what conditions should (or can) a patient be hospitalized remotely?

Remark 1 (Absolute versus relative call-in thresholds). Although we formulate the call-in decision using a relative threshold $a(x)$ that depends on the initial condition, the resulting admission rule is effectively equivalent to an absolute threshold on the patient's health score. Specifically, as will be shown later when characterizing the optimal call-in threshold (Proposition 3), patients are admitted once their health score reaches a fixed absolute level. The relative formulation is used mainly for modeling convenience.

3.1. Stochastic Dynamics of the Individual Health Score

We model the evolution of a patient's health conditions through negative-drift BMs, which capture the average rates of recovery during hospitalization as well as the randomness in the actual recovery. Specifically, the health score of a patient that begins remote hospitalization with initial health score $x > 0$ 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 eventual discharge, randomness allows the health score to potentially increase at any point, 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. The patient's LOS in remote hospitalization thus can be described as the hitting time at which the patient's severity reaches either health condition 0 (discharge) or health condition $x+a$ (called-in):

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

Accordingly, the call-in likelihood, $\mathbb{P}\left(\mathcal{B}^R(\tau_R(x,a)) = x+a\right)$, is the probability that a remotely hospitalized patient with initial health condition x will deteriorate so as to necessitate call-in before they are discharged. The expected LOS of remote hospitalization is $\mathbb{E}\left[\tau_R(x,a)\right]$, and both these mean and the call-in probability are readily available. We have

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

where we define $\rho := 2\theta_R/\sigma_R^2 > 0$. In terms of this probability, the mean LOS is given by

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

If a patient is called in to the hospital (whether at first virtual triage or after some initial home hospitalization), they must travel from their home, and, naturally, their health condition may further degrade while traveling. During this transition period, we assume that the patient's health score changes according to a random variable $Z(x,a,T)$, supported on $(-x-a,\infty)$. This formulation captures the inherent uncertainty during transport: depending on individual circumstances, the patient's condition may deteriorate, remain stable, or even slightly improve (e.g., due to spontaneous recovery). Nevertheless, clinical studies suggest that, on average, transport exposes patients to physiological risks that make deterioration more likely than improvement (Fanara et al. 2010, Parmentier-Decrucq et al. 2013, Droogh et al. 2015, Jia et al. 2016). Therefore, we model the expected value of $Z(x,a,T)$ as $T\theta_T$, where $\theta_T \geq 0$, reflecting the expected deterioration during travel, which may vary across settings and patient populations and can be small or even negligible under optimal transport conditions.

The model dynamics of a patient's severity at the hospital are similar to that of the remote case, but with the important difference that 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

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

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 $x + a + Z$ to hit zero. The expected LOS at the hospital is therefore

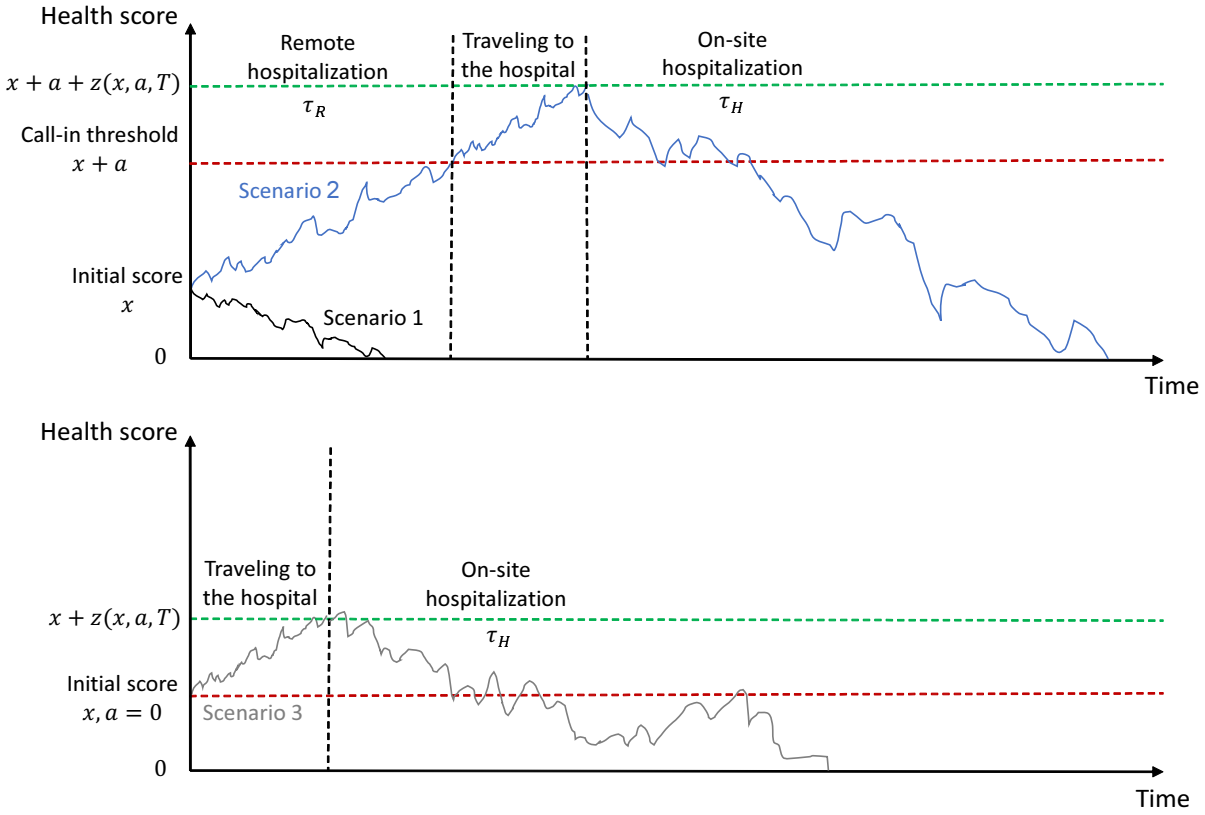
$$\mathbb{E}\left[\tau_H(x, a, Z)\right] = \mathbb{E}\left[\mathbb{E}\left[\tau_H(x, a, Z) \mid Z\right]\right] = \frac{1}{\theta_H}\mathbb{E}\left[(x + a + Z(a, x, T))\right] = \frac{1}{\theta_H}\left(x + a + T\theta_T\right). \quad (2)$$

Finally, we complete the patient-level 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 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 $\bar{A} = \left(0 \vee (\bar{S} - x - T\theta_T)\right)$, 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.

To support the practical relevance of this clinical constraint, we note that distance and travel time are indeed considered in hybrid and virtual-hospital decision making. Our clinical collaborators at Sheba Beyond report that physicians routinely account for a patient's distance from the hospital when determining whether remote hospitalization is appropriate and when a deteriorating patient should be called in. This practice aligns with empirical evidence showing that travel time meaningfully affects clinical presentation and access to care. For example, Kelly et al. (2014) document distance-related delays in reaching hospital services in rural settings, and Clark et al. (2025) show that longer travel times are associated with more severe disease at presentation and higher resource utilization. Broader reviews similarly identify travel time and distance as major determinants of access to healthcare services Mseke et al. (2024). In addition, explicit travel-time thresholds have been linked to increased clinical risk in emergency care Jang et al. (2021) and trauma settings Berry (2025). These findings collectively support modeling distance-dependent deterioration and a maximum allowable delay before hospital admission.

Figure 2 depicts three sample-path scenarios, all of which commence with a patient's health score at x and involve a travel time of T (if necessitated by the health progression). 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.

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



REMARK 2. [LOS distribution and parameter estimation] Although we are interested in this BM model for its within-care representation of the patient's health progression, let us note that its hitting-time-based LOS actually reproduces the inverse Gaussian distribution already common in healthcare modeling (e.g. Whitmore 1975, Hashimoto et al. 2023). In Appendix D, we provide more details on this connection and use it to explain how the BM model parameters can be estimated from real-world healthcare data, motivating the data requirements of our model. □

3.2. A Static Planning Problem for Multi-Patient Hybrid Health Networks

The discussion so far has centered around the primary *modeling* focus of this paper, meaning the health progression of an individual patient. Nevertheless, the foremost scope for *decision making* in this paper is the design of novel hybrid health networks that offer both at-home and on-site hospitalization to many different patients, where these multiple patient types may vie for the hybrid hospital's limited resources. Let us now connect the individual-level to the system-level in terms of both the health network's objectives and its limitations.

Average cost-of-care objective. Because the negative drifts inherently capture health conditions that eventually improve, we will take the average cost of care as the primary metric by which these system-level decisions are assessed. Building first from a single patient type for the

sake of clarity, let h_R and h_H denote the holding cost rate for remote and on-site hospitalization, respectively, for this focal patient type. Similarly, h_T denotes the cost rate during patient travel. Additionally, suppose that patients (of this particular type) arrive according to a Poisson process with rate $\lambda > 0$. Then, the total long run average cost can be written

$$\begin{aligned}&\lambda\left(h_R\mathbb{E}\left[\tau_R(x,a)\right]+\left(h_TT+h_H\mathbb{E}\left[\tau_H(x,a,Z)\right]\right)p_x(a)\right)\\&\quad=\lambda\left(\frac{h_R}{\theta_R}\left(\left(1-p_x(a)\right)x-p_x(a)a\right)+p_x(a)\left(h_TT+\frac{h_H}{\theta_H}\left(x+a+\theta_TT\right)\right)\right).\end{aligned}\tag{3}$$

We refer to (3) as the *average cost of care* for this patient type under call-in threshold a. A full justification for interpreting (3) as the system's long-run average cost appears in Appendix A. In the main text, we take this expression as the core objective in our static planning problem.

Before introducing the optimization constraints and multi-type problem, we first define condensed notation by rewriting the value function (3) as:

$$\lambda\left(\frac{h_R}{\theta_R}\left(\left(1-p_x(a)\right)x-p_x(a)a\right)+p_x(a)\left(h_TT+\frac{h_H}{\theta_H}\left(x+a+\theta_TT\right)\right)\right)=\lambda\left(\alpha+\beta p_x(a)+\gamma p_x(a)a\right),\tag{4}$$

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

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

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

In addition to the shorter expression, each of α, β, and γ offer interpretation to the call-in decision. 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.

Workload across care modalities. The hospital allocates resources, primarily medical staff, between two groups: the on-site team treating in-hospital patients and the virtual team managing remote patients. Focusing on a single patient type for clarity, we first define the offered workload for each group. The on-site workload for this patient type is given by:

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

and likewise denote the remote workload as

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

Finally, let the total workload therefore be

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

Naturally, total care workload is subject to resource constraints, which must be shared across multiple patient types.

Notice that the workloads per patient for on-site and remote care receive the same weight in the definition of the total workload. In practice however, the amount of resources consumed per hour might be different for these two types of care, requiring the total workload to be defined as a weighted sum: $W_S(a) := dW_H(a) + (1-d)\,W_R(a)$ for some $d \in (0,1)$. In Appendix B, we discuss how our model and results apply to this case. While the formulation is slightly modified to incorporate d, the structure of the optimal policy, its properties, and the main insights remain unchanged.

Minimal cost-of-care subject to limited shared resources across many patient types. Thus far, we have described the stochastic model and health score dynamics at the individual patient level. In practice, however, health systems must allocate medical resources across heterogeneous patient types, who compete for limited capacity.

Suppose the hybrid hospital serves patient types indexed by $k = 1, \ldots, K$. We augment the prior notation with superscripts for each type: x^k is the initial health score, T^k is the travel time, and θ_H^k is the hospital recovery rate for type k. Let λ^k denote the Poisson arrival rate for type k, with independent arrival processes across types. This structure accommodates heterogeneity in both severity and distance: some types may share travel times with varying initial conditions, others may share severity but differ in distance, and others may differ in both.

Each patient type has its own call-in threshold a^k, and we use the vector $\vec{a} = (a^1, \ldots, a^K)$ to denote the collection of thresholds, with feasible set $\vec{\mathcal{A}} = \left[0, \bar{A}^1\right] \times \cdots \times \left[0, \bar{A}^K\right]$.

Given a total resource capacity C shared across modalities and patient types, the hybrid hospital's static planning problem is:

$$\begin{aligned} \min_{\vec{a} \in \vec{\mathcal{A}}} V(\vec{a}) &= \sum_{k=1}^{K} \lambda^k \left(\alpha^k + \beta^k p_x^k(a^k) + \gamma^k p_x^k(a^k) a^k\right) \\ \text{s.t. } & \sum_{k=1}^{K} W_S^k(a^k) \le C. \end{aligned} \tag{5}$$

For each type k, we denote the optimal threshold by $a^*_{C,k}$ to emphasize its dependence on total resources. The solution balances the total workload across on-site and remote care, ensuring that $\sum_k W_T^k(a^k)$ does not exceed C. The behavior of W_H^k and W_R^k as functions of a^k determines which thresholds are feasible, directly reflecting the impact of resource scarcity, as further analyzed in Section 4.1.

Finally, the existence of a feasible solution depends on both λ^k and C. Even with one patient type, a large arrival rate λ^1 combined with limited capacity C may render the workload constraint infeasible for any $a \in \mathcal{A}$. Section 4.1 characterizes the feasibility region in terms of (λ, C) pairs.

Remark 3 (Workload and waiting-time interpretation). While our analysis focuses on finite resource allocation, the workload constraint in (5) can also be interpreted as limiting waiting times in the hybrid network. For example, replacing C with a smaller $\tilde{C} \leq C$ yields the equivalent constraint $1/(C - W_S(a)) \leq 1/(C - \tilde{C})$, bounding the inverse idleness. This term appears in utilization-based approximations for mean waiting times in $GI/G/C$ queues (see, e.g., Whitt 1993), which, though imprecise (Gupta et al. 2010), capture the well-known exponential growth of waiting with utilization. Thus, (5) can be interpreted as minimizing cost subject to implicit constraints on waiting times or desired staffing regimes. □

The static planning problem in (5) is generally complex and not analytically transparent. We therefore first analyze the case of a single patient type to build intuition before turning to the multi-type setting. When $K = 1$, the problem reduces to:

$$\begin{aligned} \min_{a \in \mathcal{A}} V(a) &= \lambda \left(\alpha + \beta p_x(a) + \gamma p_x(a) a\right) \\ \text{s.t.} \quad & W_S(a) \leq C. \end{aligned} \tag{6}$$

This single-type planning problem will be the focus of the next two sections. As before, we drop the k superscripts when focusing on a single patient type.

4. Minimizing the Cost-of-Care for a Single Patient Type

To identify the optimal call-in threshold and the resulting allocation between on-site and home care, we first analyze how system performance depends on this threshold. This section examines the total workload for a single patient type and characterizes the feasible region of problem (6).

The pivotal factor influencing this behavior is the ratio of relative recovery rates: θ_H/θ_R. Additionally, let $\Delta > 0$ be defined as

$$\Delta = \frac{\rho \theta_T T}{\rho x - 1 + e^{-\rho x}}. \tag{7}$$

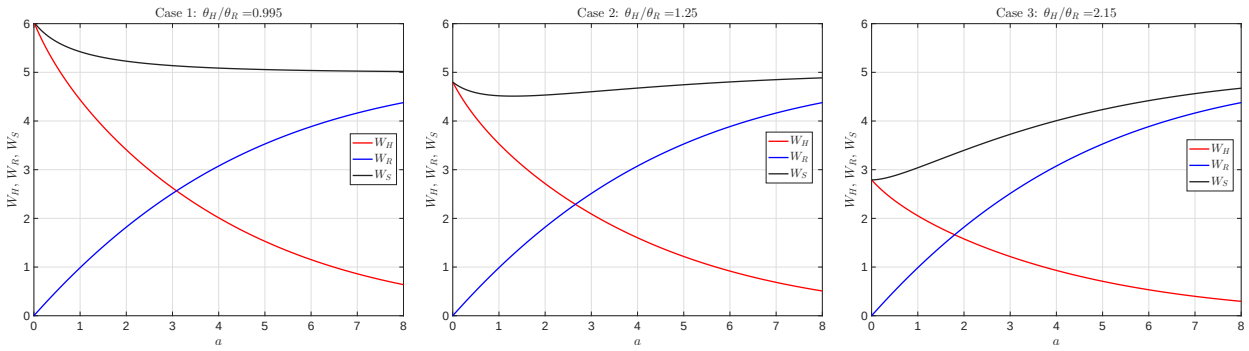
Through these two quantities, we can classify the shape of the workload as a function of the call-in threshold.

Proposition 1. *The total workload $W_S(a)$ satisfies the following:*

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

Figure 3 illustrates the three cases in Proposition 1.

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



Given the problem context, Cases 2 and 3 appear more realistic than Case 1, with Case 2 being particularly relevant. It is unlikely that home recovery would surpass the full capabilities of hospital care, and the case where the hospital is only marginally better is of particular managerial interest. The insights from Proposition 1 will be useful in the next section when analyzing the feasibility region, and will further support the capacitated analysis in Section 4.3.

4.1. 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 (6). The feasibility region of (6) is defined as:

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

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

$$a_{\min} = \arg\min_{a \in \mathcal{A}} W_S(a). \tag{8}$$

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, there exists $a \in \mathcal{A}$ such that $W_S(a) \leq C$ if and only if the solution to (8) is such that $W_S(a_{\min}) \leq C$. Since $W_S(a_{\min})/\lambda$ does not depend on λ, and $a_{\min}$ minimizes it as well, we are essentially looking for (λ, C) pairs such that $\lambda\,(W_S(a_{\min})/\lambda) \leq C$. Using this observation in tandem with Proposition 1, we obtain the following characterization of the feasibility region.

PROPOSITION 2. *The feasibility region of the optimization problem (6) is given by:*

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

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 $W_S(a)$ 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 $a_{\min} = 0$.*

Propositions 1 and 2 lay the groundwork for analyzing the full optimization problem, starting with the single-type case in (6). We first study the unconstrained problem with $C = \infty$, then build on this solution to address the finite C case. The insights from the single-type setting will also guide our extension to multiple patient types.

4.2. Optimal Call-In Structure (Unlimited Resources)

Using the notation from (4), we seek to minimize the expected cost rate subject only to the clinical bounds on patient health:

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

Proposition 3 characterizes the uncapacitated optimal call-in threshold a_∞^*, which may take values in $[0, \bar{A}]$, where $\bar{A} = \bar{S} - T\theta_T - x$. If $a_\infty^* = 0$, on-site hospitalization is always preferred; if $a_\infty^* = \bar{A}$, remote care is fully utilized until the clinical limit. For any $a_\infty^* \in (0, \bar{A})$, hospitalization begins remotely, but patients are called in before reaching the maximum allowable severity.

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*

$$e^{-\rho \tilde{a}} = (1 - \beta\rho/\gamma - \rho\tilde{a})\, e^{\rho x}. \tag{9}$$

which can be expressed by

$$\tilde{a} = \frac{1}{\rho}\left(1 + \mathsf{W}\left(-e^{-\rho x + \beta\rho/\gamma - 1}\right)\right) - \frac{\beta}{\gamma}, \tag{10}$$

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

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

We find that γ, the marginal cost difference between on-site and remote hospitalization, plays a central role in determining where to hospitalize patients and when to call them in if initially treated remotely. If $\gamma \ge 0$, remote hospitalization is always preferred. When $\gamma < 0$, on-site care is marginally more cost-effective, but remote hospitalization may still be beneficial depending on β, which captures the cost impact of immediate transfer. We next examine the role of patient travel.

Figure 4 illustrates the optimal call-in threshold and probability across transfer times and initial conditions. The parameters are based on values reported in the literature: in-hospital care is set to $h_H = 100$ (reflecting daily costs of \$2,400–\$4,000), remote care to $h_R = 60$ (reflecting 40–60% cost reductions Levine et al. 2020, The Commonwealth Fund 2020, Zychlinski et al. 2024), travel cost to \$15 per mile (based on US Government Accountability Office 2012), recovery rates to $\theta_H = 0.04$ and $\theta_R = 0.02$, and travel deterioration rate to $\theta_T = 0.2$ (motivated by Fanara et al. 2010, Parmentier-Decrucq et al. 2013, Droogh et al. 2015, Jia et al. 2016).

Remote hospitalization is not cost-effective for patients living very close or very far from the hospital, with $a_\infty = 0$ indicating direct on-site admission. The call-in threshold is non-monotonic in T: it initially increases, reaching a maximum at $\hat{T}$ (constant across severity levels), then decreases back to zero. The call-in probability exhibits the opposite pattern: decreasing before rising again. As initial severity x increases, the threshold decreases, and the distance range where remote care is viable narrows. These properties are formalized in Theorem 1.

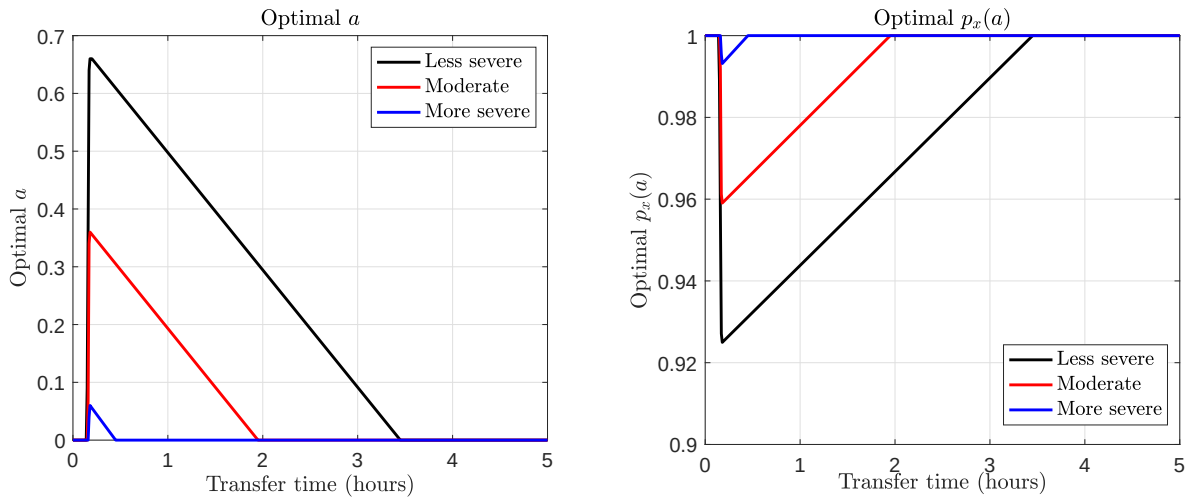
In rural areas, average travel times to hospitals typically range between 30 and 90 minutes (Texas A&M University Health Science Center 2023, Hearn et al. 2024, Rural Health Information Hub 2025b,a). According to Figure 4, home hospitalization is most effective for more severe patients who reside within approximately 10–25 minutes from the hospital. For patients with moderate initial conditions, home hospitalization remains effective for transfer times of up to two hours. For patients with milder initial conditions, the effective range extends up to 3.5 hours, making remote hospitalization particularly advantageous for this group.

Specifically, to establish 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 following Equation (4), we can recognize 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*

$$T_{LB} = -\frac{\gamma}{\eta}\left(x - \frac{1}{\rho}\left(1-e^{-\rho x}\right)\right) \qquad \text{and} \qquad T_{UB} = \frac{\bar{S}-x}{\theta_T}. \tag{11}$$

Figure 4 **Optimal call-in threshold and call-in probability as a function of travel time for different initial health scores.**



Then, if the marginal hospitalization cost is higher at the hospital ($\gamma \geq 0$), $a^*_\infty > 0$ if and only if $T < T_{UB}$.

Furthermore, if the marginal hospitalization cost is higher at home ($\gamma < 0$), then the $\hat{T}$ which is the unique solution to

$$\bar{S} = \frac{1}{\rho}\left(1 + \mathsf{W}\left(-e^{\eta\rho\hat{T}/\gamma - 1}\right) + \left(\theta_T - \frac{\eta}{\gamma}\right)\hat{T}\right), \tag{12}$$

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

$$\frac{\partial a^*_\infty}{\partial T} = -\frac{\eta}{\gamma}\frac{1 - \mathsf{W}\left(-e^{\rho\eta T/\gamma - 1}\right)}{1 + \mathsf{W}\left(-e^{\rho\eta T/\gamma - 1}\right)} > 0, \tag{13}$$

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

$$\frac{\partial a^*_\infty}{\partial T} = -\theta_T < 0, \tag{14}$$

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

Theorem 1 reveals that even when remote hospitalization has lower marginal cost, it may not be clinically appropriate for patients who live far from the hospital. This is due to a medical constraint: the combination of a maximum allowable severity level ($\bar{S}$) and the deterioration that occurs during travel ($\theta_T T \geq 0$) imposes a stricter threshold on who can safely remain at home. This effect becomes more pronounced for sicker patients: as initial severity increases, the lower bound T_{LB} rises, the upper bound T_{UB} falls, while the slope of a^*_∞ with respect to T remains unchanged. Given evidence that rural populations often present with greater severity (Lewis 2022), this interaction between distance and severity may restrict access to remote care precisely for the patients who need it most.

Notably, these clinical constraints apply even when resources are assumed to be abundant. The next section shows that incorporating capacity limits preserves this structure—and may even amplify it. Under realistic parameters, limited capacity further reduces the feasibility of remote care for patients with high severity or long travel distances, reinforcing the need to account for both operational and clinical considerations in hybrid hospital design.

Appendix E analyzes the impact of recovery variability on the optimal call-in policy and shows that higher variability consistently leads to more permissive call-in thresholds and expands the range in which remote hospitalization is effective.

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

We now return to the capacitated problem in (6), with two objectives: determining the call-in policy under finite resources and allocating resources between on-site and remote hospitalization.

To begin, Theorem 2 characterizes the solution of the capacitated problem (6).

THEOREM 2. *Assume that $W_S(a_{\min}) \le C$ for $a_{\min} \in \mathcal{A}$ as defined in (8) (i.e., the feasibility region is not empty). Then, problem (6) has a unique solution $a_C^* \in \mathcal{A}$, such that:*

- *If $W_S(a_{\min}) = C$, then $a_C^* = a_{\min}$.*
- *If $W_S(a_{\min}) < C$, then:*
  - *If $W_S(a_\infty^*) \le C$, then $a_C^* = a_\infty^*$,*
  - *If $W_S(a_\infty^*) > C$, 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 $W_S(a) = C$.*

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

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 cost-of-care function V on holding costs, we denote it as $V(h_R, h_H, a)$. Recall the uncapacitated minimization problem,

$$\min_{a \in \mathcal{A}} V(h_R, h_H, a), \tag{15}$$

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

$$\begin{aligned} &\min_{a \in \mathcal{A}} V(h_R, h_H, a) \\ &\text{s.t. } W_T(a) \le C, \end{aligned} \tag{16}$$

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

$$\Gamma = \begin{cases} -\frac{V'(h_R, h_H, a_C^*)}{W_S'(a_C^*)}, & \text{if } W_S(a_{\min}) < C \text{ and } W_S(a_\infty^*) > C \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

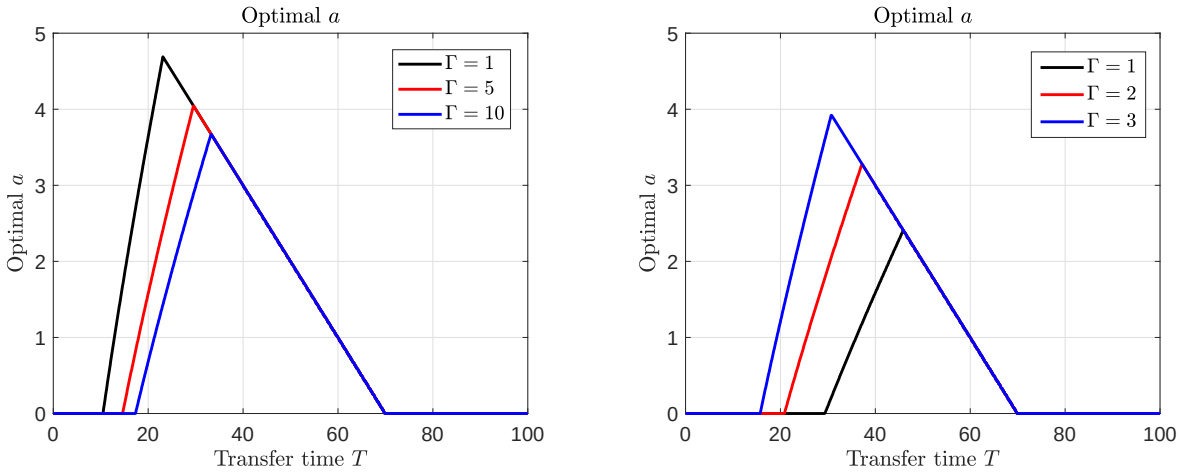
Note that $\Gamma \ge 0$, since in the case where $W_S(a_{\min}) < C$ and $W_S(a_\infty^*) > C$, $V'(h_R, h_H, a_C^*)$ and $W_S'(a_C^*)$ must be non-zero and with opposite signs (see the proofs of Proposition 1 and Lemma 4). Now, consider a similar uncapacitated optimization problem with Γ-modified costs:

$$\min_{a \in \mathcal{A}} V(h_R + \Gamma, h_H + \Gamma, a). \tag{18}$$

Proposition 4 establishes the equivalence between the solutions of (16) and (18). 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 patient travel distance on the optimal call-in policy, as indicated in Theorem 1, remain consistent.

PROPOSITION 4. *Assume that $W_S(a_{\min}) < C$ (i.e. the feasibility region of (16) contains more than one point). Then, the problem (18) has a unique solution in $\mathcal{A}$ which equals a_C^*.*

Figure 5 **Shifted call-in policy for different Γ and T. The parameters are $\theta_T = 0.1$, $h_R = 5.1$, $h_T = 2$, $x = 8$, $\bar{S} = 15$, $\lambda = \sigma_R = 1$. In the left plot, $\theta_H = 0.05$, $\theta_R = 0.06$, $h_H = 1$; on right, $\theta_H = 0.1$, $\theta_R = 0.05$, $h_H = 7$.**



The parameter Γ captures the effect of resource scarcity and provides immediate managerial insight. Under finite capacity, both remote and on-site costs increase simultaneously by Γ, shifting the call-in threshold depending on the underlying cost and recovery rates. Specifically, the adjusted cost difference is:

$$\gamma(\Gamma) = \frac{-(h_R + \Gamma)}{\theta_R} + \frac{h_H + \Gamma}{\theta_H} = \gamma + \Gamma\left(\frac{1}{\theta_H} - \frac{1}{\theta_R}\right). \tag{19}$$

Since $\Gamma \geq 0$, $\gamma(\Gamma)$ may increase or decrease depending on the relation between θ_H and θ_R. Per Proposition 3, the sign of $\gamma(\Gamma)$ determines the optimal call-in threshold and whether remote hospitalization is preferred.

For example, if $\gamma > 0$ but $\gamma(\Gamma) < 0$, patients who would stay remotely hospitalized under abundant resources may be called in earlier or admitted directly on-site under limited resources. Similarly, when $\theta_H > \theta_R$ and $\gamma < 0$, the range of distances where remote care is viable shrinks further as Γ grows, particularly for patients with worse initial health (x).

The right plot in Figure 5 illustrates this case: as Γ increases, the optimal call-in threshold declines, and the distance range for remote hospitalization contracts. Conversely, the left plot shows the opposite behavior when $\theta_H < \theta_R$, where $\gamma(\Gamma) > \gamma$. For instance, at $T = 25$, patients are called in immediately when $\Gamma = 1$, but remain under remote care for higher Γ. In both cases, Proposition 4 shows that the capacitated solution mirrors the uncapacitated problem with Γ-adjusted costs ($h_R + \Gamma$ and $h_H + \Gamma$).

Figures 6 and 7 illustrate the optimal capacitated solution for varying resource levels and travel times. Figure 6 corresponds to Case 3 in Proposition 1. In the top plots ($T = 2$), $\theta_H/\theta_R = 2.5 > \Delta + 1 = 2.14$. With ample resources ($C \geq 4$), $a^*_\infty \approx 4$. As resources become scarce, $W_S(a)$—strictly increasing in this case—forces a^*_C to decrease until $C \approx 2.5$, marking the feasibility boundary.

Note that this case arises when the on-site recovery rate is significantly higher than in remote hospitalization. In such settings, admitting patients on-site at an earlier stage of deterioration allows the system to conserve total resources.

The bottom plots ($T = 8$) show a smaller feasibility region, ending at $C \approx 3.3$, reflecting the need for more resources as distance grows.

The right plots present resource allocation between W_H and W_R. Under scarcity, most resources (e.g., 80% at $T = 2$) are allocated to on-site care; as capacity increases, more resources shift to remote care, reaching 82% allocation when resources are abundant.

Figure 7 corresponds to Case 1 in Proposition 1, where $\theta_H/\theta_R = 0.83 < 1$. Here, $W_S(a)$ is strictly decreasing. With ample resources, $a^*_\infty \approx 2$; as capacity tightens, a^*_C increases. In contrast to Figure 6, increasing resources here shifts allocation away from remote care and toward on-site hospitalization.

Remark 4. [Call-in optimization with dedicated resources] In some settings, the resources allocated to on-site and remote hospitalization are dedicated, meaning they cannot be shared across modes. This introduces two separate resource constraints—one for each hospitalization mode—which leads to the following optimization problem. We solve this problem in Appendix C.

$$\begin{aligned} \min_{a \in \mathcal{A}} & V(a) = \lambda \left(\alpha + \beta p_x(a) + \gamma p_x(a) a\right) \\ \text{s.t.} \quad & W_H(a) \leq C_H, \\ & W_R(a) \leq C_R. \end{aligned} \tag{20}$$

Figure 6 **Optimal capacitated solution. The parameters are** $\theta_H = 0.5$, $\theta_R = 0.2$, $\theta_T = 0.1$, $h_H = 2.65$, $h_R = 1.4$, $h_T = 2$, $x = 1$, $\bar{S} = 15$, $\lambda = 1$, $\sigma_R = 1$.

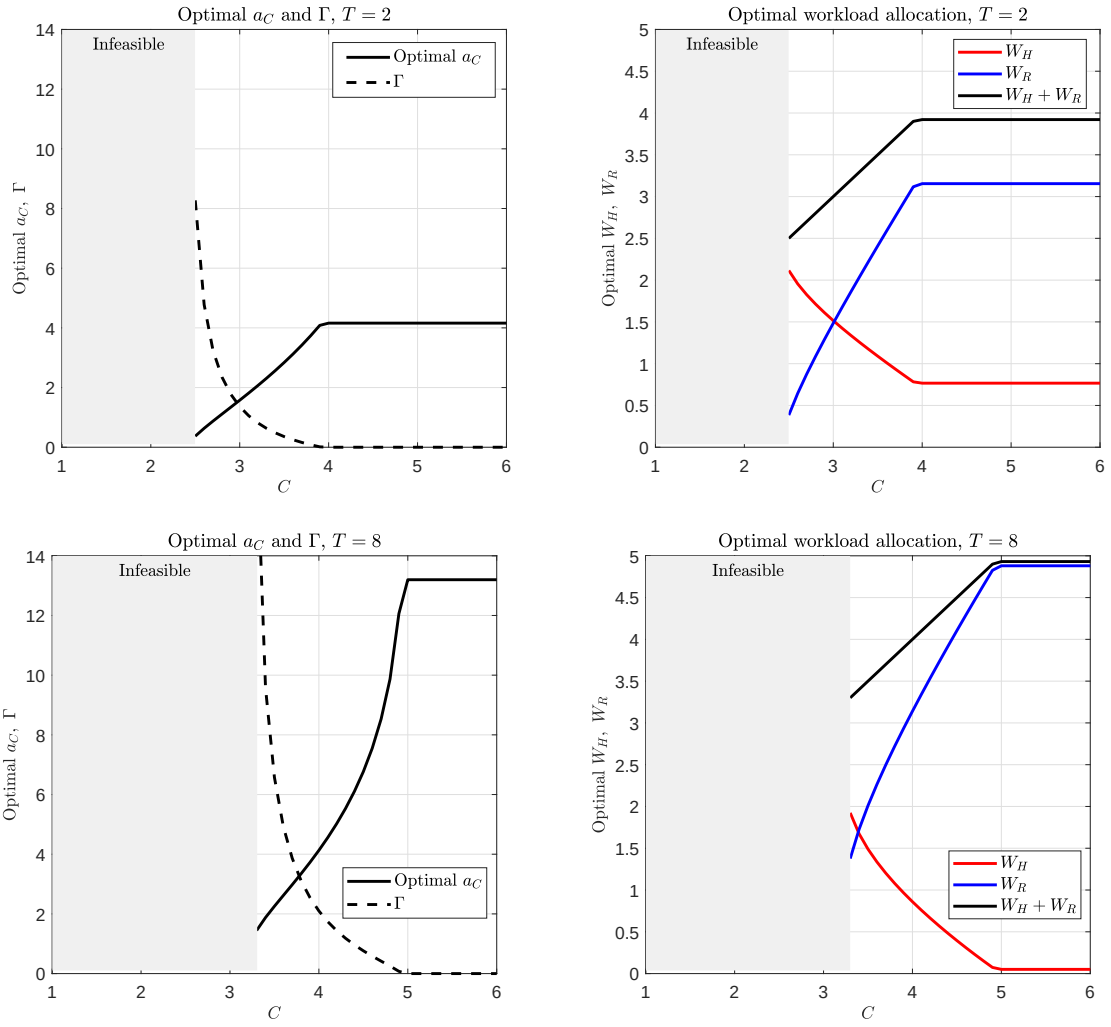
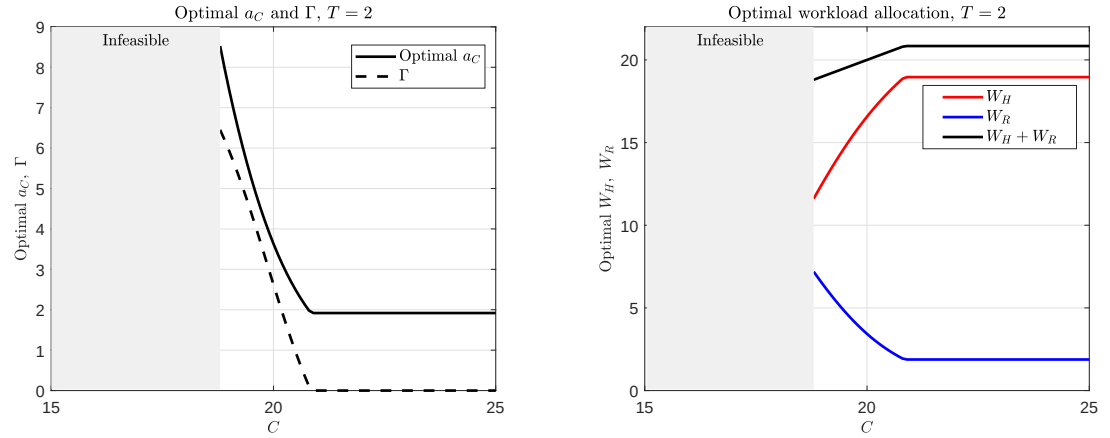


Figure 7 **Optimal capacitated solution. The parameters are** $\theta_H = 0.05$, $\theta_R = 0.06$, $\theta_T = 0.1$, $h_H = 2.65$, $h_R = 5.1$, $h_T = 2$, $x = 1$, $\bar{S} = 10$, $\lambda = 1$, $\sigma_R = 1$.



In this case, similar results to the pooled-resources case hold; however, due to the separate capacity constraints that shape the feasibility region, at most one hospitalization mode (on-site or remote) is penalized.

5. Generalization to Multiple Patient Types

With the preceding results, we return to the full static planning problem for multiple patient types defined in (5). As in the single-type case, the solution (if it exists) minimizes $V(\vec{a})$ while balancing the total on-site and remote workloads W_H^k and W_R^k across all types $k = 1, \ldots, K$, ensuring their sum does not exceed C. The dependence of W_H^k and W_R^k on a^k, established in Section 4, determines the feasible thresholds satisfying the resource constraint. Existence of a solution depends on the problem parameters, particularly $\vec{\lambda}$ and C: if arrival rates are large and capacity small, no feasible solution may exist for any $\vec{a} \in \vec{\mathcal{A}}$.

5.1. Characterizing the Feasibility Region in Multiple Patient Dimensions

Based on the given problem parameters, we characterize the feasibility region, meaning the $(\vec{\lambda}, C)$ for which there exists an $\vec{a} \in \vec{\mathcal{A}}$ satisfying the constraint in the optimization problem (5). Formally, this is defined as:

$$\mathcal{C}_{FR}^K = \left\{ (\vec{\lambda}, C) \in \mathbb{R}_+^{K+1} : \exists \vec{a} \in \vec{\mathcal{A}} \text{ s.t. } \sum_{k=1}^{K} W_S^k(a^k) \leq C \right\}.$$

We can now make a series of observations for the multi-type workload that essentially mirror what we saw for the single-type feasibility region in Section 4.1. Let $a_{\min}^k$, $k = 1, \ldots, K$, denote the value of $a^k \in \mathcal{A}^k$ for which the total workload is minimal, i.e.,

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

Note that Proposition 1 guarantees that $a_{\min}^k$ is unique. However, relative to the a_0^k in Proposition 1, each $a_{\min}^k$ is restricted to the range $\mathcal{A}^k = [0, \bar{A}^k]$, whereas $a_0^k \in \mathbb{R}_+$. By definition, there exists $\vec{a} \in \vec{\mathcal{A}}$ such that $\sum_{k=1}^K W_S^k(a^k) \leq C$ if and only if $\sum_{k=1}^K W_S^k(a_{\min}^k) \leq C$. Since $W_S^k(a_{\min}^k)/\lambda^k$ does not depend on λ^k, and $a_{\min}^k$ minimizes it as well, we are essentially looking for $(\vec{\lambda}, C)$ such that $\sum_{k=1}^K \lambda^k \left(W_S^k(a_{\min}^k)/\lambda^k\right) \leq C$.

Using this sequence of observations and Proposition 1, we obtain the following characterization of the multi-type feasibility region. Like we first saw in Proposition 2, the structure of the minimal workload for each type will depend on the comparison of the type-specific recovery rates, θ_H^k and θ_R^k, and the type-specific distance-dependent threshold, $\Delta^k = \rho^k \theta_T^k T^k / (\rho^k x^k - 1 + \exp(-\rho^k x^k))$, as generalized from (7).

PROPOSITION 5. *The feasibility region of the optimization problem (5) is given by:*

$$\mathcal{C}_{FR}^{K} = \left\{ (\vec{\lambda}, C) \in \mathbb{R}_{+}^{K+1} : \sum_{k=1}^{K} W_S^k(a_{\min}^k) \leq C \right\},$$

where for each $k = 1, \ldots, K$:

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

Much like Proposition 2, Proposition 5 reduces the multi-type feasibility region to the workload-minimizing call-in thresholds, characterized by the relationships between θ_H^k, θ_R^k, and Δ^k. This sets the stage for analyzing the optimal call-in thresholds in the general multi-type, resource-limited setting.

5.2. Scaffolding to the Optimal Multi-Type Solution

Throughout this section, we consider cases where the feasibility region is non-empty and infinite. Due to the non-convexity of both the objective and constraint in (5), finding a global optimum is challenging. Nevertheless, the problem's structure allows us to derive key insights, beginning with the existence of a global optimal solution.

LEMMA 1. *If the feasibility region is not empty, then there exists a feasible vector $\vec{a}_C^*$ which is the solution to the optimization problem (5).*

Next, we differentiate between two cases: when the vector of optimal call-in thresholds with unlimited resources is within the feasibility region and when it is not. For the first case, we have the following result.

LEMMA 2. *If $\vec{a}_\infty^* \in \mathcal{C}_{FR}^K$, then $\vec{a}_\infty^*$ is the unique solution of the optimization problem.*

Lemma 2 states the intuitive and desirable fact that if we can select the optimal solution for each patient type while still satisfying the resource constraint, then this would be the optimal outcome. However, if this threshold choice is not feasible, we must adjust the thresholds for some or all types, reducing their total workloads until the capacity constraint is met. Much like how Theorem 2 and Proposition 4 identified the structure of the optimal call-in threshold for the single-type setting, Theorem 3 characterizes the solution for the second case, in which the abundant resource solution is not attainable.

THEOREM 3. *Suppose $\vec{a}_\infty^* \notin \mathcal{C}_{FR}^K$ and let $\vec{a}_C^*$ be a solution to the optimization problem (5). Then:*

1. *Each threshold in* $\vec{a}_C^*$ *must lie between (and can be equal to)* $a_{\min}^k$ *and* $a_{\infty,k}^*$ *of the corresponding type.*
2. *The resource constraint at* $\vec{a}_C^*$ *is active, i.e., the sum of the total workloads at* $\vec{a}_C^*$ *is exactly* C*.*
3. *Denote by* E *the set of all indices for which* $[\vec{a}_C^*]_k \notin \{a_{\min}^k, a_{\infty,k}^*\}$*. Then:*
   - *There exist a unique* $\Gamma > 0$ *such that* $\Gamma = -V'_k([\vec{a}_C^*]_k)/W_T^{k\prime}([\vec{a}_C^*]_k)$*, for all* $k \in E$*.*
   - *Specifically,* $\vec{a}_C^*$ *restricted to the entries in* E *is the unique solution to the unconstrained optimization problem:* $\min_{\vec{a}\in\vec{\mathcal{A}}_E} \sum_{k\in E} V_k\left(h_R+\Gamma, h_H+\Gamma, a_k\right)$*, where* $\vec{\mathcal{A}}_E$ *is the boundary restricted to the entries in* E*.*

Theorem 3 shows that for patient types whose optimal thresholds do not lie at the boundaries $a_{\min}^k$ or $a_{\infty,k}^*$, the solution mirrors the unconstrained case but with costs shifted by a common Γ. As in the single-type setting, Γ captures resource scarcity by uniformly increasing both remote and on-site costs, shifting call-in thresholds up or down depending on each type's cost and recovery rates. The next section presents numerical examples illustrating this effect.

5.3. Numerical Examples for the Multi-Type Problem

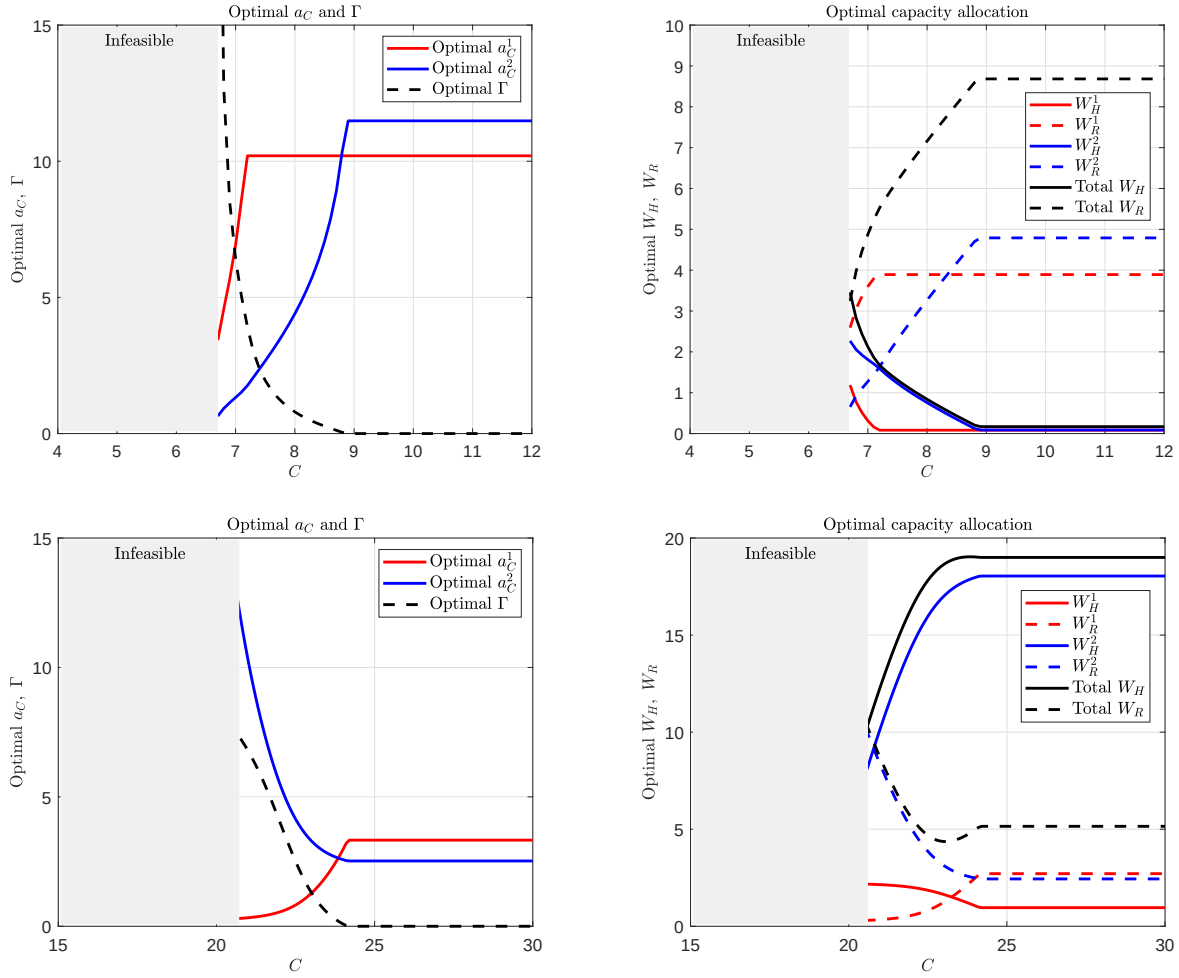
We now present two examples illustrating the optimal multi-type solution—both call-in thresholds and resource allocation—for two patient types. In the first example (top plots of Figure 8), the types differ, among other factors, in distance: Type 1 is farther from the hospital, Type 2 is closer. As expected from the single-type analysis in Section 4.2, when resources are ample, Type 1 has a lower call-in threshold than Type 2. However, as resources tighten, the optimal mix cannot be maintained for both types. Since both types fall into Case 3 of Proposition 1 (where $W_S(a)$ is strictly increasing), thresholds decrease. Notably, when $C \approx 8.8$, the thresholds cross: the distant patient (Type 1) stays longer in remote care, while the closer patient is called in earlier. As thresholds drop, on-site workload grows while remote workload declines.

In the second example (bottom plots of Figure 8), we consider two additional patient types. As resources become scarce, Type 1's threshold decreases (patients are called in earlier), while Type 2's threshold increases (patients stay longer at home). This occurs because the types fall into different cases of Proposition 1: Type 1 belongs to Case 1 ($W_S(a)$ strictly decreasing), while Type 2 belongs to Case 3 ($W_S(a)$ strictly increasing). Thus, resource scarcity affects each type differently, with call-in thresholds adjusting up or down depending on their structural case. As a result, total workload allocation between on-site and remote care may also be non-monotonic, as allocations for each type shift in opposite directions.

6. Discussion and Conclusion

The hybrid hospital model constitutes a service network design problem, where the decision to admit a patient remotely or on-site governs resource allocation across hospitalization modes. We

Figure 8 **Optimal capacitated solution with two patient types. In the top plots:** $(\theta_H^1, \theta_H^2) = (0.35, 0.5)$, $(\theta_R^1, \theta_R^2) = (0.25, 0.2)$, $(\theta_T^1, \theta_T^2) = (0.1, 0.1)$, $(h_H^1, h_H^2) = (2.65, 3)$, $(h_R^1, h_R^2) = (1.4, 1.4)$, $(h_T^1, h_T^2) = (2, 2)$, $(x^1, x^2) = (1, 1)$, $(\bar{S}^1, \bar{S}^2) = (12, 15)$, $(\lambda^1, \lambda^2) = (1, 1)$, $(\sigma_R^1, \sigma_R^2) = (1, 1)$, $(T^1, T^2) = (8, 5)$. **In the bottom plots:** $(\theta_H^1, \theta_H^2) = (0.5, 0.05)$, $(\theta_R^1, \theta_R^2) = (0.2, 0.06)$, $(\theta_T^1, \theta_T^2) = (0.1, 0.1)$, $(h_H^1, h_H^2) = (2.65, 2.65)$, $(h_R^1, h_R^2) = (1.4, 5.1)$, $(h_T^1, h_T^2) = (2, 2)$, $(x^1, x^2) = (1, 1)$, $(\bar{S}^1, \bar{S}^2) = (15, 15)$, $(\lambda^1, \lambda^2) = (1, 1)$, $(\sigma_R^1, \sigma_R^2) = (1, 1)$, $(T^1, T^2) = (2, 2)$.



adopt a stochastic modeling framework that captures the dynamic progression of individual health conditions during care and travel. Optimal system design centers on determining the call-in threshold that minimizes total operational costs, shaping resource allocation across settings and patient types.

Managerially, our results both guide resource allocation and caution that remote hospitalization may not serve distant patients as effectively as intended. The main results (Theorems 1, 2, 3) share a common analytical structure that enables simple managerial "spot checks" based on the difference in marginal costs, γ, and recovery rates. With unlimited resources, γ determines the shape of the optimal call-in threshold. Under capacity constraints, the optimal threshold adjusts as

if h_H and h_R are both shifted by $\Gamma \geq 0$. When $\theta_R < \theta_H$, (19) implies that resource scarcity further narrows the range of patients and distances for which home hospitalization remains viable. This effect extends directly to the multi-type setting, where limited resources lead to similar shrinking in the feasible hybrid care region.

Beyond healthcare, similar structural insights may apply to other public services. For example, online education in rural areas faces analogous trade-offs, further complicated by disparities in internet access (Lai and Widmar 2021). As with remote hospitalization, recent evidence on learning losses during the COVID-19 pandemic highlights potential risks of over-reliance on remote formats (Halloran et al. 2021, Goldhaber et al. 2022).

Modeling-wise, our framework links the stochastic progression of patient health directly to hybrid hospital design. Rather than treating service times as generic random variables, we model recovery explicitly as drifted Brownian motion, enabling analysis of how variability in recovery interacts with patient travel. This distinguishes our approach from most queueing and healthcare operations models, where randomness typically enters only via service durations.

One possible limitation is the assumption that all patients eventually recover, as the negative drift ensures finite hitting times to full recovery. While conservative, this assumption highlights the fragility of hybrid hospital design: even when failure only results in added cost, remote hospitalization remains viable only for limited patient distances and severities, even under optimal operations.

This modeling framework also opens several avenues for future research. One direction is to develop dynamic policies that allow patient swapping when capacity constraints prevent call-ins. Another is to extend the model to hybrid protocols that start hospitalization on-site and transition patients to remote care as their condition improves.

Declarations

Funding. N.Z. received partial financial support from ISF Grant 277/21, the Israel National Institute for Health Policy Research (Grant 2021/160/R), and the Bernard M. Gordon Center for Systems Engineering at the Technion.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Acknowledgements

The authors sincerely thank the Editor-in-Chief, Michel Mandjes, as well as the anonymous Associate Editor and reviewers, for their constructive feedback and valuable suggestions that significantly improved the paper.

References

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

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* .

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.

Berry C (2025) Disparities in access to time-sensitive emergency trauma care. *JAMA surgery* 160(3):321–321.

Bertsekas D (1997) *Nonlinear Programming* (Taylor & Francis).

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.

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.

Clark N, Hernandez A, Bertalan M, Wang V, Greenberg S, Ibrahim A, Stewart B, Scott J (2025) Travel time as an indicator of poor access to care in surgical emergencies. *JAMA Network Open* 8(1):e2455258–e2455258.

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.

Droogh J, Smit M, Absalom A, Ligtenberg J, Zijlstra J (2015) Transferring the critically ill patient: Are we there yet? *Critical care* 19:1–7.

Fanara B, Manzon C, Barbot O, Desmettre T, Capellier G (2010) Recommendations for the intra-hospital transport of critically ill patients. *Critical Care* 14:1–10.

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* .

Gupta V, Harchol-Balter M, Dai JG, Zwart B (2010) On the inapproximability of M/G/K: why two moments of job size distribution are not enough. *Queueing Systems* 64:5–48.

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.

Hashimoto E, Ortega E, Cordeiro G, Cancho V, Silva I (2023) The re-parameterized inverse gaussian regression to model length of stay of Covid-19 patients in the public health care system of Piracicaba, Brazil. *Journal of Applied Statistics* 50(8):1665–1685.

Hearn M, Pinto C, Moss J (2024) Evaluating the connection between rural travel time and health: A cross-sectional analysis of older adults living in the Northeast United States. *Journal of Primary Care & Community Health* 15:21501319241266114.

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.

Jang W, Lee J, Eun S, Yim J, Kim Y, Kwak M (2021) Travel time to emergency care not by geographic time, but by optimal time: a nationwide cross-sectional study for establishing optimal hospital access time to emergency medical care in south korea. *PloS One* 16(5):e0251116.

Jia L, Wang H, Gao Y, Liu H, Yu K (2016) High incidence of adverse events during intra-hospital transport of critically ill patients and new related risk factors: A prospective, multicenter study in china. *Critical Care* 20:1–13.

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.

Leff B, Burton L, Mader S, Naughton B, Burl J, Inouye S, Greenough W, Guido S, Langston C, Frick K (2005) Hospital at home: feasibility and outcomes of a program to provide hospital-level care at home for acutely ill older patients. *Annals of internal medicine* 143(11):798–808.

Levine D, Ouchi K, Blanchfield B, Saenz A, Burke K, Paz M, Diamond K, Pu C, Schnipper J (2020) Hospital-level care at home for acutely ill adults: a randomized controlled trial. *Annals of internal medicine* 172(2):77–85.

Lewis T (2022) People in rural areas die at higher rates than those in urban areas. *Scientific American* .

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.

Mseke E, Jessup B, Barnett T (2024) Impact of distance and/or travel time on healthcare service access in rural and remote areas: A scoping review. *Journal of Transport & Health* 37:101819.

Murata M, Nakagawa N, Kawasaki T, Yasuo S, Yoshida T, Ando K, Okamori S, Okada Y (2022) Adverse events during intrahospital transport of critically ill patients: a systematic review and meta-analysis. *The American journal of emergency medicine* 52:13–19.

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

Parmentier-Decrucq E, Poissy J, Favory R, Nseir S, Onimus T, Guerry MJ, Durocher A, Mathieu D (2013) Adverse events during intrahospital transport of critically ill patients: Incidence and risk factors. *Annals of intensive care* 3:1–10.

Rural Health Information Hub (2025a) Healthcare access in rural communities. URL https://www.ruralhealthinfo.org/topics/healthcare-access, accessed May 2025.

Rural Health Information Hub (2025b) Transportation in rural America. URL https://www.ruralhealthinfo.org/topics/transportation, accessed May 2025.

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).

Stirzaker D (2005) *Stochastic processes and models* (OUP Oxford).

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

Texas A&M University Health Science Center (2023) Travel time for health care has increased for both rural and urban americans. URL https://vitalrecord.tamu.edu/travel-time-for-health-care-has-increased-for-both-rural-and-urban-americans/, accessed May 2025.

The Commonwealth Fund (2020) Hospital at home programs improve outcomes, lower costs but face resistance. URL https://www.commonwealthfund.org/publications/newsletter-article/hospital-home-programs-improve-outcomes-lower-costs-face-resistance, accessed May 2025.

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

US Government Accountability Office (2012) Ambulance providers: Costs and medicare margins. Technical Report GAO-13-6, GAO, URL https://www.gao.gov/assets/gao-13-6.pdf, accessed May 2025.

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.

Whitmore G (1975) The inverse gaussian distribution as a model of hospital stay. *Health Services Research* 10(3):297.

Whitt W (1993) Approximations for the GI/G/m queue. *Production and Operations Management* 2(2):114–161.

Zychlinski N, Fluss R, Goldberg Y, Zubli D, Barkai G, Zimlichman E, Segal G (2024) Tele-medicine controlled hospital-at-home for acutely ill patients is associated with better clinical outcomes when compared to in-hospital stay. *PloS One* 19(8):e0309077.

Appendix A: Justification of Total Long-Run Average Cost

Let $A(t)$ be a renewal process of incoming patients of a single type. The patients' health scores evolve according to the description in Section 3.1. We assume that patients are independent, and thus their health scores evolve according to independent copies of B_R, Z, B_H, indexed by k. Specifically, let

$$\begin{aligned}
&\mathcal{B}^R(k,t) = x + \sigma_R B^R(k,t) - \theta_R t,\\
&\tau_R(k,x,a) = \inf\{t \geq 0 : \mathcal{B}^R(k,t) = 0 \text{ or } \mathcal{B}^R(k,t) = a + x\},\\
&\mathcal{B}^H(k,t) = x + a + Z(k,x,a,T) + \sigma_H B^H(k,t) - \theta_H t,\\
&\tau_H(k,x,a,Z(k,x,a,T)) = \inf\{t \geq 0 : \mathcal{B}^H(k,t) = 0\}.
\end{aligned}$$

For simplicity, we write $\tau_R(k) = \tau_R(k,x,a)$ and $\tau_H(k,Z(k)) = \tau_H(k,x,a,Z(k,x,a,T))$. Thus, the cost of the k-th patient is given by:

$$V_k(a) = h_R \tau_R(k) + 1_{\{\mathcal{B}^R(k,\tau_R(k)) = a+x\}} \big(h_T T + h_H \tau_H(k, Z(k))\big),$$

where $\{V_k(a)\}$ are i.i.d. with

$$\mathbb{E}[V_k(a)] = \mathbb{E}[V(a)] = h_R \mathbb{E}\left[\tau_R(x,a)\right] + \big(h_T T + h_H p_x(a) \mathbb{E}\left[\tau_H(x,a,Z)\right]\big).$$

The average cost at time t is then defined as:

$$V(a,t) := \frac{1}{t} \sum_{k=1}^{A(t)} V_k(a).$$

By the Renewal-Reward Theorem (e.g., Equation (31) on page 82 in Stirzaker 2005), the long run average cost is given by:

$$V(a) = \lim_{t\to\infty} V(a,t) = \lambda \mathbb{E}[V(a)] = \lambda \left(h_R \mathbb{E}\left[\tau_R(x,a)\right] + \left(h_T T + h_H p_x(a) \mathbb{E}\left[\tau_H(x,a,Z)\right]\right)\right),$$

justifying Equation (3) as the system's total long-run average cost.

Appendix B: The Total Workload as a Weighted Sum

Consider the case where the total workload is defined as the following weighted sum:

$$W_S(a) := d W_H(a) + (1-d)\, W_R(a), \quad \text{for some } d \in (0,1).$$

We now discuss how this modification affects the results in the paper. In summary, the analysis remains unchanged, but the parameter d alters the feasibility region and adjusts the relative weight that on-site and remote care receive when interpreting their cost increases as equivalent to the presence of resource constraints.

For a single patient type, this modification does not alter the individual workloads $W_H(a)$ and $W_R(a)$. Consequently, their properties, established in Lemma 3, remain valid. Recalling that

$$\begin{aligned}
W_H(a) &= \frac{\lambda p_x(a)}{\theta_H} \left(x + a + T\theta_T\right),\\
W_R(a) &= \frac{\lambda}{\theta_R} \left((1 - p_x(a))x - p_x(a)a\right),
\end{aligned}$$

the new weights can be effectively absorbed into the constants $1/\theta_H$ and $1/\theta_R$, respectively. Therefore, Proposition 1, which characterizes $W_S(a)$ as a function of a, remains unchanged—except that the cases are now determined by the ratio $\frac{\theta_H}{\theta_R} \cdot \frac{1-d}{d}$ instead of simply $\frac{\theta_H}{\theta_R}$. The same holds for the feasibility region characterized in Proposition 2.

The cost function V is not affected by the incorporation of the parameter d, which plays a role only in the constraint of the optimization problem, not in the cost function itself. Hence, Proposition 3 and Theorem 1, both of which pertain to the unconstrained optimization problem, remain valid.

Theorem 2, which concerns the existence, uniqueness, and general properties of the solution to the constrained optimization problem, also remains valid. Its conclusions rely on the structure of V (which is unchanged) and the behavior of $W_S(a)$ as a function of a (which also remains unchanged).

The equivalence to an unconstrained optimization problem established in Proposition 4 still holds, now expressed as

$$\min_{a\in\mathcal{A}} V\big(h_R + (1-d)\Gamma,\, h_H + d\Gamma,\, a\big).$$

That is, the "price" of limited resources is no longer split evenly between the two modes of care but is distributed according to the weights in a straightforward and intuitive manner.

Finally, all these results extend to the multi–patient type setting as before.

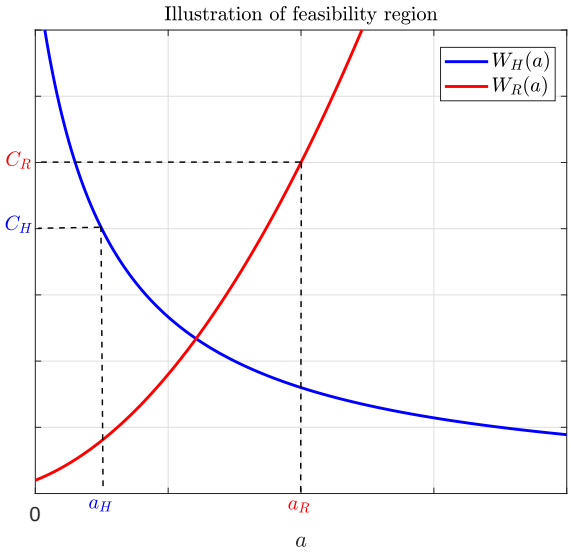
Appendix C: Optimal Call-In Structure with Dedicated Resources

Recall that $W_H(a)$ is decreasing in a and $W_R(a)$ is increasing in a. Let

$$a_H := \inf_{a\geq 0}\{a : W_H(a) \leq C_H\} \quad \text{and} \quad a_R := \sup_{a\geq 0}\{a : W_R(a) \leq C_R\},$$

where we define $a_H = \infty$ and $a_R = -\infty$ in the case of empty sets, which corresponds to cases where at least one of the workloads always exceeds the available capacity. A typical case is illustrated in Figure 9.

Figure 9 **Illustration of feasibility region in the case of dedicated resources.**



$$\mathcal{C}_{FR} = \left\{(\lambda, C_H, C_R) \in \mathbb{R}^3_+ : \exists a \in \mathcal{A} : W_H(a) \le C_H, W_R(a) \le C_R\right\}$$
$$= \begin{cases} \emptyset, & \text{if } a_H > \min\{a_R, \bar{A}\} \\ \left\{(\lambda, C_H, C_R) \in \mathbb{R}^3_+ : a_H \le a \le a_R \wedge \bar{A}\right\}, & \text{otherwise.} \end{cases}$$

Theorem 4 characterizes the solution of the capacitated problem (20).

THEOREM 4. *Assume that $\mathcal{C}_{FR} \neq \emptyset$. Then, problem (20) has a unique solution $a_C^* \in \mathcal{A}$, such that:*

- *If $a_\infty^* \in \mathcal{C}_{FR}$, then $a_C^* = a_\infty^*$.*
- *If $a_\infty^* < a_H$, then $a_C^* = a_H$.*
- *If $a_\infty^* > a_R$, then $a_C^* = a_R \wedge \bar{A}$.*

Similar to the pooled resource case, we now interpret the structure of the dedicated-capacity problem and make its solution explicit by establishing an equivalence between the capacitated and uncapacitated problems. To emphasize the dependence of the cost-of-care function V on the holding costs, we write it as $V(h_R, h_H, a)$.

Recall the uncapacitated minimization problem:

$$\min_{a \in \mathcal{A}} V(h_R, h_H, a), \tag{21}$$

which, by Proposition 3, admits a unique solution $a_\infty^* \in \mathcal{A}$.

Next, consider the capacitated problem:

$$\begin{aligned} \min_{a \in \mathcal{A}} \quad & V(h_R, h_H, a) \\ \text{s.t.} \quad & W_H(a) \le C_H, \\ & W_R(a) \le C_R, \end{aligned} \tag{22}$$

which, under the assumption $\mathcal{C}_{FR} \neq \emptyset$, has a unique solution $a_C^* \in \mathcal{A}$ by Theorem 4.

To bridge these two formulations, define:

$$\Gamma_H = \begin{cases} -\dfrac{V'(h_R, h_H, a_C^*)}{W_H'(a_C^*)}, & \text{if } W_H(a_\infty^*) > C_H, \\ 0, & \text{otherwise,} \end{cases} \tag{23}$$

and

$$\Gamma_R = \begin{cases} -\dfrac{V'(h_R, h_H, a_C^*)}{W_R'(a_C^*)}, & \text{if } W_R(a_\infty^*) > C_R, \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

First, note that at most one of Γ_H or Γ_R are non-zero due to the structure of W_H and W_R. Second, if one of them is non-zero, it must be positive because $W_H' > 0$, $W_R' < 0$, and V' has the opposite sign at a_C^*.

Now, consider the following modified uncapacitated problem, with adjusted holding costs:

$$\min_{a \in \mathcal{A}} V(h_R + \Gamma_R, h_H + \Gamma_H, a). \tag{25}$$

The next proposition establishes the equivalence between the capacitated problem (22) and the modified uncapacitated problem (25). This implies that the structural properties of the uncapacitated problem extend to the capacitated case. In particular, the policy structure characterized by the parameters α, β, γ (as in Proposition 3), and the role of travel distance in shaping the optimal call-in threshold (as in Theorem 1), remain valid.

PROPOSITION 6. *Assume that the feasibility region of (22) contains more than one point. Then, the modified problem (25) admits a unique solution in $\mathcal{A}$, which coincides with the unique solution a_C^* of the original capacitated problem (22).*

Appendix D: Data Requirements and Parameter Estimation

We model the evolution of patients' health scores as negative-drift Brownian motions. However, instead of estimating the drift and variance directly, we can utilize station-level LOS data, which hospitals typically collect consistently and systematically. Specifically, we need data on the duration of remote and on-site hospitalizations for patients.

Under our model, the LOS follows an inverse Gaussian distribution – a distribution that has been widely used to model LOS in healthcare models for many years (see Whitmore 1975 and the more recent Hashimoto et al. 2023). Thus, the required data includes patients' LOS at each hospitalization stage. From this data, we can estimate the parameters of the distribution, which correspond to the expectancy and variance of τ_R and τ_H. This estimation can be accomplished using simple maximum-likelihood estimation.

We then estimate the health scores at call-in/discharge and subsequently build estimates for θ_H and θ_R. For patients who were called in, the necessary data includes their health scores before and after traveling, allowing us to estimate θ_T and $p_x(a)$.

Appendix E: The Effect of Recovery Variability on the Optimal Call-In Policy

For a fixed call-in threshold a, the call-in probability $p_x(a)$ increases with the variance of the drifted BM recovery process, σ, since greater variability raises the chance of crossing the threshold. We now examine how this variability affects the optimal call-in policy.

The left panel of Figure 10 depicts the optimal call-in threshold a as a function of transfer time T for three values of σ. As shown, for a given T, the optimal threshold a increases with σ. Intuitively, while higher variability raises the risk of deterioration, it also increases the chance of spontaneous improvement, allowing the system to tolerate slightly worse patient conditions before calling them into the hospital. Furthermore, as σ increases, the feasible range of transfer times for which home hospitalization remains effective modestly expands.

The right panel presents the corresponding hitting probabilities $p_x(a)$, evaluated at the respective optimal thresholds. As expected, the call-in probabilities remain high, ensuring timely hospital transfers despite the more permissive thresholds.

These results highlight that moderate increases in recovery variability may support somewhat broader applicability of remote hospitalization, both in terms of patient condition and acceptable transfer distances.

Appendix F: Proofs

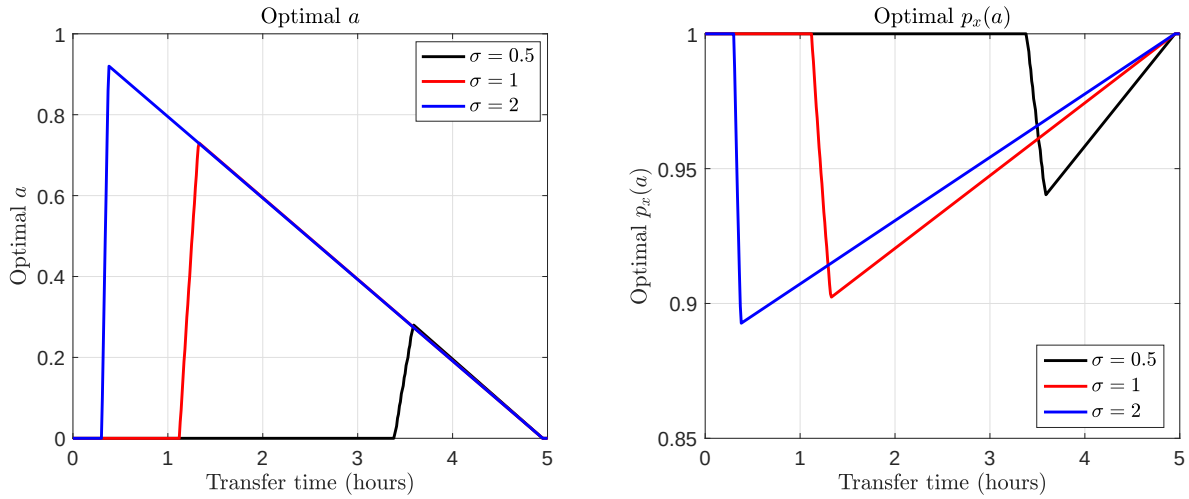
Proof of Lemma 3. Recall that $p_x(a)$ is the hitting probability given by

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

and therefore, its first derivative with respect to a is

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

Figure 10 **Impact of recovery variance on the optimal call-in policy: (left) optimal call-in threshold a as a function of transfer time T for different values of σ; (right) corresponding hitting probabilities $p_x(a)$ at the optimal thresholds. The parameters are as in Figure 4.**



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 $p_x'(a) \neq 0$,

$$W_H'(a) = \frac{\lambda}{\theta_H}\left(p_x'(a)(a + x + \theta_T T) + p_x(a)\right) = \frac{\lambda}{\theta_H} p_x'(a)\left(a + x + \theta_T T + \frac{p_x(a)}{p_x'(a)}\right). \tag{26}$$

Now,

$$\frac{p_x(a)}{p_x'(a)} = -\frac{e^{\rho a} - e^{-\rho x}}{\rho e^{\rho a}} = -\frac{1}{\rho}(1 - e^{-\rho(a+x)}) > -(a + x), \tag{27}$$

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

$$a + x + \theta_T T + \frac{p_x(a)}{p_x'(a)} > \theta_T T \quad \Rightarrow \quad a + x + \frac{p_x(a)}{p_x'(a)} > 0. \tag{28}$$

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:

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

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

$$W_R'(a) = -\frac{\lambda}{\theta_R}(p_x'(a)(a + x) + p_x(a)) = -\frac{\lambda}{\theta_R} p_x'(a)\left(a + x + \frac{p_x(a)}{p_x'(a)}\right) > 0, \tag{29}$$

where the inequality is from (28) and since $p_x'(a) < 0$. Q.E.D.

Proof of Proposition 1. We begin by separately characterizing the respective dependence of the on-site and remote hospitalization workloads on a.

LEMMA 3. *$W_H(a)$ is a strictly decreasing function of a; $W_R(a)$ is a strictly increasing function of a.*

The intuition behind Lemma 3 is that as a increases, patients spend more time at home by design, increasing $W_R(a)$ and reducing $W_H(a)$. We next analyze $W_S(a)$.

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

$$\begin{aligned} W_S'(a) = W_H'(a) + W_R'(a) &\overset{(26),(29)}{=} \frac{\lambda}{\theta_H} p_x'(a)\left(a + x + \theta_T T + \frac{p_x(a)}{p_x'(a)}\right) - \frac{\lambda}{\theta_R} p_x'(a)\left(a + x + \frac{p_x(a)}{p_x'(a)}\right) \\ &\overset{(7)}{=} \frac{\lambda p_x'(a)}{\theta_H}\left((1-r)\left(a + x + \frac{p_x(a)}{p_x'(a)}\right) + \theta_T T\right) \\ &\overset{(27)}{=} \frac{\lambda |p_x'(a)|}{\rho \theta_H}\left((1-r)(1-\rho(a+x) - e^{-\rho(a+x)}) - \theta_T T \rho\right), \end{aligned}$$

where the absolute value is used 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,

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

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

$$\begin{aligned} &h(0) = \rho x - 1 + e^{-\rho x} > 0 \qquad \text{because } e^{-x} > 1 - x \text{ for } x > 0, \\ &h'(a) = \rho - \rho e^{-\rho(a+x)} = \rho(1 - e^{-\rho(a+x)}) > 0, \\ &h(a) \xrightarrow{a \to \infty} \infty. \end{aligned} \tag{30}$$

Thus, $W_S'(a)$ 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

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

where Δ is defined in (7).

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 (30), it is clear that $W_S'(0) < 0$, and that there exists $a_0 > 0$ such that $W_S'(a) < 0$ for $a \in [0, a_0)$, $W_S'(a_0) = 0$, and $W_S'(a) > 0$ for $a \in (a_0, \infty)$. This concludes the proof. Q.E.D.

Proof of Proposition 2. Per section 4.1, the feasibility region of optimization problem (6) is

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

All that is left is to characterize $a_{\min}$. By Proposition 1, if $\theta_H/\theta_R \le 1$, then $W_S(a)$ is strictly decreasing 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 $W_S(a)$ is strictly increasing 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 $W_S'(a) < 0$ for $a \in [0, a_0)$, $W_S'(a_0) = 0$, and $W_S'(a) > 0$ 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 $W_S(a)$ 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 4. *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 4, $\gamma \geq 0$ 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 4 provides that there is a unique optimal solution. Moreover, because the definition of $\tilde{a}$ in Equation (9) is precisely the first order condition in Equation (33) within the proof of Lemma 4, 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 4, 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 (10) 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 (33) 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 (10). Rearranging (33) 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 $\mathsf{W}(ze^z) = z$ 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 (10) 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 (33) 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 4. 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

$$V'(a) = p'_x(a)\left(\beta + \gamma a + \gamma\frac{p_x(a)}{p'_x(a)}\right) \overset{(27)}{=} p'_x(a)(\beta + \gamma a - \gamma\frac{1}{\rho}(1 - e^{-\rho(a+x)}))$$
$$= \frac{|p'_x(a)|}{\rho}\left(\gamma(1 - e^{-\rho(x+a)}) - \beta\rho - \gamma\rho a\right). \tag{31}$$

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

$$\left(\gamma(1 - e^{-\rho(x+a)}) - \rho\beta - \gamma\rho a\right)' = -\gamma\rho\left(1 - e^{-\rho(x+a)}\right), \tag{32}$$

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

$$e^{-\rho a} = (1 - \rho\beta/\gamma - \rho a)\, e^{\rho x}. \tag{33}$$

Now, let us notice that, as functions of a, the left-hand side of Equation (33) 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 (33) at $a = 0$ is strictly greater than the right-hand side of (33) 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 (31) and recalling that $\gamma > 0$ in this case, Equation (32) then shows that $V'(a) < 0$ for all $a \in \mathbb{R}_+$. For $\gamma = 0$, we can quickly recognize from Equation (31) that, again, $V'(a) < 0$ for all a.

Let us now suppose that $\gamma < 0$. Through Equation (32), 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 (33) starts above the exponential in the left-hand side of (33), 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 (32), 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 (10). 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 (10) 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 (10) itself simplifies to

$$x + \tilde{a} = \frac{1}{\rho}\left(1 + \mathsf{W}\left(-e^{\frac{\eta\rho}{\gamma} T - 1}\right)\right) - \frac{\eta}{\gamma} T. \tag{34}$$

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 (34), 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 (12). 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. $a_{\min} := \arg\min_{a\in\mathcal{A}} W_S(a)$
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 $W_S(a)$ behaves as a function of a.
   (a) $W_S(a)$ is strictly increasing, then $a_{\min} = 0$. Importantly and in particular, $W_S(a)$ is strictly increasing to the right of $a_{\min}$.
   (b) $W_S(a)$ is strictly decreasing, then $a_{\min} = \bar{A}$. Importantly and in particular, $W_S(a)$ is strictly decreasing to the left of $a_{\min}$. Meaning, as we decrease a, starting from $a_{\min}$, the value of $W_S(a)$ increases.
   (c) $W_S(a)$ has a unique minimum in $(0, \bar{A})$, it strictly decreases before it and strictly increases after.
5. The conclusion from the item above is that if we pick any $\hat{a} \in \mathcal{A}$ which satisfies $\hat{a} \neq a_{\min}$ (but both $\hat{a} > a_{\min}$ and $\hat{a} < a_{\min}$ are possible), then if we move from $\hat{a}$ to $a_{\min}$, the value of $W_S(a)$ **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 $\hat{a} \in \mathcal{A}$, $V(a)$ **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 $W_S(a_{\min}) = C$, 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 $W_S(a_{\min}) < C$. If $W_S(a^*_\infty) \leq C$, 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 $W_S(a_{\min}) < C$ and $W_S(a^*_\infty) > C$. In particular, we must have $a_{\min} \neq a^*_\infty$. By the properties listed above, when we start at a^*_∞ and go towards $a_{\min}$, $W_S(a)$ must strictly decrease and $V(a)$ must strictly increase. Since $W_S(a)$ is continuous, there must be a value for a, call it $\hat{a}$, strictly between a^*_∞ and $a_{\min}$ for which $W_S(\hat{a}) = C$, which also means $\hat{a}$ is feasible. Additionally, any other value for a before we reach $\hat{a}$ must have $W_S(a) > C$ and hence is not feasible. Any value of a after $\hat{a}$ 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 $W_S(a) = C$. Q.E.D.

Proof of Proposition 4. Throughout this proof we assume that $W_S(a_{\min}) < C$. First, if $W_S(a^*_\infty) \leq C$, then $\Gamma = 0$, and problems (15) and (18) are identical and their solution is a^*_∞. Theorem 2 assures us that in this case, the solution to (16) satisfies that $a^*_C = a^*_\infty$, which proves the desired result.

We turn to the case where $W_S(a^*_\infty) > C$. 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 $W_S(a) = C$. In particular, a^*_C must be an internal point in $\mathcal{A}$ and $W'_S(a^*_C) \neq 0$.

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

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

and, therefore,

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

Namely,

$$V(h_R + \Gamma, h_H + \Gamma, a) = V(h_R, h_H, a) + \Gamma W_S(a). \tag{35}$$

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

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

Clearly, this derivative equals zero for $a = a^*_C$. By (35), 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 (16), which concludes the proof. Q.E.D.

Proof of Proposition 5. This proof follows very similarly to that of Proposition 2. In the discussion of section 5.1, we establish that the feasibility region of optimization problem (5) is

$$\mathcal{C}^K_{FR} = \left\{ (\vec{\lambda}, C) \in \mathbb{R}^{K+1}_+ : \sum_{k=1}^{K} W^k_S(a^k_{\min}) \leq C \right\}.$$

All that is left is to characterize $a^k_{\min}$ for $k = 1, \ldots, K$. By Proposition 1, if $\theta^k_H / \theta^k_R \leq 1$, then $W^k_S(a^k)$ is strictly decreasing in a^k. Thus, in this case, $a^k_{\min} = 0$, which proves the first item. Next, by Proposition 1, if $\theta^k_H / \theta^k_R \geq 1 + \Delta^k$, then $W^k_S(a^k)$ is strictly increasing in a^k. Thus, in this case, $a^k_{\min} = \bar{S}^k - x^k - T^k \theta^k_T$, which proves the second item.

Finally, in the case where $1 < \theta^k_H / \theta^k_R < 1 + \Delta^k$, by Proposition 1, there exists $a^k_0 > 0$ such that $W^{k\prime}_S(a^k) < 0$ for $a^k \in [0, a^k_0)$, $W^{k\prime}_S(a^k_0) = 0$, and $W^{k\prime}_S(a^k) > 0$ for $a^k \in (a^k_0, \infty)$. If $a^k_0 < \bar{S}^k - x^k - T^k \theta^k_T$, then $a^k_{\min} = a^k_0$. Otherwise, $\bar{S}^k - x^k - T^k \theta^k_T \in (0, a^k_0]$, and since $W^k_S(a^k)$ is strictly decreasing in this interval, we have $a^k_{\min} = \bar{S}^k - x^k - T^k \theta^k_T$ which completes the proof of the third item and the proposition. Q.E.D.

Proof of Lemma 1. If the feasibility region contains exactly one vector, then this vector is optimal. Otherwise, since the total workloads are continuous functions, the feasibility region contains an infinite number of vectors. We begin by proving that in this case, the feasibility region is a compact set, by proving it is bounded and closed. The set $\mathcal{C}_{FR}^K$ is bounded because

$$\|\vec{a}-\vec{b}\|_2 \leq \|\vec{a}\|_2 + \|\vec{b}\|_2 \leq 2K \max_k \left\{\bar{A}^k\right\}, \quad \forall a,b \in \mathcal{C}_{FR}^K.$$

Next, let $\vec{c}$ be a limit point of $\mathcal{C}_{FR}^K$. Assume by contradiction that $\vec{c} \notin \mathcal{C}_{FR}^K$. First, consider the case where there exists an entry k such that $[\vec{c}]_k \notin [0, \bar{A}^k]$, meaning, it is outside of the hypercube $[0,\bar{A}^1] \times \ldots \times [0,\bar{A}^K]$. Clearly, there is a small enough neighbourhood of $\vec{c}$ with no points that belong to $\mathcal{C}_{FR}^K$, which contradicts the fact that $\vec{c}$ is a limit point. Therefore, $\vec{c}$ must be in the hypercube.

We are left with the case where $\sum_{k=1}^K W_T^k(c^k) > C$, meaning that there exists $\delta > 0$ such that $\sum_{k=1}^K W_T^k(c^k) = C + \delta$. Since $\{W_T^k\}$ are continuous functions, so is $\sum_{k=1}^K W_T^k(c^k)$ as a function from $\mathbb{R}_+^K$ to $\mathbb{R}_+$. Hence, there exists a small enough neighborhood of $\vec{c}$ such that for every point $\vec{a}$ in it we have $\sum_{k=1}^K W_T^k(a^k) > C + \delta/2$. Thus, no points in this neighborhood belong to $\mathcal{C}_{FR}^K$ which, again, is a contradiction. Hence, $\mathcal{C}_{FR}^K$ is closed.

Since $\mathcal{C}_{FR}^K$ is compact and the cost function is continuous, by the Extreme Value Theorem the infimum is attained and there exists an optimal solution. Q.E.D.

Proof of Lemma 2. By Proposition 3, $[\vec{a}_\infty^*]_k$ is the unique minimizer of $V_k([\vec{a}]_k)$. Since $V(\vec{a}) = \sum_{k=1}^K V_k([\vec{a}]_k)$, $\vec{a}_\infty^*$ is the unique minimizer of $V(a)$. Q.E.D.

Proof of Theorem 3. For each type, we know from Propositions 1 and 3 and their proofs that V_k and W_T^k each can exhibit one of the following behaviours in the allowable threshold interval: (1) strictly decreasing, (2) strictly increasing or (3) strictly decreasing, attains a minimum and strictly increasing afterwards. We use the symbols $\searrow$, $\nearrow$, and $\searrow\!\!\nearrow$ to refer to these cases respectively. Thus, there are nine possible combinations of how V_k and W_T^k behave for a specific type. For example, there is a possibility that V_k is strictly increasing while W_T^k is strictly decreasing. We will use the notation $V_k \searrow W_T^k \nearrow$ for this case. We now prove that for each combination, it is impossible or necessarily sub-optimal to choose the threshold not between $a_{\min}^k$ and $a_{\infty,k}^*$.

1. $V_k \nearrow W_T^k \nearrow$: In this case $a_{\min}^k = a_{\infty,k}^* = 0$. Any other choice of threshold would result in a higher cost and a higher workload and therefore is sub-optimal.
2. $V_k \nearrow W_T^k \searrow$: In this case $a_{\min}^k = \bar{A}^k$ and $a_{\infty,k}^* = 0$ and therefore any choice of threshold has to be between the two.
3. $V_k \nearrow W_T^k \searrow\!\!\nearrow$: In this case $a_{\infty,k}^* = 0$ and $a_{\min}^k$ is an interior point. A choice of a threshold to the right of $a_{\min}^k$ results in a higher cost and a higher workload than choosing, for example, $a_{\min}^k$, and therefore is sub-optimal.
4. $V_k \searrow W_T^k \nearrow$: Similar to case 2.
5. $V_k \searrow W_T^k \searrow$: Similar to case 1.

6. $V_k \searrow W_T^k \searrow\nearrow$: Similar to case 3.
7. $V_k \searrow\nearrow W_T^k \nearrow$: In this case $a_{\min}^k = 0$ and $a_{\infty,k}^*$ is an interior point. Choosing a threshold to the right of $a_{\infty,k}^*$ results in a higher cost and a higher workload than choosing, for example, $a_{\infty,k}^*$, and therefore is sub-optimal.
8. $V_k \searrow\nearrow W_T^k \searrow$: Similar to case 7.
9. $V_k \searrow\nearrow W_T^k \searrow\nearrow$: In this case both $a_{\infty,k}^*$ and $a_{\min}^k$ are interior points. If they are equal we are done. If they are not, assume that $a_{\infty,k}^* < a_{\min}^k$. Choosing a threshold to the right of $a_{\min}^k$ results in a higher cost and a higher workload than choosing, for example, $a_{\min}^k$, since both V_k and W_S^k are strictly increasing to the right of $a_{\min}^k$, making this a sub-optimal choice. The same holds for choosing a threshold to the left of $a_{\infty,k}$. The case where $a_{\infty,k}^* > a_{\min}^k$ is similar.

This concludes the proof of the first item in Theorem 3. For the second item, assume by way of contradiction that the capacity constraint at $\vec{a}_C^*$ is inactive, i.e., $\sum_k W_T^k([\vec{a}_C^*]_k) < C$. Since we are under the assumption that $\vec{a}_\infty^* \notin \mathcal{C}_{FR}^K$, by the first item in Theorem 3, there must be at least one type $\bar{k}$ for which $[\vec{a}_C^*]_{\bar{k}}$ is between $a_{\infty,\bar{k}}^*$ and $a_{\min}^{\bar{k}}$ but not equal to $a_{\infty,\bar{k}}^*$. In all of the nine combinations we considered for the behavior of V_k and W_T^k, and as we show in the proof of Theorem 2, when we move from $a_{\min}^{\bar{k}}$ towards $a_{\infty,\bar{k}}^*$ the cost strictly decreases and the total workload strictly increases. Thus, given that $\sum_k W_T^k([\vec{a}_C^*]_k) < C$, we have some slack, and there exists a threshold between $[\vec{a}_C^*]_{\bar{k}}$ and $a_{\infty,\bar{k}}^*$ such that if we choose it instead of $[\vec{a}_C^*]_{\bar{k}}$ and leave all other entries of $\vec{a}_C^*$ the same we get a feasible solution with a strictly smaller cost than with the choice of $\vec{a}_C^*$. This contradicts the fact that $\vec{a}_C^*$ is a globally optimal solution. This concludes the proof of the second item in Theorem 3.

For the third item, recall that E denotes the set of indices for which the corresponding entries of the optimal solution $\vec{a}_C^*$ are not equal to $a_{\min}^k$ or $a_{\infty,k}^*$. Denote by $\vec{a}_{C,E}^*$ the solution $\vec{a}_C^*$ restricted to the entries in E. Consider the following optimization problem:

$$\begin{aligned} &\min_{\vec{a}\in\vec{\mathcal{A}}_E} V(\vec{a}) \\ &\text{s.t.} \sum_{k\in E} W_T^k(a^k) = C - \sum_{k\in E^c} W_T^k([\vec{a}_C^*]_k), \end{aligned} \tag{36}$$

namely, we fix the entries in E^c, and optimize over the rest.

LEMMA 5. *$\vec{a}_{C,E}^*$ is an optimal solution for the optimization problem (36).*

Proof of Lemma 5. Assume by way of contradiction that there exist $\vec{b} \in \vec{\mathcal{A}}_E$ for which $V(\vec{b}) < V(\vec{a}_{C,E}^*)$. Then if we take the elements of $\vec{a}_C^*$ in the indices that belong to E^c with the corresponding elements of b we get a feasible vector for the original optimization problem but with a lower cost than that of $\vec{a}_C^*$. This contradicts the fact that $\vec{a}_C^*$ is an optimal solution. Q.E.D.

Continuing with the proof of the third item in Theorem 3, we now prove that $\vec{a}_{C,E}^*$ is regular. By the first item in Theorem 3 and the definition of E, the entries of $\vec{a}_{C,E}^*$ must be strictly between the corresponding $a_{\min}^k$ and $a_{\infty,k}^*$ and therefore must be interior points. Thus, all of the boundary constraints aside from the capacity constraint are inactive. $\vec{a}_{C,E}^*$ is a feasible solution, so all that is left is to verify is that at least one

of the derivatives $W_S^{k\prime}([\vec{a}^*_{C,E^c}]_k)$ is not zero. But, by Proposition 1, these derivatives can only be zero once, at the corresponding $a^k_{\min}$. By the definition of E, $[\vec{a}^*_{C,E}]_k \neq a^k_{\min}$, for all $k \in E$. Thus, we conclude that $\vec{a}^*_{C,E}$ is regular.

Now, $\vec{a}^*_{C,E}$ is an optimal solution for the optimization problem (36), and it is regular. In addition, all of the inequality constraints are inactive. Thus, by the KKT sufficient conditions for optimality (e.g., Proposition 3.3.1. in Bertsekas 1997), we obtain that there exists a unique Lagrange multiplier $\Gamma \geq 0$ such that $V'_k([\vec{a}^*_C]_k) + \Gamma W_T^{k\prime}([\vec{a}^*_C]_k) = 0$, for all $k \in E$. Moreover, by the definition of E and the properties of $\{V_k\}$ and $\{W_S^k\}$, none of these derivatives are zero. Thus, Γ must be strictly positive.

Lastly, we wish to prove that $\vec{a}^*_{C,E}$ is the unique optimal solution of the un-capacitated optimization problem $\min_{\vec{a}\in\vec{\mathcal{A}}_E} \sum_{k\in E} V_k(h_R+\Gamma, h_H+\Gamma, a_k)$, where $\Gamma > 0$ is the previously considered unique Lagrange multiplier for which $\Gamma = -V'_k([\vec{a}^*_C]_k)/W_T^{k\prime}([\vec{a}^*_C]_k)$ for all $k \in E$. The proof now follows the exact same steps as that of Proposition 4. Namely, for each $k \in E$, $[\vec{a}^*_C]_k$ must be the unique minimum of $V_k(h_R+\Gamma, h_H+\Gamma, a_k)$ in the allowed interval. Thus, $\vec{a}^*_{C,E}$ is the unique minimizer of their sum. Q.E.D.

Proof of Theorem 4. We begin by recalling the relevant notation and previously established results:

1. $a^*_\infty := \arg\min_{a\in\mathcal{A}} V(a)$.
2. $a_H := \inf\{a \geq 0 : W_H(a) \leq C_H\}, \quad a_R := \sup\{a \geq 0 : W_R(a) \leq C_R\}$.
3. The values a^*_∞, a_H, and a_R are all unique.
4. The function $W_H(a)$ is strictly decreasing, while $W_R(a)$ is strictly increasing.
5. From the analysis in the proof of Proposition 3, the function $V(a)$ can exhibit one of the following three behaviors:
   (a) strictly increasing;
   (b) strictly decreasing;
   (c) unimodal: decreasing up to a unique minimum in $\mathbb{R}_+$, then strictly increasing.
6. From the last item, it follows that any deviation from a^*_∞ strictly increases the value of $V(a)$. Moreover, $V'(a)$ has at most one zero, and if it exists, it corresponds to the unique minimizer.

We now distinguish between two cases:

Case 1: $a^*_\infty \in \mathcal{C}_{FR}$. In this case, a^*_∞ is feasible and, by definition, minimizes $V(a)$ over the entire domain $\mathcal{A}$. Therefore, it is also the unique minimizer over the feasible region, i.e., $a^*_C = a^*_\infty$.

Case 2: $a^*_\infty \notin \mathcal{C}_{FR}$, i.e., either $a^*_\infty < a_H$ or $a^*_\infty > a_R$. By the properties above, moving away from a^*_∞ toward the feasibility region causes $V(a)$ to strictly increase.

- If $a^*_\infty < a_H$, then the first feasible point encountered as a increases is a_H, and since $V(a)$ is strictly increasing in this direction, the optimal feasible solution is $a^*_C = a_H$.
- If $a^*_\infty > a_R$, then the first feasible point when decreasing a toward the feasibility region is $\min(\bar{A}, a_R)$, which is thus the optimal solution. Hence, $a^*_C = a_R \wedge \bar{A}$, as stated. Q.E.D.

Proof of Proposition 6. We begin by considering the case where $W_H(a^*_\infty) \le C_H$ and $W_R(a^*_\infty) \le C_R$. In this case, it follows from definitions (23) and (24) that $\Gamma_H = 0$ and $\Gamma_R = 0$, respectively. Therefore, problems (21) and (25) are identical, and their unique solution is a^*_∞. Moreover, Theorem 4 guarantees that in this case $a^*_C = a^*_\infty$, which proves the desired result.

Next, we consider the case where $W_H(a^*_\infty) > C_H$ and $W_R(a^*_\infty) \le C_R$, implying that $\Gamma_H > 0$ and $\Gamma_R = 0$. Theorem 4 ensures that a^*_C is the unique solution to the constrained problem (22).

To prove the proposition, we now leverage a structural property of the objective function $V(a)$. Recall that:

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

and hence,

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

That is,

$$V(h_R, h_H + \Gamma_H, a) = V(h_R, h_H, a) + \Gamma_H W_H(a). \tag{37}$$

Differentiating the right-hand side of (37) with respect to a, and using the definition of Γ_H, we obtain:

$$(V(h_R, h_H, a) + \Gamma_H W_H(a))' = V'(h_R, h_H, a) - \frac{V'(h_R, h_H, a^*_C)}{W'_H(a^*_C)} W'_H(a).$$

By construction, this expression equals zero when $a = a^*_C$. Therefore, by (37), we conclude that the derivative of $V(h_R, h_H + \Gamma_H, a)$ is also zero at $a = a^*_C$.

From the analysis in the proof of Proposition 3, we know that for any fixed $h_R, h_H > 0$, the function $V'(h_R, h_H, a)$ has at most one zero in $\mathbb{R}_+$, and if such a zero exists in $(0, \bar{A})$, it corresponds to a unique global minimizer. Hence, a^*_C is the unique solution to the modified uncapacitated problem (25), concluding the proof for this case.

The proof for the only remaining case where $W_H(a^*_\infty) \le C_H$ and $W_R(a^*_\infty) > C_R$, implying that $\Gamma_H = 0$ and $\Gamma_R > 0$ follows the same line of arguments and is hence omitted. Q.E.D.