informs
https://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 00, No. 0, Xxxx 0000, pp. 000–000
ISSN 1523-4614, EISSN 1526-5498

Submitted to *Manufacturing & Service Operations Management*

Hospital versus Home Care: Trading off Pre- and Post-Discharge Infection and Mortality Risks

Mor Armony
NYU—Stern, marmony@stern.nyu.edu

Galit B. Yom-Tov
Technion—Israel Institute of Technology, gality@technion.ac.il

Abstract. *Problem definition:* Determining the optimal length of stay (LOS) and post-treatment location is critical for hematology-oncology (blood cancer) patients, who are highly vulnerable to life-threatening infections. Early discharge to home care reduces infection risk, while extended hospital observation minimizes mortality risks if an infection occurs. We address this trade-off by developing LOS optimization models tailored to these patients.

Methodology/results: We develop a Newsvendor-type model to explore how infection and mortality risks influence optimal LOS of individual patients. We further consider the social optimization problem in which capacity constraints limit the ability of hospitals to keep patients for the entirety of their optimal LOS. We find that, in the optimal solution to the fluid model used to approximate the original stochastic system, each type of patient is discharged at at most two discrete time points, one of which might be equal to zero or to the optimal uncapacitated length of stay. Based on this analysis, we propose an online index-based speedup policy (ISP) to guide patient discharge decisions.

Managerial implications: Our model enables physicians to personalize LOS based on patients' risk profiles and dynamically adapt to hospital capacity constraints. In a case study, we show that around 75% of the patients need *some* observation, and a speedup-only policy, that discharges all patients at the same discrete time point, is optimal for 90% of patient types during high demand. Adopting ISP can reduce patient mortality rate by 27.7% compared to current practice.

Key words: Healthcare Operations, Hospitalization, Home Care, Mortality, Infection Risk, Discharge Policy

1. Introduction

Cancer is a leading cause of death in the US, and imposes significant health and financial burdens. In 2014, cancer-related healthcare costs in the US totaled $87.8 billion, with 27% attributed to hospital inpatient stays (AHRQ 2014). Cancer inpatients typically face higher hospital costs and longer lengths of stay (LOS)

compared to non-cancer inpatients (Suda et al. 2006). Beyond treating the disease itself, hospitals manage complications such as healthcare-associated infections (HAI), which are more common among cancer inpatients and significantly increase both LOS and mortality (Cornejo-Juárez et al. 2016).

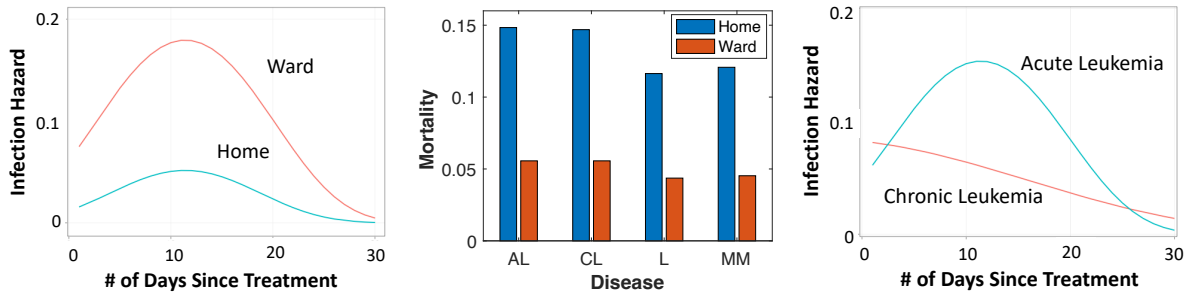
Hematology-oncology malignancies, including Acute Leukemia (AL), Chronic Leukemia (CL), Lymphoma (L), and Multiple Myeloma (MM)), are particularly challenging due to treatments that significantly weaken the immune system. In a large retrospective study of over 41,000 cancer patients admitted for suspected infection, mortality rates among those who were treated for leukemia, lymphoma, and myeloma were as high as 14.3%, 8.9%, and 8.2%, respectively (Kuderer et al. 2006). These mortality risks following infection are so high that it may be best to keep patients at the hospital for the sake of monitoring and rapid intervention in case of infection (Carmen et al. 2019). Determining a patient's "optimal" LOS thus hinges on balancing infection and mortality risks (among other factors), making it important to understand the factors influencing the timing of discharge, especially for cancer patients.

Modeling papers in Operations Management typically assume that, everything else being equal, it is better to keep patients hospitalized for as long as medically indicated. In particular, in the absence of cost considerations or capacity constraints, there should be no rush to send a patient home. However, this perspective often overlooks the significant risk of HAI that has been observed empirically (e.g. Hauck and Zhao (2011), Bichescu and Hilafu (2023)). According to Magill et al. (2018), in 2015, 3.2% of hospitalized patients developed HAI. These risks are much higher for Hematology patients. For instance, Carmen et al. (2019) report that 41.6% of chemotherapy cycles for hematology patients ended with infection during hospitalization. This highlights an additional critical benefit of early discharge beyond cost or capacity savings: reducing infection risk.

Deciding when to send a patient home aligns with the growing trend of shifting treatments from hospitals to outpatient clinics or home care (Clarke et al. 2021, Americal Hospital Association 2023); both settings can support physician and nurse visits as well as diagnostic tests (Song et al. 2022). These solutions have been shown to be safe for some hematology-oncology patients (van Tiel et al. 2005). However, their implementation requires balancing the risks and benefits of hospital versus out-of-hospital alternatives for each patient. Key factors include the patient's health status and home environment, such as proximity to the hospital, availability of an isolated space, and access to personal and medical support. This work develops a systematic approach that incorporates these considerations, to determine the optimal LOS, shedding light on how in-hospital and home-care observations should be combined.

Our work builds on an empirical study by Carmen et al. (2019) conducted in a hematology ward (HW) of an Israeli hospital, which shows that the risk of infection is higher in the hospital than at home, and both risks follow similar time-dependent trends (see Figure 1(a)). Conversely, patients who develop infections in the hospital have higher survival rates due to immediate access to care (see Figure 1(b)). This tradeoff highlights that the decision of when to send a patient home should be based on location-dependent infection

and mortality risks. Further complicating this decision, the time-to-infection hazard rate varies by patient-specific factors, such as their underlying disease (see Figure 1(c)). Given this trade-off, the LOS decision is relevant even in the absence of cost or capacity constraints. Thus, we examine both the uncapacitated and the capacitated cases.



(a) Infection hazard

(b) Mortality risk given infection for four disease types

(c) Infection hazard as a function of disease type for in-ward patients

Figure 1 Infection Hazard-rate Functions and Mortality Risk (Carmen et al. 2019) (AL, CL, L, and MM)

In the uncapacitated case (§3), we focus on individual patients, assuming minimal interaction between them. We formulate the decision of when to transition a patient to home care as a Newsvendor-type problem, aiming to maximize the total expected survival reward minus observation costs. While the optimal time to move a patient to home care is monotone in most system parameters, we find the surprising result that it is *not* monotone in the at-home infection hazard rate.

In the capacitated case (§4), a key question is determining the capacity needed to meet system demand. This is related to the system's offered load—the expected number of patients under unlimited capacity. However, due to high costs stemming from the need for isolated rooms and specialized medical staff, as well as increased demand driven by advancements in treatment and patient survival (LLSC 2016), capacity typically falls short and hospitals cannot handle all the offered load. For example, data from the Technion SEELAB show that the hematology unit at Rambam Medical Center, a large tertiary hospital in Israel operates at an average of 97% occupancy. This highlights the need to consider capacity constraints and establish rules for determining which patients should be moved to home care and optimize the transfer timing.

Traditionally, studies on offered load in queueing systems treat service time as exogenous (Whitt 2013). For example, in a stationary multiserver G/G/N queue, the offered load is the product of the arrival rate and the expected service time. What distinguishes our setting is that the service time here is *a result of an optimization*—specifically, the optimal hospital LOS is an output of the Newsvendor-type problem formulation in the uncapacitated case.

Optimizing the timing of patients discharge in an overloaded system raises some important questions. Ideally, the optimal thing to do would be to send every patient home at the time prescribed by the uncapacitated case analysis. But, with limited capacity, we must consider whether it is better to use an equitable policy that sends *all* patients home earlier than optimal, or to send some patients home at the optimal time while sending others home immediately. Alternatively, a more complex policy may be needed, with multiple thresholds applied to different patient groups.

While the space of possible policies is large, we find that the capacitated problem reduces to dividing each patient group into at most *two* classes, each with its own threshold for discharge time. In the heterogeneous patient case, we prove that most patient types remain as single classes, with at most one patient type divided into two. The conditions that characterize these thresholds reveal an intuitive index that guides the LOS decision. This is formalized through the development of the practical Index-based Speedup Policy (ISP) for online discharge decisions (§4.3).

In a case study (§5) based on real patient data, we demonstrate how the optimal discharge threshold varies by patient profile. This analysis suggests that 75% of the patient population needs *some* hospital observation period (§5.1). We compare the resulting bed occupancy with actual hospital data, suggesting a capacity shortage that might limit full implementation of the patient-optimal policy. We then examine how the discharge policies should adjust under capacity constraints (§5.2). We apply both homogeneous (§5.2.1) and heterogeneous (§5.2.2) solutions developed in Sections 4.2 and 4.3, respectively, and show that while some patient groups require a two-threshold policies, a single-threshold speedup is most commonly optimal.

The ISP algorithm developed in Section 4.3 estimates that the studied hospital can reduce patient mortality rate by 27.7%. This reduction is split between capacity-based mortality (4.2%) and optimization-based mortality (23.5%), suggesting that optimizing LOS accounts for $3/4$ of the effect. This case study provides evidence that even under capacity constraints, hospitals can significantly reduce mortality by choosing wisely which patients should be observed at the hospital and which ones should be moved to home-care observation and when. Finally, we analyze the sensitivity of our results to location-based differences in observation quality, measured by the gap in survival probabilities between hospital and home care (§D.2).

The paper is organized as follows: Section 2 provides background on the motivating medical context and surveys the relevant literature. In Section 3, we formulate and solve the single-patient uncapacitated problem, exploring the structural properties of the solution. Section 4 formulates the hospital capacitated problem, solved using fluid approximation, and introduces the ISP algorithm. Section 5 details the results of our numerical experiments. Finally, Section 6 summarizes our results and suggests directions for further research.

2. Medical Context and Literature Review

Our paper relates to three key research streams: optimization of medical decisions, optimization of patient flow, and asymptotic approximations of queueing systems.

Our research is relevant to all cancer treatments and was specifically motivated by HW data (Carmen et al. 2019). According to the World Health Organization, cancer is a leading cause of death globally, responsible for an estimated 10 million deaths in 2020, or 1 in 6 deaths worldwide (https://www.who.int/news-room/fact-sheets/detail/cancer). Among cancer types, patients with hematological malignancies are particularly susceptible to infections due to a weakened immune system caused by the disease or its treatment, leading to considerable infection-related mortality (Cornely et al. 2015, Taccone et al. 2009). Despite significant advances in treatment strategies for hematological cancers over the past decade, infection prevention and effective treatment for infected patients remain critical challenges.

Our paper is inspired by growing empirical evidence on the impact of physical location of treatment or observation on health outcomes. Specifically, Carmen et al. (2019) demonstrated that the choice of post-treatment observation location (dedicated ward, general ward, or home) impacts infection and mortality rates among hematology patients. More broadly, patient off-placement found to impact LOS, mortality, and readmission (Song et al. 2020). Overall, these studies highlight the critical role of location choice in determining treatment outcomes.

Patients' LOS has also been linked to health outcomes (e.g., Hauck and Zhao 2011). Reducing LOS (speedup) in response to high load has been shown to increase mortality (Kc and Terwiesch 2009, Bartel et al. 2020) and readmission rates (Kc and Terwiesch 2012, Bichescu and Hilafu 2023). Conversely, the longer patient's LOS, the higher the risk of hospital-acquired conditions, such as pressure ulcers and infections (Berry Jaeker and Tucker 2017). These findings underline the importance of incorporating the influence of LOS on health outcomes in models designed to optimize it. For example, Chan et al. (2012) formulated a model to support ICU discharge decisions, while Shi et al. (2021) proposed a hospital-wide optimization method for patient-discharge decisions.

Research also shows that post-discharge followup can effectively reduce mortality risk (Leschke et al. 2012). Recent developments show that, for some patient types, telemedicine followup can be as effective as in-person followups (Marquez-Algaba et al. 2022). These studies support the notion that home care and home observation can be safe when adequate follow-up resources are provided, compensating for shorter LOS.

Most LOS optimization studies assume that the hospital is the best location for patients until recovery, with early discharges driven mostly by capacity or cost constraints. However, in hematology, there is an additional delicate tradeoff between location, LOS, and health due to the high infection risk and its reduction

in home care. Therefore, in this context, discharge decisions should not be driven only by capacity and cost considerations.

Cancer chemotherapy is typically administered in treatment cycles (averaging 8.68 cycles per patient, see Carmen 2017). Our study focuses on a single cycle[1]. Each cycle starts with chemotherapy treatment (of 1–10 days) followed by a 30-day recovery stage to allow the immune system to recuperate. Treatments lasting over 7 hours are conducted in the HW, while shorter protocols are given in an outpatient clinic. The recovery stage can be done either at the HW protective isolation or through home care. The highest risk during this stage is developing infection. During recovery, patients are monitored for infection and immune recovery via temperature and blood tests. Therefore, home-care recovery requires rest and tests to be done either at home or at the outpatient clinic. If signs of infection (e.g., fever) arise, patients are instructed to immediately visit the hospital ED for tests and treatment. Infection treatment necessitates hospitalization, usually at a general ward. The flexible component of patient LOS, during the treatment cycle, is the post-procedure, pre-infection recovery stage, which serves as the focus of our optimization model.

Over-congestion is a common challenge faced by hospitals worldwide. Various strategies have been used in practice or proposed in the literature to manage over-congestion, including early discharge (speedup), admission control (blocking), and prioritization. For example, in ED services, blocking can be done by ambulance diversion (Allon et al. 2013). In our context, sending a patient home immediately after treatment can be viewed as a form of blocking. The tradeoff between admission control and speedup has been explicitly studied in the literature. This tradeoff has been analyzed in Markovian settings, including single-server queueing systems (Adusumilli and Hasenbein 2010, Ata and Shneorson 2006), multi-server queueing systems (Lee and Kulkarni 2014, Yom-Tov and Chan 2021), and multi-class queueing systems (Ulukus et al. 2011). In our capacitated model (§4), we extend these approaches by generalizing the underlying assumptions, considering general service time distributions and general risk functions in a multi-server multi-class setting.

Dynamically changing patient severity has been considered in various papers. For example, Mills et al. (2013) addressed the prioritization of patient evacuation in mass-casualty events, and Deo et al. (2013) focused on allocating doctor appointments for chronic patients whose medical state may depend on the allocation decisions. Ouyang et al. (2020) considered the tradeoff between speedup and blocking in an off-placement scenario, where patients may be moved between the ICU and general wards within the hospital. Their model considered a single patient class, with health status modeled by a two-state Markov chain. They formulated an MDP to minimize patient mortality and developed heuristics for more elaborate health-status Markov chains. Some of their heuristics are closely related to the ISP algorithm we propose in Section 4, as

[1] In addition to chemotherapy, HW also provides other treatments such as bone marrow transplants and radiotherapy, but this study focuses on chemotherapy patients.

well as the myopic policy benchmark proposed in Section 5.2.2. We explain the mathematical similarities and differences in Section 4 and provide a numerical comparison in Section 5.2.2.

Mathematically, our analysis of the capacitated model builds on the literature of heavy-traffic approximations for general queueing systems. The most closely related framework is that developed by Whitt (2006), which uses a fluid model to approximate an overloaded G/G/n+GI queueing system. This framework was further utilized by Bassamboo and Randhawa (2016), who studied prioritization policies in a queueing model with general abandonment and service times (G/G/n+GI). Their focus was on which customers to serve next, considering that some might abandon the queue if forced to wait too long before their service starts. In their framework, service is uninterruptible once started. In contrast, our focus is on deciding who to send home next, which, in queueing theory terms, translates to deciding whose service time should be truncated and by how much.

3. The Uncapacitated Case: A Single-Patient Perspective

We start by studying the uncapacitated case, focusing on a single patient with no capacity constraints. The goal is to determine the optimal hospital LOS that maximizes the patient's expected survival reward while minimizing treatment costs associated with hospitalization and home care. Our approach mimics the Newsvendor model, where the decision of how many newspapers to buy is replaced by the decision of how long to keep the patient in the hospital.

We model the decision of when to send a patient home as a problem of pre-selecting a patient-specific *stay-up-to threshold*. The patient stays in the ward until this threshold, unless they develop an infection before that time. If infection occurs, the patient will remain in the hospital—typically in a general ward—where they are treated for the infection. Conversely, if the patient remains uninfected until that threshold, they are discharged to home care at that time. The resulting HW observation LOS is thus the minimum of the time until infection and the stay-up-to threshold. There are costs associated with both keeping the patient in the hospital for too long (overage cost) and discharging the patient prematurely (underage cost). Determining the optimal stay-up-to threshold mirrors the Newsvendor problem, where decisions balance the trade-off between overage and underage costs, with some important differences that we will discuss later.

Our formulation assumes a finite decision horizon T for the recovery stage of a specific treatment cycle. This aligns with the working assumption for hematological patients that if a patient does not develop an infection within the first 30 days after treatment, any subsequent infection is unlikely to be related to that specific chemotherapeutic treatment (Carmen et al. 2019).

We next introduce relevant notations. Let $r_w(t)$ ($r_h(t)$) denote the hazard-rate function of developing an infection at time t, given that the patient is in the ward (at home) at that time. That is, $r_w(t)$ ($r_h(t)$) is the risk of infection in the ward (at home) at time t, conditional on the patient remaining infection-free up to

that point. These functions, r_w and r_h, are of general form. To illustrate their empirical behavior, Figure 1(c) displays these functions for two patient types: CL and AL. For CL patients, the hazard rate is monotone decreasing, while for AL patients, the hazard rate follows a non-monotonic pattern, initially increasing and then decreasing.

For simplicity, we assume that the hazard-rate functions depend on the time elapsed since the completion of the treatment and the patient location at that time, and *not* on the duration the patient has been in a particular location. For example, $r_h(t)$ is the risk of developing an infection exactly t time units after treatment (provided no infection has occurred earlier), given that the patient is at home at time t and *independently* of when the patient was transitioned from the hospital to home care.

If a patient develops an infection, their survival probability is p_w if the infection occurs at the hospital, or p_h if it occurs at home. (Clearly, the survival probability is complementary to the mortality risk presented in Section 1, and is used instead for convenience.) Our model also includes treatment costs for both hospital and home care. Let c_w (c_h) denote the per-unit-time cost of caring for a patient in the ward (or at home). Additionally, let c_I denote the cost of treating an infection. This infection treatment cost includes all hospitalization expenses incurred from the onset of infection until recovery or death and is assumed to be independent of when or where the infection started. For convenience, we formulate c_I as part of the reward received if no infection occurs by time T. Finally, let R denote the reward associated with surviving a treatment cycle. Without loss of generality, we normalize $R = 1$, and assume that all cost parameters (c_w, c_h, and c_I) are scaled accordingly.

We assume the following properties for the system parameters:

ASSUMPTION 1.

1. *Infection risk at the hospital is higher than home care, i.e.,* $0 \le r_h(t) \le r_w(t), \forall t$.
2. *Survival probability at the hospital is higher than at home, i.e.,* $0 \le p_h \le p_w < 1$.
3. *Cost of care at the hospital is higher than at home, i.e.,* $c_w \ge c_h \ge 0$.
4. *The cost associated with treating an infection is higher than the per-time-unit cost of hospital care (and, consequently, home care as well), i.e.,* $c_I \ge c_w \ge 0$.

As noted in the introduction, these assumptions, and especially items (1) and (2), align with the empirical findings of Carmen et al. (2019). The model parameters may be influenced by the patient's medical condition as well as exogenous factors. For example, the patient's home environment, such as their ability to maintain a semi-sterile setting, may impact the infection risk at home (r_h). The level of support available at home may impact both medical risks and the cost of home care (c_h).

Denote by τ the stay-up-to threshold, and let G_τ be the cumulative distribution function (CDF) of the time until infection for a patient who is sent home at τ if no infection develops by that time. By the connection between the CDF and the hazard rate, we have:

$$G_\tau^c(t) := 1 - G_\tau(t) = e^{-\left(\int_0^{\tau \wedge t} r_w(u)du + \int_\tau^{\tau \vee t} r_h(u)du\right)}. \tag{1}$$

The expected reward associated with a threshold τ, denoted by J_τ, can be expressed as:

$$J_\tau = p_w G_\tau(\tau) + p_h \left(G_\tau(T) - G_\tau(\tau)\right) - c_w \int_0^\tau G_\tau^c(u)du - c_h \int_\tau^T G_\tau^c(u)du + (1+c_I)G_\tau^c(T), \quad (2)$$

where the first term is the product of the probability of developing infection in the ward (before time τ) and the probability of surviving it (p_w). The second term is the product of the probability of developing infection at home (between time τ and T) and the probability of surviving it (p_h). The third and fourth terms are the hospitalization and home-care costs, respectively, and the fifth term is the reward obtained if no infection has occurred up to time T.

Let τ_{opt} be the *minimal optimal threshold* for the reward function J_τ. That is, τ_{opt} is the smallest value of τ that maximizes J_τ. The existence and uniqueness of τ_{opt} follows from the continuity of J_τ over the compact set $[0,T]$.

The formulation (2) differs from the well-known Newsvendor problem due to the nonlinearity of the overage and underage costs. This key difference means that, in general, J_τ is not concave and a closed-form solution for the optimal threshold τ_{opt} cannot be readily derived. Instead, the solution typically requires numerical methods or analytical approaches tailored to the structure of J_τ.

Next, we analyze a specific case where J_τ exhibits a simple structure, allowing the resulting optimal policy to be derived directly. Specifically, we assume that the risk functions are constant over time, enabling us to find a closed-form solution, as characterized in the following proposition.

PROPOSITION 1. *Assume that Assumption 1 holds and that, in addition,* $0 < r_h < r_w$ *and both are constant over time. Let*

$$\tau_0 := T + \frac{1}{r_h} \ln \frac{(p_w - p_h)r_w + c_h \cdot r_w/r_h - c_w}{(r_w - r_h)(1 + c_I - p_h + c_h/r_h)}.$$

Then,

$$\tau_{opt} = \begin{cases} \tau_0, & \text{if } 0 \le \tau_0 \le T; \\ 0, & \text{if } \tau_0 < 0; \\ T, & \text{if } \tau_0 > T. \end{cases}$$

All the proofs for this section appear in Appendix A. Additional closed-form solutions, assuming that Assumption 1 does not hold, appear in an online supplement.

Having established the tractability of solving for the optimal threshold (at least numerically) under any given set of system and patient parameters, we now turn to analyze the structural properties of τ_{opt}. The reward formulation (2) allows us to uncover key monotonicity properties of the τ_{opt} with respect to various parameters.

PROPOSITION 2. *Under Assumption 1, the minimal optimal threshold* τ_{opt} *is monotone decreasing in* c_w, c_I, *and* p_h, *and monotone increasing in* c_h *and* p_w. *It is also monotone increasing in* r_h, *assuming that* $c_h = 0$.

While most of these properties align with one's intuitive expectations, we discover a counterintuitive result: the optimal threshold, τ_{opt}, is not monotone with respect to the in-hospital risk of developing an infection, r_w. It is intuitive to expect τ_{opt} to be monotone decreasing in r_w, implying that a higher in-hospital infection risk would prompt earlier transfer to home care. Surprisingly, this intuition does not always hold. As illustrated in the (discrete-time) counterexample in Table EC.1 in Appendix A, the interplay of system parameters may result in non-monotonic behavior of τ_{opt} with respect to r_w.

4. The Capacitated Model: A Hospital Ward Perspective

So far, we have examined the problem of determining the optimal time to send a patient to home care assuming that the hospital has ample capacity. In practice, hospitals often face capacity constraints, where limited number of beds, equipment, or medical personnel necessitate discharging patients earlier than would be ideal under an uncapacitated policy. This section focuses on the question of how should the time of sending patients to home care be adjusted when capacity is limited?

We start with a general discussion of our modeling approach and the mathematical method used to study this problem. Then, Section 4.1 introduces the queueing model and formalizes a discharge policy. In Section 4.2, we introduce a fluid approximation for the queueing model under a general policy structure, assuming a homogeneous patient population. We prove that the fluid optimal solution can take only five possible structures and derive conditions for optimality for each. Finally, in Section 4.3, we expand the analysis to heterogeneous patient populations, identify necessary conditions for optimality, and propose a practical Index-based Speedup Policy (ISP), which is evaluated numerically in the next section.

The first question in the capacitated case is how to determine when the hospital is indeed capacity-constrained. To address this, we model the hospital ward as a queueing system with n beds. Let λ be the patient arrival rate into the ward and S be the random variable of patient LOS. Define the offered load inflicted on the system by its patients as $U = \lambda E[S]$. Then, to determine whether the system is in under- or over-loaded states, compare U to n: if $U << n$, capacity constraints are minimal; if $U \geq n$, they are substantial. A specific challenge in this context is that the service time, S, and consequently its expected value, $E[S]$, is not exogenous. Instead, the patient LOS is an *outcome* of optimization and may be impacted by the system's capacity.

To disentangle patient LOS from system capacity, we consider an alternative definition of offered load: the expected number of busy servers in an infinite-server queue. Recall that τ_{opt} is the minimal optimal threshold for a patient in the uncapacitated case. In particular, in a system with an infinite number of servers, a patient will stay in the hospital until time τ_{opt} and then sent home, unless they developed an infection earlier. We will refer to this as a *full stay*. Let $S_{\tau_{opt}}$ be the patient LOS given a full stay. The system's offered load is then defined as $U_{\tau_{opt}} = \lambda E[S_{\tau_{opt}}]$, and the system is considered overloaded if $U_{\tau_{opt}} >> n$.

In our study, the overloaded regime is particularly relevant for two reasons: (1) It is exactly in this regime that the tradeoff between utilizing capacity and optimizing hospital LOS is critical. (2) The hospital ward

motivating this study operates in a highly-loaded state, with an average occupancy of 97%. Such high load is typical due to the high costs of hematology hospitalizations (stemming mostly from the need for patient isolation during and after treatment) and increased demand (stemming from advancements in cancer treatment in recent decades).

We assume that patients who cannot be hospitalized for observation in the ward due to capacity constraints are sent to home care immediately. Therefore, our model will have blocking dynamics. While in practice, a patient without an available ward bed might be admitted to another ward, this is generally undesirable for hematology patients. Such off-placements not only increase exposure to hospital-acquired infections (Carmen et al. 2019) but may also lead to inferior patient care (Song et al. 2020). In fact, Carmen et al. (2019) showed that Internal wards may be strictly inferior to home care for hematology patients. Thus, assuming that no patients are sent to a ward other than the HW and that the same threshold τ_{opt} is used to determine when to send a patient home for all patients who are not blocked, we can model this system as a $G/G/n/n$ loss system in the overload regime, with service time distribution as $S_{\tau_{opt}}$.

REMARK 1. The n beds in the HW queue are assumed to be a fixed capacity dedicated to patient observation. However, in practice this assumption may be violated in two ways: (1) infected patients may remain in the HW, occupying some of this "dedicated" capacity, and (2) ward beds may also be used for other purposes, such as patient treatments. To address (1), one could use an argument analogous to Chan et al. (2012), which incorporates readmission load into the cost parameters rather than explicitly modeling patient returns. This approach was further validated by Armony et al. (2018) using a high-fidelity simulation. To bypass (2), we examine hospital data to uncover the average number of beds allocated to observation in practice and use this as the baseline capacity in our numerical study (Section 5).

The complexity of the underlying stochastic process makes exact analysis to identify the optimal policy in the capacitated case prohibitively complex. Instead, we approximate the system using a fluid model, where a discrete flow of patients is modeled as a continuous flow and discrete-time is replaced with continuous-time. This fluid model approach has been successfully applied in queueing theory by Whitt (2006), Kang and Ramanan (2010), and Zhang (2013) for systems with general service times and abandonment distributions, with rigorous justification provided in those settings. Our approach mimics the methodology of Bassamboo and Randhawa (2016), who used fluid approximations to optimize scheduling in an overloaded queueing system with impatient customers, albeit with a focus on prioritization rather than patient service times, as we do here. While we do not rigorously establish that the fluid model is the limiting behavior of the underlying stochastic system, we adopt it as a reasonable and practical approximation.

4.1. The $G/G/n/n$ Fluid Model

Consider a $G/G/n/n$ system with generally distributed service time (LOS) $S_{\tau_{opt}}$ and an offered load $U_{\tau_{opt}} = \lambda E\left[S_{\tau_{opt}}\right]$. Assume that the ward has n beds, and let $\rho = \frac{U_{\tau_{opt}}}{n} = \frac{\lambda E\left[S_{\tau_{opt}}\right]}{n}$. We assume that the

system is overloaded, thus $\rho > 1$. In an overloaded system, arriving patients frequently encounter a full ward. In such cases, the following two options can be considered for handling a new patient arrival:

1. Blocking: Send the new patient directly to home care.
2. Speedup: Send to home care the patient who has been observed in the ward the longest (before their optimal stay-up-to threshold τ_{opt}) to make room for the new patient.

Note that the Speedup policy may be considered more fair as it treats all patients the same, in distribution. More generally, a policy can involve a combination of blocking and speedup, with variations in the choice rule of which patient to discharge when applying speedup and under what conditions. The fluid model can be utilized to approximate the reward function for any given policy. We proceed with formulating this fluid model.

Consider an arbitrary discharge policy π where discharge times do not exceed τ_{opt}. Let S_π be the service time of patients under the policy π, with corresponding CDF F_π, mean service time m_π, and service rate μ_π. In particular, we have:

$$m_\pi = \frac{1}{\mu_\pi} = E[S_\pi] = \int_0^{\tau_{opt}} F_\pi^c(x)dx, \tag{3}$$

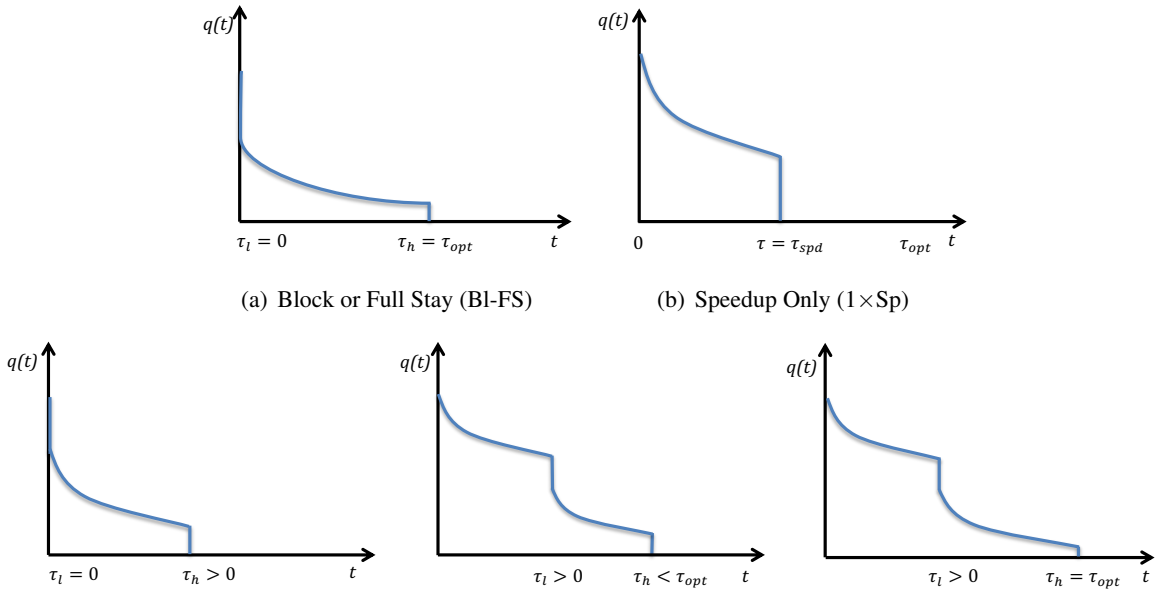
where $F_\pi^c := 1 - F_\pi$. To describe the fluid model in steady state, we adapt the characterization provided by Whitt (2006) to blocking (loss) systems.

The G/GI/n/n Fluid Model in Steady State. For an arbitrary policy π, the G/GI/n/n fluid model with service-time distribution F_π that operates under overloaded conditions, where $\rho_\pi := \frac{\lambda E[S_\pi]}{n} > 1$, has a unique steady state $q(t)$, characterized as: $q(0^-) = \bar{\lambda}$, $q(0) = \mu_\pi$, and $q(t) = \mu_\pi F_\pi^c(t), t \geq 0$, where $\bar{\lambda} := \lambda/n$. The fluid blocking rate is given by $\bar{\lambda} - \mu_\pi$.

Loosely speaking, $q(t)$ describes the fluid content in steady state of all of the fluid that arrived exactly t time units ago. Patients may stay up the the optimal LOS, τ_{opt}. Figures 2(a)–2(b) depicts the fluid model for the two specific policies described above (among other cases - as described later in Corollary 1): (a) *Block or Full Stay*: Patients are either blocked or complete their full stay if admitted; (b) *Speedup Only*: Patients are discharged early to accommodate new arrivals. While in the stochastic system the Speedup Only policy is not necessarily a threshold policy, it becomes a threshold policy under the (deterministic) fluid model. This is because, in steady state, all the fluid with the longest system time arrived at *exactly* the same moment. Let τ_{spd} denote the threshold under the Speedup Only policy, representing the maximal LOS of patients under a single-threshold being used when capacity constraints prevent everyone from staying until τ_{opt}. To compute the Speedup Only threshold, τ_{spd}, solve:

$$1/\bar{\lambda} = \int_0^{\tau_{spd}} G_T^c(x)dx, \tag{4}$$

where G_T is the CDF of the time until infection if the patient remains hospitalized for the full horizon T, as given by Equation (1). The solution for Equation (4) exists and is unique due to the intermediate-value theorem, the overload assumption, and the continuity of the distribution G_T.



(a) Block or Full Stay (Bl-FS) (b) Speedup Only (1×Sp)

(c) Block or Speedup (Bl-Sp) (d) 2-level Speedup (2×Sp) (e) Speedup or Full Stay (Sp-FS)

Figure 2 Fluid Content ($q(t)$) as a Function of LOS (t) for the Five Possible Solutions to the Capacitated Optimization Problem (7)

LEMMA 1. *Assuming the system is overloaded, τ_{spd} is: (1) monotone increasing in the risk of developing infection in the ward, r_w, and in the number of ward beds, n. (2) monotone decreasing in the patient arrival rate, λ. (3) independent of all other system parameters.*

The proof follows directly from Eq. (4) and the definition of G_T^c using the hazard-rate function r_w. Lemma 1 shows that under the Speedup Only policy, in an overloaded system, LOS is determined by the infection risk at the hospital and the arrival rate per bed. For a fixed number of beds, a higher hospital infection risk increases the "allowed" patient LOS in the hospital. Interestingly, τ_{spd} is independent of the home-care parameters as long as the system is overloaded.

4.1.1. General Policies. Up to now, we have focused on single-threshold policies, where patients are sent home at the specified threshold ($\tau = 0$ under blocking or $\tau = \tau_{spd}$ under speedup only) unless they develop an infection beforehand. In general, we might consider a broader family of policies π, in which, for all $x > 0$, a *fraction* $\psi_\pi(x)$ of patients who have been in the ward for x time units are sent home at time x, where $0 \le \psi_\pi(\cdot) \le 1$. For $x = 0$, $\psi_\pi(0)$ refers to the fraction of patients who are blocked upon arrival. We follow the mathematical convention that at any point in time, speedups occur before blocking. In the overloaded regime $\psi_\pi(0)$ is determined by the values of $\psi_\pi(x)$, $x > 0$, and the ward capacity. That is, $\psi_\pi(0)$ is the smallest number such that the system is not overcapacity. A threshold policy τ ($\tau \ge 0$) is a special case of this family, with $\psi_\pi(\tau) = 1$ and $\psi_\pi(x) = 0$ for all $0 < x < \tau$. For simplicity, we focus on policies

where the set of time points x such that $\psi_\pi(x) > 0$ is *finite*.[2] We refer to this finite set of $K+2$ thresholds as $\vec{\tau} = \{\tau_k,\ k \in \{0,1,...,K,K+1\}\}, K < \infty$, where $0 = \tau_0 < \tau_1 < ... < \tau_K < \tau_{K+1} = \tau_{opt}$.

In the fluid model, this policy may equivalently be described as a pair $(\vec{\tau}, \vec{\delta})$, where $\vec{\tau}$ is the set of thresholds, matched by $\vec{\delta} = \{\delta_k \geq 0,\ k \in \{0,1,...,K,K+1\}\}$—its corresponding set of non-negative mass values. At each threshold τ_k the fluid content decreases instantaneously by a mass of δ_k. The steady-state fluid content under this general policy is described as follows: the process starts at $q(0^-) = \bar{\lambda}$, then decreases instantly to $q(0) = \bar{\lambda} - \delta_0$, and at time t, the fluid content is given by

$$q(t) = \left(\bar{\lambda} - \sum_{\{k:\tau_k \leq t\}} \frac{\delta_k}{G_T^c(\tau_k)} \right) G_T^c(t), \quad \text{for } t \geq 0. \tag{5}$$

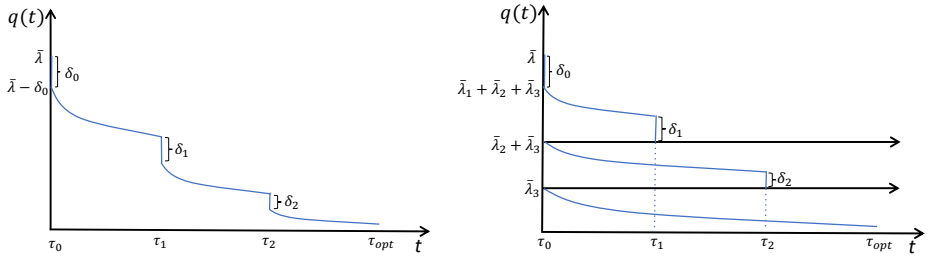
For this policy to be admissible, the following condition must be satisfied: $\sum_{k=0}^{K+1} \frac{\delta_k}{G_T^c(\tau_k)} = \bar{\lambda}$.

To show that the two representations of the general policy—$(\vec{\tau}, \vec{\delta})$ and ψ—are equivalent, note that we have that for all $x > 0$,

$$\psi(x) = \begin{cases} \frac{\delta_k}{q(\tau_k)+\delta_k} & \text{if } x = \tau_k \text{ for some } k, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Given the above policy description, it is natural to seek to optimize over all the admissible pairs $(\vec{\tau}, \vec{\delta})$ to identify the pair that maximizes the overall value function in steady state. To facilitate this, we now provide an alternative and equivalent description of the policy, which is particularly well-suited for evaluating the value function. Following Bassamboo and Randhawa (2016), we observe that the admissible policy $\pi = (\vec{\tau}, \vec{\delta})$ may alternatively be described as a partition of the patient population into $K+2$ classes, each with an arrival rate of $\bar{\lambda}_k = \frac{\delta_k}{G_T^c(\tau_k)}$, for $k = 0,1,...,K+1$. For class k, we apply the single-threshold policy $\pi_k = \tau_k$. Figure 3 illustrates the equivalence between the two policy representations. Thus, a general admissible policy may be described equivalently as $\pi = (\vec{\tau}, \vec{\lambda})$, since, given $\vec{\tau}$, there is a one-to-one correspondence between $\vec{\delta}$ and $\vec{\lambda}$. With this policy characterization, we are now ready to formalize our fluid-level optimization problem.

[2] Given the continuity of the value function J_τ, we may approximate any general policy to the desired level of accuracy using a finite set of thresholds (see Remark 1 in Bassamboo and Randhawa 2016).



(a) Fluid dynamics (b) Fluid dynamics in a multi-class system

Figure 3 **Fluid Content ($q(t)$) as a Function of LOS (t) Viewed as (a) a Single Class with Multiple Thresholds Versus (b) Multiple Classes with a Single Threshold Each**

4.2. The Fluid-level Capacitated Optimization Problem

To find the optimal policy for hospital LOS in the capacitated case at the fluid level, we seek to divide the fluid-level patient population $\bar{\lambda}$ into $K+2$ classes of sizes $\bar{\lambda}_k$ $(k=0,1,...,K+1)$ and determine a set of $K+2$ discharge thresholds that maximize the total value gained by the entire patient population. The optimization problem is given by

$$\begin{aligned} \sup_{K\in\mathbb{N};\ (\vec{\lambda},\vec{\tau})\in\mathbb{R}_+^{K+2}\times\mathbb{R}_+^{K+2}} & \sum_{k=0}^{K+1}\bar{\lambda}_k J_{\tau_k} \\ \text{s.t.} & \sum_{k=1}^{K+1}\bar{\lambda}_k m_{\tau_k}\leq 1, \\ & \sum_{k=0}^{K+1}\bar{\lambda}_k=\bar{\lambda}, \\ & 0=\tau_0\leq\tau_1\leq\tau_2...\leq\tau_K\leq\tau_{K+1}=\tau_{opt}, \end{aligned} \tag{7}$$

where J_τ is the value function associated with a threshold policy τ as defined in (2), and m_{τ_k} is the expected time in the ward for a patient of class k:

$$m_\tau=\frac{1}{\mu_\tau}=\int_0^\tau x g_T(x)dx+\int_\tau^\infty \tau g_T(x)dx=\int_0^\tau x g_T(x)dx+\tau G_T^c(\tau),$$

recalling that G_T is the CDF of the time until infection for a patient remaining in the hospital until time T, and g_T its PDF. Note that, m_τ is differentiable with respect to τ and that $m'_\tau=G_T^c(\tau)$.

We next argue that, in optimality, the system is critically loaded; that is, there exists an optimal solution to (7) where the first constraint is obtained as an equality.

LEMMA 2. *If the system is overloaded when all patients are sent home according to the threshold policy τ_{opt}, then if an optimal solution to (7) exists, then there exists a solution to (7) where the constraint $\sum_{k=1}^{K+1}\bar{\lambda}_k m_{\tau_k}\leq 1$ is obtained as an equality.*

All the proofs for this section appear in Appendix B.

By Lemma 2, problem (7) may be equivalently rewritten as

$$\begin{aligned} \sup_{K\in\mathbb{N};\ (\vec{\lambda},\vec{\tau})\in\mathbb{R}_+^{K+2}\times\mathbb{R}_+^{K+2}} & \sum_{k=0}^{K+1}\bar{\lambda}_k J_{\tau_k} \\ \text{s.t.} & \sum_{k=1}^{K+1}\bar{\lambda}_k m_{\tau_k}=1, \\ & \sum_{k=0}^{K+1}\bar{\lambda}_k=\bar{\lambda}, \\ & 0=\tau_0\leq\tau_1\leq\tau_2...\leq\tau_K\leq\tau_{K+1}=\tau_{opt}. \end{aligned} \tag{8}$$

Observe that for a fixed value of K and a given set $\vec{\tau}$, the problem (8) reduces to a linear program in $\vec{\lambda}$ with two constraints. Consequently, a basic solution will have at most two non-zero values of λ_k. Therefore, as in Bassamboo and Randhawa (2016), we conclude that there exists a solution to (8) where at most two classes are non-empty. This result is formalized in the next proposition.

PROPOSITION 3. *There exists an optimal solution to (7) (and (8)) with at most two non-empty classes.*

The proof is essentially identical to the proof of Proposition 1 in Bassamboo and Randhawa (2016) and is hence omitted. By Proposition 3, optimization problem (8) may be reduced to the following:

$$\begin{aligned} \sup_{0\le\bar{\lambda}_l\le\bar{\lambda};0\le\tau_l\le\tau_h\le\tau_{opt}} \quad & \bar{\lambda}_l J_{\tau_l} + (\bar{\lambda}-\bar{\lambda}_l) J_{\tau_h} \\ \text{s.t.} \quad & \bar{\lambda}_l m_{\tau_l} + (\bar{\lambda}-\bar{\lambda}_l) m_{\tau_h} = 1. \end{aligned} \tag{9}$$

Given τ_l and τ_h, $\bar{\lambda}_l$ is uniquely determined by solving the equality constraint in (9), as $\bar{\lambda}_l = \frac{\bar{\lambda}-\mu_{\tau_h}}{1-\mu_{\tau_h}/\mu_{\tau_l}}$ if $\tau_l \neq \tau_h$, and $\bar{\lambda}_l = 0$, otherwise. Furthermore, for $\bar{\lambda}_l$ to satisfy the conditions of the optimization problem (9), the following must hold:

$$\mu_{\tau_{opt}} \le \mu_{\tau_h} \le \bar{\lambda} \le \mu_{\tau_l} \le \infty. \tag{10}$$

Recalling the definition of τ_{spd} in (4), it follows from (10) that any feasible solution must satisfy $\tau_l \le \tau_{spd} \le \tau_h$. Since $\mu_{\tau_h} > 0$, we have $\bar{\lambda}_l \le \bar{\lambda}$ and the optimization problem (9) simplifies to

$$\sup_{0\le\tau_l\le\tau_{spd}\le\tau_h\le\tau_{opt}} \bar{\lambda}_l(\tau_l,\tau_h) J_{\tau_l} + (\bar{\lambda}-\bar{\lambda}_l(\tau_l,\tau_h)) J_{\tau_h}, \tag{11}$$

where $\bar{\lambda}_l(\tau_l,\tau_h) = \frac{\bar{\lambda}-\mu_{\tau_h}}{1-\mu_{\tau_h}/\mu_{\tau_l}}$, if $\mu_{\tau_l} \neq \mu_{\tau_h}$, and $\bar{\lambda}_l(\tau_l,\tau_h) = 0$, otherwise.

The culmination of this discussion is summarized in the following corollary, which states that the solution of the capacitated problem reduces to *five* simple and mutually exclusive cases, as depicted in Figure 2.

COROLLARY 1. *There exists an optimal solution to the capacitated problem (7) with up-to-two threshold structure as in (11), with the following exhaustive and mutually exclusive cases:*

1. *Block or Full Stay (Bl-FS):* $\tau_l = 0,\ \tau_h = \tau_{opt}$,
2. *Speedup only (1×Sp):* $0 < \tau_l = \tau_h \le \tau_{opt}$,
3. *Block or Speedup (Bl-Sp):* $\tau_l = 0,\ 0 < \tau_h < \tau_{opt}$,
4. *Two-level Speedup (2×Sp):* $0 < \tau_l < \tau_h < \tau_{opt}$,
5. *Speedup or Full Stay (Sp-FS):* $0 < \tau_l < \tau_h = \tau_{opt}$.

Now that we have identified the five possible cases for the solution of the capacitated problem, we turn to the question of identifying the specific solution given certain characteristics of the problem primitives. We begin by specifying some *necessary* conditions for the optimal solution of (11).

PROPOSITION 4 **(Necessary optimality conditions).** *If the function J_τ is differentiable with respect to τ, with derivative*

$$J'_\tau = G^c_\tau(\tau)\left((p_w - p_h)r_w(\tau) - (c_w - c_h)\right) + G^c_\tau(T)\left(r_h(\tau) - r_w(\tau)\right)(1 + c_I - p_h) - c_h\left(\left(r_h(\tau) - r_w(\tau)\right)\int_\tau^T G^c_\tau(u)du\right),$$

then,

(a) An optimal solution to (11) *of the form* (τ_l, τ_h) *with* $0 < \tau_l < \tau_{spd} < \tau_h < \tau_{opt}$ *(the 2×Sp policy) must satisfy*

$$\frac{J_{\tau_h} - J_{\tau_l}}{m_{\tau_h} - m_{\tau_l}} = \frac{J'_{\tau_l}}{1 - G_T(\tau_l)} = \frac{J'_{\tau_h}}{1 - G_T(\tau_h)}. \tag{12}$$

(b) An optimal solution to (11) *of the form* $(0, \tau_h)$ *with* $\tau_{spd} < \tau_h < \tau_{opt}$ *(Bl-Sp policy) must satisfy*

$$\frac{J_{\tau_h} - J_0}{m_{\tau_h}} = \frac{J'_{\tau_h}}{1 - G_T(\tau_h)} \geq \frac{J'_0}{1 - G_T(0)} \equiv J'_0. \tag{13}$$

(c) An optimal solution to (11) *of the form* (τ_l, τ_{opt}) *with* $0 < \tau_l < \tau_{spd} < \tau_{opt}$ *(Sp-FS policy) must satisfy*

$$\frac{J_{\tau_{opt}} - J_{\tau_l}}{m_{\tau_{opt}} - m_{\tau_l}} = \frac{J'_{\tau_l}}{1 - G_T(\tau_l)} \leq \frac{J'_{\tau_{opt}}}{1 - G_T(\tau_{opt})}. \tag{14}$$

(d) An optimal solution to (11) *of the form* $(0, \tau_{opt})$ *(Bl-FS policy) must satisfy*

$$J'_0 \equiv \frac{J'_0}{1 - G_T(0)} \leq \frac{J_{\tau_{opt}} - J_0}{m_{\tau_{opt}}} \leq \frac{J'_{\tau_{opt}}}{1 - G_T(\tau_{opt})}. \tag{15}$$

Proposition 4 establishes that the ratio $\xi(\tau) := J'_\tau/(1 - G_T(\tau))$ plays a critical role in characterizing the solution. This ratio expresses the marginal increase in patient's value from raising the threshold relative to the corresponding marginal increase in expected service time. While the former potentially benefits the patients, the latter hurts the patient population by reducing available capacity.

Examining the necessary conditions of (14) and (15), we note that if τ_{opt} is an internal maximum of J_τ, these conditions will never be satisfied. This is because, in such a case $\xi(\tau_{opt}) = J'_{\tau_{opt}} = 0$. In particular, in such cases, it is never optimal to keep patients in the hospital for their full stay (FS). This observation is formalized in the next corollary:

COROLLARY 2. *If τ_{opt} is an internal maximum point of J_τ, that is, if $0 < \tau_{opt} < T$, then the policies Sp-FS or Bl-FS are not optimal. In particular, under these conditions, it is never optimal to keep a patient in the hospital for their full stay when the hospital ward is overloaded.*

While Proposition 4 provides necessary conditions for optimality of (11), it would also be desirable to establish sufficient conditions for optimality. Such conditions are provided in Lemma EC.8. However, as it turns out, the conditions presented in that lemma are practically unrealistic (see Figure 5 and the accompanying discussion). Hence, we include the lemma in the Appendix for completeness.

4.3. The Multi-Patient Type Case

We now extend the capacitated ward problem to the heterogeneous-patient case, accounting for multiple patient types, each with its own characteristics. Let type-i patients ($i=1,...,I$), $r_h^i(\cdot)$ and $r_w^i(\cdot)$ denote their infection hazard-rate functions at home and in the ward, respectively, and let p_h^i and p_w^i be the corresponding probabilities of surviving that infection. These characteristics result in a type-dependent reward function J_τ^i and a type-specific optimal threshold τ_{opt}^i. The arrival rate for type-i patients is denoted by λ^i. In this heterogeneous-patient setting, the hospital needs to determine optimal stay-up-to thresholds for each type, while accounting for its limited capacity.

To model overload in this context, let $S^i_{\tau^i_{opt}}$ be the LOS of a type-i patient under a full stay policy with threshold τ_{opt}^i. Define the system's offered load as $U=\sum_{i=1}^I \lambda^i E\left[S^i_{\tau^i_{opt}}\right]$, where n is the number of beds and $\rho=\frac{U}{n}$ is the ward load. The system is considered overloaded if $\rho>1$.

Similarly to the single-patient type case, we use a fluid-model approximation to evaluate the system's workload in steady state. We again consider a generic policy that divides each patient type i into K^i+2 classes. For class k ($k=0,...,K^i+1$) of patient type i, the arrival rate is $\bar\lambda_k^i$ and a single threshold τ_k^i is applied. The thresholds satisfy $0=\tau_0^i\le\tau_1^i\le...\le\tau_{K^i}^i\le\tau_{K^i+1}^i=\tau_{opt}^i$. The corresponding multi-type optimization problem is given by:

$$\begin{aligned}\sup_{\vec K\in\mathbb{N}^I;\ (\vec\lambda^i,\vec\tau^i)\in\mathbb{R}_+^{K^i+2}\times\mathbb{R}_+^{K^i+2},\ i=1,...,I} & \sum_{i=1}^I\sum_{k=0}^{K^i+1}\bar\lambda_k^i J^i_{\tau_k^i} \\ \text{s.t.} \quad & \sum_{i=1}^I\sum_{k=1}^{K^i+1}\bar\lambda_k^i m^i_{\tau_k^i}=1, \\ & \sum_{k=0}^{K^i+1}\bar\lambda_k^i=\bar\lambda^i,\quad i=1,...,I, \\ & 0=\tau_0^i\le\tau_1^i\le...\le\tau_{K^i}^i\le\tau_{K^i+1}^i=\tau_{opt}^i,\quad i=1,...,I,\end{aligned} \tag{16}$$

where the expected LOS of a patient of type i and class k is

$$m^i_{\tau_k^i}\equiv\frac{1}{\mu^i_{\tau_k^i}}=\int_0^{\tau_k^i}xg_T^i(x)dx+\int_{\tau_k^i}^\infty\tau_k^i g_T^i(x)dx=\int_0^{\tau_k^i}xg_T^i(x)dx+\tau_k^i(1-G_T^i(\tau_k^i)). \tag{17}$$

The first constraint above should be $\sum_{i=1}^I\sum_{k=1}^{K^i+1}\bar\lambda_k^i m^i_{\tau_k^i}\le 1$ as opposed to $\sum_{i=1}^I\sum_{k=1}^{K^i+1}\bar\lambda_k^i m^i_{\tau_k^i}=1$, but by an argument analogous to that of Lemma 2, the two formulations are equivalent.

For fixed values of $K^1,...,K^I$ and $\vec\tau^1,...,\vec\tau^I$, the above is a linear program in $\vec\lambda^1,...,\vec\lambda^I$ with $\sum_{i=1}^I(K^i+2)$ variables and $I+1$ constraints. By a similar argument to Proposition 3, there exists an optimal solution to (16) with at most $I+1$ non-empty patient classes. This is formalized next.

PROPOSITION 5. *There exists an optimal solution to (16) that creates at most $I+1$ patient classes. In that solution, at least $I-1$ of the patient types have a single type-specific threshold applied to them and up to one patient type has up to two type-specific thresholds applied to it.*

We conclude that handling the multi-patient type capacitated case is not as onerous as one might think. It involves determining which single patient type should be assigned two thresholds and identifying the corresponding $I+1$ thresholds. The $I-1$ classes, each with the single threshold, apply a speedup only policy (Figure 2(b)), where the speedup threshold can also be 0. Unlike the homogeneous-patient case, the speedup threshold here cannot be simply characterized as in (4), because the relationship between the patient types influence the optimal thresholds.

By Proposition 5, the optimization problem (16) can be equivalently written as

$$\begin{aligned} \sup_{(\lambda_l^i,\tau_l^i,\tau_h^i)\in\mathbb{R}^3_+,\ i=1,\ldots,I} \quad & \sum_{i=1}^{I}\left(\bar{\lambda}_l^i J^i_{\tau_l^i}+(\bar{\lambda}^i-\bar{\lambda}_l^i)J^i_{\tau_h^i}\right) \\ \text{s.t.} \quad & \sum_{i=1}^{I}\left(\bar{\lambda}_l^i m^i_{\tau_l^i}+(\bar{\lambda}^i-\bar{\lambda}_l^i)m^i_{\tau_h^i}\right)=1, \\ & 0\le\bar{\lambda}_l^i\le\bar{\lambda}^i, \quad i=1,\ldots,I, \\ & 0\le\tau_l^i\le\tau_h^i\le\tau_{opt}^i, \quad i=1,\ldots,I. \end{aligned} \tag{18}$$

Referring back to Proposition 5, we note that, since it is unknown a priori which patient type will require two thresholds and which type one, the formulation allows up to two thresholds per patient type for all I patient types. Our next result establishes necessary conditions for the optimality of a solution to the problem (18).

PROPOSITION 6 **(Necessary optimality conditions).** *If the functions J_τ^i, $i=1,\ldots,I$, are all differentiable as a function of τ, with derivatives $J_\tau^{i'}$, respectively, then an optimal solution to* (18) *must satisfy for all $i,j\in\{1,2,\ldots,I\}$, $\tau^i\in\{\tau_l^i,\tau_h^i\}$, $\tau^j\in\{\tau_l^j,\tau_h^j\}$,*

$$\frac{J^{i'}_{\tau^i}}{1-G_T^i(\tau^i)}=\frac{J^{j'}_{\tau^j}}{1-G_T^j(\tau^j)},\ 0<\tau^k<\tau^k_{opt},\ k\in\{i,j\}, \tag{19}$$

$$J_0^{i'}\le\frac{J^{j'}_{\tau^j}}{1-G_T^j(\tau^j)},\ \tau^i=0,\ 0\le\tau^j\le\tau^j_{opt}, \tag{20}$$

and

$$\frac{J^{i'}_{\tau^i_{opt}}}{1-G_T^i(\tau^i_{opt})}\ge\frac{J^{j'}_{\tau^j}}{1-G_T^j(\tau^j)},\ \tau^i=\tau^i_{opt},\ 0\le\tau^j\le\tau^j_{opt}. \tag{21}$$

Proposition 6 highlights the pivotal role of the index $\xi^i(\tau):=\frac{J_\tau^{i'}}{1-G_T^i(\tau)}$ in determining the speedup and blocking thresholds for various patient types. Specifically, it states that: (1) if it is optimal to use blocking for a certain patient class (i.e., setting a speedup threshold $\tau=0$), then the index $\xi^i(\tau)$ is *minimal* at time 0 (see (20)). (2) if it is optimal to allow a full stay for a class (i.e., setting a speedup threshold $\tau=\tau_{opt}$), then the index $\xi^i(\tau)$ is *maximal* at time τ_{opt} (see (21))[3]. Intuitively, the index $\xi^i(\cdot)$ expresses the tradeoff between

[3] Similarly to Corollary 2, we will see that under practically relevant conditions, this case will not be optimal unless $\tau_{opt}=T$. This is formalized in Corollary 3.

the *marginal value* of extending the patient's stay and the corresponding *marginal* expected service time (or LOS). These two necessary conditions align with a policy that prioritizes speedups for the patient type that minimizes this index. This motivates us to propose the *dynamic Index-based Speedup Policy (ISP)*, which is reminiscent of the well-known generalized Cμ rule (van Mieghem 1995, Mandelbaum and Stolyar 2004).

The Dynamic Index-based Speedup Policy (ISP). Building on Proposition 6, we propose a dynamic discharge policy for hospital wards operating under capacity constraints. Under this policy, when the hospital ward is at full capacity and a new patient arrives, the system identifies and discharges the patient with the lowest index $\xi(\tau) := J'_\tau/(1-G_T(\tau))$, where τ is the *current* hospitalization time. If a patient remains in the ward until their optimal threshold τ_{opt}, they are automatically discharged at that time. Algorithm 1 provides a discrete-time implementation of the ISP algorithm. Note that this algorithm is sufficiently flexible to allow for a dynamic update of the index functions ξ based on real-time patient status changes. This algorithm will be evaluated numerically in the next section, demonstrating its performance in realistic hospital scenarios.

Algorithm 1: The Discrete-Time ISP Algorithm

When a new patient arrives with $\tau^i_{opt} > 0$, if there is a bed available, admit the patient to the ward.

If there are no beds available, calculate the following index for all patients in the ward:

1. For hospitalized patients, calculate $\xi^i(\tau) = J^{i'}_\tau/(1-G^i_T(\tau)) \approx \frac{J^i(\tau+1)-J^i(\tau)}{1-G^i_T(\tau+1)}$, where τ is their current LOS.
2. For the new patient, calculate $\xi^i(0) = J^{i'}_0/(1-G^i_T(0)) \approx \frac{J^i(1)-J^i(0)}{1-G^i_T(1)}$.

Find the patient with the lowest index.

If the lowest index corresponds to the new patient, block that patient. If the lowest index corresponds to a currently hospitalized patient, send that patient to home care and admit the new one.

REMARK 2. The ISP algorithm is more elaborate than the heuristic proposed by Ouyang et al. (2020), though it shares a common structure: in both cases, when a patient arrives to a full unit, an index-based policy chooses whether the new patient is admitted or another patient is sped up. The key difference is the proposed indexes. Ouyang et al. (2020) analyzed a single-class case where all patients' health conditions and LOS were governed by the same Markov chain (with six health-stages). They evaluated two policies (among others): (a) The Greedy policy uses the mortality risk difference as an index, which, in our notation, corresponds to $J_{\tau_{opt}} - J_\tau$ (assuming costs are zero); (b) The Ratio policy applies the ratio of the mortality risk difference to the patient's LOS (given their medical state) as the index, which translates to $\frac{J_{\tau_{opt}}-J_\tau}{m_{\tau_{opt}}}$ in our framework. In contrast, the ISP algorithm uses a data-driven index of $\xi(\tau) = \frac{J'_\tau}{(1-G_T(\tau))}$, which reflects the predicted *marginal* increase in mortality risk divided by the predicted expected *change* in the patient LOS (based on their current state). Furthermore, the ISP algorithm accommodates heterogeneous patients with varying risk functions, making it more versatile.

We end this section with Corollary 3, observing that if the reward functions J are increasing up to their optimal threshold τ_{opt}, it is never optimal to keep a patient for its full stay unless their $\tau_{opt} = T$. This follows straightforwardly from Proposition 6.

COROLLARY 3. *If the functions J^i are differentiable and $J^{i'}(\tau) > 0$ for all $\tau < \tau^i_{opt}$, and if patient type i obtains its maximum reward, J^i, at an internal point $0 < \tau^i_{opt} < T$, then the full stay policy is never optimal for patient type i in the overloaded regime.*

5. Numerical Analysis

In this section, we explore the practical relevance of our theoretical results through numerical simulations based on real-world scenarios inspired by Carmen et al. (2019). First, we explore the uncapacitated, single-patient case (§5.1) and then investigate the impact of capacity constraints (§5.2). For consistency with actual hospital practices, which involve daily decision-making, and the prediction models of Carmen et al. (2019), we conduct all simulations in daily resolution.

Carmen et al. (2019) provides risk models based on patient profile including: demographics (e.g., age, disease), treatment details (e.g., chemotherapy treatment length), health status (e.g., white blood cell (WBC) counts), and location (e.g., post-treatment location). We integrate their risk function into our model to evaluate the performance of our policies in practice.

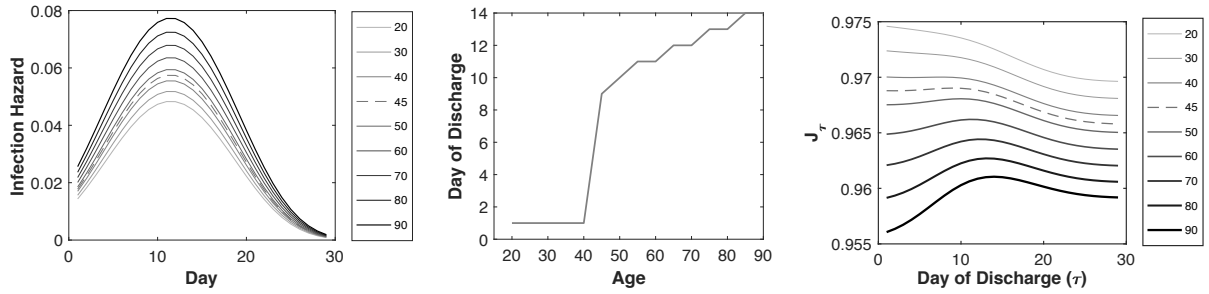
5.1. The Uncapacitated Case

In this section, we explore the optimal LOS for hematology patients in the hospital ward under uncapacitated conditions, focusing on how patient-specific factors influence the decision. Our analysis includes the four main types of hematological cancers (AL, CL, L, and MM). For each disease type, we vary key patient characteristics such as age and treatment length, and then compute the optimal effective threshold using the single-patient optimization model presented in Section 3[4]. Consistent with typical clinical practices, the maximal hospital LOS is set to 30 days ($T = 30$). In addition, for ease of exposition, we assume zero hospitalization costs (i.e., $c_w = c_h = c_I = 0$).

The infection risk functions used across the investigated cases can be categorized into two main types, as shown in Figure 1(c): monotone decreasing functions, where the infection risk consistently declines over time, and increasing-decreasing functions, where the infection risk initially rises, reaches a peak, and then declines. Notably, for each patient type, the infection risk functions for home and hospital settings share the same shape, differing only in their respective risk level. We find that the increasing-decreasing risk functions generally result in no observation period unless the patient has a significant infection history, such as more than eight prior infections, and is over 40 years old. In contrast, monotone decreasing risk functions consistently lead to non-zero observation periods. (See Appendix D.1.)

[4] The numerical results rely on the discrete-time MDP model presented in E.1, which is equivalent to a discrete-time version of the continuous-time Newsvendor-type model discussed in Section 3.

Considering a specific case study, presented in Figure 4, which analyzes the discharge policy based on age for patients with a significant history of infections (9 prior episodes) but in a good post-treatment condition (WBC > 1000). The infection risk function here is of the increasing-decreasing type, peaking on day 12 regardless of age, while overall risk levels increase with age (see Figure 4(a)). Figure 4(b) highlights a sharp policy shift: no observation is recommended for patients aged 40 or younger, while for those above 45, observation is recommended until *around* the risk function peak. This abrupt shift between the two policies cannot be readily inferred from the incremental effect of age on the risk function, which is relatively small. This change arises from the behavior of the expected reward function J_τ, shown in Figure 4(c). For younger patients, J_τ is decreasing, favoring immediate discharge, whereas for older patients, J_τ is increasing-decreasing, justifying observation. Finally, Figure 5(a) presents the ISP index (ξ) for different age groups, along with the corresponding τ_{opt} values for each group. A closer look at the decision-relevant range, shown in Figure 5(b), focuses on age groups where $\tau_{opt} > 0$ and considers only days up to τ_{opt}. Within this range, ξ remains positive (as assumed in Lemma EC.8 in the Appendix), but it is not necessarily monotone—it increases and then decreases over time.

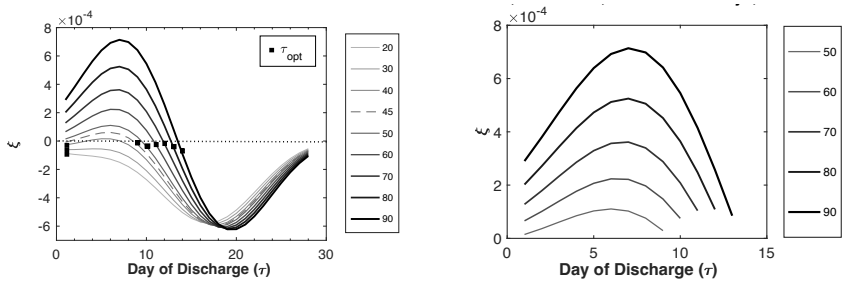


(a) Ward-acquired infection risk as a function of days since treatment and age

(b) Optimal day of discharge, τ_{opt}, as a function of age

(c) J_τ as a function of day of discharge and age

Figure 4 Policy as a Function of Age (AL, High WBC, 9–10-Days Protocol, 9 Past Infections)



(a) ξ as a function of day of discharge and age

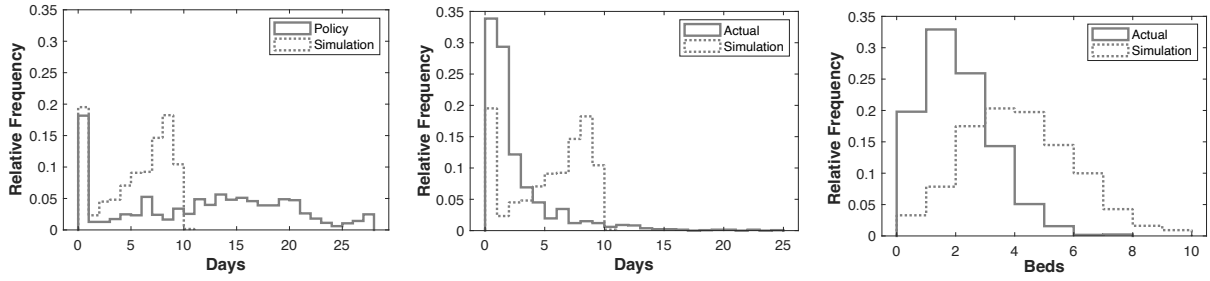
(b) ξ from 0 to τ_{opt} as a function of day of discharge and age

Figure 5 Policy as a Function of Age (AL, High WBC, 9–10-Days Protocol, 9 Past Infections)

We now turn to examining the interaction between the discharge policy, system occupancy, and patient LOS using data of 1332 patients who were candidates for observation during a specific period in the HW (dataset details in Carmen et al. (2019)). Figure 6(a) shows a histogram of the optimal policy for observation time, τ_{opt}, as determined by our uncapacitated model (*'Policy'*, solid line). While τ_{opt} is the maximal LOS under our model, actual observation LOS can be shorter if patients get an infection during their stay. To provide a realistic depiction of actual observation LOS under the optimal policy, we simulated infection incidents and the resulting observation times for each patient, assuming the optimal policy was implemented and sufficient capacity was always available (*'Simulation'*, dotted line). Note that we rely on *simulated* infection events rather than the historical data of patient's infection time, because many patients were discharged earlier than τ_{opt}, providing no information on whether or not they would have gotten infection had they not been released that early. The results indicate that under the optimal policy with no capacity constraints, only 18.2% of the patients should not stay for observation, while the remainder would benefit from some hospital stay. Furthermore, a substantial fraction of the patients (50.4%) would experience a shorter LOS than optimal due to hospital-acquired infections.

In practice, the hospital may have capacity constraints, leading to either speeding up currently-observed patients or blocking new patients from being observed. Figure 6(b) displays actual LOS of patients in our sample (*'Actual'*, solid line), based on observed data, alongside the simulated LOS from Figure 6(a) for comparison. We observe that, in the data, 33.9% of the patients had zero observation days, with 26.6% discharged immediately after treatment and 7.3% contracting an infection right away. Additionally, the actual LOS is statistically shorter than the simulated LOS assuming infinite capacity (*'Simulation'*, dotted line). However, the actual stay exhibits a slightly heavier right tail. Besides limited capacity, another possible explanation for this gap in LOS is the use of a different discharge policy by physicians, recommending shorter LOS for some patients.

Figure 6(c) displays the histogram of observation bed occupancy for both the optimal uncapacitated policy and the actual policy used at the hospital. The optimal uncapacitated policy would have occupied



(a) Optimal hospitalization (b) Optimal vs. actual LOS (c) Number of occupied beds

Figure 6 Hospital LOS Histogram and the Resulting Bed Occupancy Distribution under the Optimal Policy, Simulated Policy assuming Infinite Capacity, and Actual Data (Sample of 1332 Patients Treated in HW)

up to 10 observation beds simultaneously, while the actual hospital policy used up to 8 beds. On average, the optimal uncapacitated policy would have needed 4.1 beds (SD$=1.83$), while in practice, the average number of patients in observation was 1.6 (SD$=1.24$).

Our data also allows for comparison between the actual policy and the optimal uncapacitated policy with respect to patient outcomes. By simulating the mortality rate under the optimal uncapacitated policy, we found that, while the actual mortality rate in the data was 4.43%, it could have been reduced by 27.5% to 3.2% with the optimal uncapacitated policy. This suggests that patient outcomes could be improved by changing the discharge policy and increasing capacity. In the next section, we apply our capacitated models to explore the effect of limited capacity.

5.2. The Capacitated Case

Next, we examine how capacity constraints affect the optimal policy and study the performance of the ISP algorithm. In Section 5.2.1, we analyze the homogeneous-patient case, and in Section 5.2.2 the heterogeneous case. Our main performance measure is the expected reward, J, with focus on the survival rate, which is obtained from J by setting hospitalization and home-care costs to zero.

5.2.1. The Capacitated Case with Homogeneous Patients We start by examining how the shape of the expected reward functions affects the impact of limited capacity. As demonstrated in Figure 4(c), the reward function, J, is generally flat around the optimal threshold τ_{opt}, with more significant changes occurring as we move farther from τ_{opt}. Therefore, strict capacity constraints will likely result in higher mortality rate (lower survival rates), while less stringent constraints are unlikely to change mortality rate dramatically.

Next, we examine the optimal policy types for patients in our data, considering their profile and system load under the assumption of a *homogeneous* patient population. According to Corollary 1, the optimal policy may involve up to two thresholds and fall into one of five possible types. We aim to empirically test which of these types is more prevalent in practice. For this analysis, we assume the system is overloaded, with system loads ranging from moderately to highly overloaded, corresponding to values of ρ equal 1.02, 1.05, 1.1, or 1.2 (i.e., the arrival rate $(\bar{\lambda})$ is $\mu_{\tau_{opt}} \times \rho$).

We solve the optimization problem for 19,200 combinations of disease (4), WBC level (2), age (15), number of past infections (10), treatment protocol length (4), and system load (4). For these combinations, we identify the parameter sets that result in the optimality of each of the five policy types depicted in Figure 2. Table 1 presents the frequency of each policy type. Because τ_{spd} is rarely an integer, exact 1×Sp policies are infrequent due to discretization. Therefore, we define "1×Sp-type" policy as one where $\tau_h - \tau_l = 1$, meaning the two groups that are sped up are discharged within two consecutive days. This category is further divided into two subcases: one where the upper threshold $\tau_h = \tau_{opt}$ (which is also an Sp-FS policy) and another where the upper threshold $\tau_h < \tau_{opt}$ (which is also a 2×Sp policy). These two subcases are

labeled in Table 1 as "1×Sp or Sp-FS" and "1×Sp or 2×Sp", respectively. Notably, we observed no patient's parameter set with pure Sp-FS policy solutions but did encounter some parameter sets with pure 2×Sp policy solutions.

Table 1 Frequency of solution types as a function of patient characteristics

Case		Bl-Sp	2×Sp	1×Sp or 2×Sp	1×Sp or Sp-FS	Sp-FS	Bl-FS
All		2.5%	7.2%	45.1%	45.0%	0.0%	0.1%
Age	20	1.8%	4.6%	38.8%	54.9%	0.0%	0.0%
	30	1.6%	4.7%	40.3%	53.0%	0.0%	0.4%
	40	2.4%	5.5%	41.6%	50.1%	0.0%	0.4%
	50	2.3%	6.6%	43.2%	47.9%	0.0%	0.0%
	60	3.0%	7.6%	45.3%	43.5%	0.0%	0.7%
	70	2.4%	8.4%	47.5%	41.7%	0.0%	0.0%
	80	3.0%	9.1%	49.8%	38.2%	0.0%	0.0%
	90	2.9%	10.1%	51.6%	35.3%	0.0%	0.0%
Disease	AL	7.9%	0.0%	62.9%	28.7%	0.0%	0.4%
	CL	0.0%	10.9%	32.5%	56.6%	0.0%	0.0%
	L	0.0%	17.4%	35.2%	47.4%	0.0%	0.0%
	MM	0.0%	2.5%	44.4%	53.1%	0.0%	0.0%
Load	1.02	0.0%	8.7%	16.9%	74.2%	0.0%	0.2%
	1.05	0.2%	11.2%	34.0%	54.4%	0.0%	0.2%
	1.1	0.9%	8.8%	56.6%	33.5%	0.0%	0.2%
	1.2	8.7%	0.2%	73.0%	18.1%	0.0%	0.0%

Our numerical analysis leads to the following observations: a) Policies of 1×Sp type account for approximately 90% of our examples. Only 7.2% correspond to a pure 2×Sp type, 2.5% to a Bl-Sp type, and 0.1% to a Bl-FS type. In practice, Bl-Sp and Bl-FS policy types should be hardly used. b) As the system load increases, the 1×Sp policy type becomes more prevalent compared to a pure 2×Sp type. However, the Bl-Sp policy type is more common under extremely high loads (e.g., when $\rho = 1.2$). c) The prevalence of policy types across different cancer types is similar, with the exception of AL, which shows a higher percentage of Bl-Sp and "1×Sp or 2×Sp" policy types. The dominance of the 1×Sp type over all other cases demonstrates that optimality and fairness often align, so that one does not need to sacrifice one goal for the other.

Comparing the total reward (which is equivalent to the total survival rate because of the zero hospitalization and home-care costs assumption) under the uncapacitated and capacitated policies, we find that the decrease in survival rate due to capacity limitation is relatively small. Specifically, the reduction is less than 1.5% when $\rho = 1.2$, 0.7% when $\rho = 1.1$, and 0.25% when $\rho = 1.05$.

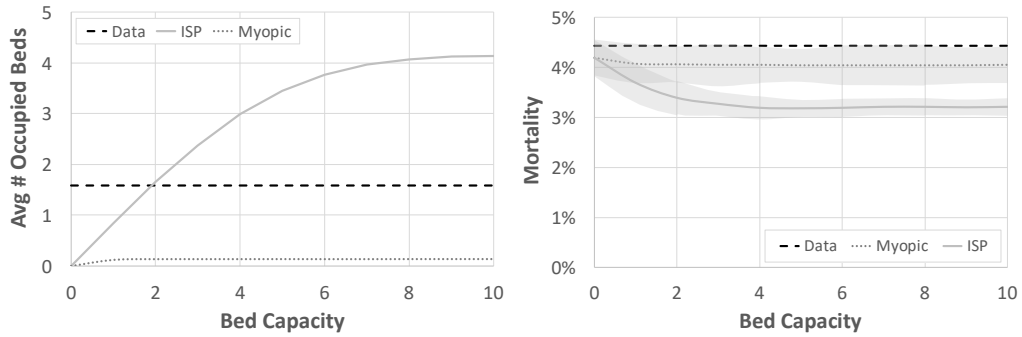
The high prevalence of 1×Sp-type policies naturally raises the question: how much would the total survival rate decrease if we restricted decisions to this simple and fair policy? To examine this, we compared the expected reward (with zero hospitalization and home-care costs) under the 1×Sp policy versus the optimal capacitated policy. Across all examined combinations, the reduction in survival rate remained below

0.5%. Notably, Proposition 5 established that in a multi-class setting, the optimal policy follows a $1\times$Sp structure for all but one patient type. Hence, this analysis provides an empirical upper bound on the potential loss if one were to impose a single-threshold constraint per class in the multi-class optimization.

5.2.2. The Capacitated Case with Heterogeneous Patients Thus far, we have assumed a limited capacity setting with homogeneous patients. We now extend our analysis to the multi-type patient scenario. Section 4.3 introduced the ISP algorithm, which implements a multi-threshold policy by selecting patients for early discharge or blocking using the index ξ from Proposition 6. To evaluate the algorithm's performance, we conduct a data-driven simulation using the same 1332-patient sample from Section 5.1.

In the simulation, patient arrivals followed the actual chemotherapy completion dates. Each patient infection and mortality risk functions were based on their actual profile at admission. We calculated the optimal uncapacitated LOS threshold and allowed patients to stay for that duration if observation beds were available and no infection was detected. If no bed was available upon arrival, we calculated the index $\xi(\tau) = J_\tau'/(1-G_T(\tau))$ (or its discrete-time equivalent), as outlined in Algorithm 1, for both the new patient and for all currently hospitalized patients. The patient with the lowest index was sent to home care (which was done by either blocking the new arrival or expediting an existing patient's discharge, depending on whose index is lower). We varied the number of hospital beds dedicated to observation from 0 to 10; note that we chose to stop at 10 beds because it was the maximal number of occupied beds under infinite capacity (see Figure 6(c)).

Figure 7 presents the ISP simulation results ('ISP', solid lines), and the actual hospital data ('Data', dashed lines). It also includes results for a myopic policy ('Myopic', dotted line), which we will describe later. Figure 7(a) shows that the average number of occupied beds under ISP does not exceed 4.1, which occurs when capacity is 10, compared to 1.6 in practice. Naturally, occupancy increases as bed capacity increases. Notably, with capacity of only 2 available beds, the ISP algorithm and the actual hospital data exhibit the same average occupancy.



(a) Average load (b) Mortality rate

Figure 7 Number of Occupied Beds and Mortality Rate as a Function of Hospital Capacity

Figure 7(b) displays the mortality rate as a function of available observation beds (3-std confidence intervals are marked with gray shading). With 10 beds, the ISP algorithm reduces mortality rate to 3.2%, a 27.7% improvement over the data baseline of 4.43%. To distinguish between the benefits of ISP to those of added capacity, we compare the mortality rate in the data to the mortality under ISP policy in the case of equal average capacity usage. As seen in Figure 7(a), this occurs when ISP is used with two beds. Figure 7(b) shows that, with a capacity of two available beds, the ISP algorithm achieves a 23.5% reduction in mortality rate (from 4.43% to 3.39%). Therefore, the remaining 4.2% improvement is attributable to increased capacity. However, this comparison is not exactly "apples to apples" because, in practice, beds availability for observation changes dynamically and physicians also have access to discretionary real-time information about patients' medical status and home-care suitability. Our simulation lacks access to such discretionary information, which could further refine decision-making. For example, Kim et al. (2015) estimated that discretionary information improved ICU admission decisions more than a small capacity increase.

As an additional benchmark, we compare ISP to a myopic policy that evaluates hospitalization versus home care based on short-term mortality risk. Chan et al. (2012) established a 50% performance guarantee for a myopic discharge policy and demonstrated that it typically performs better in practice. The myopic approach mimics a physician's real-time decision-making by comparing immediate risks rather than long-term consequences. Specifically, for each patient observed for t days, we calculate the index $r_h(t+1)(1-p_h)/r_w(t+1)(1-p_w)$, which compares the next-period mortality risk at home versus at the hospital. An index greater than 1 means that risk at home care is greater than at the hospital. Under limited capacity, the same index determines which patient is prioritized for discharge, by identifying the patient with the lowest index.

Figure 7 presents the performance of this myopic policy ('Myopic', dotted lines). It uses very little capacity and achieves an 8.8% reduction in mortality rate compared to observed hospital data. Still, this is significantly lower than the 27.7% reduction achieved by ISP. In particular, with ample capacity, ISP reduces mortality by 20.5% compared to the baseline of the myopic policy.

REMARK 3. As explained in Remark 2, the policies suggested by Ouyang et al. (2020) can be interpreted using our J and τ_{opt} notations. We specifically examined the analogy of the Ratio policy (RP), which outperformed the Greedy policy in Ouyang et al. (2020). Unlike our setting, Ouyang et al. (2020) does not have an underlying uncapacitated optimization problem. To ensure a 'fair' comparison, we implemented RP assuming that patient LOS is first optimized according to Section 4. Our findings indicate that, unlike the myopic policy, RP achieves similar performance to ISP, when used as a decision index instead of ξ. This suggests that ISP's advantage stems primarily from its look-ahead quality and its integration of both the value function and the optimal LOS, rather than from its specific functional form.

6. Conclusions

Our paper developed methodologies to optimize LOS for hematology patients, balancing hospital-acquired infection risk (which incentivizes early discharge of high-risk patients) and home-acquired infection risk (which favors extended observation at the hospital to enable timely infection treatment), while considering their resulting mortality risks. Using newsvendor-type formulation, we explore how infection risk dynamics shape optimal observation policies for individual patients. We then extend this analysis to the social optimization problem where capacity constraints limit adherence to the unconstrained optimal solution. Our analysis covers a wide range of risk function dynamics that occur in practice, providing actionable insights for hospital observation policies.

A key aspect of our model is its incorporation of patient heterogeneity, reflecting the reality that HWs treat multiple patient types simultaneously. Our numerical analysis of the single-patient type suggests that restricting ourselves to the speedup-only policy has a minimal impact on patient risk, simplifying implementation in the multi-type patient setting. Even when this simplification is suboptimal, our multi-type capacitated model reveals that at most one patient type will need more than a single discharge threshold.

HWs are inherently small due to the need to keep patients isolated (which is expensive), and hence are typically overloaded. But, our model also applies for other medical units, where capacity may be less constrained. In such units, patient volumes may fluctuate over time and infection risk may depend on ward occupancy, influencing the hospitalization versus home-care decision. Such dependencies are natural to consider because infection risk may increase with the number of people a patient encounters during her stay. Further work could explore these dependencies.

Another promising direction is optimizing discharge timing when patient risk assessments evolve dynamically. We note that models for predicting infection risk of Hematology patients *dynamically over time* do not exist in the literature yet. Hence, following Carmen et al. (2019), we assume that the infection hazard rates are fixed at treatment completion, with no real-time updates during observation (expect for the elapsed time and whether or not the patient got an infection). For all practical purposes, our current models are the best that can be implemented with current technology. But we believe that new information that is received during observation may change the infection hazard rate of a specific patient and thus the optimal observation time may need to be updated dynamically. Our uncapacitated (single-patient) model can naturally extend to this setting and our ISP algorithm remains applicable with dynamic risk functions. However, further research is needed to determine if the index we proposed performs well in this case. This will be important to address as new models for predicting patient risks dynamically become available.

Our analysis assumes that patients are either observed in the HW or at home care. In practice, however, patients without an available HW bed may instead be hospitalized in the general ward rather than be sent home. While this alternative prevents immediate discharge, prior research (Carmen et al. 2019) indicates that dedicated HW care is superior to general-ward care, in terms of both infection and mortality risk. This

suggests that our analysis may be utilized to incorporate the option of general-ward hospitalization versus home care in a nested fashion.

Additionally, insurance reimbursement policies may penalize readmissions. For example, in the US, Medicare contracts with quality reviewers who assess whether hospital discharge planning was *adequate* and may deny payment for certain readmissions. Our uncapacitated model can provide justification for discharge recommendations in such settings. In the capacitated case, readmission costs are explicitly captured by the parameter c_I, which accounts for hospitalization costs following an infection. If reimbursement rules impose higher costs when infections occur at home rather than in the hospital, this distinction can be easily incorporated into our model.

To summarize, our paper explored the question of where a patient should be observed following cancer treatment, considering that both hospital care and home care have their pros and cons. Our framework allows one to find a solution that strikes the right balance between the two locations, effectively achieving the "best of both worlds." In particular, we show that, in many cases, smartly timing the transition between hospital and home care considering both clinical and capacity factors can significantly improve patient outcomes. Beyond healthcare, our framework may be applied to other service systems, where there are inherent trade-offs between professional service and self-service and a fine balance needs to be achieved.

References

Adusumilli KM, Hasenbein JJ (2010) Dynamics admission and service rate control of a queue. *Queueing Systems* 66:131–154.

AHRQ (2014) Medical expenditure panel survey, URL https://meps.ahrq.gov/data_stats/tables_compendia_hh_interactive.jsp, last accessed Feb, 18, 2019.

Allon G, Deo S, Lin W (2013) The impact of hospital size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research* 61(3):544–562.

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

Armony M, Chan CW, Zhu B (2018) Critical care capacity management: Understanding the role of a step down unit. *Production and Operations Management* 27(5):859–883.

Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic systems* 5(1):146–194.

Ata B, Shneorson S (2006) Dynamic control of an M/M/1 service system with adjustable arrival and service rates. *Management Science* 52(11):1778–1791.

Bartel AP, Chan CW, 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.

Bassamboo A, Randhawa RS (2016) Scheduling homogeneous impatient customers. *Management Science* 62(7):2129–2147.

Berry Jaeker JA, Tucker AL (2017) Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science* 63(4):1042–1062.

Bichescu B, Hilafu H (2023) Effects of hospital-acquired conditions on readmission risk: The mediating role of length of stay. *Manufacturing & Service Operations Management* 25(4):1603–1621.

Carmen R (2017) *Resource Efficiency Improvements in Hospitals*. Ph.D. thesis, KU Leuven.

Carmen R, Yom-Tov GB, van Nieuwenhuyse I, Foubert B, Ofran Y (2019) The role of specialized hospital units in infection and mortality risk reduction among patients with hematological cancers. *PLoS ONE* 14(3):e0211694.

Chan CW, Farias VF, Bambos N, J GE (2012) Optimizing ICU discharge decisions with patient readmissions. *Operations Research* 60(6):1323–1342.

Clarke DV, Newsam J, Olson DP, Adams D, Wolfe AJ, Fleisher LA (2021) Acute hospital care at home: The CMS waiver experience. *NEJM Catalyst Innovations in Care Delivery* 2(6):324–331.

Cornejo-Juárez P, Vilar-Compte D, García-Horton A, López-Velázquez M, Ñamendys-Silva S, Volkow-Fernández P (2016) Hospital-acquired infections at an oncological intensive care cancer unit: Differences between solid and hematological cancer patients. *BMC Infectious Diseases* 16(1):274.

Cornely OA, Gachot B, Akan H, Bassetti M, Uzun O, Kibbler C, et al (2015) Epidemiology and outcome of fungemia in a cancer cohort of the infectious diseases group (IDG) of the European Organization for Research and Treatment of Cancer (EORTC 65031). *Clinical Infection Diseases* 61(3):324–331.

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.

Hauck K, Zhao X (2011) How dangerous is a day in hospital? a model of adverse events and length of stay for medical inpatients. *Medical Care* 49(12):1068–1075.

Kang W, Ramanan K (2010) Fluid limits of many-server queues with reneging. *The Annals of Applied Probability* 20(6):2204–2260.

Kc D, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* 55(9):1486–1498.

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.

Kim SH, Chan CW, Olivares M, Escobar GJ (2015) ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science* 61(1):19–38.

Kuderer NM, Dale DC, Crawford J, Cosler LE, Lyman GH (2006) Mortality, morbidity, and cost associated with febrile neutropenia in adult cancer patients. *Cancer* 106(10):2258–2266.

Lee N, Kulkarni VG (2014) Optimal arrival rate and service rate control of multi-server queues. *Queueing Systems* 76:37–50.

Leschke J, Panepinto JA, Nimmer M, Hoffmann RG, Yan K, Brousseau DC (2012) Outpatient follow-up and rehospitalizations for sickle cell disease patients. *Pediatric Blood & Cancer* 58(3):406–409.

LLSC (2016) Facts and statistics:. Technical report, The Leukemia & Lymphoma Society of Canada, URL https://www.llscanada.org/disease-information/facts-and-statistics.

Magill SS, O'Leary E, Janelle SJ, Thompson DL, Dumyati G, Nadle J, Wilson LE, Kainer MA, Lynfield R, Greissman S, et al. (2018) Changes in prevalence of health care–associated infections in US hospitals. *New England Journal of Medicine* 379(18):1732–1744.

Mandelbaum A, Stolyar AL (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research* 52(6):836–855.

Marquez-Algaba E, Sanchez M, Baladas M, España C, Dallo HS, Requena M, Torrella A, Planas B, Raventos B, Molina C, Ribo M, Almirante B, Len O (2022) Covid-19 follow-app. mobile app-based monitoring of covid-19 patients after hospital discharge: A single-center, open-label, randomized clinical trial. *Journal of Personalized Medicine* 12(1).

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

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

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

Song H, Andreyeva E, David G (2022) Time is the wisest counselor of all: The value of provider-patient engagement length in home health care. *Management Science* 68(1):420–441.

Song H, Tucker AL, Graue R, Moravick S, Yang JJ (2020) Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science* 66(9):3825–3842.

Suda KJ, Motl SE, Kuth JC (2006) Inpatient oncology length of stay and hospital costs: implications for rising inpatient expenditures. *Journal of Applied Research* 6(2):126–132.

Taccone FS, Artigas AA, Sprung CL, Moreno R, Sakr Y, Vincent JL (2009) Characteristics and outcomes of cancer patients in European ICUs. *Critical Care* 13(1):R15.

Ulukus MY, Güllü R, Örmeci L (2011) Admission and termination control of a two-class loss system. *Stochastic Models* 27(1):2–25.

van Mieghem JA (1995) Dynamic scheduling with convex delay costs: The generalized cμ rule. *The Annals of Applied Probability* 5(3):809–833.

van Tiel FH, Harbers MM, Kessels AG, Schouten HC (2005) Home care versus hospital care of patients with hematological malignancies and chemotherapy-induced cytopenia. *Annals of Oncology* 16(2):195–205.

Whitt W (2006) Fluid models for multiserver queues with abandonments. *Operations Research* 54(1):37–54.

Whitt W (2013) OM forum: Offered load analysis for staffing. *Manufacturing & Service Operations Management* 15(2):166–169.

Yom-Tov GB, Chan CW (2021) Balancing admission control, speedup, and waiting in service systems. *Queueing systems* 97:163–219.

Zhang J (2013) Fluid models of many-server queues with abandonment. *Queueing Systems* 73:147–193.

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Appendix A: Proofs and Additional Details for Section 3

PROOF OF PROPOSITION 1: Assume that Assumption 1 holds and that, in addition, $0 < r_h < r_w$ and both are constant over time. In this case, we have that

$$J'_\tau = e^{-r_w\tau}\left((p_w-p_h)r_w-(c_w-c_h)+e^{-r_h(T-\tau)}(r_h-r_w)(1+c_I-p_h)-c_h\frac{r_h-r_w}{r_h}\left(1-e^{-r_h(T-\tau)}\right)\right).$$

Solving for $J'_\tau = 0$ we obtain

$$e^{-r_h(T-\tau)} = \frac{(p_w-p_h)r_w+c_h\frac{r_w}{r_h}-c_w}{(r_w-r_h)(1+c_I-p_h+c_h/r_h)},$$

So that the unique value of τ_0 that solves for $J'_\tau = 0$ satisfies

$$\tau_0 := T+\frac{1}{r_h}\ln\frac{(p_w-p_h)r_w+c_h\frac{r_w}{r_h}-c_w}{(r_w-r_h)(1+c_I-p_h+c_h/r_h)}.$$

Thus,

$$\tau_{opt} = \min\left\{T,\max\left\{0,T+\frac{1}{r_h}\ln\frac{(p_w-p_h)r_w+c_h\frac{r_w}{r_h}-c_w}{(r_w-r_h)(1+c_I-p_h+c_h/r_h)}\right\}\right\}.$$

To confirm that τ_{opt} is indeed a global maximum of the function J_τ we take the second derivative of J_τ with respect to τ and evaluate it at τ_0, as follows:

$$J''_\tau = -r_w J'_\tau + r_h e^{-r_w\tau}e^{-r_h(T-\tau)}(r_h-r_w)(1+c_I-p_h+c_h/r_h),$$

so that

$$J''_{\tau_0} = r_h e^{-r_w\tau_0}e^{-r_h(T-\tau_0)}(r_h-r_w)(1+c_I-p_h+c_h/r_h) < 0.$$

□

We next prove Proposition 2 by breaking it into smaller more specific lemmas each concentrating on the monotonicity with respect to a different parameter. But first we prove a structural lemma that will be used throughout.

LEMMA EC.1 **(Monotonicity of the survival function).** *Under Assumption 1, we have that:*

(a) The survival function $G^c_\tau(\cdot)$ is decreasing in τ.

(b) The expected time in the ward $\int_0^\tau G^c_\tau(u)du$, is increasing in τ.

(c) The expected time of being observed at home $\int_\tau^T G^c_\tau(u)du$ is decreasing in τ.

(d) The probability that a patient will develop infection at home, $G_\tau(T)-G_\tau(\tau)$, is decreasing in τ.

(e) The probability that a patient will develop infection at the hospital, $G_\tau(\tau)$, is increasing in τ.

PROOF: Let $\tau_1 < \tau_2$.

(a) We wish to show that $G^c_{\tau_1}(t) \geq G^c_{\tau_2}(t)$, for all t.

$$G^c_{\tau_1}(t) = \exp\left(-\int_0^{\tau_1\wedge t} r_w(u)du-\int_{\tau_1}^{\tau_1\vee t} r_h(u)du\right) \geq \exp\left(-\int_0^{\tau_2\wedge t} r_w(u)du-\int_{\tau_2}^{\tau_2\vee t} r_h(u)du\right) = G^c_{\tau_2}(t),$$

where the inequality follows from Assumption 1 and specifically from assuming that the risk of developing infection is higher at the hospital than at home.

(b) We wish to show that $\int_0^{\tau_1} G^c_{\tau_1}(t)dt \leq \int_0^{\tau_2} G^c_{\tau_2}(t)dt$.

$$\int_0^{\tau_1} G^c_{\tau_1}(t)dt = \int_0^{\tau_1}\exp\left(-\int_0^t r_w(u)du\right)dt \leq \int_0^{\tau_2}\exp\left(-\int_0^t r_w(u)du\right)dt = \int_0^{\tau_2} G^c_{\tau_2}(t)dt.$$

(c) We wish to show that $\int_{\tau_1}^{T} G^c_{\tau_1}(t)dt \geq \int_{\tau_2}^{T} G^c_{\tau_2}(t)dt$.

$$\int_{\tau_2}^{T} G^c_{\tau_2}(t)dt \leq \int_{\tau_2}^{T} G^c_{\tau_1}(t)dt \leq \int_{\tau_1}^{T} G^c_{\tau_1}(t)dt,$$

where the first inequality follows from item (a).

(d) We first observe that $G_\tau(T) - G_\tau(\tau) = G^c_\tau(\tau) - G^c_\tau(T)$. By definition,

$$G^c_\tau(\tau) - G^c_\tau(T) = \exp\left(-\int_0^\tau r_w(u)du\right)\left(1 - \exp\left(-\int_\tau^T r_h(u)du\right)\right),$$

which is decreasing in τ as a product of two function who are both decreasing in τ.

(e) The result follows by observing that $G_\tau(\tau) = (G_\tau(\tau) - G_\tau(T)) + G_\tau(T)$ which is a sum of two functions that are increasing in τ by items (d) and (a).

□

LEMMA EC.2 (**Monotonicity in** c_w). *Under Assumption 1, the effective threshold τ_{opt} is monotone decreasing in c_w. That is, as the hospitalization cost, c_w, increases the patient will be discharged home earlier.*

PROOF: Recall the expression for the expected reward function in (2)

$$\begin{aligned} J_\tau(c_w) &= p_w G_\tau(\tau) + p_h\left(G_\tau(T) - G_\tau(\tau)\right) - c_w\int_0^\tau G^c_\tau(u)du - c_h\int_\tau^T G^c_\tau(u)du + (1+c_I)G^c_\tau(T) \\ &= f(\tau) - c_w\int_0^\tau G^c_\tau(u)du, \end{aligned}$$

where we explicitly write the dependence of J on the parameter c_w, and $f(\cdot)$ is a generic function which is a function of τ but not of c_w. Let $c_w^1 < c_w^2$, and let τ^1_{opt} and τ^2_{opt} be the optimal thresholds associated with c_w^1 and c_w^2, respectively. We wish to show that $\tau^1_{opt} \geq \tau^2_{opt}$.

Let $\tau > \tau^1_{opt}$, then it is sufficient to show that $J_\tau(c_w^2) \leq J_{\tau^1_{opt}}(c_w^2)$, because this would imply that τ cannot be the optimal threshold with respect to c_w^2. The latter inequality indeed holds since

$$\begin{aligned} J_{\tau^1_{opt}}(c_w^2) - J_\tau(c_w^2) &= f(\tau^1_{opt}) - f(\tau) + c_w^2\left(\int_0^\tau G^c_\tau(u)du - \int_0^{\tau^1_{opt}} G^c_{\tau^1_{opt}}(u)du\right) \\ &\geq f(\tau^1_{opt}) - f(\tau) + c_w^1\left(\int_0^\tau G^c_\tau(u)du - \int_0^{\tau^1_{opt}} G^c_{\tau^1_{opt}}(u)du\right) \\ &= J_{\tau^1_{opt}}(c_w^1) - J_\tau(c_w^1) \geq 0, \end{aligned}$$

where the first inequality follows from Lemma EC.1 (b) and the fact that $c_w^1 \leq c_w^2$, and the second inequality follows from the optimality of the threshold τ^1_{opt}. □

LEMMA EC.3 (**Monotonicity in** c_h). *Under Assumption 1, the effective threshold τ_{opt} is monotone increasing in c_h. That is, as home-care cost, c_h, increases the patient will be discharged home later.*

PROOF: Recall the expression for the expected reward function in (2)

$$\begin{aligned} J_\tau(c_h) &= p_w G_\tau(\tau) + p_h\left(G_\tau(T) - G_\tau(\tau)\right) - c_w\int_0^\tau G^c_\tau(u)du - c_h\int_\tau^T G^c_\tau(u)du + (1+c_I)G^c_\tau(T) \\ &= f(\tau) - c_h\int_\tau^T G^c_\tau(u)du, \end{aligned}$$

where we explicitly write the dependence of J on the parameter c_h, and $f(\cdot)$ is a function of τ but not of c_h. Let $c_h^1 < c_h^2$, and let τ_{opt}^1 and τ_{opt}^2 be the optimal thresholds associated with c_h^1 and c_h^2, respectively. We wish to show that $\tau_{opt}^1 \leq \tau_{opt}^2$.

Let $\tau > \tau_{opt}^2$, then it is sufficient to show that $J_\tau(c_h^1) \leq J_{\tau_{opt}^2}(c_h^1)$, because this would imply that τ cannot be the optimal threshold with respect to c_h^1. The latter indeed holds since

$$\begin{aligned} J_{\tau_{opt}^2}(c_h^1) - J_\tau(c_h^1) &= f(\tau_{opt}^2) - f(\tau) + c_h^1 \left(\int_\tau^T G_\tau^c(u)du - \int_{\tau_{opt}^2}^T G_{\tau_{opt}^2}^c(u)du \right) \\ &\geq f(\tau_{opt}^2) - f(\tau) + c_h^2 \left(\int_\tau^T G_\tau^c(u)du - \int_{\tau_{opt}^2}^T G_{\tau_{opt}^2}^c(u)du \right) \\ &= J_{\tau_{opt}^2}(c_h^2) - J_\tau(c_h^2) \geq 0, \end{aligned}$$

where the first inequality follows from Lemma EC.1 (c) and the fact that $c_h^1 \leq c_h^2$, and the second inequality follows from the optimality of the threshold τ_{opt}^2. □

LEMMA EC.4 (**Monotonicity in** c_I). *Under Assumption 1, the effective threshold τ_{opt} is monotone decreasing in c_I. As the cost of hospitalization after infection, c_I, increases, the patient will be sent home earlier so as to reduce the probability of developing an infection.*

PROOF: Once more, recall the expression for the expected reward function in (2)

$$\begin{aligned} J_\tau(c_I) &= p_w G_\tau(\tau) + p_h \left(G_\tau(T) - G_\tau(\tau) \right) - c_w \int_0^\tau G_\tau^c(u)du - c_h \int_\tau^T G_\tau^c(u)du + (1+c_I) G_\tau^c(T) \\ &= f(\tau) + (1+c_I) G_\tau^c(T), \end{aligned}$$

where we explicitly write the dependence of J on the parameter c_I, and $f(\cdot)$ is a function of τ but not of c_I. Let $c_I^1 < c_I^2$, and let τ_{opt}^1 and τ_{opt}^2 be the optimal thresholds associated with c_I^1 and c_I^2, respectively. We wish to show that $\tau_{opt}^1 \geq \tau_{opt}^2$.

Let $\tau > \tau_{opt}^1$, then it is sufficient to show that $J_\tau(c_I^2) \leq J_{\tau_{opt}^1}(c_I^2)$ because this would imply that τ cannot be the optimal threshold with respect to c_I^2. The latter indeed holds since

$$\begin{aligned} J_{\tau_{opt}^1}(c_I^2) - J_\tau(c_I^2) &= f(\tau_{opt}^1) - f(\tau) + (1+c_I^2) \left(G_{\tau_{opt}^1}^c(T) - G_\tau^c(T) \right) \\ &\geq f(\tau_{opt}^1) - f(\tau) + (1+c_I^1) \left(G_{\tau_{opt}^1}^c(T) - G_\tau^c(T) \right) \\ &= J_{\tau_{opt}^1}(c_I^1) - J_\tau(c_I^1) \geq 0, \end{aligned}$$

where the first inequality follows from Lemma EC.1 (a) and the fact that $c_I^1 \leq c_I^2$, and the second inequality follows from the optimality of the threshold τ_{opt}^1. □

LEMMA EC.5 (**Monotonicity in** p_h). *Under Assumption 1, the effective threshold is monotone decreasing in p_h. As the survival probability in case of infection at home, p_h, increases, the threshold will be lower; hence, the patient will be sent home earlier.*

PROOF: Once more, recall the expression for the expected reward function in (2)

$$\begin{aligned} J_\tau(p_h) &= p_w G_\tau(\tau) + p_h \left(G_\tau(T) - G_\tau(\tau) \right) - c_w \int_0^\tau G_\tau^c(u)du - c_h \int_\tau^T G_\tau^c(u)du + (1+c_I) G_\tau^c(T) \\ &= f(\tau) + p_h (G_\tau(T) - G_\tau(\tau)), \end{aligned}$$

where we explicitly write the dependence of J on the parameter p_h, and $f(\cdot)$ is a function of τ but not of p_h. Let $p_h^1 < p_h^2$, and let τ_{opt}^1 and τ_{opt}^2 be the optimal thresholds associated with p_h^1 and p_h^2, respectively. We wish to show that $\tau_{opt}^1 \geq \tau_{opt}^2$.

Let $\tau > \tau_{opt}^1$, then it is sufficient to show that $J_\tau(p_h^2) \leq J_{\tau_{opt}^1}(p_h^2)$, because this would imply that τ cannot be the optimal threshold with respect to p_h^2. The latter indeed holds since

$$\begin{aligned} J_{\tau_{opt}^1}(p_h^2) - J_\tau(p_h^2) &= f(\tau_{opt}^1) - f(\tau) + p_h^2 \left(G_{\tau_{opt}^1}(T) - G_{\tau_{opt}^1}(\tau_{opt}^1) - (G_\tau(T) - G_\tau(\tau)) \right) \\ &\geq f(\tau_{opt}^1) - f(\tau) + p_h^1 \left(G_{\tau_{opt}^1}(T) - G_{\tau_{opt}^1}(\tau_{opt}^1) - (G_\tau(T) - G_\tau(\tau)) \right) \\ &= J_{\tau_{opt}^1}(p_h^1) - J_\tau(p_h^1) \geq 0, \end{aligned}$$

where the first inequality follow from EC.1 (d) and the second inequality follows from the optimality of τ_{opt}^1 with respect to p_h^1. □

LEMMA EC.6 (**Monotonicity in** p_w). *Under Assumption 1, the effective threshold is monotone increasing in p_w. As the survival probability in case of infection at the hospital, p_w, increases, the threshold will be higher; hence, the patient will be sent home later.*

PROOF: Once more, recall the expression for the expected reward function in (2)

$$\begin{aligned} J_\tau(p_w) &= p_w G_\tau(\tau) + p_h \left(G_\tau(T) - G_\tau(\tau) \right) - c_w \int_0^\tau G_\tau^c(u)du - c_h \int_\tau^T G_\tau^c(u)du + (1+c_I) G_\tau^c(T) \\ &= f(\tau) + p_w (G_\tau(\tau)), \end{aligned}$$

where we explicitly write the dependence of J on the parameter p_w, and $f(\cdot)$ is a function of τ but not of p_w. Let $p_w^1 < p_w^2$, and let τ_{opt}^1 and τ_{opt}^2 be the optimal thresholds associated with p_w^1 and p_w^2, respectively. We wish to show that $\tau_{opt}^1 \leq \tau_{opt}^2$.

Let $\tau > \tau_{opt}^2$, then it is sufficient to show that $J_\tau(p_w^1) \leq J_{\tau_{opt}^2}(p_w^1)$, because this would imply that τ cannot be the optimal threshold with respect to p_w^1. The latter indeed holds since

$$\begin{aligned} J_{\tau_{opt}^2}(p_w^1) - J_\tau(p_w^1) &= f(\tau_{opt}^2) - f(\tau) + p_w^1 \left(G_{\tau_{opt}^2}(\tau_{opt}^2) - G_\tau(\tau) \right) \geq f(\tau_{opt}^2) - f(\tau) + p_w^2 \left(G_{\tau_{opt}^2}(\tau_{opt}^2) - G_\tau(\tau) \right) \\ &= J_{\tau_{opt}^2}(p_w^2) - J_\tau(p_w^2) \geq 0, \end{aligned}$$

where the first inequality follows from Lemma EC.1 (e) and the fact that $p_w^1 \leq p_w^2$, and the second inequality follows from the optimality of the threshold τ_{opt}^2. □

LEMMA EC.7 (**Monotonicity in** r_h). *Under Assumption 1 and assuming that $c_h = 0$, the effective threshold is monotone increasing in r_h. As the risk of getting infection at home care, r_h, increases, the threshold will be higher; hence, the patient will be sent home later.*

PROOF: Once more, recall the expression for the expected reward function in (2):

$$\begin{aligned} J_\tau(r_h) &= p_w G_\tau(\tau) + p_h \left(G_\tau(T) - G_\tau(\tau) \right) - c_w \int_0^\tau G_\tau^c(u)du - c_h \int_\tau^T G_\tau^c(u)du + (1+c_I) G_\tau^c(T) \\ &\overset{c_h=0}{=} p_w G_\tau(\tau) + p_h (1 - G_\tau^c(T)) - p_h G_\tau(\tau) - c_w \int_0^\tau G_\tau^c(u)du + (1+c_I) G_\tau^c(T) \\ &= f(\tau) + (1 + c_I - p_h) G_\tau^c(T) \\ &= f(\tau) + (1 + c_I - p_h) e^{-\left(\int_0^\tau r_w(x)dx + \int_\tau^T r_h(x)dx \right)}, \end{aligned}$$

where we explicitly write the dependence of J on the parameter r_h, and $f(\cdot)$ is a function of τ but not of r_h. Let $r_h^1 \le r_h^2$, and let τ_{opt}^1 and τ_{opt}^2 be the optimal thresholds associated with r_h^1 and r_h^2, respectively. We wish to show that $\tau_{opt}^1 \le \tau_{opt}^2$.

Let $\tau > \tau_{opt}^2$, then it is sufficient to show that $J_\tau(r_h^1) \le J_{\tau_{opt}^2}(r_h^1)$, because this would imply that τ cannot be the optimal threshold with respect to r_h^1. Assuming $c_h = 0$, the latter indeed holds since

$$
\begin{aligned}
&J_{\tau_{opt}^2}(r_h^1) - J_\tau(r_h^1) \\
&= f(\tau_{opt}^2) + (1 + c_I - p_h) e^{-\int_0^{\tau_{opt}^2} r_w(x)dx} e^{-\int_{\tau_{opt}^2}^{T} r_h^1(x)dx} - f(\tau) - (1 + c_I - p_h) e^{-\int_0^{\tau} r_w(x)dx} e^{-\int_{\tau}^{T} r_h^1(x)dx} \\
&= f(\tau_{opt}^2) - f(\tau) + (1 + c_I - p_h) e^{-\int_0^{\tau_{opt}^2} r_w(x)dx} \left(e^{-\int_{\tau_{opt}^2}^{T} r_h^1(x)dx} - e^{-\int_{\tau_{opt}^2}^{\tau} r_w(x)dx} e^{-\int_{\tau}^{T} r_h^1(x)dx} \right) \\
&= f(\tau_{opt}^2) - f(\tau) + (1 + c_I - p_h) e^{-\int_0^{\tau_{opt}^2} r_w(x)dx} e^{-\int_{\tau}^{T} r_h^1(x)dx} \left(e^{-\int_{\tau_{opt}^2}^{\tau} r_h^1(x)dx} - e^{-\int_{\tau_{opt}^2}^{\tau} r_w(x)dx} \right) \\
&\ge f(\tau_{opt}^2) - f(\tau) + (1 + c_I - p_h) e^{-\int_0^{\tau_{opt}^2} r_w(x)dx} e^{-\int_{\tau}^{T} r_h^2(x)dx} \left(e^{-\int_{\tau_{opt}^2}^{\tau} r_h^2(x)dx} - e^{-\int_{\tau_{opt}^2}^{\tau} r_w(x)dx} \right) \\
&= J_{\tau_{opt}^2}(r_h^2) - J_\tau(r_h^2) \ge 0,
\end{aligned}
$$

where the first inequality follows from Assumption 1 and our assumption that $r_h^1 < r_h^2$ by which $e^{-r_h^1(x)} > e^{-r_h^2(x)} \ge e^{-r_w(x)}$. The second inequality follows from the optimality of the threshold τ_{opt}^2. □

□

We next provide an example that shows that the optimal threshold is *not* necessarily monotone in r_w. This is illustrated by the discrete-time counter-example shown in Table EC.1. This table compares two scenarios, 1 and 2: in Scenario 1 the patient has a lower risk of infection at the hospital, r_w, compared to Scenario 2. At the same time, it is optimal for the patient to be sent home sooner under Scenario 1. The example demonstrates that if the survival probability from an infection at the hospital, p_w, is large enough compared to the utility of sending a patient home, then as the infection risk at the hospital, r_w, increases, the overall survival rate at the hospital increases and, therefore, keeping the patient longer at the hospital could be advantageous.

Table EC.1 An Example of Non-Monotonicity in the Hospital Infection Probability (r_w)

	Scenario 1		Scenario 2	
Day (t)	$r_w(t)$	Decision	$r_w(t)$	Decision
1	0.38	home	0.48	ward
2	0.37	home	0.47	ward
3	0.36	home	0.46	home
4	0.35	home	0.45	home
5	0.34	home	0.44	home
6		home		home

Note. $T = 6, c_w = 0.2, c_h = 0, c_I = 0.1, p_h = 0.1, p_w = 0.7, r_h = 0.28$.

Appendix B: Proofs for Section 4

PROOF OF LEMMA 2: Suppose that the system is overloaded under the threshold τ_{opt} and consider an optimal solution $(K, \vec{\lambda}, \vec{\tau})$ to (7). Suppose that $\sum_{k=1}^{K+1} \bar{\lambda}_k m_{\tau_k} < 1$. Then, by the overload assumption, there is at least one class k_0 such that $\tau_{k_0} < \tau_{opt}$. Recall that $\tau_{K+1} := \tau_{opt}$. Clearly, $m_{k_0} < m_{K+1}$, and, by the optimality of τ_{opt}, we have that $J_{\tau_{opt}} \ge J_{\tau_{k_0}}$. We now construct a modified solution $(K, \tilde{\lambda}, \tilde{\tau})$ as follows:

- $\tilde{\tau}_k = \tau_k$, for all k.
- $\tilde{\lambda}_k = \bar{\lambda}_k$, for all $k \neq k_0$ and $k \neq K+1$.
- For $0 < \epsilon \leq \bar{\lambda}_{k_0}$, set $\tilde{\lambda}_{k_0} = \bar{\lambda}_{k_0} - \epsilon$ and $\tilde{\lambda}_{K+1} = \bar{\lambda}_{K+1} + \epsilon$.

Then as long as ϵ is small enough we have that for the new solution the two constraints of (7) are still satisfied and the objective function is not smaller than in the original solution. Repeat this process until $\sum_{k=1}^{K+1} \bar{\lambda}_k m_{\tau_k} = 1$. □

PROOF OF PROPOSITION 4: The proof follows a straightforward approach relying on first-order necessary conditions for optimality. First note that the Lagrangian corresponding to the optimal solution to problem (9) (which is equivalent to (11)) is

$$L(\tau_l, \tau_h, \bar{\lambda}_l) = -\bar{\lambda}_l J_{\tau_l} - (\bar{\lambda} - \bar{\lambda}_l) J_{\tau_h} + \alpha \left(\bar{\lambda}_l m_{\tau_l} + (\bar{\lambda} - \bar{\lambda}_l) m_{\tau_h} - 1 \right) - \beta \tau_l \\ + \delta(\tau_h - \tau_{opt}) + \epsilon(\tau_l - \tau_h) - \gamma_l \bar{\lambda}_l + \gamma_u (\bar{\lambda}_l - \bar{\lambda}). \quad \text{(EC.1)}$$

By the KKT conditions, the following must hold at a (local) maximum.

- **Stationarity:**

$$\frac{\partial L}{\partial \tau_l} = -\bar{\lambda}_l J'_{\tau_l} + \alpha \bar{\lambda}_l (1 - G_T(\tau_l)) - \beta + \epsilon = 0, \quad \text{(EC.2)}$$

$$\frac{\partial L}{\partial \tau_h} = -(\bar{\lambda} - \bar{\lambda}_l) J'_{\tau_h} + \alpha (\bar{\lambda} - \bar{\lambda}_l)(1 - G_T(\tau_h)) + \delta - \epsilon = 0, \quad \text{(EC.3)}$$

$$\frac{\partial L}{\partial \bar{\lambda}_l} = -J_{\tau_l} + J_{\tau_h} + \alpha \left(m_{\tau_l} - m_{\tau_h} \right) - \gamma_l + \gamma_u = 0, \quad \text{(EC.4)}$$

- **Primal Feasibility:**

$$\bar{\lambda}_l m_{\tau_l} + (\bar{\lambda} - \bar{\lambda}_l) m_{\tau_h} = 1,$$

$$0 \leq \tau_l \leq \tau_h \leq \tau_{opt},$$

$$0 \leq \bar{\lambda}_l \leq \bar{\lambda}.$$

- **Dual Feasibility:**

$$\beta, \gamma_l, \gamma_h, \delta, \epsilon \geq 0.$$

- **Complimentary Slackness:**

$$-\beta \tau_l + \delta(\tau_h - \tau_{opt}) + \epsilon(\tau_l - \tau_h) - \gamma_l \bar{\lambda}_l + \gamma_u (\bar{\lambda}_l - \bar{\lambda}) = 0.$$

Focusing on solutions where $0 < \bar{\lambda}_l < \bar{\lambda}$ and where $\tau_l < \tau_h$, we can immediately conclude by complimentary slackness that $\epsilon = \gamma_l = \gamma_u = 0$. Now considering the four cases described in the statement of the lemma we have that

(a) If $0 < \tau_l < \tau_h < \tau_{opt}$, then by complimentary slackness we have that $\beta = \delta = 0$. Thus, by (EC.2), we get

$$\alpha = \frac{J'_{\tau_l}}{1 - G_T(\tau_l)}.$$

Further, by (EC.3), we get

$$\alpha = \frac{J'_{\tau_h}}{1 - G_T(\tau_h)}.$$

Finally, by (EC.4), we get

$$\alpha = \frac{J_{\tau_h} - J_{\tau_l}}{m_{\tau_h} - m_{\tau_l}}.$$

Thus, item (a) in the lemma follows.

(b) Similarly, if $\tau_l = 0$ and $\tau_h < \tau_{opt}$, we have that, by complimentary slackness, $\delta = 0$. Thus, by (EC.2), we get

$$\alpha \geq \frac{J_0'}{1 - G_T(0)} = J_0'.$$

Further, by (EC.3), we get

$$\alpha = \frac{J_{\tau_h}'}{1 - G_T(\tau_h)}.$$

Finally, by (EC.4), we get

$$\alpha = \frac{J_{\tau_h} - J_0}{m_{\tau_h}}.$$

Thus, item (b) in the lemma follows.

(c) Analogously, if $\tau_l > 0$ and $\tau_h = \tau_{opt}$, we have that, by complimentary slackness, $\beta = 0$. Thus, by (EC.2), we get

$$\alpha = \frac{J_{\tau_l}'}{1 - G_T(\tau_l)}.$$

Further, by (EC.3), we get

$$\alpha \leq \frac{J_{\tau_{opt}}'}{1 - G_T(\tau_{opt})}.$$

Finally, by (EC.4), we get

$$\alpha = \frac{J_{\tau_{opt}} - J_{\tau_l}}{m_{\tau_{opt}} - m_{\tau_l}}.$$

Thus, item (c) in the lemma follows.

(d) Finally, if $\tau_l = 0$ and $\tau_h = \tau_{opt}$, by (EC.2), we get

$$\alpha \geq \frac{J_0'}{1 - G_T(0)}.$$

Further, by (EC.3), we get

$$\alpha \leq \frac{J_{\tau_{opt}}'}{1 - G_T(\tau_{opt})}.$$

Finally, by (EC.4), we get

$$\alpha = \frac{J_{\tau_{opt}} - J_0}{m_{\tau_{opt}}}.$$

Thus, item (d) in the lemma follows.

□

PROOF OF COROLLARY 2: Recall the definition of τ_{opt} as the minimal threshold that maximizes J_τ. Suppose that $0 < \tau_{opt} < T$. In that case, necessarily, $J_{\tau_{opt}}' = 0$, and $J_{\tau_{opt}} > J_\tau$, for all $\tau < \tau_{opt}$. In particular, $\xi(\tau_{opt}) = 0$ and $(J_{\tau_{opt}} - J_\tau)/(m_{\tau_{opt}} - m_\tau) > 0$ for all $\tau < \tau_{opt}$. Thus inequalities (14) and (15) cannot be satisfied, and hence the policies Sp-FS and Bl-FS cannot be optimal. □

PROOF OF PROPOSITION 6: Consider the family of policies with up to two thresholds per customer type $\{\tau^i, i = 1, ..., I\}$ with $0 \leq \tau_l^i \leq \tau_h^i \leq \tau_{opt}^i$ as in the problem formulation (18). The proof follows a straightforward approach relying on first-order necessary conditions for optimality. First note that the Lagrangian corresponding to the optimal solution to problem (18) is

$$L(\vec{\tau}_l, \vec{\tau}_h, \vec{\bar{\lambda}}_l) = -\sum_{i=1}^{I} \left(\bar{\lambda}_l^i J_{\tau_l^i}^i + (\bar{\lambda}^i - \bar{\lambda}_l^i) J_{\tau_h^i}^i \right) + \alpha \left(\sum_{i=1}^{I} \left(\bar{\lambda}_l^i m_{\tau_l^i}^i + (\bar{\lambda}^i - \bar{\lambda}_l^i) m_{\tau_h^i}^i \right) - 1 \right)$$
$$+ \sum_{i=1}^{I} \left(-\beta^i \tau_l^i + \delta^i (\tau_h^i - \tau_{opt}^i) + \epsilon^i (\tau_l^i - \tau_h^i) - \gamma_l^i \bar{\lambda}_l^i + \gamma_u^i (\bar{\lambda}_l^i - \bar{\lambda}^i) \right).$$

By the KKT conditions, the following must hold at a (local) maximum.

- **Stationarity:**

$$\frac{\partial L}{\partial \tau_l^i} = -\bar{\lambda}_l^i J_{\tau_l^i}^{i'} + \alpha\bar{\lambda}_l^i(1-G_T^i(\tau_l^i)) - \beta^i + \epsilon^i = 0, \quad i=1,...,I,$$
$$\frac{\partial L}{\partial \tau_h^i} = -(\bar{\lambda}-\bar{\lambda}_l^i)J_{\tau_h^i}^{i'} + \alpha(\bar{\lambda}-\bar{\lambda}_l^i)(1-G_T^i(\tau_h^i)) + \delta^i - \epsilon^i = 0, \quad i=1,...,I,$$
$$\frac{\partial L}{\partial \bar{\lambda}_l^i} = -J_{\tau_l^i}^i + J_{\tau_h^i}^i + \alpha\left(m_{\tau_l^i}^i - m_{\tau_h^i}^i\right) - \gamma_l^i + \gamma_u^i = 0, \quad i=1,...,I.$$

- **Primal Feasibility:**

$$\sum_{i=1}^{I}\left(\bar{\lambda}_l^i m_{\tau_l^i}^i + (\bar{\lambda}^i-\bar{\lambda}_l^i)m_{\tau_h^i}^i\right) = 1,$$
$$0 \le \tau_l^i \le \tau_h^i \le \tau_{opt}^i, \quad i=1,...,I,$$
$$0 \le \bar{\lambda}_l^i \le \bar{\lambda}^i, \quad i=1,...,I.$$

- **Dual Feasibility:**

$$\beta^i, \gamma_l^i, \gamma_h^i, \delta^i, \epsilon^i \ge 0, \quad i=1,...,I.$$

- **Complimentary Slackness:**

$$\sum_{i=1}^{I}\left(-\beta^i\tau_l^i + \delta^i(\tau_h^i - \tau_{opt}^i) + \epsilon^i(\tau_l^i - \tau_h^i) - \gamma_l^i\bar{\lambda}_l^i + \gamma_u^i(\bar{\lambda}_l^i - \bar{\lambda}^i)\right) = 0.$$

We first establish the following convention: Without loss of generality, if $\bar{\lambda}_l^i = 0$ for some i, then $\tau_l^i = 0$, and if $\bar{\lambda}_l^i = \bar{\lambda}^i$ for some i, then $\tau_h^i = \tau_{opt}^i$. Also, if for some i we have that $0 < \bar{\lambda}_l^i < \bar{\lambda}^i$, and $\tau_l^i = \tau_h^i$, then we can equivalently consider a solution where $\bar{\lambda}_l^i = \bar{\lambda}^i$ and $\tau_h^i = \tau_{opt}^i$. This convention guarantees that for all i, $\tau_l^i < \tau_h^i$ unless $\tau_l^i = \tau_h^i = \tau_{opt}^i$.

Consider an optimal solution to (18) with i such that $\tau_l^i < \tau_h^i$, then by complimentary slackness we have that $\epsilon^i = 0$. Thus, if $\tau_l^i > 0$, $\alpha = J_{\tau_l^i}^{i'}/(1-G_T^i(\tau_l^i))$. Similarly, if $\tau_h^i < \tau_{opt}^i$, $\alpha = J_{\tau_h^i}^{i'}/(1-G_T^i(\tau_h^i))$. Additionally, if $\tau_l^i = 0$, then $\alpha \ge J_{\tau_l^i}^{i'}/(1-G_T^i(\tau_l^i))$, and if $\tau_h^i = \tau_{opt}^i$, then $\alpha \le J_{\tau_h^i}^{i'}/(1-G_T^i(\tau_h^i))$. This completes the proof of the proposition. □

PROOF OF COROLLARY 3: Recall the definition of τ_{opt}^i as the minimal threshold that maximizes J_τ^i. Suppose that, for some customer type i, $0 < \tau_{opt}^i < T$. In that case, necessarily, $J_{\tau_{opt}^i}^{i'} = 0$, and thus $\xi^i(\tau_{opt}^i) = 0$. If, by contradiction, τ_{opt}^i is an optimal threshold for some patients of type i, then by (21), we have that necessarily,

$$0 = \xi^i(\tau_{opt}^i) \ge \xi^j(\tau^j) \ge 0,$$

where j is a customer type whose one of its optimal thresholds is τ^j. Thus, τ^j must be equal to τ_{opt}^j for all j. But from the overloaded assumption, we know that such a solution is not feasible. Thus, we reach a contradiction. □

Appendix C: Sufficient Conditions for Capacitated Case

While Proposition 4 outlines necessary conditions for the optimality of the 2×Sp, Bl-Sp, Sp-FS, and Bl-FS policies, it does not cover the "boundary" policy of 1×Sp. We are especially interested in characterizing conditions under which this *simplest* and most *equitable* policy that uses the same speedup threshold for all patients (1×Sp) is optimal. In the next lemma we will provide sufficient conditions for the optimality of this boundary policy, as well as that of Bl-Sp and Bl-FS, under the assumption that the function J is increasing up to the optimal threshold τ_{opt}. In our empirically based numerical analysis in Section 5, we observe that virtually all the cases we encounter in our data indeed satisfy this assumption. However, they do not satisfy the assumption that $\xi(\cdot)$ is an increasing function. Thus we suspect that the practical value of Lemma EC.8 is limited.

LEMMA EC.8 **(Sufficient conditions for optimality)**. *Assume that the function J_τ is differentiable as a function of τ and that $J'_\tau > 0$ for all $0 \le \tau \le \tau_{opt}$. Define $\xi(\tau) := \frac{J'_\tau}{1-G_T(\tau)}$. Then $\xi(\tau) > 0$. In addition, assume that $\xi(\tau)$ is strictly increasing. Then,*

(a) The policies 2×Sp, Bl-Sp, and Sp-FS are not optimal.

(b) If $\frac{J_\tau - J_0}{m_\tau} < \xi(\tau)$ for all $\tau_{spd} \le \tau \le \tau_{opt}$ then the Bl-FS policy is optimal.

(c) If $\frac{J_\tau - J_0}{m_\tau} > \xi(\tau)$ for all $\tau_{spd} \le \tau \le \tau_{opt}$ then the policy 1×Sp is optimal.

PROOF OF LEMMA EC.8:

(a) Suppose that $\xi(\tau)$ is strictly increasing; then, if $\tau_l < \tau_h$ we have that

$$\frac{J'_{\tau_l}}{1-G_T(\tau_l)} \equiv \xi(\tau_l) < \xi(\tau_h) \equiv \frac{J'_{\tau_h}}{1-G_T(\tau_h)}$$

which is in contradiction to we Equation (12). Thus, the policy 2×Sp cannot be optimal.

Now consider (13). If $\xi(\tau)$ is strictly increasing then

$$\frac{J_{\tau_h} - J_0}{m_{\tau_h}} = \frac{J_{\tau_h} - J_0}{m_{\tau_h} - m_0} = \frac{\int_0^{\tau_h} J'_\tau d\tau}{\int_0^{\tau_h} m'_\tau d\tau} = \frac{\int_0^{\tau_h} \xi(\tau) m'_\tau d\tau}{\int_0^{\tau_h} m'_\tau d\tau} < \frac{\int_0^{\tau_h} \xi(\tau_h) m'_\tau d\tau}{\int_0^{\tau_h} m'_\tau d\tau} = \xi(\tau_h)$$

which is a contradiction to (13). Thus, the policy Bl-Sp cannot be optimal.

Finally consider (14). If $\xi(\tau)$ is strictly increasing then

$$\frac{J_{\tau_{opt}} - J_{\tau_l}}{m_{\tau_{opt}} - m_{\tau_l}} = \frac{\int_{\tau_l}^{\tau_{opt}} J'_\tau d\tau}{\int_{\tau_l}^{\tau_{opt}} m'_\tau d\tau} = \frac{\int_{\tau_l}^{\tau_{opt}} \xi(\tau) m'_\tau d\tau}{\int_{\tau_l}^{\tau_{opt}} m'_\tau d\tau} > \frac{\int_{\tau_l}^{\tau_{opt}} \xi(\tau_l) m'_\tau d\tau}{\int_{\tau_l}^{\tau_{opt}} m'_\tau d\tau} = \xi(\tau_l).$$

Explanations: The equalities hold because the expressions are negative so when we increase the numerator the expressions decrease. The resulting inequality is in contradiction to (14), which implies that the policy Sp-FS cannot be optimal.

(b) We start by observing that by Corollary 1 an optimal solution to (11) exists and it belongs to one of the five cases outlined in the corollary. In the case of the current lemma, since the policies 2×Sp, Bl-Sp, and Sp-FS have been eliminated in part (a), the only viable options are policies of type 1×Sp and Bl-FS.

Suppose that $\frac{J_\tau - J_0}{m_\tau} < \xi(\tau)$ for all $\tau_{spd} \le \tau \le \tau_{opt}$ and consider τ such that $\tau_{spd} < \tau < \tau_{opt}$. Then, we have that $\frac{J_\tau - J_0}{\mu_\tau} < \xi(\tau)$ if and only if

$$m'_\tau (J_\tau - J_0) - J'_\tau \mu_\tau < 0. \tag{EC.5}$$

Now note that for policies of type Bl-Sp, 1×Sp, and Bl-FS, the objective function of (11) at a speedup threshold of τ is equal to

$$V_{cap}(\tau) := \lambda_l(\tau) J_0 + (\bar{\lambda} - \lambda_l(\tau)) J_\tau,$$

where $\lambda_l(\tau) = \bar{\lambda} - \mu_\tau$ and, in particular, $\lambda_l = 0$ in 1×Sp (because then $\bar{\lambda} = \mu_\tau$). Then, the derivative of $V_{cap}(\tau)$ is

$$V'_{cap}(\tau) = \lambda'_l(\tau) J_0 + \bar{\lambda} J'_\tau - \lambda'_l(\tau) J_\tau - \lambda_l(\tau) J'_\tau = (\bar{\lambda} - \lambda_l(\tau)) J'_\tau + \lambda'_l(\tau)(J_0 - J_\tau) =$$

$$\mu_\tau J'_\tau + \lambda'_l(\tau)(J_0 - J_\tau) = \frac{J'_\tau}{m_\tau} - \frac{m'_\tau}{m_\tau^2}(J_\tau - J_0) = \frac{1}{m_\tau^2}\left(J'_\tau m_\tau - m'_\tau (J_\tau - J_0)\right).$$

Consider an arbitrary threshold τ and let $\Delta > 0$. Then,

$$V_{cap}(\tau + \Delta) - V_{cap}(\tau) = \lambda_l(\tau + \Delta) J_0 + (\bar{\lambda} - \lambda_l(\tau + \Delta)) J_{\tau+\Delta} - (\lambda_l(\tau) J_0 + (\bar{\lambda} - \lambda_l(\tau)) J_\tau),$$

which, since $\lambda_l(\tau) = \bar{\lambda} - \mu_\tau$, is equal to

$$\begin{aligned}
&(\bar{\lambda} - \mu_{\tau+\Delta})J_0 + (\mu_{\tau+\Delta})J_{\tau+\Delta} - ((\bar{\lambda} - \mu_\tau)J_0 + (\mu_\tau)J_\tau) \\
&= -\mu_{\tau+\Delta}J_0 + \mu_{\tau+\Delta}J_{\tau+\Delta} + \mu_\tau J_0 - \mu_\tau J_\tau + \mu_\tau J_{\tau+\Delta} - \mu_\tau J_{\tau+\Delta} \\
&= (\mu_{\tau+\Delta} - \mu_\tau)(J_{\tau+\Delta} - J_0) + \mu_\tau(J_{\tau+\Delta} - J_\tau),
\end{aligned}$$

where the latter is positive for Δ small enough, by (EC.5). where the latter is positive if and only if (EC.5) holds. Therefore, the τ that maximize $V_{cap}(\tau)$ is the largest threshold possible in the range $[\tau_{spd}, \tau_{opt}]$, i.e., τ_{opt}, which implies that the Bl-FS policy is optimal here.

(c) The proof is entirely analogous to the proof of (b), expect for the reverse inequality of Eq. (EC.5) which implies that $V_{cap}(\tau)$ is decreasing in τ.

□

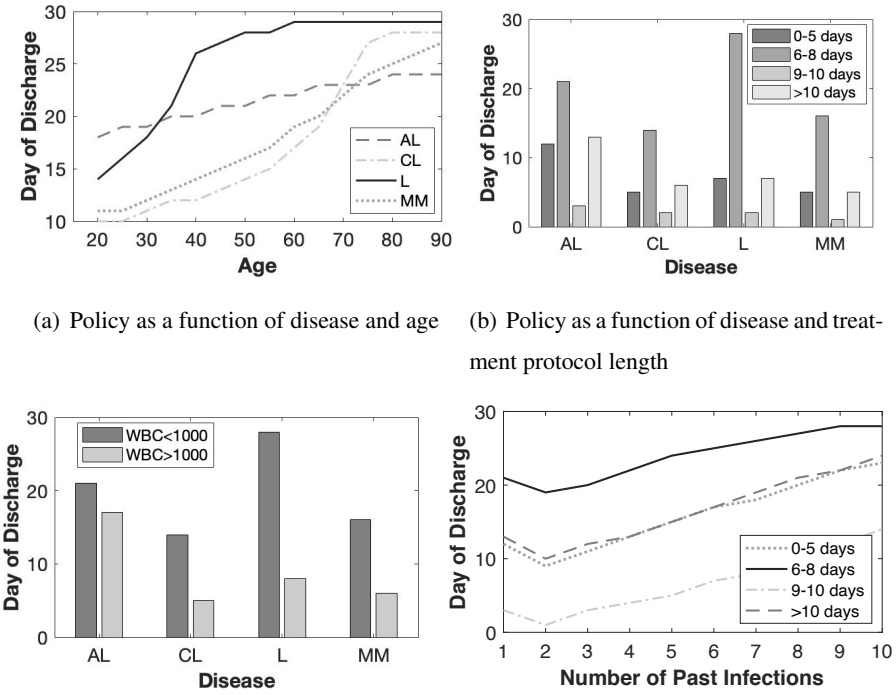
Appendix D: Case Study - Supplemental Results

D.1. Optimal Policy as a Function of Patient Characteristics - Further Details for Section 5.1

As explained in Section 5.1, we investigate how various factors affect the optimal observational hospital LOS for hematology patients in the presence of ample capacity. Figure EC.1 shows how the optimal effective threshold changes as a function of patient characteristics such as type (disease, age), current medical state (treatment protocol, WBC (white blood cell count) at the end of the protocol treatment), and medical history (number of past infections). We observe that, in general, as age increases, it is optimal for the patient to stay longer in the HW (Figure 1(a)). When examining the influence of the length of protocol on the length of hospital observation (Figure 1(b)), we notice that patients with a medium-length protocol (6–8 days), that are the most aggressive treatment protocols, should stay longer in the ward than patients with short- or long-length protocol. The state of the patient at the end of treatment is an important factor too, as observed in Figure 1(c): high-risk patients whose WBC at the end of their treatment is low need a much longer in-hospital observation period. Finally, patient history also impacts risk and the optimal observation time, as we observe in Figure 1(d). The non-monotonicity around the low number of past infections likely follows from the fact that the first couple of treatment cycles have a higher infection risk than later cycles (see Table 2 in Carmen et al. 2019).

D.2. Sensitivity to the Infection Survival Probability

An important aspect of our model and analysis is its potential to serve as a decision-support tool for hospitals, helping them assess how changes in patient management protocols impact individual treatment recommendations and outcomes. For example, a hospital may implement improved home-care procedures to reduce the probability of infection-related mortality at home, bringing it closer to the mortality level observed at the hospital. Even a simple post-discharge follow-up can effectively lower mortality risks (Leschke et al. 2012), and recent studies show that, for some patient types, telemedicine follow-ups via mobile-app can be as effective as in-person follow-ups (Marquez-Algaba et al. 2022). Evidence from Carmen et al. (2019) highlights the potential for such improvements: while infection-related mortality risk at home is reported to be 11%–15%, the corresponding mortality risk for infections acquired at the



(a) Policy as a function of disease and age

(b) Policy as a function of disease and treatment protocol length

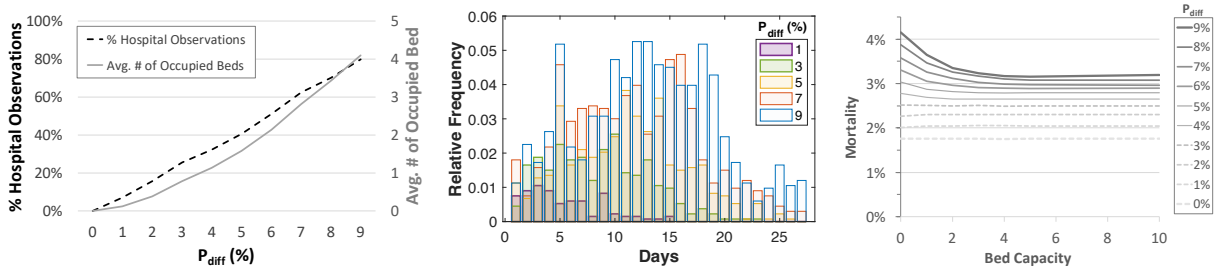
(c) Policy as a function of disease and WBC count at the end of treatment

(d) Policy as a function of treatment protocol length and history of past infections

Figure EC.1 Optimal Effective Threshold (Based on Infection and Mortality Risks Given in Carmen et al. 2019)

hospital is much lower, at 4%–6% — an average difference of 9%. This discrepancy is partly attributed to delays in recognizing and treating infections among home-care patients. Hospitals could mitigate this gap by (a) improving infection identification time and (b) reducing the infection-to-antibiotic treatment time through better access-to-treatment process once patients return to the hospital.

To quantify the impact of reducing the mortality gap between home-care and hospital observation, we conduct a counterfactual numerical experiment. Our data exhibits an average of 9% difference in infection mortality risk between home care and hospital ($p_{diff} = (1 - p_h) - (1 - p_w)$). In our experiment, we progressively reduce this gap from 9%



(a) Average number of occupied beds and percent of hospital observations

(b) Distribution of $\tau_{opt}|\tau_{opt} > 0$

(c) Mortality rate using ISP algorithm

Figure EC.2 Capacity Requirements and Performance as a Function of p_{diff}

to 0% in jumps of 1%, by keeping p_w fixed while changing p_h accordingly. Figure EC.2 illustrates the implications of this reduction. As the difference diminishes, hospital observation becomes redundant for most patients: the proportion of patients requiring hospitalization decreases (Figure EC.2(a), '% Hospital Observations'), and the optimal hospital observation time, τ_{opt}, also decreases (Figure EC.2(b)). As a result, the occupancy needed to treat these patients reduces (Figure EC.2(a), 'Avg. # of Occupied Beds'). Furthermore, Figure EC.2(c) shows how overall the mortality rate evolves as a function of bed capacity for different p_{diff} values under the ISP algorithm. If the mortality risk at home and at the hospital were equal, the mortality rate would drop to 1.75%, and all patients would be sent to home care right away. Notably, our experiment shows that using ISP with ample capacity achieves the same mortality reduction as lowering p_{diff} to 5.5% in a zero-capacity scenario—both yielding a mortality rate of 3.2%. This suggests two primary levers to improve patient health: (1) expanding hospital bed capacity for observation to align with the τ_{opt} policy and (2) reducing the excess mortality risk associated with home care. Our findings indicate that the second lever—improving home-care outcomes—has the greater potential to reduce mortality.

Appendix E: Discrete-time Formulation of the Uncapacitated Single Patient Problem

In Section 3, we discussed the problem of optimally determining the time to move a patient to home care after treatment, in the lack of capacity constraints. That formulation assumed that a patient can be sent to home care at any time along a continuum. In practice, discharges from the hospital tend to occur once a day, typically in the afternoon (Armony et al. 2015). In such cases, a discrete-time problem formulation may be more appropriate. In this section, we describe the discrete-time model that we formulate as a Markov-Decision Problem (MDP). We use this formulation in our numerical examples throughout the paper.

E.1. MDP Formulation for a Single Patient

We propose a discrete-time MDP formulation in which the physician makes a decision every period (day) on whether to keep the patient in the hospital or to discharge her to be cared for at home. Here too, we focus on the decision of when to send a patient to home care after the treatment and *prior* to developing infection and within the 30-day cycle time. If and when an infection occurs it is clear that the patient should be hospitalized. Thus, we refer to a state wherein a patient has developed an infection as an absorbing state. We also make the realistic assumption that once a patient has been sent to home care, she will not return to the hospital unless she has developed an infection. Thus, sending a patient to home care will also result in a transition to an absorbing state. The states and transitions of the MDP are defined as follows:

- Define $\mathcal{S}$, the set of system states. We interpret the state s, for $s \in \mathcal{S}$, $s \in \{1, ..., T\}$, as being at the beginning of day s after completing $s-1$ hospitalization days. In addition, we have an absorbing state, Δ (indicating that the patient was either discharged or infected). Hence, $\mathcal{S} = \{1, ..., T, \Delta\}$. The initial state is $s = 1$ and the final state is $s = \Delta$.
- Define $\mathcal{A}$ as the set of admissible actions. In general, $\mathcal{A} = \{w, h\}$, where w stands for *ward* (we use "ward" and "hospital" interchangeably) and h for *home*. We denote by $a(s)$ the action taken in state s. Hence, $a(s) = w$ means that the patient stays in the hospital ward in state s, and $a(s) = h$ means that the patient is discharged at state s. It is assumed that at the beginning of the horizon (time 0) the patient is at the hospital.
- Let $t = 0, ..., T$ denote the time period. We use a subscript t to denote the state or action at time t.

- Define P as the probability transition matrix. The entry $P(s,s',a)$ in the matrix P describes the probability of moving from state s to s' given choice of action a. Hence, $P(s,s',a) = Pr(s_{t+1} = s' | s_t = s, a_t = a)$. If the patient is discharged (i.e., $a(s) = h$), she moves from state $s \in \{1,...,T\}$ to Δ. Hence, $P(s,\Delta,h) = 1$ for all $s \in \{1,...,T\}$.

If the patient stays at the hospital for observation for another day (i.e., $a(s) = w$) then she may move to state $s+1$ if she develops no infection during that period, or move to Δ if she does. Hence, $P(s,s+1,w)$ and $P(s,\Delta,w)$ for all $s \in \{1,...,T\}$ are determined by the hazard rate of developing an infection in state s at the hospital. We define by $r_w(s)$ (by $r_h(s)$) the risk function[5] of developing an infection at time s given that the patient is in the ward (at home) at the beginning of that period and has not developed an infection. Then, $P(s,\Delta,w) = r_w(s)$ and $P(s,s+1,w) = 1 - r_w(s)$ for all $s \in \{1,...,T-1\}$. Formally, we assume that $r_w(T) = r_h(T) = 0$.

Once we get to state Δ, we stay there indefinitely. Hence, $P(\Delta,\Delta,a) = 1$ for all $a \in \{h,w\}$. At the end of the horizon, T, any uninfected patient who is still at the hospital is sent home. Thus, formally, we have that $P(T,\Delta,w) = P(T,\Delta,h) = 1$.

Note that the functions r_w and r_h are of general form, as is demonstrated in Figure 1(c).

It is important to note that we assume that the hazard-rate functions depend on the time that has elapsed since the completion of the treatment and on the location of the patient at that time. Importantly, the infection hazard-rate function does not depend on the time that the patient was placed in that location. Thus, for example, $r_h(t)$ is the risk of developing an infection exactly t time units after treatment (provided that no infection has developed prior to that time) given that the patient is at home at that time and independently of when the patient was sent home from the hospital.

- Define a reward matrix R. Its elements $R(s,s',a)$ denote the reward gained from making the transition from state s to state s', given action a. To spell out the elements of the reward function we need to first define the specific gains and costs realized by each action in every state.

We normalize the rewards such that the patient receives a reward of 1 if she survives a cycle and 0 otherwise. A positive reward can be accumulated in one of two cases: either 1) the patient survived until state T without developing an infection (recall the assumption that one cannot develop an infection in state T or thereafter, i.e., $r_w(T) = r_h(T) = 0$.); or 2) the patient developed an infection and survived. A patient who has developed an infection survives with probability p_w (or p_h) if the infection developed while the patient was at the hospital (or home).

We also incur costs in each state. Denote by c_w (c_h) the hospitalization cost at the hospital (at home) per day for days $1,...,T-1$. (Note that the patient does not incur a hospitalization cost on day T since in the last day discharge occurs automatically if the patient has not developed an infection).

Denote by c_I the cost of treating infections (I for infection). This infection treatment cost includes all hospitalization costs incurred during the period that the patient is hospitalized from the beginning of the infection until recovery/death. The infection treatment cost is assumed to be independent of the time or the location at which the infection started. In the context of our formulation, the cost c_I is incurred once we move to state Δ, unless an infection did not occur at all up to time T. For convenience, we formulate this cost as a reward that is received if no infection occurred up to time T. We assume that $c_I >> c$. This is a reasonable assumption because, in expectation, treating an infection requires more than one day of hospitalization.

[5] We use risk function and hazard-rate function interchangeably.

We next compute $R(s,\Delta,h)$ for all $s<T$, which is the total expected reward to go when the patient is *discharged* in state s. For ease of notation we denote this function by $R_h(s):=R(s,\Delta,h)$. This reward function takes into account the probabilities of getting an infection and of recovering at or after time s until the end of the horizon. This reward function can be computed by backward induction as follows:

$$\begin{aligned} R_h(T) &= 1+c_I \\ R_h(s) &= -c_h + r_h(s)p_h + (1-r_h(s))R_h(s+1) \quad \forall s\in\{1,...,T-1\}. \end{aligned} \tag{EC.6}$$

To compute the reward gained from keeping a patient in the ward for an extra day, we need to consider two options: a) the patient gets infected on that day, or b) the patient remains uninfected for another day. In the former case, the expected reward is p_w-c_w: the system incurs a one-day hospitalization cost and receives a reward for the expected patient survival. In particular, $R(s,\Delta,w)=p_w-c_w$ for all $s\in\{1,...,T-1\}$. If the patient remains uninfected, the system incurs only the daily hospitalization cost, c_w. Thus, $R(s,s+1,w)=-c_w$, $\forall s<T$. If the patient reaches state T without getting an infection the reward is $1+c_I$. Hence, $R(T,\Delta,w)=1+c_I$.

In sum, the immediate reward function R for all $s\in\{1,..,T,\Delta\}$ is

$$R(s,s',a)=\begin{cases} -c_w, & \text{if } s<T,\ s'=s+1,\ a(s)=w; \\ p_w-c_w, & \text{if } s<T,\ s'=\Delta,\ a(s)=w; \\ R_h(s), & \text{if } s<T,\ s'=\Delta,\ a(s)=h; \\ 1+c_I, & \text{if } s=T,\ s'=\Delta,\ \text{for all } a; \\ 0, & \text{if } s=\Delta,\ s'=\Delta,\ \text{for all } a. \end{cases} \tag{EC.7}$$

E.1.1. The MDP formulation. Define a policy π such that $\pi_t(s)$ is the action the physician takes at time t if the state is s. Denote by $V_T(\pi)$ the expected reward over the finite horizon T if policy π is used. In particular,

$$V_T(\pi):=E^{\pi}\left[\sum_{t=1}^{T-1}R(s_t,s_{t+1},a_t)+R(s_T,\Delta,a_T)\right].$$

Let $v_t(s)$ be the optimal reward-to-go function from time t onward, given that the state at time t is s. (Note that, by definition, at time t, the state s can be either t or Δ.) By Equation (EC.7), since no reward is gained once state Δ is reached we have that

$$v_t(\Delta)=0,\ \ \forall t=1,...,T.$$

For all $s\neq\Delta$, we have that

$$v_T(s)=1+c_I,\quad s=T,$$

and for $t=1,...,T-1$ and $s=t$, we have that

$$v_t(s)=\max\begin{cases} R_h(s), & a_t=h; \\ p_w\cdot r_w(t)+v_{t+1}(s+1)\cdot(1-r_w(t))-c_w, & a_t=w. \end{cases} \tag{EC.8}$$

Let $R_w(s)$ be the reward-to-go if the patient is kept at the hospital at time $t=s$, $s\neq\Delta$, and the optimal action is taken from time $t+1$ onward. Then,

$$R_w(s)=p_w\cdot r_w(s)+v_{s+1}(s+1)\cdot(1-r_w(s))-c_w. \tag{EC.9}$$

Now we can rewrite Equation (EC.8) more compactly as

$$v_s(s)=\max\{R_h(s),R_w(s)\},\quad s\neq\Delta. \tag{EC.10}$$

Note that the problem at hand is an unconstrained MDP with finite state and action spaces. Thus, we can conclude that there exists an optimal policy that is non-randomized, using standard MDP theory.