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

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

Submitted to *Manufacturing & Service Operations Management*

Managing Scarce MRI Capacity in Overloaded Queueing Systems

Zhenghang Xu
Rotman School of Management, University of Toronto, zhenghang.xu@Rotman.utoronto.ca

Adam Diamant
Schulich School of Business, York University, adiamant@schulich.yorku.ca

Andre A. Cire
Operations Management and Statistics, University of Toronto, andre.cire@Rotman.utoronto.ca

Eugene Furman
Operations and Decision Sciences, American College of Greece, efurman@alba.acg.edu

Opher Baron
Rotman School of Management, University of Toronto, opher.baron@Rotman.utoronto.ca

Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and are not intended to be a true representation of the article's final published form. Use of this template to distribute papers in print or online or to submit papers to another non-INFORM publication is prohibited.

Abstract. *Problem definition.* Many service systems operate for long periods under severe congestion, where demand persistently exceeds available capacity. In these environments, expanding capacity is often impractical due to physical, financial, or workforce constraints. For these regimes, we study how service systems should dynamically allocate scarce capacity across multiple customer classes with different waiting costs. We focus on three design choices: how to prioritize customers over time, whether to optimize for individual versus system-level performance, and how to pool resources/demand. *Methodology/Results.* We model a congested, multi-class service system with flexible servers and time-varying demand as a finite-horizon control problem that captures the cumulative cost of waiting. For a broad class of waiting costs, we characterize the structure of the optimal capacity allocation policy, including when additional customer classes should receive service and how capacity should be shared across them. We then study how pooling can improve equitability of access. We apply our theoretical results to a case study of over seven million MRI encounters spanning ten years. *Managerial Implications.* We find that commonly used performance measures, such as the fraction of customers exceeding a waiting-time target, provides a poor description of performance in persistently congested systems. Instead, average waiting time among served customers, end-of-horizon queue length, and the distribution of capacity across classes better capture operational outcomes. Our results further show that system-level objectives substantially reduce service starvation compared to policies that focus on serving only the highest-cost customers, and that partial pooling can significantly improve overall performance.

1. Introduction

Service systems frequently undergo long periods during which inflows exceed effective processing capacity. Whether triggered by demand spikes, staffing constraints, or interruptions to routine operations, these periods can generate substantial queues that are slow to dissipate. The COVID-19 pandemic is one such example, where widespread closures and restrictions created large backlogs of postponed surgeries, diagnostic tests, and elective medical procedures (e.g., Aggarwal et al., 2020; Carr et al., 2021). Queued patients faced prolonged waits, which had adverse effects on health outcomes and increased overall costs (Jain et al., 2020). Similar dynamics arise in other capacity-constrained settings, such as transportation networks, where tasks cannot be abandoned and delay costs accumulate over time (e.g., Evler et al., 2022).

Magnetic resonance imaging (MRI), which provides physicians with essential information for diagnosing and monitoring a patient's condition, is a salient example of such a system. Timely imaging is critical: early diagnoses improve clinical outcomes, and substantial delays are associated with higher morbidity, mortality, and system-wide costs (Fraser Institute, 2014; Czeisler et al., 2020; Canadian Medical Association, 2020; Medicine Net, 2020; BMJ, 2021). For instance, Cournane et al. (2016) report that the cost of treatment more than doubles when imaging delays fall in the 75th rather than the 25th percentile of the wait-time distribution. In Ontario, Canada, the median time from referral to MRI between 2014 and 2019 was approximately two months, which was shorter than the historical average of 8–12 months (Ontario Association of Radiologists, 2014). However, due to COVID-19, roughly 50% fewer MRIs were performed, causing the backlog of requests to grow to more than 475,000 patients (CTV, 2021). Estimates indicated that clearing these queues would require more than ten months of continuous operation at 120% of nominal capacity, with costs exceeding $300 million (Canadian Medical Association, 2020). Figure 1 illustrates the aggregate growth in queue length over a 700-day horizon for one health region in the MRI appointment system subsequently analyzed in this study, highlighting the inherent instability in the system.

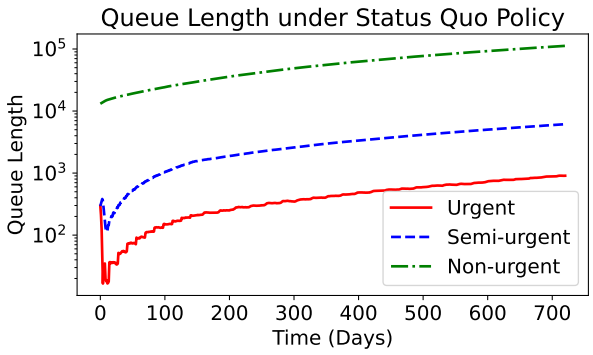


Figure 1 **Queue lengxth plot under status quo policy in one of the regions in our system. Lines represent patient classes with different urgency levels.**

In this work, we study how scarce capacity should be allocated in multi-class service systems experiencing persistent over-congestion, while accounting for the negative externalities of long delays. We model the flow of time-varying MRI requests using a queueing network with multiple classes and non-preemptive priority. Each queue corresponds to a hospital or a cluster of hospitals that pool demand and scanner capacity, a common arrangement in regional radiology networks designed to balance workloads and mitigate local bottlenecks. When all MRIs are busy, arriving patients join an effectively infinite-capacity waiting list, consistent with practice where patients are scheduled rather than turned away, and where balking or abandonment is extremely rare due to the diagnostic importance of MRI scans. To represent the heterogeneity in clinical urgency and the consequences of delayed imaging, we introduce class-dependent service times and cost functions that capture the accumulating health risks and financial burdens associated with waiting. We then formulate a finite-horizon optimal control problem that captures fluid dynamics and congestion costs, enabling dynamic capacity sharing across patient classes as queues evolve.

Our main contribution is to explore three operational levers for capacity management – namely, (i) adopting a system-level (versus individual) objective function, (ii) dynamic prioritization, and (iii) resource pooling – to mitigate service starvation and reduce excessively long wait times. Specifically:

1. *Non-Stable Service Systems: Fluid Model, Optimal Policies, and Performance Measures* (Sections 3-4). We study persistently congested, non-stable service systems with large initial queues and dynamic priority assignment without abandonment, expanding on work such as Van Mieghem (1995), Mandelbaum and Stolyar (2004), Gurvich and Whitt (2009) and Akan et al. (2012). Our framework integrates time-varying demand and accumulating priority in a multi-server, multi-class setting (Yu et al., 2025), with waiting costs modeled by smooth and increasing functions that capture both the severity of delays and their scale (i.e., number affected). Using a fluid approximation, we formulate a finite-horizon optimal control problem with a system-level objective that balances immediate delay costs and future congestion. We then characterize the structure of the optimal fluid policy, showing that the set of customer classes receiving service expands monotonically over time and that, once a class is activated and allocated positive capacity, it is served for the remainder of the planning horizon. We further show that commonly used metrics, such as the fraction of patients exceeding wait-time targets (FET), are ill-suited for over-congested systems, and propose alternative measures that offer deeper insight into system performance.
2. *Resource Pooling* (Section 5). We study hospital pooling within regional clusters to account for heterogeneity in patient arrival rates and local MRI capacity. We propose an optimization-based pooling model that is amenable to a decomposition-based algorithm, and show that selective pooling configurations can improve both equity and system performance in the presence of travel disutility across regions. Empirically, we find that pooling decisions are relatively robust to the specification of waiting cost function, suggesting that they can be made with a long-term, equity-oriented perspective.

3. *Case Study* (Section 6). We validate our results using a dataset of over seven million MRI appointments spanning ten years. Under the observed system parameters, the myopic benchmark (status quo policy) leads to steadily increasing queue lengths over time, reflecting persistent congestion (see Figure 1). However, comparing our dynamic priority policy with the myopic benchmark, we show substantial reductions in service starvation and improved system-level performance across a range of cost structures. We further demonstrate that a discrete-time, online implementation of the optimal fluid policy achieves near-optimal performance while remaining computationally efficient for practical deployment.

Section 2 presents related work while Section 3 formalizes our queuing model and optimal control problem. Section 4 presents the structure of the optimal fluid policy and analyzes cost-function choices that capture the desired trade-offs of the system. Section 5 develops a pooling model and solution approach. We then present a numerical study in Section 6 and conclude in Section 7 with managerial insights.

2. Related Work

We model system dynamics using a deterministic fluid approximation of a multi-server queue with time-varying demand, prioritized service, and infinitely patient jobs with heterogeneous service requirements. Patients who arrive to a fully occupied system join an infinite-capacity queue and incur waiting costs at a class-specific and potentially nonlinear, time-varying rate. Because of this behavior, our paper is closely related to the literature that studies queues where jobs can dynamically change their priority.

Several works investigate admission control policies that account for the time-varying waiting costs. One approach is to myopically admit jobs that incur the largest penalties, including settings where priority monotonically increases or decreases (e.g, Chaudhry and Choudhary, 1997; Anteneodo, 2009). Another stream of literature focuses on the performance of the non-preemptive earliest deadline scheduling (EDF) rule where jobs are prioritized based on their completion times (e.g., Frederickson, 1983; Jeffay et al., 1991) Unfortunately, this policy performs poorly in overloaded systems where a "domino effect" occurs, i.e., all jobs miss their deadline after a single job misses its deadline. Further, its complexity increases in the number of servers making it difficult to employ in our multi-server setting. Our methodology addresses the drawbacks of both policies in that we propose a dynamic (non-myopic) service policy for an arbitrary fixed time horizon (e.g., Armony and Ward, 2010) while also accounting for time-dependent changes in the priority level of different jobs (e.g., Avi-Itzhak and Levy, 2004; Sandmann, 2006).

Queues with static priorities have been extensively studied and pioneering work dates back to teletraffic operations (e.g., Cobham, 1954; Jaiswal, 1968). Queues with static priorities have primarily been analyzed using Markov chains (e.g., Fayolle et al., 1982; Latouche and Ramaswami, 1999). However, for a large number of servers and priority classes, such analysis is difficult due to the large dimensionality of the system. For instance, Osogami et al. (2004) propose a technique to reduce a high-dimensional Markov chain to two dimensions. The tractability of such an approach diminishes as the number of servers and

priority classes increases, and its suitability is restricted to systems with static priorities and preemption. Wang et al. (2015) present an exact approach for computing the sojourn time distribution of a preemptive $M/M/c$ queue with two priority classes and different service times. Using Queueing and Markov Chain Decomposition, Abouee-Mehrizi et al. (2012) derive exact solutions for a single server $M/G/1$ queueing system that employs a simple dynamic priority policy for inventory rationing; their analytical approach is extended to a double-sided queuing setting in Diamant and Baron (2019). Finally, Baron et al. (2019) derive heavy traffic results for an $M/M/c$ queue with two priority classes, where capacity can be rationed. Although static prioritization is more analytically tractable than systems with time-varying priority policies, they can lead to service starvation (Pattara-Aukom et al., 2002; Sun et al., 2014).

Initial work on dynamic, or delay-dependent, service priority can be traced to Jackson (1960) who study a single-server Markovian queue without preemption and introduce a stochastic priority function as the difference between an urgency number drawn from a general distribution and the current waiting time; as jobs continue to wait in queue, their priority score monotonically increases. Kleinrock (1964) and Netterman and Adiri (1979) extend these single-server results to a more general class of functions with an arbitrary number of priority classes, while Bagchi and Sullivan (1985) consider a similar system with generally distributed service times. Wang (2004) and Gómez-Corral et al. (2005) analyze a multi-server setting with two priority classes where jobs can dynamically change their priority from low to high. In general, service systems where a customer's priority increases as their waiting time increases are known as queues with self-generating priorities or priority accumulation (Krishnamoorthy and Narayanan, 2003; Krishnamoorthy et al., 2008). Smith (1956) demonstrates the optimality of the $c\mu$-scheduling rule if the cost of waiting is linear, while Van Mieghem (1995) shows that this policy is asymptotically optimal (under heavy traffic) for a $G/G/1$ queue with convex costs. Mandelbaum and Stolyar (2004) and Gurvich and Whitt (2009) extend these results to a parallel-server processing network while Harrison and Zeevi (2004) incorporates abandonment and show that the optimal service policy has a "*bang-bang*" structure.

We add to the literature by deriving a time-dependent capacity allocation policy for a multi-class, multi-server transient control problem with a nonlinearly increasing delay-cost function. The system is persistently congested, arrivals are time-varying, service priorities are dynamic, and priority accumulation (i.e., the cost of waiting) reflects both average waiting time and the number of patients experiencing that delay.

The dynamic assignment of jobs based on their priority level has evolved into research that combines queueing models with various interpretations of fairness. Avi-Itzhak and Levy (2004) define a fairness measure that is proportional to the second moment of the waiting-time distribution, while Sandmann (2006) propose an expected discrimination frequency to determine a fair allocation of capacity amongst multiple customer classes. In a call center setting, Armony and Ward (2010) analyze the trade-off between minimizing a customers' waiting time and fairly dividing the workload amongst a set of servers with heterogeneous service rates. In particular, they minimize the steady-state waiting time subject to a fairness constraint on

the proportion of time each server is idle. Finally, Teymoori et al. (2015) consider a set of weighted queues and formulate a network utility maximization problem to obtain a fair allocation of capacity amongst them.

Generally, evaluating fairness in a queueing system is typically done by means of defining time-varying functions whose aggregate values (e.g., total cost) must be balanced amongst jobs of different types. This approach is similar to the early literature on dynamic priority assignment. Mendelson and Whang (1990) and Kim and Mannino (2003), for instance, study single-server systems where the cost of waiting increases linearly with time; Gavirneni and Kulkarni (2016) introduce a waiting cost that follows a Burr distribution. Because cost functions traditionally include waiting times as their main component, several studies derive the waiting time distribution for queues with priority accumulation. To this end, Stanford et al. (2014) and Sharif et al. (2014) determine the waiting time distribution for single and multi-server queues, respectively, with multiple customer classes when the service time is exponentially distributed. Further, Li and Stanford (2016) extend such results to systems with heterogeneous servers and linearly increasing costs.

Our approach links the literature on dynamic priority assignment, queues with priority accumulation, and queueing fairness, while incorporating an optimization framework. Similar to papers with priority accumulation, patients gain priority through a general class of cost functions that depend on the waiting times of all queued patients. Consistent with research on fair queueing, we evaluate the equity associated with capacity assignments by comparing the aggregate costs accrued by different patient types over the planning horizon, minimizing any excessive cost accumulation. We also study a practically relevant yet understudied setting: a chronically overloaded queue (Yu et al., 2025). Our analysis uses a fluid approximation of the underlying stochastic system (Mandelbaum et al., 1998; Bassamboo and Randhawa, 2010; Pender et al., 2017) that preserves its essential features, i.e., time-varying demand, multiple servers, and heterogeneous patient types, while enabling the derivation of time-dependent capacity-rationing policies.

We analyze a fundamental issue in the provisioning of medical services, i.e., waitlist management and timely access to care for congested systems. Culyer and Cullis (1976) present a review of early work while several qualitative/empirical studies analyze the relationship between hospital policies and waitlist length (e.g., Coyle, 1984; Wolstenholme, 1993). Other empirical work suggests interventions to clear a backlog of patients waiting for medical care (Habtamu et al., 2011; Habtamu and Burton, 2015). Several scheduling policies aim to reduce service backlogs. For example, Jiang et al. (2020) propose a weight-accumulation policy, where priority increases with waiting time, and a priority-promotion policy, where patients advance to a higher priority class after exceeding a wait threshold. Our study extends these approaches by introducing a system-level objective from which both policies arise as special cases. We also derive analytical results and assess how alternative cost functions influence key performance measures.

Finally, we examine the benefits of pooling resources in settings where demand exceeds capacity. While prior work has shown the advantages of pooling under linear cost structures (e.g., Barron and Baron, 2022), it is unrealistic to expect that the pooling of resources will somehow unlock large amounts of idle capacity

due to the substantial congestion. Instead, our focus is on optimizing cluster assignments to strategically manage the nonlinear costs of prolonged waiting. Pooling heterogeneous service classes can reshape queue distributions and alter patient prioritization (e.g., Mandelbaum et al., 1998), making patient mix selection is important for improving performance (vanDijk and van derSluis, 2008; Vanberkel et al., 2012). To this end, we formulate a mixed-integer linear program that uses a pointwise stationary approximation of system dynamics (Bassamboo et al., 2006, 2009) to identify the optimal clustering of hospitals. We then develop a solution method for large-scale instances based on a dual-conic Benders decomposition algorithm.

3. Queueing Model for Overload Systems

We model our stochastic system after the MRI service process which is typical of many North American hospital systems (e.g., CBC 2020). After a physician determines that an MRI is medically necessary, a referral is sent to an affiliated hospital. A radiologist then assesses the referral request and assigns it a priority level. This designation defines a wait-time target that should be adhered to based on four priority classes: emergent ($<$ 24 hours), urgent ($<$ 2 days), semi-urgent ($<$ 10 days), and non-urgent ($<$ 28 days). These targets are set at the 90^{th} percentile, i.e., the time within which 90% of patients in each class should undergo an MRI from the date of their referral. After a patient is assigned to a priority level, they are contacted by hospital staff and assigned an MRI appointment. Once completed, the images are interpreted by a radiologist and the results are sent to the referring physician who reviews the diagnosis (e.g., Lagzi et al., 2023). After a priority level is assigned, a patient's status typically remains unchanged unless subsequent diagnostic findings from other prognostic tools (e.g., lab tests) warrant reclassification.

3.1. Backlog Formulation for a Single Pool

Consider a single hospital, or pooled hospitals, offering MRI services. Let $\mathcal{I} := \{1, \ldots, I\}$ denote a set of I distinct patient types (or classes) such that type-$i \in \mathcal{I}$ patients have similar service requirements. To model the dynamics of the MRI waitlist, we analyze a queueing system with $m \in \mathbb{Z}_+$ flexible servers; each server represents an MRI machine that can serve patients from any class. Arrivals of type-i patients follow a non-homogeneous Poisson process with intensity rate $\lambda_i(t) > 0$, which we assume is a continuously differentiable function with respect to t. In accordance with our application, at $t = 0$, the service system is in a congested state for all i. We also assume the service time of a type-i patient, denoted as S_i, is generally distributed with rate parameter μ_i and is independent of the arrival process and service time of patients who have yet to arrive or are currently in the system; see Figure 2 for an example with $I = 3$ patient classes.

We capture waiting costs by a continuous, increasing, and non-negative cost function $g_i : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. For each patient class, g_i is a differentiable on $\mathbb{R}_{\geq 0}$, but not necessarily convex. A direct approach to calculating the waiting cost would be to record the time since each patient entered the system and use this duration as input to $g_i(\cdot)$. However, this approach is analytically (and practically) intractable as it requires that the time each patient spends in the system be continuously tracked (e.g., Down and Lewis, 2006). Instead, let

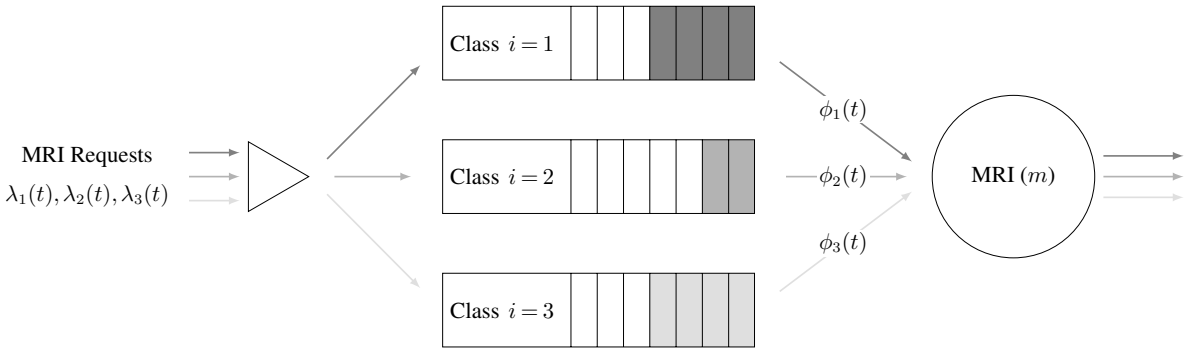


Figure 2 Service dynamics for three classes with arrival rates $\lambda_i(\cdot)$ and allocation $\phi_i(\cdot)$ for each class.

$\{W_i(t) \,|\, t \geq 0\}$ be an *indirect* estimate of the waiting time of type-i patients at time t. Confining our analysis to settings where $\dot{\lambda}(t)$ is bounded by a small constant, it is natural to consider the time-varying analogue of Little's Law (Kim and Whitt, 2013a), and thus, we set $W_i(t) := Y_i(t)/\lambda_i(t)$ where $Y_i(t)$ is a headcount stochastic process that tracks the number of queued type-i patients that are currently in the system at time t. In our application, arrival rates increase linearly over long horizons (months to years), service times are relatively short, and queued class-i patients are served in order of arrival (e.g., Wolff and Yao, 2014). Thus, there are intervals during which $\lambda_i(t)$ is effectively constant (days to weeks), and $Y_i(t)/\lambda_i(t)$ is a valid approximation of the wait time of type-i patients (Little, 2011; Wolff, 2011; Askin and Hanumantha, 2016).

The objective is to minimize the cumulative instantaneous waiting cost over all I classes, represented by $\sum_{i=1}^{I} g_i(W_i(t))Y_i(t)$ for all $t \geq 0$, where $g_i(W_i(t))$ is the marginal cost of waiting for type-i patients; multiplying by $Y_i(t)$ scales this cost by the number of patients that are currently experiencing it.

We analyze the fluid model of this system (see Zychlinski, 2023; Chan et al., 2025, and the references therein). The decision maker selects a time-dependent fraction of MRI capacity, $\phi_i(t) \in [0,1]$, to allocate to type-i patients so as to minimize the total cost over a fixed planning horizon $T > 0$. Let $\boldsymbol{\phi}(t) := (\phi_1(t), \phi_2(t), \ldots, \phi_I(t))$. We impose a single restriction, based on our over-congested setting, i.e., that capacity must be fully allocated at every t, $\sum_{i=1}^{I} \phi_i(t) = 1$. Let m be the number of servers and $\lambda_i(t)$, $i = 1, ..., I$, the arrival rate of type-i patients. Denote by $y_i(t)$ the cumulative proportion of type-i patients in the system at time t and let $\boldsymbol{y}(t) := (y_1(t), y_2(t), ..., y_I(t))$. The fluid dynamics are then given by

$$y_i(t) = y_i(0) + \int_0^t \lambda_i(u)du - \int_0^t \mu_i \left(y_i(u) \wedge \phi_i(u)m\right) du, \tag{1}$$

where and $\int_0^t \lambda_i(u)du$ and $\int_0^t \mu_i \left(y_i(u) \wedge \phi_i(u)m\right) du$ represent the total number of type-i patients who join the system and depart over the time interval $[0,t]$, respectively. Our objective is to minimize the expected

total cost over the planning horizon T, with no terminal cost to reflect the arbitrary choice of horizon length:

$$\begin{aligned}
\min_{\mathbf{y}(t),\boldsymbol{\phi}(t):\, t\in[0,T]} \quad & \sum_{i=1}^{I}\int_0^T g_i\left(\frac{y_i(u)}{\lambda_i(u)}\right) y_i(u)du, & \\
\text{s.t.} \quad & \dot{y}_i(t) = \lambda_i(t) - \mu_i\left(y_i(t)\wedge \phi_i(t)m\right), & \forall i\in\mathcal{I}, \forall t\in[0,T], \\
& \sum_{i=1}^{I}\phi_i(t) = 1, & \forall t\in[0,T], \\
& \phi_i(t)\in[0,1], & \forall i\in\mathcal{I}, \forall t\in[0,T], \\
& y_i(t)\geq 0, & \forall i\in\mathcal{I}, \forall t\in[0,T].
\end{aligned} \tag{FL}$$

The objective of the fluid model FL minimizes the total waiting-time cost of all queued type-i patients over T. The first constraint captures the queue length dynamics for class-i patients for all i, while the second and third ensure that all capacity is allocated at all times. The final two constraints enforce control bounds on $\phi_i(t)$ and the non-negativity of the queue lengths across all classes. The model assumes arrivals follow a time-dependent stochastic process (e.g., non-homogeneous Poisson), allowing for class-specific, time-varying arrival rates while also preserving analytical tractability. Further, we assume that patients do not abandon the waitlist. Due to the shortage of MRI appointments and long waiting times, cancellations are rare; e.g., in on our dataset, less than 4% of MRI requests are withdrawn for various reasons.

4. Optimal Policies and Practical Cost Functions

We next investigate properties of model FL. In Section 4.1, we characterize the structure of the optimal fluid policy for cost functions under which the Hamiltonian is convex, that is, when the incremental cost of delay increases more steeply as patients wait longer or as queues grow. In Section 4.2, we discuss cost function selection based on considerations of patient waiting times and broader healthcare settings.

4.1. Optimal Policies

We consider the following assumption about the nature of the workload in this congested system.

ASSUMPTION 1. *The total system workload $\sum_{i=1}^{I} y_i(t)$ at any $t\leq T$ and for any $\boldsymbol{\phi}(t)$ is greater than the capacity m. However, for any i and t with $\phi_i(t)=1$, $y_i(t)$ is strictly decreasing at time t.*

Given the significant backlog for MRI services seen in practice, the assumption reflects the reality that the system begins in a state of congestion and operates under these conditions for all $t\leq T$; see, e.g., our case study in Section 6. Moreover, this assumption permits scenarios in which dedicating all MRI machines to a single patient class would eventually clear its queue, which is reasonable in practice. Nevertheless, under this assumption, we now show that the optimal control problem FL admits a more tractable representation.

PROPOSITION 1. *Consider the following optimization problem*

$$\begin{aligned}
\min_{\mathbf{y}(t),\boldsymbol{\phi}(t):\, t\in[0,T]} \quad & \sum_{i=1}^{I}\int_0^T g_i\left(\frac{y_i(u)}{\lambda_i(u)}\right) y_i(u)du, & \\
\textit{s.t.} \quad & \dot{y}_i(t) = \lambda_i(t) - \mu_i\phi_i(t)m, & \forall i\in\mathcal{I}, \forall t\in[0,T], \\
& \sum_{i=1}^{I}\phi_i(t) = 1, & \forall t\in[0,T], \\
& \phi_i(t)\in[0,1], & \forall i\in\mathcal{I}, \forall t\in[0,T].
\end{aligned} \tag{S-FL}$$

Under Assumption 1, the optimal control policies for FL and S-FL are identical.

Because the system is congested (Assumption 1), S-FL simplifies the queuing dynamics by eliminating the minimum term in the fluid dynamic equations; they now become a linear differential equations with time-dependent functions. Furthermore, the constraint $y_i(t)\geq 0$ is now redundant and can be omitted from the model, since allocating capacity to empty queues is suboptimal. That is, for type-i patients, an optimal fluid policy allocates capacity such that $y_i(t)>0$ for all t unless $y_i(t)=0$ for all i.

Let $\boldsymbol{\pi}(t) \coloneqq (\pi_1(t),\pi_2(t),...,\pi_I(t))$ be the costate vector corresponding to the fluid dynamic equations in (1); e.g., Hartl et al. (1995). The Hamiltonian function $\mathcal{H}(\boldsymbol{y},\boldsymbol{\phi},\boldsymbol{\pi},t)$, which captures the instantaneous total cost and shadow value dynamics of the system, admits the following representation:

$$\mathcal{H}(\boldsymbol{y},\boldsymbol{\phi},\boldsymbol{\pi},t) \coloneqq \sum_{i=1}^{I} g_i\left(\frac{y(t)}{\lambda_i(t)}\right) y_i(t) + \sum_{i=1}^{I}\pi_i(t)\dot{y}_i(t)$$

Based on this definition of $\mathcal{H}(\cdot)$, instead of restricting $g_i(\cdot)$ to a family of convex functions, we impose a more general requirement on admissible cost functions:

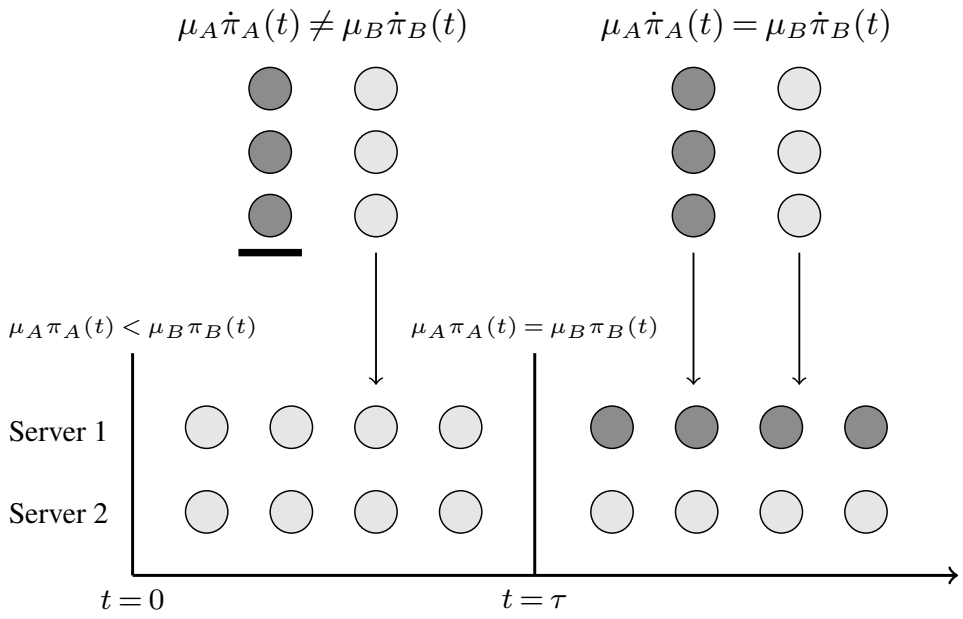
ASSUMPTION 2. *The cost functions $g_i(\cdot)$, $i\in\mathcal{I}$, are such that $\mathcal{H}(\boldsymbol{y},\boldsymbol{\phi},\boldsymbol{\pi},t)$ is convex in $(\boldsymbol{y},\boldsymbol{\phi})$.*

One implication of Assumption 2 is that the expression $z\,g_i(z)$ is convex in z. Nonetheless, the chosen cost function $g_i(\cdot)$ can be general (e.g., convex, concave, or a mixture of concave-convex). This assumption also ensures that S-FL admits a unique optimal solution (Hartl et al., 1995). As we demonstrate in Section 4.2, many practical cost functions satisfy Assumption 2. Furthermore, in our setting, convexity of the Hamiltonian also has a practical interpretation: it implies that the incremental harm of delaying patients increases more rapidly as queues grow in size or as individual patients wait longer, which reflects the accelerating clinical, operational, financial, and equity consequences of prolonged MRI delays.

We next characterize the optimal capacity allocation policy associated with S-FL. To build intuition as to its structure, we first present a two-class example illustrating how the optimal control prioritizes patients.

EXAMPLE 1. Consider the system depicted in Figure 3 with $I=2$ patient types, $\mathcal{I}=\{A,B\}$, with B (light gray) being more costly at time $t=0$ than A (dark gray). In this setting, the optimal policy initially assign all $m=2$ servers to type-B patients as they accrue the highest cost. As time progresses, the cost

Figure 3 Visual depiction of the optimal policy structure with $m = 2$ and two patient classes.



associated with type-A patients rises due to the influx of arrivals and because, as no type-A patients are served, their waiting times increase. Conversely, the cost associated with type-B patients decrease due to service completions. Suppose there exists a time $0 < \tau < T$ where the cost rate of both types are equal. Then, after time τ, capacity will be shared between both patient types for all future periods. While Figure 3 shows a scenario with equal capacity sharing, the optimal policy need not prescribe equal allocations. □

The optimal capacity allocation in Example 1 illustrates three properties: (i) service initially prioritizes a single patient class; (ii) there is no chattering, meaning no rapid switching between patient classes having exclusive access to the servers; and (iii) service is progressively extended to additional patient classes.

We formalize this structure in Theorem 1 using Pontryagin's Minimum Principle. Specifically, the optimal state and costate trajectories satisfy the Hamiltonian dynamics, and at each time, the control minimizes the Hamiltonian pointwise. At $t = 0$, the first-order (or optimality) conditions uniquely characterize the optimal control; we refer to the associated proofs for a formal definition of these equations. Thus, exactly one patient class is prioritized, implying that the optimal fluid policy initially has a bang–bang structure (Hartl et al., 1995). For sufficiently large planning horizons, however, there exists a *switching time* $\tau > 0$ where first-order conditions become degenerate and no longer uniquely determine the control (Chachuat, 2007). We show that beyond this time, the optimal solution does not exhibit arbitrarily fast switching between classes (i.e., chattering). Instead, capacity is shared amongst eligible classes, and as time progresses, additional switching times arise and the set of patient classes that receive service monotonically expands.

THEOREM 1. *Under Assumptions 1 and 2, and noting $\pi_i(t) > 0$ for all i and $t \leq T$, the optimal solution to S-FL has the following structure.*

1. ***Strict Priority:*** *There exists a time $\tau \leq T$, referred to as the minimum switching time, such that all service capacity is allocated to patients of some class i for all $0 \leq t < \tau$ if*

$$\mu_i \dot{\pi}_i(0) < \mu_j \dot{\pi}_j(0) \text{ and } \mu_i \pi_i(t) > \mu_j \pi_j(t) \quad \text{for all } j \neq i.$$

2. ***Sequential Activation:*** *For each class $j \neq i$, there exists a unique switching time $\tau_j > \tau$ at which class j first begins to receive positive capacity. That is, as time progresses, additional classes enter service at these switching times, and the set of classes receiving capacity expands monotonically over time.*
3. ***Capacity Sharing:*** *Let $\mathcal{J}$ denote the set of classes receiving a positive capacity assignment at time t. For any $j, j' \in \mathcal{J}$, the optimal allocation satisfies*

$$\mu_j \pi_j(t) = \mu_{j'} \pi_{j'}(t) \quad \text{and} \quad \mu_j \dot{\pi}_j(t) = \mu_{j'} \dot{\pi}_{j'}(t).$$

Moreover, the optimal solution does not admit chattering.

A consequence of Theorem 1 is that, unless two classes initially have identical queue lengths, arrival rates, and cost functions, a single patient class is assigned all capacity at $t = 0$. If the planning horizon is sufficiently short, this policy is optimal for all $t \in [0, T]$. However, if T is long enough, there are at most $I - 1$ switching times corresponding to instants when additional patient classes gain access to the server; they are subsequently served for the remainder of the horizon. Thus, while the bang-bang control initially minimizes cost, after the first switching time, multiple patient classes are simultaneously allocated capacity.

The proof of Theorem 1 proceeds in two parts. First, it analyzes the set of differential equations $\dot{y}_i(t) = \lambda_i(t) - \mu_i \phi_i(t) m$ for every i, which presents relevant structural properties that we exploit. Specifically, because the control variables are continuous and bounded on a closed interval and the arrival function is continuously differentiable, by the Picard-Lindelöf theorem, the dynamical system (1) has a unique solution. Furthermore, the system dynamics are described by ordinary differential equations with two forcing terms: (i) $\mu_i \phi_i(t) m$; and (ii) $\lambda_i(t) > 0$. These conditions allow us to show that any state reachable at some time $\tau > 0$ using controls $0 \leq \phi_i(t) \leq 1$ can be attained if we instead restrict $\phi_i(t)$ to a bang-bang policy. We then directly analyze the optimality conditions of the Hamiltonian $\mathcal{H}(\cdot)$, demonstrating that the control has the specific bang–bang structure above and that the optimal solution cannot exhibit chattering behavior. We next show that over a small time period after the first switching time τ, the cost of using a bang-bang policy is higher than if the control shared capacity with an additional class (e.g., Hu et al., 2022). The argument is then extended to any sufficiently short time period thereafter. Finally, as the system is congested (see Assumption 1), the optimal policy will dynamically balance the cost rates among all served patient classes.

PROPOSITION 2. *For any $t > 0$, the unique capacity allocation using the singular control can be obtained by solving a linear system of equations. Furthermore, without loss of generality, if type-1 patients are prioritized at $t = 0$, a switching time $0 < \tau \leq T$ exists if and only if there exists an $i \in \mathcal{I}$ such that $\mu_1 \dot{\pi}_1(\tau^-) < \mu_i \dot{\pi}_i(\tau^-)$ and $\mu_1 \dot{\pi}_1(\tau) = \mu_i \dot{\pi}_i(\tau)$ for $\tau^- < \tau$. This condition can be verified in $O(I)$ time.*

This result shows that the optimal allocation policy for patient classes that share capacity at time $t > 0$ can be efficiently computed using only the *current* arrival rates and queue lengths. That is, we demonstrate that the proportion of capacity allocated to each patient class with the singular control can be obtained by solving a linear system of equations, with the Generalized Legendre-Clebsch condition (Robbins, 1967) ensuring the policy minimizes the Hamiltonian; see Appendix (EC.2). This approach avoids directly solving the underlying nonlinear dynamical system, an important advantage for practical implementation, as otherwise an allocation would be require discrete time intervals in practical settings (e.g., daily). The proposition further shows that switching times can be computed online by checking when a simple analytical condition is met. Moreover, as more classes share capacity, the number of conditions that need verification decreases.

Together, these insights support an algorithm that identifies, at any time t, the set of patient types that share capacity as well as the optimal proportion that should be allocated to each class. Specifically, the algorithm allocates service capacity to patient types over time based on dynamic priority gradients. It begins by assigning all capacity to the type with the smallest value of $\mu_i \dot{\pi}_i(0)$ and initializes the active set $\mathcal{J}$ with that class. Suppose that type-1 is the sole member of $\mathcal{J}$ at $t = 0$. The algorithm records the queue lengths at time $t > 0$ and, for some fixed $\Delta > 0$, and compares them with those at the prior time $t - \Delta$. More precisely, at time $t > 0$, any type i such that $\mu_1 \dot{\pi}_1(t - \Delta) < \mu_i \dot{\pi}_i(t - \Delta)$ and $\mu_1 \dot{\pi}_1(t) \geq \mu_i \dot{\pi}_i(t)$ is added to $\mathcal{J}$. The optimal capacity allocation among types in $\mathcal{J}$ is then obtained by solving a linear system for $\{\phi_j^*(t)\}_{j \in \mathcal{J}}$ subject to the constraint that $\phi_j^*(t) \geq 0$. At each subsequent decision epoch, the active set and capacity allocations are updated until $t = T$. The algorithm's pseudocode appears in Section EC.1.

Finally, we relate our findings to the generalized $c\mu$-rule, known to be asymptotically optimal for multi-server systems in heavy traffic (e.g., Baras et al., 1985; Mandelbaum and Stolyar, 2004).

COROLLARY 1. *Suppose that $g_i(z) = \gamma_i / \lambda_i(t)$ for each i and all $t \geq 0$. Then, the optimal allocation policy is determined by prioritizing classes using the generalized $c\mu$-rule.*

While Corollary 1 is known, our formulation differs in three specific ways. First, we focus on the exact optimal solution of a transient system, whose queue length process does not converge to a fixed point, and where the goal is to optimally allocate capacity under scarcity (e.g., Yu et al., 2025). Second, the accumulation of costs expresses the average experience of all patients waiting in the system and not just the cost incurred by those that depart (e.g., Van Mieghem, 1995). Finally, the interpretation of the $c\mu$-rule differs: capacity is still allocated in proportion to the instantaneous marginal value of service, however, rather than continuing until the queue is emptied, the allocation continues until marginal costs are equalized.

4.2. Cost Function Choice

We next assess choices for the cost function $g(\cdot)$ that balances the competing priorities inherent in any service system. In particular, convex cost functions describe an environment where the cost rate intensifies the longer a patient waits; such may be the case with, for example, many cancer diagnoses (Stephens, 2020).

In contrast, concave cost functions represent a setting where backlogged patients accumulate priority at a decreasing rate, which may be applicable to certain chronic diseases (Manta et al., 2019).

Convex Priority Accumulation: We first model an environment where the health/financial costs accumulate more quickly as patients wait longer. To this end, for type-i patients, $i \in \mathcal{I}$, let

$$g_i\left(\frac{y_i(t)}{\lambda_i(t)}\right) := \gamma_i \left(\frac{y_i(t)}{\lambda_i(t)}\right)^n \tag{PLY}$$

where $\gamma_i > 0$ is a cost parameter and $n \in \mathbb{Z}_+$ is the polynomial order. Note that PLY is a common choice from the perspective of utility theory where agents are assumed to have an increasing absolute risk aversion (e.g., Pratt, 1978). For instance, quadratic utility functions ($n = 2$) form the basis of modern portfolio theory (Levy and Markowitz, 1979). In our context, a convex cost function implies that patient priority grows at an accelerating rate, increasing the likelihood that longer-waiting patients are prioritized. While other convex functions can be used, monomials ensure that the direct adjoining approach of Pontryagin's minimum principle can be applied while guaranteeing optimality. Thus, they are a tractable choice and provide representative insights into the structure of other convex functions. In subsequent analysis, we examine two instances of S-FL using the objective in PLY: FL-S for $n = 2$ and FL-Q for $n = 5$.

Concave Priority Accumulation: We also consider an environment where costs accumulate more slowly the longer patients wait. From an economic perspective, this assumes agents are risk averse and have decreasing marginal utility (e.g., Gerber and Pafum, 1998; Afèche et al., 2013). For type $i \in \mathcal{I}$ patients,

$$g_i\left(\frac{y_i(t)}{\lambda_i(t)}\right) := \gamma_i \sqrt{\frac{y_i(t)}{\lambda_i(t)}}, \tag{SQR}$$

where $\gamma_i > 0$ is a cost parameter. We refer to S-FL with the SQR objective as FL-R.

When the cost structure follows a radical function, patients who have waited for longer experience diminishing penalty increases. As opposed to convex priority, newly referred patients quickly gain priority as compared to backlogged patients, whose cost increases at a slower rate the longer they remain in the system. The radical function is practical and provides representative insights for other concave functions.

Concave-Convex (Rotated-Logit) Priority Accumulation: To strike a balance between prioritizing patients with long waiting times and preventing newly referred patients from joining the backlog, we investigate a setting where $g_i(\cdot)$ is a rotated-logit (`r-logit`) function. This class of functions is widely used in population modeling and statistics (Tsoularis and Wallace, 2002). Its shape is also similar to the standardized inverse normal CDF. The rotated-logit function $h_i(\cdot)$ is defined by

$$h_i(z) = \gamma_i \ln\left(-\frac{z}{z - \chi_i}\right),$$

where $\gamma_i \geq 0$ is a scale parameter and $\chi_i > 0$ controls the shape. Because $h_i(z)$ can take negative values as z becomes smaller, the function may violate non-negativity. Thus, we construct the shifted version

$$\begin{aligned} g_i\left(\frac{y_i(t)}{\lambda_i(t)}\right) &= h_i\left(\frac{y_i(t)}{\lambda_i(t)} + \eta_i\right) - h_i(\eta_i) \\ &= \gamma_i \ln\left(\frac{(\eta_i + y_i(t)/\lambda_i(t))\,(\chi_i - \eta_i)}{(\chi_i - \eta_i - y_i(t)/\lambda_i(t))\,\eta_i}\right), \end{aligned} \tag{RLOG}$$

where $\eta_i \in (0, \chi_i)$. This transformation shifts $h_i(z)$ to the left by η_i and up by $-h_i(\eta_i)$ ensuring $g_i(0) = 0$. Since $g_i(\cdot)$ is increasing, it is also non-negative. Parameters χ_i and η_i must be chosen such that $\frac{y_i(t)}{\lambda_i(t)} < \chi_i - \eta_i$ for all i, t. When $\eta_i < \chi_i/2$, $g_i(z)$ is concave and increasing on $z \in (0,\, \chi_i/2 - \eta_i)$, and convex and increasing on $z \in (\chi_i/2 - \eta_i,\, \chi_i - \eta_i)$. Thus, the rotated logit cost function is concave-convex, and increases exponentially when the waiting time is sufficiently small or large, which captures the behavior of PLY. Otherwise, it increases with a decreasing rate which mimics SQR. We refer to S-FL with the RLOG objective as FL-Log. We select parameters such that the convex tail of $g_i(z)$ is realized only for sufficiently long wait times (see Section 6 and Appendix EC.2 for more details).

We benchmark the above choices against practical alternatives, including the linear $\gamma_i \frac{y_i(t)}{\lambda_i(t)}$ (Anderson and Moore, 2007) and constant γ_i (Hu et al., 2022) cost functions; objectives utilized in the extant literature. We refer to S-FL with the linear and constant objectives as FL-Lin and FL-Con, respectively.

PROPOSITION 3. *For constant, linear, monomial, radical, or rotated-logit cost functions, the Hamiltonian* $\mathcal{H}(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\pi}, t)$ *is convex in* $(\boldsymbol{y}, \boldsymbol{\phi})$.

The proposition states that these cost functions all satisfy Assumption 2, and thus, the corresponding policies satisfy Theorem 1 and are globally optimal. While more complex functional forms can be selected, these functions help illustrate the tradeoffs associated with convex, concave, and concave-convex accumulating priority functions while the co-state variables admit closed-form expressions.

5. Capacity and Demand Pooling

In the previous section, we derived an optimal capacity allocation policy for a single hospital or cluster of hospitals. This framework addressed two operational levers for rationing limited resources in congested systems without mechanisms for capacity expansion: (i) adopting a system-level versus an individual objective, i.e., the objective function in FL; and (ii) dynamically adjusting prioritization, i.e., the optimal policy characterized by Theorem 1. While the analysis helps mitigate the negative externalities associated with long wait times, additional performance gains may be achieved by modifying existing operational protocols.

Specifically, in the current system, each hospital independently manages its own waitlist, and patients are assigned to a single site-specific queue. However, as suggested by prior work (e.g., Jiang et al., 2023), resource pooling – the third operational lever investigated in this work – has the potential to reduce any inequity that arises from imbalanced demand or insufficient capacity at individual sites. In this section, we

develop a computational approach to group hospitals into clusters that share demand and MRI capacity. Once a cluster is defined, one can leverage the optimal policy from the previous section to prioritize patient classes. Our pooling model is introduced in Section 5.1, and an iterative computational solution method based on Benders decomposition (Benders, 1962) is presented in Section 5.2.

5.1. Pooling Model

Let $\mathcal{J} := \{1, \ldots, J\}$ be a set of hospitals. We consider a disutility function that represents the assignment cost of hospitals to a cluster k. Although there are many satisfactory interpretations, we assume for simplicity that the disutility cost is the maximum pairwise distance $d(j, j')$ between hospitals j and j', $j, j' \in \mathcal{J}$, assigned to the same cluster. A naïve approach to identify an optimal pooling strategy would be to enumerate every possible clustering configuration and evaluate its cost using the objective in S-FL together with the scaled disutility cost. However, this approach is not feasible for large systems. We consider two approximations that enable an optimization-based clustering approach.

Linearization: We assume $g_i(\cdot)$ is linear, which is a common first-order approximation technique used to represent complex, non-linear systems (Kelly et al., 1997; Elhedhli, 2006).

Simple Stationary Approximation: We use the average stationary arrival rate (Green and Kolesar, 1991; Green et al., 1991) which allows us to solve for the optimal stationary control policy. This approach is commonly used in the literature (Bassamboo et al., 2006; Yom-Tov et al., 2021) and is particularly effective for arrival rates that vary slowly with time (Bassamboo et al., 2009).

Based on the above approximation techniques, we formulate a combined clustering and capacity allocation problem. The input parameters are the hospital capacities m_j (collected in $\boldsymbol{m}$), the initial queue lengths q_{ij} (collected in $\boldsymbol{q}_i$), and the stationary arrival rates λ_{ij} (collected in $\boldsymbol{\lambda}_i$). The decision variables include binary assignment variables x_{jk} indicating whether hospital j is assigned to cluster k (with $\boldsymbol{x}_k$ denoting the corresponding assignment vector), cluster activation variables z_k, and capacity allocation variables $\phi_{ik} \geq 0$ representing the average proportion of capacity in cluster k allocated to class-i patients over the horizon T. Additional variables include $w_{jj'k}$ to enforce joint cluster assignments, ρ_k to capture the maximum pairwise distance within cluster k, and y_{ik} denoting the queue length of class-i patients assigned to cluster k.

For a given waiting cost parameter $\gamma_i > 0$ for each class i, and a user-defined parameter ε that balances total waiting costs against patients' travel time, we obtain the following optimization problem:

$$\min_{w_{jj'k}, z_k, x_{jk}, \phi_{ik}, y_{ik}, \rho_k} \sum_{k=1}^{J} \sum_{i=1}^{I} \gamma_i \left(\frac{y_{ik}}{1 - z_k + \boldsymbol{\lambda}_i^{\mathsf{T}} \boldsymbol{x}_k} \right) y_{ik} + \varepsilon \sum_{k=1}^{J} \rho_k \tag{POOL}$$

$$\sum_{i=1}^{I} \phi_{ik} = z_k, \qquad \forall k, \tag{2}$$

$$\sum_{k=1}^{J} x_{jk} = 1, \qquad \forall j, \tag{3}$$

$$z_k \geq x_{jk}, \qquad \forall j,k, \tag{4}$$

$$z_k \leq \sum_{j=1}^{J} x_{jk}, \qquad \forall k, \tag{5}$$

$$\rho_k \geq d(j,j')w_{jj'k} \qquad \forall j,j',k, \tag{6}$$

$$w_{jj'k} \leq x_{jk}\, x_{j'k} \qquad \forall j,j',k \tag{7}$$

$$y_{ik} = T\left(\sum_{j=1}^{J} \lambda_{ij}x_{jk} - \mu_i\phi_{ik}\sum_{j=1}^{J} m_j x_{jk}\right) + \sum_{j=1}^{J} q_{ij}x_{jk} \qquad \forall i,k \tag{8}$$

$$\phi_{ik}, \rho_k, y_{ik} \geq 0, \qquad \forall i,j,k, \tag{9}$$

$$w_{jj'k}, x_{jk}, z_k, \in \{0,1\}, \qquad \forall j,j,',k. \tag{10}$$

The objective of POOL minimizes a time-independent counterpart of S-FL where $g_i(\cdot) := \gamma_i\left(\frac{y_{ik}}{1-z_k+\boldsymbol{\lambda}_i^{\mathsf{T}}\boldsymbol{x}_k}\right)$, augmented by a cluster-specific linear cost that captures the disutility patients incur when they must travel long distances. We include $1-z_k$ in the denominator of the first term to ensure the objective is well-defined for any assignment. That is, if no hospitals are assigned to cluster k, $y_{ik}=0$ for all i (there are no queues), $\boldsymbol{\lambda}_i^{\mathsf{T}}\boldsymbol{x}_k = 0$ (there are no arrivals), and $z_k = 0$, which means that the term in the objective is zero for that cluster. Otherwise, $y_{ik} > 0$ for all i, $\boldsymbol{\lambda}_i^{\mathsf{T}}\boldsymbol{x}_k > 0$, and $z_k = 1$ which implies that the term reflects the true cost of its assignment. Furthermore, when $z_k = 1$, the ratio now represents a time-independent estimate of waiting time, as given by Little's law. Constraints (4)-(5) ensure z_k is one when at least one hospital assigned to cluster k. The objective is subject to (2), which defines a valid capacity allocation control policy for each cluster, while (3) ensures that every hospital is assigned to exactly one cluster. Finally, due to the y_{ik}^2 terms in the objective function, the model is a quadratic optimization problem and convex in $\boldsymbol{x}_k$.

The queue length equation is given by defining y_{ik} as in (8). To ensure queues do not become negative over the planning horizon T, ϕ_{ik} must be selected appropriately in (9). Constraints (7) acts as an indicator for the inclusion of hospitals j and j' in cluster k; this occurs whenever $w_{jj'k} = 1$. Thus, ρ_k must be at least the maximum pairwise distance $d(j,j')$ among hospitals assigned to cluster k. Minimizing the objective ensures that (6) is tight at optimality, so ρ_k equals the min–max distance for each open cluster.

When $\varepsilon \to 0$, the optimal solution will include fewer clusters as the complete pooling case tends to result in the smallest waiting time, and subsequently, cost. Conversely, as $\varepsilon \to \infty$, the optimal solution will feature a larger number of clusters until the "no pooling" solution dominates; this is driven by the significant weight attributed to travel. Typically, inducing some degree of pooling is desirable. In practice, a decision-maker will choose ε in such a way that it results in the creation of a small number of clusters. Regardless, the optimization problem is challenging to solve as it is a non-convex, fractional integer programming problem with polynomial terms (e.g., Ahmadi-Javid and Hoseinpour, 2022; Marand and Hoseinpour, 2024).

5.2. Pooling Optimization Approach

In this section, we exploit the special structure of POOL to develop an iterative cutting-plane computational method based on Benders decomposition. We first introduce a set of results that motivate the iterative solution algorithm used to determine the optimal pooling strategy. Define $\boldsymbol{z} := \{z_k\}_{\forall k}$ and $\boldsymbol{x} := \{\boldsymbol{x}_k\}_{\forall k}$.

PROPOSITION 4. *For POOL, the following properties are satisfied:*

1. *The objective function is non-decreasing and convex in x_{jk} for a fixed ϕ_{ik}.*
2. *For a fixed pooling structure $(\boldsymbol{z}, \boldsymbol{x})$, the objective function is non-increasing and jointly convex in ϕ_{ik}.*

The results in Proposition 4 establish important structural properties. In particular, joint convexity of the subproblem (the second property) ensures that the Lagrangian dual function of the system is well-defined. Moreover, since the objective function is convex in each dimension of x_{jk}, there exist supporting hyperplanes that provide lower bounds to the objective function. To proceed, first define

$$\tilde{g}_i(z_k, \boldsymbol{x}_k, \phi_{ik}) := \frac{\gamma_i y_{ik}^2}{1 - z_k + \boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k},$$

which is the cost contribution of cluster k to the objective when $|\boldsymbol{x}_k|_1 \geq 1$. Then, define the problem

$$G(\boldsymbol{z}, \boldsymbol{x}) := \min_{\phi_{ik} \geq 0} \sum_{k=1}^{J} \sum_{i=1}^{I} \tilde{g}_i(z_k, \boldsymbol{x}_k, \phi_{ik}) \text{ with (2), (8) and } y_{ik} \geq 0 \text{ for all } i, k, \tag{SUB}$$

which is referred to as the Benders subproblem. Given the definition of $G(\cdot)$, we establish a generalized Benders procedure that solves POOL to optimality and works iteratively as follows:

1. Solve a relaxed, tractable version of POOL to obtain a given assignment $(\boldsymbol{z}^{(\ell)}, \boldsymbol{x}^{(\ell)})$.
2. Evaluate its performance by solving the subproblem $G(\boldsymbol{z}^{(\ell)}, \boldsymbol{x}^{(\ell)})$.
3. Using the optimal solution of the subproblem, compute a supporting hyperplane at $(\boldsymbol{z}^{(\ell)}, \boldsymbol{x}^{(\ell)})$, yielding a valid cut that is a lower bound of $G(\cdot)$ to any $\boldsymbol{z}$ and $\boldsymbol{x}$. The cut will be tight at $(\boldsymbol{z}^{(\ell)}, \boldsymbol{x}^{(\ell)})$.
4. Add this cut to the relaxed allocation problem, re-solve it, and return to Step 2.

This iterative process tightens the relaxed allocation problem at each iteration, and convergence follows when successive solutions yield identical cluster assignments or when the improvement in the objective becomes negligible; we refer to Rahmaniani et al. (2017) for a formal details.

We next develop Steps 2 and 3. Let $\mathcal{J}_k \subseteq \mathcal{J}$ denote the subset of hospitals in cluster k, and define the aggregate type-i arrival rate, type-i queue length, and service capacity of cluster k as

$$\tilde{\lambda}_{ik} := \sum_{j \in \mathcal{J}_k} \lambda_{ij}, \qquad \tilde{m}_k := \sum_{j \in \mathcal{J}_k} m_j, \qquad \tilde{q}_{ik} := \sum_{j \in \mathcal{J}_k} q_{ij}.$$

Notice that problem SUB is separable in k. Thus, let $\left(\tilde{\lambda}_{\cdot k}^{(\ell)}, \tilde{q}_{\cdot k}^{(\ell)}, \tilde{m}_k^{(\ell)}\right)$ be the aggregate values associated with the current assignment $\mathcal{J}_k^{(\ell)}$, $\phi_{ik}^{(\ell)}$ be the corresponding optimal allocation policy in iteration ℓ, and define $\nu_k^{(\ell)}$ as the dual multiplier of the coupling constraint $\sum_{i=1}^{I} \phi_{ik} = z_k$.

By Danskin's theorem for parametric convex programs, a subgradient of SUB at the point $\left(\tilde{\lambda}_{\cdot k}^{(\ell)}, \tilde{q}_{\cdot k}^{(\ell)}, \tilde{m}_k^{(\ell)}, z_k^{(\ell)}\right)$ is given by the following partial derivatives of the objective evaluated at the optimal $\phi_{ik}^{(\ell)}$, together with the multiplier $\nu_k^{(\ell)}$, and using $\bar{y}_{ik}$ for a given solution $(\boldsymbol{x}_k^{\ell}, \phi_{ik}^{\ell})$:

$$g_{i,k}^{L,(\ell)} := \frac{\partial G_k}{\partial \tilde{\lambda}_{ik}}\Big|_{(\ell)} = \frac{\gamma_i \bar{y}_{ik}}{1 - z_k^{(\ell)} + \tilde{\lambda}_{ik}^{(\ell)}} \left(2T - \frac{\bar{y}_{ik}}{1 - z_k^{(\ell)} + \tilde{\lambda}_{ik}^{(\ell)}}\right), \quad g_{i,k}^{Q,(\ell)} := \frac{\partial G_k}{\partial \tilde{q}_{ik}}\Big|_{(\ell)} = \frac{2\gamma_i \bar{y}_{ik}}{1 - z_k^{(\ell)} + \tilde{\lambda}_{ik}^{(\ell)}},$$

$$g_k^{z,(\ell)} := \frac{\partial G_k}{\partial z_k}\Big|_{(\ell)} = \sum_{i=1}^{I} \gamma_i \left(\frac{\bar{y}_{ik}}{1 - z_k^{(\ell)} + \tilde{\lambda}_{ik}^{(\ell)}}\right)^2 - \nu_k^{(\ell)}, \qquad g_k^{M,(\ell)} := \frac{\partial G_k}{\partial \tilde{m}_k}\Big|_{(\ell)} = \sum_{i=1}^{I} \frac{-2\gamma_i \mu_i T \bar{y}_{ik}\, \phi_{ik}^{(\ell)}}{1 - z_k^{(\ell)} + \tilde{\lambda}_{ik}^{(\ell)}}.$$

Let β_k be the first term in the objective function of POOL for cluster k. Using the partial derivatives defined above, the supporting hyperplane at the current point now gives the constraint

$$\beta_k \geq G_k^{(\ell)} + \sum_{i=1}^{I} \left[g_{i,k}^{L,(\ell)}\left(\tilde{\lambda}_{ik} - \tilde{\lambda}_{ik}^{(\ell)}\right) + g_{i,k}^{Q,(\ell)}\left(\tilde{q}_{ik} - \tilde{q}_{ik}^{(\ell)}\right)\right] + g_k^{M,(\ell)}\left(\tilde{m}_k - \tilde{m}_k^{(\ell)}\right) + g_k^{z,(\ell)}\left(z_k - z_k^{(\ell)}\right), \quad (11)$$

where $G_k^{(\ell)}$ is the value of the objective function of SUB at the current solution. The above inequality is a valid Benders cut as it linearizes the exact projected cost at the current solution. Implementing (11) after solving SUB produces a Benders-type algorithm for POOL. That is, suppose $\mathcal{J}_k^{(\ell)}$ is the set of hospitals assigned to cluster k in iteration ℓ. The optimal capacity allocation ϕ_{ik}^* for $\boldsymbol{x}_k^{(\ell)}$ is obtained by solving SUB. A relaxation of POOL, known as the Benders master problem, is then given by

$$\min_{w_{jj'k}, z_k, x_{jk}, \beta_k, \rho_k} \sum_{k=1}^{J} \beta_k + \varepsilon \sum_{k=1}^{J} \rho_k \quad \text{(MST)}$$

$$\text{s.t.} \quad (3) - (7), (11),$$

$$w_{jj'k} \in \{0,1\}, z_k \in \{0,1\}, x_{jk} \in \{0,1\}, \beta_k \geq 0, \rho_k \geq 0, \qquad \forall j, k,$$

where (11) are the Bender's cut added to MST after each iteration. Notice that the master problem contains only the assignment variables $(\boldsymbol{x}, \boldsymbol{z})$ and variables β_k, while the allocation of capacity is delegated to an oracle. Given the incumbent solution $(\boldsymbol{x}^{(\ell)}, \boldsymbol{z}^{(\ell)})$, we calculate the aggregate values $\left(\tilde{\lambda}_{\cdot k}^{(\ell)}, \tilde{q}_{ik}^{(\ell)}, \tilde{m}_k^{(\ell)}\right)$ for each cluster k, compute the optimal allocation $\phi_{ik}^{(\ell)}$ via a quadratic program, and recover the multiplier $\nu_k^{(\ell)}$ of the coupling constraint $\sum_{i=1}^{I} \phi_{ik} = z_k$. By Danskin's theorem, this provides valid subgradients of the objective at the current point, i.e., $g_{i,k}^{L,(\ell)}, g_{i,k}^{Q,(\ell)}, g_k^{M,(\ell)}, g_k^{z,(\ell)}$. Repeating this process - adding a cut whenever it is violated - yields a sequence of tighter master problems due to the convexity of SUB.

Finally, we note that POOL can be reformulated using a second-order conic (SOC) representation of the clustering cost (e.g., Ahmadi-Javid and Hoseinpour, 2022). This results in a mixed-integer SOC program that can be solved directly with optimization solvers. We describe such a model in Section EC.6. However, experiments with the SOC model yield lower computational performance than the Benders algorithm above.

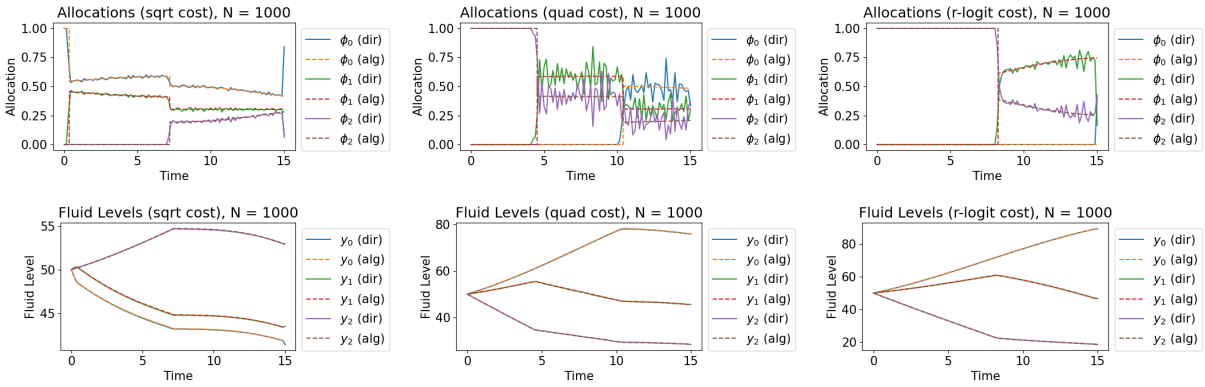


Figure 4 **Comparison of optimal allocations and fluid levels under FL-S (quad), FL-R (sqrt), and FL-Log (r-logit). Solid lines represent solutions from using the direct optimization approach, while dashed lines are solutions from employing Algorithm EC.1 with $N = 1000$ equally-spaced steps.**

6. Numerical Study

In this section, we evaluate the dynamic policies that incorporate priority accumulation. Section 6.1 presents synthetic examples that validate the optimal fluid solution from Theorem 1. Section 6.2 applies our methodology to an case study calibrated with 7.7 million MRI encounters. Section 6.3 then uses the approach from Section 5.2 to show how regional pooling can improve system performance.

6.1. Synthetic Experiments

To verify the appropriateness of Algorithm EC.1, we first compare its output against solutions obtained by solving S-FL using a standard nonlinear optimization solver. As a representative example of our experiments, we consider a synthetic setting with $I = 3$ patient classes, a horizon length of $T = 15$, and time-varying arrival rates represented by sinusoidal functions. That is, for patients of class i,

$$\lambda_i(t) = \Lambda_i \left(1 + \frac{1}{2}\sin\left(\frac{\pi t}{T}\right)\right),$$

with $(\Lambda_0, \Lambda_1, \Lambda_2) = (2, 1, 1/2)$. The arrival rates are concave and symmetric around $t = T/2$, we assume $m = 5$ identical servers with service rates $\boldsymbol{\mu} = (1.2, 1.0, 0.8)$, and set the initial fluid mass to be $\boldsymbol{y}(0) = (50, 50, 50)$. We evaluate the objectives FL-S, FL-R, and FL-Log using coefficients $\boldsymbol{\gamma} = (1.2, 1.0, 0.8)$. For FL-Log, we set $\boldsymbol{\chi} = (300, 300, 300)$ and $\boldsymbol{\eta} = (30, 30, 30)$ to ensure $y_i(t)/\lambda_i(t) \leq \chi_i - \eta_i$ for all i and t.

Using the `minimize` function in `scipy` with 100 discretized time steps (referred to as *dir*), we compare optimal allocations and fluid levels of S-FL with those generated by Algorithm EC.1 (referred to as *alg*); results are shown in Figure 4. We observe minor fluctuations using direct optimization, particularly under quadratic costs due to errors associated with discretization. Nevertheless, the allocations and fluid levels (Figure 4) as well as costs (Table EC.5) from both approaches are closely aligned. In terms of efficiency, Algorithm EC.1 runs in under one second on a laptop with 16GB RAM, while direct optimization takes

more than 60 seconds. Additional results in Appendix EC.4 examine the effect of the number of discretized time steps, showing that as N increases, the fluid levels produced by Algorithm EC.1 converge to those from direct optimization; for $N \geq 100$, the discrepancy between the two is negligible.

We next evaluate performance in a stochastic event-driven simulator, comparing the allocation produced by Algorithm EC.1 with that of a myopic policy that prioritizes patients with the highest delay cost. In this experiment, we consider a larger initial backlog, $\boldsymbol{y}(0) = (100, 100, 100)$, and extend the horizon to $T = 75$, having verified that our online algorithm is sufficiently efficient. Time is discretized into 75 steps, approximating daily capacity allocation over 2.5 months. As shown in Figure 5, the myopic policy violates the single–switching–time property of Proposition 1, exhibiting multiple switches in its capacity allocations. Consequently, the two policies generate markedly different queue-length trajectories.

Finally, we examine how different cost functions affect average waiting times. To capture aggregate performance changes, we define the weighted waiting time as $\tilde{W} := \sum_{i=1}^{I} \gamma_i \bar{W}_i$, where $\bar{W}_i$ is the average waiting time of class i. The relative improvement is then given by, $\frac{\tilde{W}_{MYOP} - \tilde{W}_{FL}}{\tilde{W}_{MYOP}}$, which measures the reduction in weighted waiting time of the optimal fluid policy relative to the myopic policy. Figure 6 presents the results of 50 replications. We observe that the relative improvement increases as a function of the right-tailed curvature of the cost function. Constant costs perform worse than the myopic policy, whereas all other cost functions match or improve upon it, with the rotated-logit cost yielding the greatest improvement. Nevertheless, as the system is persistently congested, we expect small differences in average system times.

6.2. Case Study

We now apply our approach to a real-world setting: the utilization of MRI machines within five Local Health Integration Units (LHINs) surrounding and including the City of Toronto, called the Greater Toronto Area (GTA). We parameterize arrival rates, service rates, and regional capacities, based on data from

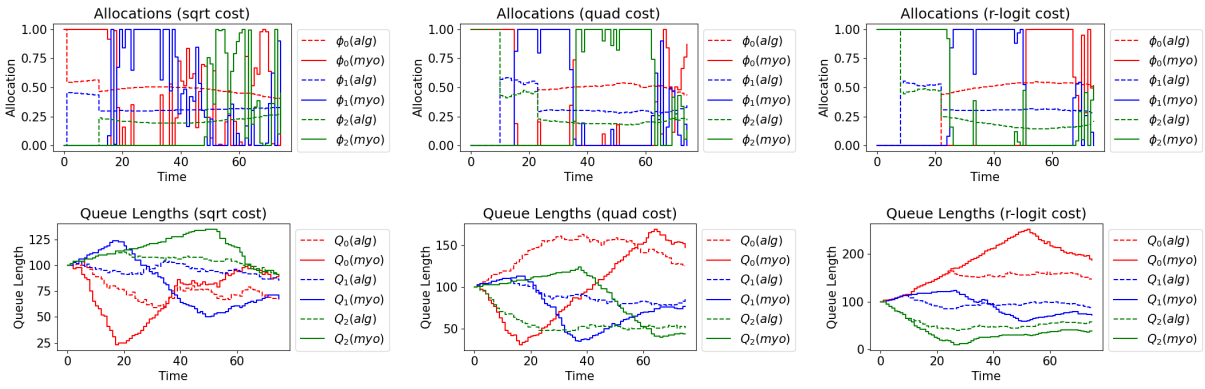


Figure 5 Comparison of optimal allocations and queue lengths under FL-S (quad), FL-R (sqrt), and FL-Log (r-logit). Solid lines represent solutions under the myopic policy, while dashed lines are solutions from employing Algorithm EC.1 with $N = 75$ equally-spaced steps.

7,728,056 MRI encounters. The regional breakdown is as follows: Toronto Central (1,410,775), Central (783,801), Central East (705,326), Central West (337,993), Mississauga Halton (539,223) and outside the GTA (3,950,938). The province distinguishes between 10 categories of MRI procedures (e.g., extremities, abdomen, spine, head) across four priority levels (emergent, urgent, semi-urgent, and non-urgent). Since emergency departments typically reserve MRI machines, we allocate capacity to emergent patients equal to the number of emergency departments in each LHIN. This reserved capacity is subtracted from the total to determine availability for non-emergent patients, i.e., urgent, semi-urgent, and non-urgent, which is given in Table 1. In regions without reserved capacity for emergent patients, they can instead be modeled as an additional priority level without loss of generality. Accordingly, our simulation experiments define 30 patient types (10 procedure categories $\times$ 3 non-emergent priority levels), as summarized in Table EC.1. Finally, we obtained another dataset containing post-COVID backlog estimates for all 30 classes within each LHIN.

Because the demand for diagnostic services is linearly increasing over time, we fit the arrival functions $\lambda_i(t)$ using a linear regression with an endogenous arrival rate and time as the predictor. In cases where the regression slope is not statistically significant ($\alpha > 0.1$), we assume a static or stationary arrival rate over the planning horizon. Further, we fix MRI durations, $1/\mu_i$, for each patient class at the 75^{th} percentile of their distribution. The values for the fitted arrival functions and service rates are presented in Appendix EC.2. Finally, we calibrate the cost coefficients γ_i to reflect health regulations (Office of the Auditor General of Ontario, 2018) by encoding relative wait-time targets across patient types. Specifically, we set γ_i by normalizing each class's target by the maximum wait-time target for any class, yielding coefficient values of 14 for urgent patients, 2.8 for semi-urgent patients, and 1.0 for non-urgent patients.

The fluid policies arise from a continuous-time control problem, and thus, we implement them by discretizing time over a planning horizon of $T = 360$ days. We also use $T = 180$, 540 and 720 for robustness.

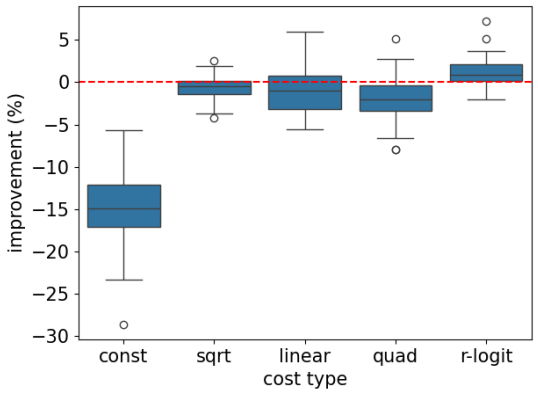


Figure 6 **Relative improvement in weighted waiting times of the optimal fluid allocation as compared with the myopic policy. The horizontal axis shows different cost functions. The vertical axis shows percentage improvement. Results are based on 50 independent replications.**

Table 1 Total Capacity Levels (MRI Machines) For Non-Emergent Patients

Toronto Central	Central	Central East	Central West	Mississauga Halton
14	5	5	2	4

At the start of each day, capacity is allocated to patient types according to the policy, held fixed throughout the day, and then updated for the following day. We compare these policies to a myopic benchmark that always prioritizes the patient class with the highest cost, as defined by the objective functions of FL-Con (`const`), FL-R (`sqrt`), FL-Lin (`linear`), FL-S (`quad`), FL-Q (`poly5`), and FL-Log (`r-logit`). Note that, as demonstrated by Corollary 1, FL-Con (`const`) corresponds to the generalized $c\mu$-rule.

At the start of each simulation, we initialize the queueing system with the current backlog. We then compare two allocation policies: the optimal solution of the fluid model (S-FL) with the myopic policy that always prioritizes the patient with largest waiting cost (denoted as MYOP). For S-FL, we use the discrete implementation of the optimal fluid policy (see Appendix EC.1). At the end of the planning horizon, we record waiting time, accumulated cost, queue length, number of service completions and fraction of service capacity assigned to each class. For each LHIN, we present the average of each performance metric over 20 simulations. In the following subsection, we pick Central region as an example to demonstrate the policy, which has five MRI machines (see Table 1) and many non-zero slope estimates (Table EC.2).

6.2.1. Choosing Performance Metrics: Why Not FET? Prior studies evaluate performance using the *fraction of patients exceeding their wait-time targets* (FET), which captures whether backlog reduction compromises timely access for new referrals (e.g., Jiang et al., 2023). However, based on the arrival rates and service capacities reported in Appendix EC.2, utilization in all five LHINs exceeds one. As a result, the queues are unstable and the system is persistently congested, causing queue lengths to grow indefinitely over time. Consequently, for a fixed waiting-time target, FET will inevitably increase because it is a function of the planning horizon (T). Table 2 illustrates this pattern across all urgency levels and cost functions for $T = 180, 360, 540$, and 720. As shown, FET becomes structurally uninformative; it converges to one as T grows regardless of policy. Consequently, it loses discriminatory power and cannot support meaningful policy comparisons, which may lead to misleading interpretations of system performance.

To address these issues, we focus on three complementary performance metrics that continue to distinguish between policies and reveal trade-offs even as T grows: (i) the *average waiting time among patients who complete service*, (ii) the *end-of-horizon queue length*, and (iii) the *normalized capacity allocation* across urgency levels. The first metric measures the delay experienced by patients who complete service, representing the realized quality of care for those who have been treated. The second metric quantifies the residual backlog accumulated over time and serves as an indicator of system congestion and unmet demand. The final metric characterizes the system's average resource distribution policy, highlighting potential imbalances in service prioritization. Together, these measures provide a comprehensive assessment of

both operational performance (through waiting and backlog) and policy behavior (through capacity allocation) under transient or unstable conditions. Moreover, neither converges to one as T increases.

	S-FL			MYOP		
	U	SU	NU	U	SU	NU
const	0.03	0.00	0.80	0.00	0.00	0.84
sqrt	0.08	0.71	0.83	0.00	0.57	0.84
linear	0.22	0.87	0.84	0.85	0.85	0.84
quad	0.51	0.44	0.82	0.97	0.91	0.84
poly5	0.51	0.44	0.82	0.98	0.93	0.84
r-logit	0.43	0.36	0.78	0.98	0.94	0.84

(a) $T = 180$

	S-FL			MYOP		
	U	SU	NU	U	SU	NU
const	0.02	0.00	0.87	0.00	0.00	0.92
sqrt	0.12	0.56	0.91	0.00	0.79	0.92
linear	0.22	0.77	0.91	0.93	0.92	0.92
quad	0.77	0.73	0.91	0.99	0.95	0.92
poly5	0.77	0.73	0.91	0.99	0.96	0.92
r-logit	0.65	0.61	0.83	0.99	0.97	0.92

(b) $T = 360$

	S-FL			MYOP		
	U	SU	NU	U	SU	NU
const	0.02	0.00	0.89	0.00	0.00	0.94
sqrt	0.11	0.45	0.94	0.04	0.86	0.94
linear	0.18	0.78	0.94	0.95	0.95	0.94
quad	0.85	0.83	0.94	0.99	0.97	0.94
poly5	0.85	0.83	0.94	0.99	0.97	0.94
r-logit	0.81	0.75	0.84	0.99	0.98	0.94

(c) $T = 540$

	S-FL			MYOP		
	U	SU	NU	U	SU	NU
const	0.01	0.00	0.90	0.00	0.00	0.96
sqrt	0.17	0.36	0.95	0.32	0.90	0.96
linear	0.23	0.72	0.94	0.97	0.96	0.96
quad	0.89	0.87	0.95	0.99	0.98	0.96
poly5	0.89	0.87	0.95	1.00	0.98	0.96
r-logit	0.85	0.82	0.84	1.00	0.98	0.96

(d) $T = 720$

Table 2 Central region: FET under horizon lengths $T = 180, 360, 540, 720$ for different cost functions under the optimal fluid and myopic policies. We denote *U* for urgent, *SU* for semi-urgent, and *NU* for non-urgent patients with a wait-time target of 1, 14 and 24 days, respectively.

6.2.2. Selecting a Model and Cost Function. We now use the three performance metrics introduced above to better understand system dynamics and policy performance. To facilitate our understanding, Figures 7 and 8 report the average waiting time for served patients and queue length at $T = 360$, along with 95% confidence intervals. Figure 9 presents the normalized capacity allocation across urgency levels.

What is the Effect of Curvature? As the right-tailed curvature of the cost functions increases, both optimal and myopic policies allocate more capacity to non-urgent patients, who have the longest queues and highest arrival rates. Specifically, as the cost function changes from `const` to `sqrt`, `linear`, `quad`, and `poly5`, the marginal cost associated with longer waiting times is larger, prompting both policies to shift more capacity to patients with longer delays, i.e., non-urgent patients. This shift is depicted in Figure 9, where higher-curvature cost functions assign a larger share of capacity to the non-urgent class while increasing the queue lengths and waiting times of the urgent and semi-urgent classes (Figures 7 and 8).

How do the Optimal Fluid and Myopic Policies Differ? Under certain cost functions – `constant`, `linear`, and `sqrt` – the optimal fluid policy *weakly dominates* the myopic policy. Weak dominance implies that one policy achieves shorter average waiting times and queue lengths across all urgency levels. This arises because the optimal fluid policy allocates capacity by prioritizing patients with shorter expected

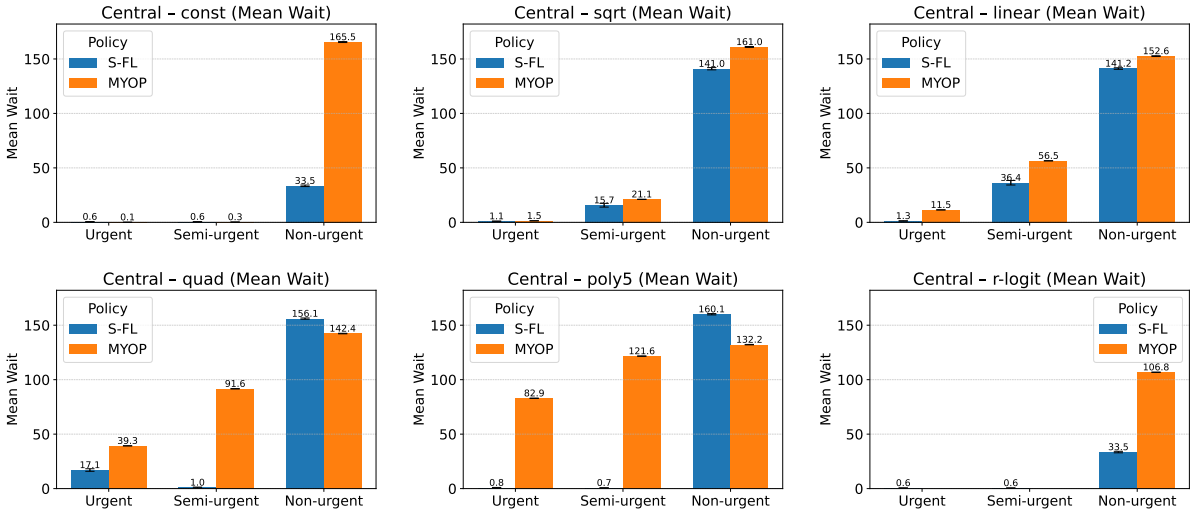


Figure 7 **Average waiting time of patients who complete service in the Central LHIN with $T = 360$. Numbers indicate mean values, and error bars show 95% confidence intervals from 20 independent replications. Missing bars denote that no patients were served.**

service durations – behavior consistent with the shortest processing time rule – which increases throughput within each urgency level. Across all cost functions, the myopic policy never weakly dominates the optimal fluid policy, although its allocation decisions sometimes differ. This is because the optimal fluid policy accounts for the entire system's congestion, not just the experience of customer's at the head of each queue. By doing so, the optimal fluid policy anticipates future demand fluctuations and proactively adjusts capacity.

We also find that the myopic policy may *starve* certain urgency classes, whereas the optimal fluid policy does not. For example, when the cost function is `r-logit`, Figure 9 shows that the myopic policy serves no urgent or semi-urgent patients. Although non-urgent patient achieve shorter queue lengths and waiting times (as shown in Figures 7 and 8), they does so at the cost of indefinitely delaying higher-priority patients. In contrast, under the same cost function, the optimal fluid policy maintains a more balanced allocation, assigning roughly 20% of capacity to urgent and semi-urgent classes. Across all cost functions, it consistently serves a positive share of each urgency level, ensuring more equitable access to capacity.

6.3. Resource Pooling in Overloaded Systems

To evaluate the performance of POOL, we enumerate all possible clustering configurations of the five LHINs surrounding and including the City of Toronto. For each candidate cluster, we apply the optimal capacity-rationing policy from Proposition 1 with $T = 360$. Although we test multiple cost functions, we find that for a fixed ε, all produce the same clusters. This suggests that clustering, as a high-level strategic decision, is far less sensitive to the choice of cost function than the resulting capacity-allocation decisions.

Table 3 compares the clustering solutions obtained from the Benders algorithm with those from full enumeration as a function of ε, which controls the weight placed on patient travel time (i.e., disutility cost). We

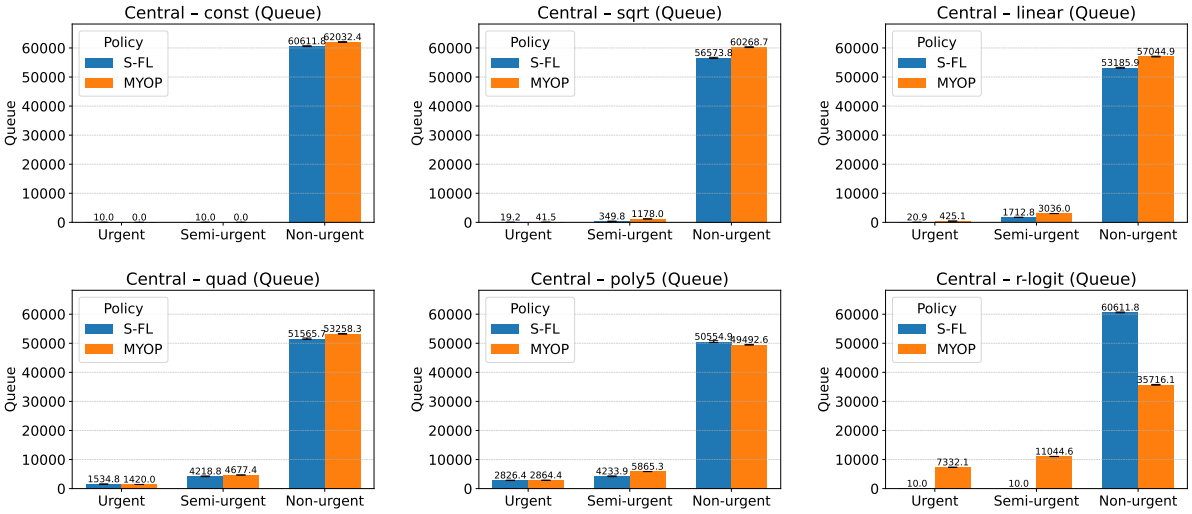


Figure 8 **Queue length at $T = 360$ for the Central LHIN. The numbers on the bars show the mean values. The error bars denote 95% confidence intervals from 20 independent replications.**

find that the two approaches produce consistent results. Specifically, for small values of ε, where waiting-time costs dominate, both approaches favor aggressive pooling, differing only in the placement of one or two adjacent LHINs. As the value of ε increases and the disutility placed on patient travel becomes a more prominent feature, both methods exhibit the same qualitative pattern of progressively disaggregating clusters until each LHIN is assigned to independent pools. Across the range of ε, the resulting pooling structures are nearly identical, with optimality gaps – computed as $(\text{cost}_{\text{Benders}} - \text{cost}_{\text{enum}})/\text{cost}_{\text{enum}}$ – below 1.5%, where the cost values are obtained by evaluating both solutions in the stochastic environment. This confirms that the Benders-based approach provides a near-optimal approximation to the enumeration benchmark.

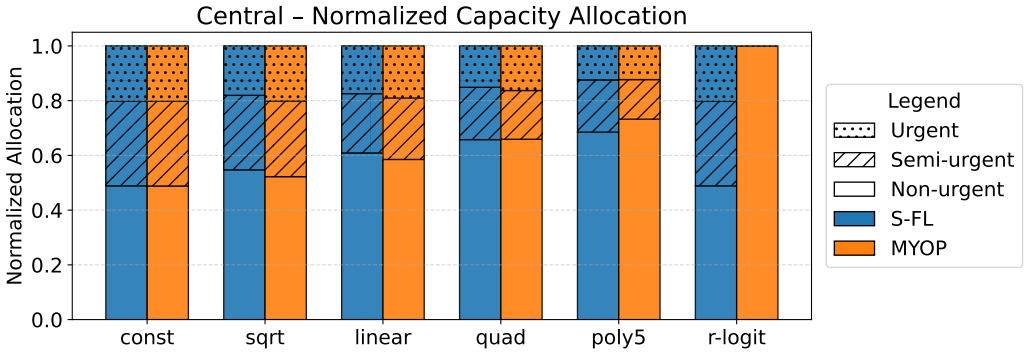


Figure 9 **Normalized capacity allocation for the Central LHIN with $T = 360$. The horizontal axis has the cost functions, while the vertical axis shows normalized capacity allocations across three urgency levels. Urgency levels e distinguished by hatching patterns, while colors differentiate the policies.**

# Clusters	ε Range ($\times 10^4$)	Optimization (Benders)	Enumeration	Cost Gap
2	≤ 2	(CW), (MH, TC, C, CE)	(TC, CE, CW), (C, MH)	0.5%
3	$3, 4$	(C), (CW), (MH, TC, CE)	(C), (TC, CE, CW), (MH)	0.5%
4	$5, 6, 7, 8$	(MH, CE), (CW), (TC), (C)	(TC), (C, CE), (CW), (MH)	1.5%
5	≥ 9	(CE), (CW), (C), (TC), (MH)	(TC), (C), (CE), (CW), (MH)	1.0%

Table 3 Comparison of Clustering Results Across ε Values: Optimization vs. Enumeration

As shown in Table 3, low values of ε yield fewer clusters than LHINs. This indicates that even in over-congested systems, pooling can improve performance by reallocating capacity across heterogeneous subsystems. Because the optimal fluid policy avoids service starvation by allocating capacity to every class, pooling promotes more equitable access to MRI machines by allowing regions with different mixes of urgent and non-urgent patients to better balance their workloads (see Figure EC.4 in Appendix EC.5).

Finally, we evaluate the scalability of the Benders algorithm using a larger experiment that pools fifteen individual hospitals located in several Northern Ontario LHINs. In this case, the algorithm remains applicable when the number of clusters is either small or large. For intermediate configurations (e.g., 6-8 clusters), the cluster with the best objective is identified within a minute, but proving its optimality (i.e., that no better cluster exists) can require runtimes of approximately 10 hours. This is generally acceptable in practice given the infrequency of clustering decisions, and that the procedure can be terminated at any iteration; for more details, see Figure EC.5 in Appendix EC.5 and the accompanying discussion.

7. Conclusions and Managerial Insights

Periods of persistent overload challenge many service systems, and the externalities associated with prolonged delays underscore the need for effective capacity allocation strategies. We study this problem in a multi-class queueing setting where service requests are time-varying, and both the number of queued jobs and their waiting times contribute to accumulating costs. Within this framework, we formulate a finite-horizon optimal control problem that captures system-level congestion. We then derive analytic expressions for the time-dependent proportion of capacity to allocate across classes, which allows us to examine how various cost functions impact performance. We use this fluid model to evaluate the benefits of resource pooling in overloaded environments after developing a solution method based on Benders decomposition.

Methodologically, our work extends prior research on dynamic prioritization (e.g., Van Mieghem, 1995; Gurvich and Whitt, 2009) by incorporating a time-varying version of Little's Law (Kim and Whitt, 2013b) and by explicitly modeling the system state through a cost structure that reflects both waiting times and the number of affected jobs. This formulation induces dynamic prioritization policies that mitigate service starvation, an outcome that is common in systems with static priorities and those that myopically serve the highest-cost individuals. We also show that a commonly used performance metric – the fraction of requests exceeding wait-time targets (Jiang et al., 2023) – is ill-suited for congested systems, as it cannot

discriminate between policies for sufficiently long horizons. Instead, we propose evaluating performance using average waiting times among served jobs, queue lengths at the end of the horizon, and time-averaged capacity allocations. Across these measures, cost-function curvature plays a crucial role in shaping capacity allocation: flatter functions mimic the limitations of static priority rules, whereas steeper functions improve responsiveness to long queues but may diminish the influence of initial urgency classifications. In addition, although dynamic prioritization can reduce excessive waiting for some classes, these gains necessarily result from reallocating capacity, thereby making less capacity available to other classes. While resource pooling can partially alleviate these pressures by increasing effective capacity and reshaping the distribution of demand in ways that can promote more equitable service, its benefits depend on the pooling configuration.

Our analysis highlights the structural consequences of scarcity. Even well-designed rationing policies can improve outcomes only by reallocating capacity; they cannot address fundamental shortfalls. While resource pooling can moderate some of the strain by changing the demand distribution in multi-class settings, nonlinear delay costs mean that indiscriminate pooling configurations may exacerbate imbalances. Thus, in overloaded systems, dynamic prioritization together with targeted pooling should be viewed as stopgap measures for extracting performance gains until meaningful capacity expansion can occur. However, as additional resources come online, it is imperative that decision-makers apply the insights learned in this study to accelerate backlog reduction. Because a return to routine operations will still typically involve operating at utilization levels near one (CBC, 2023), poor capacity management risks entrenching congestion and prolonging overload, precisely the outcome system managers seek to avoid.

References

Abouee-Mehrizi H, Balcıoğlu B, Baron O (2012) Strategies for a centralized single product multiclass M/G/1 make-to-stock queue. *Operations Research* 60(4):803–812.

Afèche P, Baron O, Kerner Y (2013) Pricing time-sensitive services based on realized performance. *Manufacturing & Service Operations Management* 15(3):492–506.

Aggarwal S, Jain P, Jain A (2020) COVID-19 and cataract surgery backlog in Medicare beneficiaries. *Journal of cataract and refractive surgery* .

Ahmadi-Javid A, Hoseinpour P (2022) Convexification of queueing formulas by mixed-integer second-order cone programming: An application to a discrete location problem with congestion. *INFORMS Journal on Computing* 34(5):2621–2633.

Akan M, Ata Bş, Olsen T (2012) Congestion-based lead-time quotation for heterogenous customers with convex-concave delay costs: Optimality of a cost-balancing policy based on convex hull functions. *Operations Research* 60(6):1505–1519.

Anderson BD, Moore JB (2007) *Optimal control: linear quadratic methods* (Courier Corporation).

Anteneodo C (2009) Exact results for the barabási queuing model. *Physical Review E* 80(4):041131.

Armony M, Ward AR (2010) Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research* 58(3):624–637.

Askin RG, Hanumantha GJ (2016) Approximate queueing network analysis for nonstationary demand. *IFAC-PapersOnLine* 49(12):1502–1507.

Avi-Itzhak B, Levy H (2004) On measuring fairness in queues. *Advances in applied probability* 919–936.

Bagchi U, Sullivan RS (1985) Dynamic, non-preemptive priority queues with general, linearly increasing priority function. *Operations Research* 33(6):1278–1298.

Baras J, Ma DJ, Makowski A (1985) K competing queues with geometric service requirements and linear costs: The μc-rule is always optimal. *Systems & Control Letters* 6(3):173–180.

Baron O, Lu T, Wang J (2019) Priority, capacity rationing, and ambulance diversion in emergency departments. *Available at SSRN 3387439* .

Barron Y, Baron O (2022) On dedicated versus pooled service in the presence of triage errors. *Available at SSRN 4147843* .

Bassamboo A, Harrison JM, Zeevi A (2006) Design and control of a large call center: Asymptotic analysis of an lp-based method. *Operations Research* 54(3):419–435.

Bassamboo A, Harrison JM, Zeevi A (2009) Pointwise stationary fluid models for stochastic processing networks. *Manufacturing & Service Operations Management* 11(1):70–89.

Bassamboo A, Randhawa RS (2010) On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations research* 58(5):1398–1413.

Benders J (1962) Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik* 4(1):238–252.

BMJ (2021) COVID-19: Failings in imaging services have put cancer patients at risk, watchdog says. https://www.bmj.com/content/374/bmj.n1749.

Bryson AE (2018) *Applied optimal control: optimization, estimation and control* (Routledge).

Canadian Medical Association (2020) The cost to return wait times to pre-pandemic levels. https://rb.gy/hilyij.

Carr A, Smith JA, Camaradou J, Prieto-Alhambra D (2021) Growing backlog of planned surgery due to covid-19. *Bmj* 372.

CBC (2020) Alberta doctors raise alarm about long waits for MRI and CT scans. https://rb.gy/pdjxr0.

CBC (2023) Ontario expanding number and range of surgeries offered at for-profit clinics. https://rb.gy/5ymjbx.

Chachuat B (2007) Nonlinear and dynamic optimization: From theory to practice. *Automatic Control Laboratory, EPFL, Switzerland* .

Chan TC, Huang SY, Sarhangian V (2025) Dynamic control of service systems with returns: Application to design of postdischarge hospital readmission prevention programs. *Operations Research* 73(4):2242–2263.

Chaudhry S, Choudhary A (1997) Time dependent priority scheduling for guaranteed qos systems. *Proceedings of Sixth International Conference on Computer Communications and Networks*, 236–241 (IEEE).

Cobham A (1954) Priority assignment in waiting line problems. *Journal of the Operations Research Society of America* 2(1):70–76.

Cournane S, Conway R, Creagh D, Byrne D, Sheehy N, Silke B (2016) Radiology imaging delays as independent predictors of length of hospital stay for emergency medical admissions. *Clinical radiology* 71(9):912–918.

Coyle R (1984) A systems approach to the management of a hospital for short-term patients. *Socio-economic planning sciences* 18(4):219–226.

CTV (2021) Almost 16 million medical procedures built up in Ontario pandemic backlog. https://t.ly/RlnsS.

Culyer A, Cullis J (1976) Some economics of hospital waiting lists in the NHS. *Journal of Social Policy* 5(3):239–264.

Czeisler MÉ, Marynak K, Clarke KE, Salah Z, Shakya I, Thierry JM, Ali N, McMillan H, Wiley JF, Weaver MD, et al. (2020) Delay or avoidance of medical care because of covid-19–related concerns—united states, june 2020. *Morbidity and mortality weekly report* 69(36):1250.

Diamant A, Baron O (2019) Double-sided matching queues: Priority and impatient customers. *Operations Research Letters* 47(3):219–224.

Down DG, Lewis ME (2006) Dynamic load balancing in parallel queueing systems: Stability and optimal control. *European Journal of Operational Research* 168(2):509–519.

Elhedhli S (2006) Service system design with immobile servers, stochastic demand, and congestion. *Manufacturing & Service Operations Management* 8(1):92–97.

Estrada R, Pavlović M (2017) L'hopital's monotone rule, gromov's theorem, and operations that preserve the monotonicity of quotients. *Publications de l'Institut Mathematique* 101(115):11.

Evler J, Schultz M, Fricke H, Cook A (2022) Stochastic delay cost functions to estimate delay propagation under uncertainty. *IEEE Access* 10:21424–21442.

Fayolle G, King PJ, Mitrani I (1982) The solution of certain two-dimensional markov models. *Advances in applied probability* 14(2):295–308.

Fraser Institute (2014) The effect of wait times on mortality in canada. https://t.ly/-Nw30.

Frederickson GN (1983) Scheduling unit-time tasks with integer release times and deadlines. *Information Processing Letters* 16(4):171–173.

Gavirneni S, Kulkarni VG (2016) Self-selecting priority queues with burr distributed waiting costs. *Production and Operations Management* 25(6):979–992.

Gerber HU, Pafum G (1998) Utility functions: from risk theory to finance. *North American Actuarial Journal* 2(3):74–91.

Gómez-Corral A, Krishnamoorthy A, Narayanan VC (2005) The impact of self-generation of priorities on multi-server queues with finite capacity. *Stochastic models* 21(2-3):427–447.

Green L, Kolesar P (1991) The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science* 37(1):84–97.

Green L, Kolesar P, Svoronos A (1991) Some effects of nonstationarity on multiserver markovian queueing systems. *Operations Research* 39(3):502–511.

Gurvich I, Whitt W (2009) Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing & Service Operations Management* 11(2):237–253.

Habtamu E, Burton M (2015) Clearing the trichiasis backlog: experiences in amhara, ethiopia. *Community Eye Health* 28(90):38.

Habtamu E, Rajak SN, Gebre T, Zerihun M, Genet A, Emerson PM, Burton MJ (2011) Clearing the backlog: trichiasis surgeon retention and productivity in northern ethiopia. *PLoS neglected tropical diseases* 5(4):e1014.

Harrison JM, Zeevi A (2004) Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Operations Research* 52(2):243–257.

Hartl RF, Sethi SP, Vickson RG (1995) A survey of the maximum principles for optimal control problems with state constraints. *SIAM review* 37(2):181–218.

Hu Y, Chan CW, Dong J (2022) Optimal scheduling of proactive service with customer deterioration and improvement. *Management Science* 68(4):2533–2578.

Jackson JR (1960) Some problems in queueing with dynamic priorities. *Naval Research Logistics Quarterly* 7(3):235–249.

Jain A, Dai T BK, Myers C (2020) COVID-19 created an elective surgery backlog: how can hospitals get back on track. *Harvard Business Review* 10.

Jaiswal NK (1968) *Priority queues*, volume 50 (Academic press New York).

Jeffay K, Stanat DF, Martel CU (1991) On non-preemptive scheduling of periodic and sporadic tasks. *IEEE real-time systems symposium*, 129–139 (US: IEEE).

Jiang Y, Abouee-Mehrizi H, Diao Y (2020) Data-driven analytics to support scheduling of multi-priority multi-class patients with wait time targets. *European Journal of Operational Research* 281(3):597–611.

Jiang Y, Abouee Mehrizi H, Van Mieghem JA (2023) Geographic virtual pooling of hospital resources: Data-driven trade-off between waiting and traveling. *Manufacturing & Service Operations Management* .

Kelly M, Skorin-Kapov D, Skorin-Kapov J (1997) Lower bounds for the hub location problem. *Location Science* 1(5):60.

Kim SH, Whitt W (2013a) Estimating waiting times with the time-varying Little's law. *Probability in the Engineering and Informational Sciences* 27(4):471–506.

Kim SH, Whitt W (2013b) Statistical analysis with Little's law. *Operations Research* 61(4):1030–1045.

Kim YJ, Mannino MV (2003) Optimal incentive-compatible pricing for m/g/1 queues. *Operations Research Letters* 31(6):459–461.

Kleinrock L (1964) A delay dependent queue discipline. *Naval Research Logistics Quarterly* 11(3-4):329–341.

Krishnamoorthy A, Babu S, Narayanan VC (2008) MAP/(PH/PH)/c queue with self-generation of priorities and non-preemptive service. *Stochastic Analysis and Applications* 26(6):1250–1266.

Krishnamoorthy A, Narayanan VC (2003) On a queueing system with self generation of priorities. *Stochastic Point Processes* 212–217.

Lagzi S, Quiroga BF, Romero G, Howard N, Chan TC (2023) Negative externality on service level across priority classes: Evidence from a radiology workflow platform. *Journal of Operations Management* 69(8):1257–1281.

Latouche G, Ramaswami V (1999) *Introduction to matrix analytic methods in stochastic modeling* (SIAM).

Lejeune MA, Margot F (2025) Response time minimization for cardiac arrests. *Production and Operations Management* 34(9):2618–2640.

Levy H, Markowitz HM (1979) Approximating expected utility by a function of mean and variance. *The American Economic Review* 69(3):308–317.

Li N, Stanford DA (2016) Multi-server accumulating priority queues with heterogeneous servers. *European Journal of Operational Research* 252(3):866–878.

Little JD (2011) OR FORUM—Little's law as viewed on its 50th anniversary. *Operations Research* 59(3):536–549.

Mandelbaum A, Massey WA, Reiman MI (1998) Strong approximations for Markovian service networks. *Queueing Systems* 30(1-2):149–201.

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.

Manta A, O'Grady J, Bleakney R, Theodoropoulos J (2019) Determining the appropriateness of requests for outpatient magnetic resonance imaging of the hip. *Canadian Journal of Surgery* 62(4):224.

Marand AJ, Hoseinpour P (2024) A congested facility location problem with strategic customers. *European Journal of Operational Research* 318(2):442–456.

Medicine Net (2020) How delayed cancer care is costing lives. https://rb.gy/8qjxua.

Mendelson H, Whang S (1990) Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations research* 38(5):870–883.

Netterman A, Adiri I (1979) A dynamic priority queue with general concave priority functions. *Operations Research* 27(6):1088–1100.

Office of the Auditor General of Ontario (2018) Annual report. https://t.ly/0B_yk.

Ontario Association of Radiologists (2014) Wait times affect patient care in ontario. https://t.ly/TTzyL.

Osogami T, Wierman A, Harchol-Balter M, Scheller-Wolf A (2004) A recursive analysis technique for multi-dimensionally infinite markov chains. *ACM SIGMETRICS Performance Evaluation Review* 32(2):3–5.

Pattara-Aukom W, Banerjee S, Krishnamurthy P (2002) Starvation prevention and quality of service in wireless LANs. *The 5th International Symposium on Wireless Personal Multimedia Communications*, volume 3, 1078–1082 (IEEE).

Pender J, Rand RH, Wesson E (2017) Queues with choice via delay differential equations. *International Journal of Bifurcation and Chaos* 27(04):1730016.

Pratt JW (1978) Risk aversion in the small and in the large. *Uncertainty in Economics*, 59–79 (Elsevier).

Rahmaniani R, Crainic TG, Gendreau M, Rei W (2017) The Benders decomposition algorithm: A literature review. *European Journal of Operational Research* 259(3):801–817.

Robbins HM (1967) A generalized legendre-clebsch condition for the singular cases of optimal control. *IBM Journal of Research and Development* 11(4):361–372.

Sandmann W (2006) Analysis of a queueing fairness measure. *13th GI/ITG Conference-Measuring, Modelling and Evaluation of Computer and Communication Systems*, 1–13 (VDE).

Sharif AB, Stanford DA, Taylor P, Ziedins I (2014) A multi-class multi-server accumulating priority queue with application to health care. *Operations Research for Health Care* 3(2):73–79.

Smith W (1956) Various optimizers for single-stage production. *Naval Res. Logist* 3.

Sonneborn L, Van Vleck F (1964) The bang-bang principle for linear control systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control* 2(2):151–159.

Stanford DA, Taylor P, Ziedins I (2014) Waiting time distributions in the accumulating priority queue. *Queueing Systems* 77(3):297–330.

Stephens K (2020) Every month delayed in cancer treatment can raise risk of death by around 10%. *AXIS Imaging News* .

Sun B, Lee MH, Dudin AN, Dudin SA (2014) Queueing system with absolute priority and reservation of servers. *Mathematical Problems in Engineering* 2014.

Teymoori P, Sohraby K, Kim K (2015) A fair and efficient resource allocation scheme for multi-server distributed systems and networks. *IEEE Transactions on Mobile Computing* 15(9):2137–2150.

Tsoularis A, Wallace J (2002) Analysis of logistic growth models. *Mathematical biosciences* 179(1):21–55.

Van Mieghem JA (1995) Dynamic scheduling with convex delay costs: The generalized $c\mu$-rule. *The Annals of Applied Probability* 809–833.

Vanberkel PT, Boucherie RJ, Hans EW, Hurink JL, Litvak N (2012) Efficiency evaluation for pooling resources in health care. *OR spectrum* 34:371–390.

vanDijk NM, van derSluis E (2008) To pool or not to pool in call centers. *Production and Operations Management* 17(3):296–305.

Wang J, Baron O, Scheller-Wolf A (2015) M/m/c queue with two priority classes. *Operations Research* 63(3):733–749.

Wang Q (2004) Modeling and analysis of high risk patient queues. *European Journal of Operational Research* 155(2):502–515.

Wolff RW (2011) Little's law and related results. *Wiley encyclopedia of operations research and management science* 4:2828–2841.

Wolff RW, Yao YC (2014) Little's law when the average waiting time is infinite. *Queueing Systems* 76(3):267–281.

Wolstenholme EF (1993) A case study in community care using systems thinking. *Journal of the Operational Research Society* 44(9):925–934.

Yom-Tov GB, Yedidsion L, Xie Y (2021) An invitation control policy for proactive service systems: Balancing efficiency, value, and service level. *Manufacturing & Service Operations Management* 23(5):1077–1095.

Yu L, Iravani S, Perry O (2025) Asymptotically optimal clearing control of backlogs in multiclass processing systems. *Operations Research* 73(4):2061–2078.

Zelikin MI, Borisov VF (2012) *Theory of chattering control: with applications to astronautics, robotics, economics, and engineering* (Springer Science & Business Media).

Zychlinski N (2023) Applications of fluid models in service operations management. *Queueing Systems* 103(1):161–185.

Appendices and Proofs of Statements

EC.1. Capacity Allocation Algorithm

Input: Define the set $\mathcal{J} = \emptyset$, time step $\Delta > 0$, and time horizon $T > \Delta$.

Initialization: At time $t = 0$, assign the full amount of capacity to a class $i \in \mathcal{I}$ if $\mu_i \dot{\pi}_i(0) \leq \mu_j \dot{\pi}_j(0)$ for all $j \neq i$. Suppose, without loss of generality, that $i = 1$. Then, set $\mathcal{J} = \mathcal{J} \cup \{1\}$.

while $t \leq T$ **do**

(a) Observe the current queue lengths at time t, which is denoted by $\tilde{y}_i(t)$ for all i.

(b) Using $\tilde{y}_i(t)$, determine whether $\mu_1 \dot{\pi}_1(t) = \mu_i \dot{\pi}_i(t)$ for all $i \in \mathcal{I} \setminus \mathcal{J}$.

if this equation holds **then**

Update $\mathcal{J} = \mathcal{J} \cup \{i\}$.

(c) Compute the optimal capacity allocation policy for all $j \in \mathcal{J}$ where $\phi_j^*(t) > 0$ by solving a linear system $\boldsymbol{A}(t)\boldsymbol{\phi}(t) = \boldsymbol{b}(t)$ defined in (EC.2), which contains $|\mathcal{J}|$ equations with $|\mathcal{J}|$ unknowns.

(d) Implement the optimal allocation $\boldsymbol{\phi}(t)$ over all classes $i \in \mathcal{I}$.

if any class i attains an empty queue after allocation **then**

Distribute all remaining capacity (if any) to the non-empty class with largest $\gamma_i \mu_i$.

(e) Increment t by setting $t = t + \Delta$.

EC.2. Case Study Parameters

Table EC.1 summarizes the notation for patient types. We present the list of body parts in Column 1, as distinguished by the Ministry of Health (Office of the Auditor General of Ontario, 2018). Columns 2-4 assign the corresponding type $i \in \mathcal{I}$ to a level of priority, i.e., urgent, semi-urgent, and non-urgent, respectively. For instance, semi-urgent patients who require a brain MRI would be denoted by type $i = 15$.

For the arrival function, we fit a time-dependent linear model using the provided dataset such that

$$\lambda_i(t) = b_0 + b_1 t.$$

We present parameters b_0 and b_1 in Tables EC.2 and EC.3, respectively, per geographical region and patient type: Urgent (U), Semi-Urgent (SU), and Non-Urgent (NU). Notice that the estimated slope parameters support the validity of using the time-varying approximation to Little's Law described in Section 3. Furthermore, some patient classes are absent in certain regions (i.e., the arrival rate is zero).

Table EC.1 Summary of Patient Types

Body Part	Urgent Indices	Semi-Urgent Indices	Non-Urgent Indices
Abdomen	1	11	21
Breast	2	12	22
Cardiac	3	13	23
Extremities	4	14	24
Head (Brain)	5	15	25
Head and Neck	6	16	26
Pelvis	7	17	27
Peripheral Vascular	8	18	28
Spine	9	19	29
Thorax	10	20	30

Table EC.2 Arrival Function: Slope Estimates

Patient Type	**Abdomen**	**Breast**	**Cardiac**	**Extremities**	**Head (Brain)**	**Head and Neck**	**Pelvis**	**Peripheral Vascular**	**Spine**	**Thorax**
Toronto Central										
	1	2	3	4	5	6	7	8	9	10
U	0.002	0.000	0.000	0.000	0.008	0.000	0.000	0.000	0.004	0.000
SU	0.008	0.002	0.000	0.000	0.009	0.002	0.003	0.000	0.004	0.000
NU	0.006	0.008	0.006	0.010	0.030	0.002	0.005	0.000	0.016	0.000
Central										
	1	2	3	4	5	6	7	8	9	10
U	0.001	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.001	0.000
SU	0.000	0.000	0.000	0.000	0.001	0.000	0.001	0.000	0.000	0.000
NU	0.005	0.002	0.000	0.017	0.012	0.006	0.003	0.000	0.015	0.000
Central East										
	1	2	3	4	5	6	7	8	9	10
U	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000
SU	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.001	0.000
NU	0.004	0.003	0.000	0.011	0.014	0.001	0.002	0.000	0.013	0.000
Central West										
	1	2	3	4	5	6	7	8	9	10
U	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.001	0.000	0.000
SU	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000
NU	0.001	0.001	0.000	0.012	0.008	0.000	0.002	0.012	0.000	0.000
Mississauga Halton										
	1	2	3	4	5	6	7	8	9	10
U	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000
SU	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000
NU	0.005	0.002	0.002	0.005	0.008	0.003	0.002	0.000	0.010	0.000

Table EC.3 Arrival Function: Intercept Estimates

Patient Type	**Abdomen**	**Breast**	**Cardiac**	**Extremities**	**Head (Brain)**	**Head and Neck**	**Pelvis**	**Peripheral Vascular**	**Spine**	**Thorax**
Toronto Central										
	1	2	3	4	5	6	7	8	9	10
U	3.619	1.171	1.860	2.231	14.532	2.736	1.399	0.995	6.254	1.124
SU	16.024	4.337	1.690	20.687	18.848	7.105	3.413	1.355	8.783	1.291
NU	23.794	15.074	12.929	49.510	87.688	7.243	9.682	2.390	50.957	2.227
Central										
	1	2	3	4	5	6	7	8	9	10
U	1.685	1.115	1.074	1.331	5.316	1.436	1.112	1.072	2.299	1.057
SU	3.022	2.034	1.000	3.968	5.645	3.149	2.094	1.086	5.042	1.062
NU	7.476	2.706	1.522	70.488	47.274	4.322	5.687	1.622	53.131	1.513
Central East										
	1	2	3	4	5	6	7	8	9	10
U	1.844	1.031	0.000	1.279	5.340	1.067	1.041	0.000	2.261	0.985
SU	4.727	2.204	1.908	3.172	11.116	1.629	2.599	1.043	4.173	1.214
NU	9.193	4.090	1.765	58.629	48.547	2.012	4.665	1.059	52.327	1.473
Central West										
	1	2	3	4	5	6	7	8	9	10
U	1.497	0.000	0.000	1.170	2.530	1.745	1.044	1.569	0.958	0.000
SU	3.084	1.226	0.000	3.644	3.821	1.429	2.151	3.013	1.074	0.000
NU	4.308	1.252	1.024	28.479	15.443	2.509	2.112	22.658	1.142	0.000
Mississauga Halton										
	1	2	3	4	5	6	7	8	9	10
U	1.535	0.000	1.020	1.130	3.644	1.084	1.030	0.000	1.757	1.000
SU	2.463	1.253	0.000	2.741	4.391	1.303	2.148	0.000	2.562	1.149
NU	9.076	2.714	0.903	46.964	34.485	5.539	4.249	1.511	40.782	1.348

Table EC.4 shows the service rates for each patient class, set at the 75th percentile of their distribution. We exclude outliers as well as patients whose appointments contain multiple MRI procedures.

Table EC.4 Service Rates

Patient Type	**Abdomen**	**Breast**	**Cardiac**	**Extremities**	**Head (Brain)**	**Head and Neck**	**Pelvis**	**Peripheral Vascular**	**Spine**	**Thorax**
Toronto Central										
	1	2	3	4	5	6	7	8	9	10
U	20.400	24.158	13.701	19.957	26.229	17.654	15.300	12.750	22.390	15.300
SU	20.400	22.950	13.701	20.400	26.229	17.654	20.400	12.750	22.950	15.300
NU	20.400	22.950	13.701	27.000	26.229	19.957	20.400	15.300	22.950	10.800
Central										
	1	2	3	4	5	6	7	8	9	10
U	20.400	22.950	15.300	20.400	20.400	20.400	20.400	91.800	20.400	15.300
SU	20.400	15.300	91.800	20.400	20.400	20.400	15.300	91.800	20.400	17.486
NU	20.400	20.400	10.200	30.600	30.600	30.600	20.400	20.400	30.600	15.300
Central East										
	1	2	3	4	5	6	7	8	9	10
U	22.950	19.305	91.800	20.400	30.600	20.400	20.400	30.600	20.400	7.650
SU	22.950	18.360	11.475	20.400	30.600	20.400	20.400	30.600	20.400	20.400
NU	20.400	20.400	11.475	22.950	30.600	20.400	20.400	30.600	30.600	15.300
Central West										
	1	2	3	4	5	6	7	8	9	10
U	20.400	91.800	20.400	20.400	30.600	20.400	20.400	91.800	30.600	20.400
SU	20.400	15.300	20.400	20.400	20.400	20.400	20.400	91.800	30.600	20.400
NU	20.400	15.300	20.400	20.400	20.400	20.400	20.400	91.800	30.600	20.400
Mississauga Halton										
	1	2	3	4	5	6	7	8	9	10
U	30.600	91.800	30.600	30.600	30.600	30.600	30.600	91.800	30.600	30.600
SU	20.400	30.600	91.800	20.400	20.400	20.400	20.400	15.300	20.400	20.400
NU	20.400	30.600	15.300	30.600	30.600	30.600	20.400	30.600	30.600	20.400

EC.3. Proofs

The proofs of the statements are as follow:

Proof of Proposition 1:

Proof. Define the instantaneous cost as $J(t,\boldsymbol{\phi}(t)) \coloneqq \sum_{i=1}^{I} g_i\left(\frac{\lambda_i(t)}{y_i(t)}\right) y_i(t)$ and suppose that $\boldsymbol{\phi}^*(t)$ is the capacity allocation policy that optimizes FL. We now show that, by contradiction, $y_i(t) \geq \phi_i^*(t)m$ for all i and t. Assume that under the optimal policy $\boldsymbol{\phi}^*(t)$, there exists a time t' (or a contiguous time interval without loss of generality) where $y_i(t') < \phi_i^*(t')m$ for some $i \in \mathcal{I}$. Under Assumption 1, $\sum_i y_i(t') > m$ for all t which implies that at time t', there is both idle capacity and a queue of patients from several classes $i' \neq i$ for $i, i' \in \mathcal{I}$ waiting for service.

Now consider a policy where (i) $\tilde{\phi}_i(t) \coloneqq \phi_i^*(t)$ for all $t \neq t'$; and (ii) at time t', capacity is assigned to patient classes i and $i' \neq i$ such that $y_i(t') = \tilde{\phi}_i(t')m$ and $\tilde{\phi}_{i'}(t') = \phi_{i'}^*(t') + m\left(\phi_i^*(t') - \tilde{\phi}_i(t')\right)$ where we assume that for class i', $y_{i'}(t') \geq \tilde{\phi}_{i'}(t')m \geq 0$, which must exist under Assumption 1. Notice that by using $\tilde{\phi}_i(t)$, there must be less idle capacity because we originally had $y_i(t') < \phi_i^*(t')m$. As a consequence, with a slight abuse of notation, it follows that $\sum_i y_i(t', \tilde{\phi}_i(t')) < \sum_i y_i(t', \phi_i^*(t'))$ and $\sum_i \mu_i(y_i(t, \phi_i^*(t')) \wedge \phi_i^*(t')m) < \sum_i \mu_i \tilde{\phi}_i(t')m$ as throughput is greater when more patients are admitted to service. Moreover, $J(t', \boldsymbol{\phi}^*(t')) > J(t', \tilde{\boldsymbol{\phi}}(t'))$ because we assume the cost function $g_i(\cdot)$ is increasing in $y_i(t)$ for all i, and is at its minimum when $y_i(t) = 0$. Next, select $t'' > t'$ where the condition $y_i(t'') < \phi_i^*(t'')m$ is satisfied for some i. If there is no such time in $[0,T]$, then we set $t'' = T$. It follows that $J(t, \boldsymbol{\phi}^*(t)) = J(t, \tilde{\boldsymbol{\phi}}(t))$ for all $t < t'$ and $J(t, \boldsymbol{\phi}^*(t)) \geq J(t, \tilde{\boldsymbol{\phi}}(t))$ for all $t'' > t \geq t'$. Thus, we have that $\int_0^{t''} J(u, \boldsymbol{\phi}^*(u))du > \int_0^{t''} J(u, \tilde{\boldsymbol{\phi}}(u))du$. By induction, this argument is extended to any collection of non-contiguous time intervals on $[0,T]$ if $t'' < T$. We conclude that $\phi_i^*(t)$ cannot be optimal and, by contradiction, $y_i(t) \geq \phi_i^*(t)m$ for all i and t. If follows that the optimal solutions to FL and S-FL are identical. □

Proof of Theorem 1:

We first provide a high-level outline of the proof, followed by a formal development of each step. The argument proceeds in two stages. First, we show that an optimal policy is bang–bang with possible singular arcs, characterize its behavior on non-singular arcs up to the first switching time, and rule out chattering solutions. Second, we characterize the policy immediately after the first switching time and extend the analysis from the two-class case to the general case with $I \geq 2$ classes. More specifically:

1. *Strict priority rule.* On any non-singular arc, the Hamiltonian optimization problem has a unique minimizer. Because $\sum_i \phi_i(t) = 1$, this implies that all capacity is allocated to the class with the largest value of $\mu_i \pi_i(t)$, yielding strict priority. However, this policy is optimal only on non-singular arcs; it remains to characterize the optimal policy on singular arcs.
2. *Adjoint dynamics.* We apply Pontryagin's minimum principle to obtain the adjoint equation for each class i with terminal condition $\pi_i(T) = 0$. We show that the adjoint dynamics satisfy $\dot{\pi}_i(t) \leq 0$ for all $t \in [0,T]$. Since the adjoint is solved backward in time from T, this implies that $\pi_i(t) \geq 0$ on $[0,T]$.

3. *Initial capacity allocation.* We perform a first-order expansion of the cost over a small interval near $t = 0$. This yields the initial bang-bang choice: set $\phi_i(0) = 1$ for the class with the smallest $\mu_i \dot{\pi}_i(0)$.
4. *No chattering.* On a singular arc supported on a subset $\mathcal{J}$ of classes, we analyze switching functions and their derivatives, i.e., $\mu_j \pi_j(t) = \mu_{j'} \pi_{j'}(t)$, $\mu_j \dot{\pi}_j(t) = \mu_{j'} \dot{\pi}_{j'}(t)$ for $j, j' \in \mathcal{J}$, which uniquely determines the singular control along the arc. Hence, the singular extremal is of intrinsic order one. By standard results for first-order singular arcs (Chachuat, 2007), this rules out the accumulation of switching times and therefore excludes chattering, i.e., infinitely many switches over a finite interval.
5. *Sequential activation for $I = 2$ classes.* Define $\sigma(t) := \mu_1 \pi_1(t) - \mu_2 \pi_2(t)$ and let τ be the first time that $\sigma(\tau) = 0$ (enters a singular arc). For sufficiently small $\delta > 0$, we compare cost functions on $[\tau, \tau + \delta]$ between two policies: (i) a bang–bang control that allocates all capacity to a single class; and (ii) a singular control that shares capacity satisfying the singular-arc conditions $\sigma(t) = \dot{\sigma}(t) = 0$. Using the convexity of $y_i \, g_i(y_i/\lambda_i)$, we show the singular control yields lower cost over this interval.
6. *Extension to $I \geq 2$ classes.* We rewrite the Hamiltonian by substituting $\phi_1(t) = 1 - \sum_{i \geq 2} \phi_i(t)$, so that each control $\phi_i(t)$ for $i \geq 2$ enters linearly with coefficient $\mu_1 \pi_1(t) - \mu_i \pi_i(t)$. On any singular arc supported on an active set $\mathcal{J}$, the optimality conditions imply $\mu_j \pi_j(t) = \mu_{j'} \pi_{j'}(t)$ for all $j, j' \in \mathcal{J}$. Moreover, differentiating the switching conditions yields $\mu_j \dot{\pi} j(t) = \mu j' \dot{\pi}_{j'}(t)$ along the arc, so the active classes behave as an aggregate set. The same local comparison argument then implies such that there exists an inactive class that becomes active at the next switching time. Because of this "no-exit property" (i.e., once a class is active it remains active), this yields monotone expansion of $\mathcal{J}$ over time.

Proof. The proof proceeds in two stages. In the first stage, we establish the existence of an optimal policy consisting of a bang-bang control with singular arcs, i.e., which are solutions in which the optimality conditions on the Hamiltonian function are degenerate. We then characterize the form of this policy for non-singular arcs and derive the initial capacity allocation. This establishes the optimality of the bang–bang policy up to the first switching time and shows that the optimal policy excludes chattering solutions. In the second stage, we derive the optimal policy for a two-class system immediately after the first switching time, at $\tau + \delta$ for small $\delta > 0$, and then show that this structure extends to systems with $I \geq 2$ patient classes.

Strict Priority: Recall that type-i arrivals follow a non-homogeneous Poisson process with intensity rate $\lambda_i(t) > 0$, which we assume is everywhere continuous for each i and t. This implies $\lambda_i(t)$ is measurable and integrable. Further, we are searching for an optimal policy over a compact interval $[0, T]$ and assume that $\phi(t)$ is also measurable and integrable. Suppose, utilizing the full range of controls $\phi(t) \in [0, 1]$ for all i, that $\mathcal{K}_i$ is the set of all endpoints corresponding to the value of $x_i(t)$ at time T. It follows from Sonneborn and Van Vleck (1964) that the set $\mathcal{K}_i$ is not only compact and convex, but is identical to the set $\tilde{\mathcal{K}}_i$ which is the set of all endpoints corresponding to the value of $x_i(t)$ at time T using controls $\phi(t) \in \{0, 1\}$. Nevertheless, as the Hamiltonian is linear in the control variables $\phi_i(t)$ but nonlinear in the state variables $y_i(t)$, singular

arcs exists (see, for instance, see Chapter 8 of Bryson, 2018). Thus, the optimal policy is a bang-bang control with singular arcs and, without loss of generality, we restrict our attention to this policy class.

We now characterize the conditions determining which class is prioritized in each non-singular arc, i.e., for $t \in [0, \tau]$. From the adjoint equation associated with Pontryagin's minimum principle, we have

$$\dot{\boldsymbol{\pi}}(t) = -\frac{\partial \mathcal{H}(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\pi}, t)}{\partial \boldsymbol{y}},$$

which gives for every $i \in \mathcal{I}$,

$$\dot{\pi}_i(t) = -\left(g_i\left(\frac{y_i(t)}{\lambda_i(t)}\right) + \frac{y_i(t)}{\lambda_i(t)} g_i'\left(\frac{y_i(t)}{\lambda_i(t)}\right)\right). \tag{EC.1}$$

Notice that $\dot{\pi}_i(t) \leq 0$ for all i and t, i.e., the costate variable $\pi_i(t)$ is monotonically decreasing. This follows from Proposition 1, $y_i(t) \geq 0$ and by assumption, $\lambda_i(t) > 0$ for all i and t, and because $g_i(z)$ is continuously increasing and non-negative for any $z \geq 0$. Thus, it follows that the right-hand-side of (EC.1) is non-positive. Using the definition of $\pi_i(t)$ and the transversality condition

$$\pi_i(t) = \int_t^T \dot{\pi}_i(t)\, dt \text{ where } \pi_i(T) = 0.$$

Due to the monotonicity property of integrals and accounting for the initial condition where $\pi_i(T) = 0$, $\pi_i(t) \geq 0$ for every i over $t \in [0, T]$. According to Pontryagin's principal of minimality, in order to minimize the Hamiltonian along bang-bang control arcs, it suffices to set the control variable $\phi_i(t) = 1$ if $\mu_i \pi_i(t) > \mu_j \pi_j(t)$ for all $j \neq i \in \mathcal{I}$, which is the condition given in the statement of the Proposition.

To determine the initial capacity allocation, notice that class i is prioritized when $\mu_i \dot{\pi}_i(0) < \mu_j \dot{\pi}_j(0)$ for $j \neq i$. Assume that at time $t = 0$, there exists an i such that $g_i\left(\frac{y_i(0)}{\lambda_i(0)}\right) y_i(0) \geq g_j\left(\frac{y_j(0)}{\lambda_j(0)}\right) y_j(0)$ for all $j \neq i$ where the inequality is strict for at least one patient class. If this were not the case, then a bang-bang policy would not be optimal. Now consider a sufficiently small interval around $t = 0$. Using a perturbation of the cost function $g_i\left(\frac{y_i(t)}{\lambda_i(t)}\right) y_i(t)$ for each i gives

$$\begin{aligned}
&g_i\left(\frac{y_i(0)}{\lambda_i(0)}\right) y_i(0) + \left(g_i\left(\frac{y_i(0)}{\lambda_i(0)}\right) y_i'(0) + \frac{y_i(0)}{\lambda_i(0)} g_i'\left(\frac{y_i(0)}{\lambda_i(0)}\right)\left(y_i'(0) - \frac{y_i(0)}{\lambda_i(0)}\lambda_i'(0)\right)\right) t \\
&= g_i\left(\frac{y_i(0)}{\lambda_i(0)}\right)(y_i(0) + \lambda(0)t) + \frac{y_i(0)}{\lambda_i(0)} g_i'\left(\frac{y_i(0)}{\lambda_i(0)}\right)\left(\lambda_i(0) - \frac{y_i(0)}{\lambda_i(0)}\lambda_i'(0)\right) t + \mu_i m \dot{\pi}_i(0)\phi_i(0)t,
\end{aligned}$$

where the equality is obtained by substituting $\dot{y}_i(t)$ and simplifying. Notice that the only term containing the control variable is $\mu_i m \dot{\pi}_i(0)\phi_i(0)t$. Thus, to minimize the objective function over a sufficiently small interval around $t = 0$, we should initially select a bang-bang control that minimizes $\sum_{i=1}^I \mu_i m \dot{\pi}_i(0)\phi_i(0)t$. It follows that $\phi_i(0) = 1$ if $\mu_i \dot{\pi}_i(0) < \mu_j \dot{\pi}_j(0)$ and $\phi_j(0) = 0$ for $j \neq i$.

Chattering: To demonstrate the absence of chattering solutions, consider the part of the Hamiltonian $\mathcal{H}(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\pi}, t)$ that explicitly contains the co-state variables, i.e., $\sum_{i=1}^I \pi_i(t)\dot{y}_i(t)$. Suppose there exists a

singular arc associated with all classes $j \in \mathcal{J}$ where $\mathcal{J} \subseteq \mathcal{I}$. This means that $j \in \mathcal{J}$ if and only if $0 < \phi_j(t) < 1$ and $\sum_{j\in\mathcal{J}} \phi_j(t) = 1$. As is apparent, for all $i \notin \mathcal{J}$, $\phi_i(t) = 0$. Along a singular arc, we have that for $j, j' \in \mathcal{J}$, $\mu_j \pi_j(t) = \mu_j \pi_{j'}(t)$. It also follows that $\mu_j \dot{\pi}_j(t) = \mu_j \dot{\pi}_{j'}(t)$ and $\mu_j \ddot{\pi}_j(t) = \mu_j \ddot{\pi}_{j'}(t)$ for $j, j' \in \mathcal{J}$ (see Chapter 3.5.4 in Chachuat, 2007). The latter relation contains $\dot{y}_j(t)$ and $\dot{y}_{j'}(t)$ which means that controls $\phi_j(t)$ and $\phi_{j'}(t)$ enter linearly into $\mu_j \ddot{\pi}_j(t) = \mu_j \ddot{\pi}_{j'}(t)$ as function of system states $y_j(t)$ and $y_{j'}(t)$. Thus, we can construct a linear system with J variables and J unknowns to compute the optimal rationing policy for all $\phi_j(t)$ where $j \in \mathcal{J}$. Since this holds for every subset $\mathcal{J}$ and $\forall t$, the intrinsic order of any singular solution is one (Zelikin and Borisov, 2012) implying no chattering behavior (Chachuat, 2007).

Sequential Activation and Capacity Sharing: We demonstrate that the optimal policy exhibits a *sequential activation property*; capacity is shared amongst additional patient classes after switching times. To proceed, we first analyze the case where $I = 2$. As a consequence of Proposition 1, the Hamiltonian is:

$$\begin{aligned}\mathcal{H}(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\pi}, t) = & g_1\left(\frac{y_1(t)}{\lambda_1(t)}\right) y_1(t) + g_2\left(\frac{y_2(t)}{\lambda_2(t)}\right) y_2(t) + \pi_1(t)(\lambda_1(t) - \mu_1\phi_1(t)m) \\ & + \pi_2(t)(\lambda_2(t) - \mu_2\phi_2(t)m) \\ = & g_1\left(\frac{y_1(t)}{\lambda_1(t)}\right) y_1(t) + g_2\left(\frac{y_2(t)}{\lambda_2(t)}\right) y_2(t) + \lambda_1(t)\pi_1(t) + \lambda_2(t)\pi_2(t) \\ & - m\mu_1\pi_1(t) + m\left(\mu_1\pi_1(t) - \mu_2\pi_2(t)\right)\phi_2(t).\end{aligned}$$

We note that for simplicity, we leave the formulation of the problem in the non-autonomous form noting that we can transform it into an autonomous system and then apply Pontryagin's miniumum principle on this new problem (see Theorem 3.30 in Chachuat, 2007). For each $i = \{1, 2\}$, the adjoint equation is,

$$\dot{\pi}_i(t) = -\left(g_i\left(\frac{y_i(t)}{\lambda_i(t)}\right) + \frac{y_i(t)}{\lambda_i(t)} g_i'\left(\frac{y_i(t)}{\lambda_i(t)}\right)\right).$$

Define the switching function as $\sigma(t) := \mu_1\pi_1(t) - \mu_2\pi_2(t)$ and let $\tau > 0$ be the first time point where $\sigma(\tau) = 0$ (switching time). From the first stage, we know that optimal policy has a bang-bang structure until $t = \tau$ altough singular arcs may exist. If $\tau > T$, one class is assigned capacity for all $t \in [0, T]$ and the bang-bang control is optimal. Thus, suppose that $\tau \in [0, T]$. We now show that for all $\tau \leq t \leq T$, a bang-bang policy is *suboptimal* which implies the optimality of the singular extremal, i.e., that capacity is shared. Moreover, assume that, without loss of generality, class one patients are initially prioritized.

Consider a small time period δ immediately following τ. During $t \in [\tau, \tau + \delta]$, under the bang-bang control policy, we have switched from exclusively prioritizing class one patients (by assumption) to exclusively prioritizing class two patients. As a result, $\sigma(t) < 0$ during this period as $\phi_1(t) = 0$ and $\phi_2(t) = 1$, implying $\mu_2\pi_2(t) > \mu_1\pi_1(t) \geq 0$. Under the singular control policy, both patient classes are served simultaneously over $t \in [\tau, \tau + \delta]$. This implies that $\sigma(t) = \dot{\sigma}(t) = 0$ over this period (see Chapter 3.5.4 in Chachuat, 2007). Further, define $\alpha(t)$ as the proportion of capacity assigned to class one patients for the singular control.

Since we analyze a small time period $t \in [\tau, \tau+\delta]$, the optimal control is given by the stationary solution $\alpha(t) = \alpha$ which implies that the proportion of capacity assigned to class two patients is $1-\alpha$.

We now compute the cost difference between these two policies over $t \in [\tau, \tau+\delta]$. Define $y_i^B(t)$ and $y_i^S(t)$ as the type-i queue length under the bang-bang (B) and singular (S) control policies, respectively. As a result of our setup, $y_1^B(t) \geq y_1^S(t) \geq 0$, $y_2^S(t) \geq y_2^B(t) \geq 0$ for $t \in [\tau, \tau+\delta]$. Further, because we assume $g_i(z)$ is monotonically increasing in $z \geq 0$ for $i \in \{1,2\}$, $g_1\left(\frac{y_1^B(t)}{\lambda_1(t)}\right) \geq g_1\left(\frac{y_1^S(t)}{\lambda_1(t)}\right) > 0$ and $g_2\left(\frac{y_2^S(t)}{\lambda_2(t)}\right) \geq g_2\left(\frac{y_2^B(t)}{\lambda_2(t)}\right) > 0$. This implies $y_1^B(t) g_1\left(\frac{y_1^B(t)}{\lambda_1(t)}\right) \geq y_1^S(t) g_1\left(\frac{y_1^S(t)}{\lambda_1(t)}\right)$ and $y_2^S(t) g_2\left(\frac{y_2^S(t)}{\lambda_2(t)}\right) \geq y_2^B(t) g_2\left(\frac{y_2^B(t)}{\lambda_2(t)}\right)$. We now show that the cost associated with implementing the bang-bang control (J_B) is greater than the singular control (J_S) over the time period $t \in [\tau, \tau+\delta]$.

$$
\begin{aligned}
J_B - J_S &= \int_\tau^{\delta+\tau} \left(y_1^B(t) g_1\left(\frac{y_1^B(t)}{\lambda_1(t)}\right) + y_2^B(t) g_2\left(\frac{y_2^B(t)}{\lambda_2(t)}\right) - y_1^S(t) g_1\left(\frac{y_1^S(t)}{\lambda_1(t)}\right) - y_2^S(t) g_2\left(\frac{y_2^S(t)}{\lambda_2(t)}\right)\right) dt, \\
&= \int_\tau^{\delta+\tau} \left(y_1^B(t) g_1\left(\frac{y_1^B(t)}{\lambda_1(t)}\right) - y_1^S(t) g_1\left(\frac{y_1^S(t)}{\lambda_1(t)}\right) + y_2^B(t) g_2\left(\frac{y_2^B(t)}{\lambda_2(t)}\right) - y_2^S(t) g_2\left(\frac{y_2^S(t)}{\lambda_2(t)}\right)\right) dt, \\
&\geq \int_\tau^{\delta+\tau} \left(\left(g_1\left(\frac{y_1^S(t)}{\lambda_1(t)}\right) + \frac{y_1^S(t)}{\lambda_1(t)} g_1'\left(\frac{y_1^S(t)}{\lambda_1(t)}\right)\right)\left(y_1^B(t) - y_1^S(t)\right) \right. \\
&\qquad \left. + \left(g_2\left(\frac{y_2^S(t)}{\lambda_2(t)}\right) + \frac{y_1^S(t)}{\lambda_1(t)} g_2'\left(\frac{y_2^S(t)}{\lambda_2(t)}\right)\right)\left(y_2^B(t) - y_2^S(t)\right)\right) dt,
\end{aligned}
$$

where the inequality follows from the convexity of the function $y_i(t) g_i\left(\frac{y_i(t)}{\lambda_i(t)}\right)$. Noting that $y_2^B(t) - y_2^S(t) = -\mu_2 m t + \mu_2 m(1-\alpha)t = -\mu_2 m \alpha t$ and $y_1^B(t) - y_1^S(t) = \mu_1 m \alpha t$, we get that

$$
\begin{aligned}
J_B - J_S &\geq \alpha m \int_\tau^{\delta+\tau} t\left(\mu_1 g_1\left(\frac{y_1^S(t)}{\lambda_1(t)}\right) + \mu_1 \frac{y_1^S(t)}{\lambda_1(t)} g_1'\left(\frac{y_1^S(t)}{\lambda_1(t)}\right) - \mu_2 g_2\left(\frac{y_2^S(t)}{\lambda_2(t)}\right) - \mu_2 \frac{y_2^S(t)}{\lambda_2(t)} g_2'\left(\frac{y_2^S(t)}{\lambda_2(t)}\right)\right) dt \\
&= \alpha m \int_\tau^{\delta+\tau} t\left(\mu_2 \dot{\pi}_2(t) - \mu_1 \dot{\pi}_1(t)\right) dt.
\end{aligned}
$$

For $t \in [\tau, \tau+\delta]$, the switching function $\sigma(t) = 0$ for the singular solution (by definition) and thus, both $\mu_2 \pi_2(t) = \mu_1 \pi_1(t)$ and $\mu_2 \dot{\pi}_2(t) = \mu_1 \dot{\pi}_1(t)$. As a consequence, it follows that over this short time period, $J_B - J_S \geq 0$. However, the above argument can be inductively extended for any $n \in \mathbb{Z}_{\geq 0}$ such that $t \in [\tau + (n-1)\delta, \tau + n\delta]$. This proves the optimality result for $I = 2$. We note that, for completeness, analyzing the opposite direction does not yield the same result. In particular,

$$
\begin{aligned}
J_S - J_B &\geq \alpha m \int_\tau^{\delta+\tau} t\left(\mu_2 g_2\left(\frac{y_2^B(t)}{\lambda_2(t)}\right) + \mu_2 \frac{y_2^B(t)}{\lambda_2(t)} g_2'\left(\frac{y_2^B(t)}{\lambda_2(t)}\right) - \mu_1 g_1\left(\frac{y_1^B(t)}{\lambda_1(t)}\right) - \mu_1 \frac{y_1^B(t)}{\lambda_1(t)} g_1'\left(\frac{y_1^B(t)}{\lambda_1(t)}\right)\right) dt \\
&= \alpha m \int_\tau^{\delta+\tau} t\left(\mu_1 \dot{\pi}_1(t) - \mu_2 \dot{\pi}_2(t)\right) dt.
\end{aligned}
$$

However, with the bang-bang policy, $y_2^B(t) \geq 0$ is decreasing over $t \in (\tau, \tau+\delta]$ because $\phi_2(t) = 1$ while $y_1^B(t) \geq 0$ is increasing as $\phi_1(t) = 0$. Thus, while $\mu_2 \dot{\pi}_2(\tau) = \mu_1 \dot{\pi}_1(\tau)$, $\mu_2 \dot{\pi}_2(t) > \mu_1 \dot{\pi}_1(t)$ for $t \in (\tau, \tau+\delta]$ which follows from the convexity of $y_i^B(t) g_i\left(\frac{y_i^B(t)}{\lambda_i(t)}\right)$ for $i \in \{1,2\}$ with respect to $y_i^B(t)$.

We now discuss how to extend the above result to the case where $I > 2$. To do this, without loss of generality, we continue to assume that class one patients are initially prioritized. In this more general case, given that $\phi_1(t) = 1 - \sum_{i=2}^{I} \phi_i(t)$, the Hamiltonian function can be written as follows:

$$\begin{aligned}\mathcal{H}(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\pi}, t) &= \sum_{i=1}^{I} \left(g_i \left(\frac{y_i(t)}{\lambda_i(t)} \right) y_i(t) + \pi_i(t)(\lambda_i(t) - \mu_i \phi_i(t) m) \right), \\ &= \sum_{i=1}^{I} g_i \left(\frac{y_i(t)}{\lambda_i(t)} \right) y_i(t) + \sum_{i=1}^{I} \pi_i(t)\lambda_i(t) - m\mu_1\pi_1(t) + m \sum_{i=2}^{I} \left(\mu_1\pi_1(t) - \mu_i\pi_i(t) \right) \phi_i(t).\end{aligned}$$

It follows that, for $\phi_1(0) = 1$, $\mu_1\pi_1(0) > \mu_i\pi_i(0)$ for $i \geq 2$ and class 1 is strictly prioritized at $t = 0$. For $t > 0$, there may exist switching times $0 < \tau_j \leq T$ such that $\mu_1\pi_1(\tau_j) = \mu_j\pi_j(\tau_j)$ for $j = 2, ..., I$. Some switching times may not exist within the planning horizon, i.e., $\tau_j > T$ for some $j \in \mathcal{I}$. However, for those where $\tau_j \leq T$, we can use a similar argument as in the two-class case to prove the optimality of the singular solution. The only difference is that the singular policy for the general case consists of an *aggregation* of patient types, all of whom have server access. More specifically, for the multi-class setting, this singular policy corresponds to an aggregation of all classes that have been assigned capacity up to time τ_j. Let $\mathcal{J}$ denotes the set of such classes. Along the singular arc, the optimality conditions imply that the switching functions for all $j, j' \in \mathcal{J}$ should be equal to zero. This directly gives the third statement from the theorem,

$$\mu_j\pi_j(t) = \mu_{j'}\pi_{j'}(t), \text{ and } \mu_j\dot{\pi}_j(t) = \mu_{j'}\dot{\pi}_{j'}(t) \qquad \text{for all } j, j' \in \mathcal{J},$$

and so the aggregated group effectively behaves as a single class. Consequently, at each switching time, the same comparison applies between the patients in the aggregated class $\mathcal{J}$ and the next patient class $j'' \notin \mathcal{J}$. Consequently, the singular policy remains optimal, and for switching time $\tau_{j''}$, an additional patient class enters service. Repeating this argument for all j where $\tau_j \leq T$ establishes that the optimal policy exhibits sequential activation, with the set of classes receiving capacity expanding monotonically over time. □

Proof of Proposition 2:

Proof. Assume that the first patient class exclusively accesses the server (without loss of generality) at time $t = 0$. From the first part of Theorem 1, the intrinsic order of the control problem equals one. Thus, along a singular arc, $\mu_1\pi_1(t) = \mu_i\pi_i(t)$, $\mu_1\dot{\pi}_1(t) = \mu_i\dot{\pi}_i(t)$, and $\mu_1\ddot{\pi}_1(t) = \mu_i\ddot{\pi}_i(t)$ for all classes $i \in \mathcal{I}$ that share capacity with the first patient class at time $t > 0$. Define the following functions:

$$\begin{aligned}H\left(y_i(t), \lambda_i(t)\right) &:= \frac{m\mu_i^2}{\lambda_i(t)^2} \left(2\lambda_i(t) g_i' \left(\frac{y_i(t)}{\lambda_i(t)} \right) + y_i(t) g_i'' \left(\frac{y_i(t)}{\lambda_i(t)} \right) \right), \\ h\left(y_i(t), \lambda_i(t)\right) &:= \frac{\mu_i}{\lambda_i(t)^3} \left(\lambda_i(t)^2 - y_i(t)\lambda_i'(t) \right) \left(2\lambda_i(t) g_i' \left(\frac{y_i(t)}{\lambda_i(t)} \right) + y_i(t) g_i'' \left(\frac{y_i(t)}{\lambda_i(t)} \right) \right).\end{aligned}$$

Then, we can use the above equations and the law of total probability to define a linear system to solve for $\boldsymbol{\phi}(t)$. That is, for an $I \times I$ matrix $\boldsymbol{A}(t)$ and a vector $\boldsymbol{b}(t)$ of length I, assume that (without loss of generality) the first i' classes share capacity. Then,

$$\boldsymbol{A}(t) := \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ H\left(y_1(t),\lambda_1(t)\right) & -H\left(y_2(t),\lambda_2(t)\right) & 0 & \dots & 0 & \dots \\ H\left(y_1(t),\lambda_1(t)\right) & 0 & -H\left(y_3(t),\lambda_3(t)\right) & \dots & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots \\ H\left(y_1(t),\lambda_1(t)\right) & 0 & 0 & \dots & -H\left(y_{i'}(t),\lambda_{i'}(t)\right) & \dots \\ 0 & 0 & 0 & \dots & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

$$\boldsymbol{b}(t) := \begin{pmatrix} 1 \\ h\left(y_1(t),\lambda_1(t)\right) - h\left(y_2(t),\lambda_2(t)\right) \\ h\left(y_1(t),\lambda_1(t)\right) - h\left(y_3(t),\lambda_3(t)\right) \\ \vdots \\ h\left(y_1(t),\lambda_1(t)\right) - h\left(y_{i'}(t),\lambda_{i'}(t)\right) \\ 0 \\ \vdots \end{pmatrix} \tag{EC.2}$$

As a consequence, given the current arrival rate and queue length for all $i \in \mathcal{I}$, we solve the linear system $\boldsymbol{A}(t)\boldsymbol{\phi}(t) = \boldsymbol{b}(t)$. Since the rank of $\boldsymbol{A}(t)$ equals the number of independent variables, there always exists a unique solution to the linear system for every $t > 0$.

To compute switching times, notice that by Theorem 1, $\dot{\pi}_i(t) \leq 0$ and $\pi_i(t) > 0$ for all $t \in [0, T)$. Further, due to the adjoint condition of Pontryagin's minimum principle, and because the Hamiltonian is convex in $y_i(t)$, $\pi_i(t)$ and $\dot{\pi}_i(t)$ are decreasing in $y_i(t)$. Now suppose that at $t = 0$, without loss of generality, the optimal policy assigns class-1 patients capacity while class $i \neq 1$ patients are not assigned any, i.e., $\mu_1\dot{\pi}_1(0) < \mu_i\dot{\pi}_i(0)$ for all $i \in \mathcal{I}$. Thus, the class-$i \neq 1$ queue length is increasing over time which means $\pi_i(t)$ and $\dot{\pi}_i(t)$ are monotonically decreasing over time.

The necessary condition in the proposition statement follows from the definition of a switching time. For the sufficient condition, suppose that there exists a $t = \tau$ where for some $i \in \mathcal{I}$, $\mu_1\dot{\pi}_1(\tau^-) < \mu_i\dot{\pi}_i(\tau^-)$ and $\mu_1\dot{\pi}_1(\tau) = \mu_i\dot{\pi}_i(\tau)$. Then, by Proposition 1, $\mu_1\dot{\pi}_1(t) = \mu_i\dot{\pi}_i(t)$ for all $t > \tau$ which implies τ is a switching time and capacity will now be shared with class-i patients. Given that there are I patient classes, at most, we need to verify whether $I - 1$ equations hold with equality at time t. □

Proof of Corollary 1:

Proof. Consider the general case where $I \geq 2$. Suppose there exists a singular arc associated with all classes $j \in \mathcal{J}$ where $\mathcal{J} \subseteq \mathcal{I}$. This means that $j \in \mathcal{J}$ if and only if $0 < \phi_j(t) < 1$ and $\sum_{j\in\mathcal{J}} \phi_j(t) = 1$. As is apparent, for all $i \notin \mathcal{J}$, $\phi_i(t) = 0$. Accounting for the fact that $\sum_{i=1}^{I} \phi_i(t) = 1$ for every t and $y_i'(t) = \lambda_i(t) - \mu_i m\phi_i(t)$ by Proposition 1, the Hamiltonian can be written as follows:

$$\mathcal{H}(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\pi}, t) = \sum_{i=1}^{I} g_i\left(\frac{y(t)}{\lambda_i(t)}\right) y_i(t) + \sum_{i=1}^{I} \pi_i(t)\left(\lambda_i(t) - \mu_i m\phi_i(t)\right).$$

Suppose, without loss of generality, that class one patients are assigned exclusive priority at $t = 0$. Then, for some $t > 0$, $\phi_1(t) = 1 - \sum_{i=1, i\neq 1}^{I} \phi_i(t)$. Substituting this into the Hamiltonian, we get

$$\mathcal{H}(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\pi}, t) = \sum_{i=1}^{I} g_i\left(\frac{y(t)}{\lambda_i(t)}\right) y_i(t) + \sum_{i=1}^{I} \lambda_i(t)\pi_i(t) - \mu_1 m \pi_1(t) - m \sum_{i=1, i\neq 1}^{I} \left(\mu_1\pi_1(t) - \mu_i\pi_i(t)\right)\phi_i(t).$$

Based on the above representation of the Hamiltonian, we define the switching function for class-$i \neq 1$ patients to be $\sigma_i(t) := \mu_1\pi_1(t) - \mu_i\pi_i(t)$. Define $\tilde{\Lambda}_i(t) := \int_t^T \frac{1}{\lambda_i(s)}\, ds$, which is the antiderivative of the reciprocal of $\lambda_i(t)$. Since $g_i(z) = \gamma_i/\lambda_i(t)$, the adjoint condition is independent of $y_i(t)$ and gives

$$\dot{\pi}_i(t) = -\frac{\gamma_i}{\lambda_i(t)}$$

which evaluates to $\pi_i(t) = \gamma_i\tilde{\Lambda}_i(t)$. Thus, the switching function for class $i \neq 1$ evaluates to

$$\sigma_i(t) = \mu_1\gamma_1\tilde{\Lambda}_1(t) - \mu_i\gamma_i\tilde{\Lambda}_i(t)$$

which is the condition governing the priority policy of the generalized $c\mu$-rule. □

Proof of Proposition 3:

Proof. It suffices to show that the cost function is convex in $\boldsymbol{y}$, where we assume that both $y_i > 0$ and $\lambda_i(t) > 0$ for all i and $t > 0$ which follows from Assumption 1.

1. If $g_i(y) := \gamma_i y^n$ for $\gamma_i > 0$, then $f_i(y) := yg_i(y) = \gamma_i \frac{y^{n+1}}{\lambda_i^n(t)}$. Thus, $f_i''(y) = \frac{(n+1)\gamma_i y^n}{\lambda_i^n(t)} > 0$.
2. If $g_i(y) := \gamma_i\sqrt{y}$ for $\gamma_i > 0$, then $f_i(y) := yg_i(y) = \gamma_i y\sqrt{\frac{y}{\lambda_i(t)}}$. Thus, $f_i''(y) = \frac{3\gamma_i}{4\sqrt{y\lambda_i(t)}} > 0$.
3. If $g_i(y) := \gamma_i \ln\left(\frac{\frac{y}{\lambda_i(t)}}{\left(\chi_i - \frac{y}{\lambda_i(t)}\right)}\right)$ for $\gamma_i > 0$ and $\chi_i > 0$ such that $y/\lambda_i(t) < \chi_i$, then $f_i(y) := yg_i(y) = \gamma_i y \ln\left(\frac{\frac{y}{\lambda_i(t)}}{\left(\chi_i - \frac{y}{\lambda_i(t)}\right)}\right) = y\gamma_i \ln\left(-\frac{y}{y - \chi_i\lambda_i(t)}\right)$. Thus, $f_i''(y) = \frac{\gamma_i\chi_i^2\lambda_i(t)^2}{y(y-\chi_i\lambda_i(t))^2} > 0$.

We note that when $g_i(y_i(t)) := \gamma_i\frac{y_i(t)}{\lambda_i(t)}$ and $g_i(y_i(t)) := \frac{\gamma_i}{\lambda_i(t)}$, the Hamiltonian is either a quadratic or linear function of $y_i(t)$, respectively, implying convexity.

Consider the hessian matrix associated with all second-order partial derivatives of $\mathcal{H}(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\pi}, t)$ with respect to $(\boldsymbol{y}, \boldsymbol{\phi})$. Notice that this matrix is diagonal with non-negative entries. Thus, the matrix is positive semi-definite which implies $\mathcal{H}(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\pi}, t)$ is convex in $(\boldsymbol{y}, \boldsymbol{\phi})$ for all cases. □

Proof of Proposition 4:

Proof. To demonstrate the first property, suppose that $J = K = 1$. In this case, $\tilde{g}_i(\boldsymbol{x}_k, \phi_{ik})$ is a univariate function of x_{jk} and $z_k = 1$. Thus, $\tilde{g}_i(z_k, \boldsymbol{x}_k, \phi_{ik})$ reduces to a linear function of x_{jk} which is non-decreasing. In addition, for this case, there is no term associated with the travel cost as $d(j, j') = 0$ for $j = j'$. Now suppose that $J > 1$ and, without loss of generality, assume that $z_k = 0$. It suffices to show that given all other coordinates are fixed, $\tilde{g}_i(z_k, \boldsymbol{x}_k, \phi_{ik})$ is non-decreasing in x_{jk}. Note that if $z_k = 0$, then $x_k = 0$ for all k and the result trivially holds. As a consequence, we apply L'Hopitals Monotone Rule (see Theorem 4 in Estrada and Pavlović, 2017) provided that $z_k = 1$, as follows.

For simplicity, define

$$y_{ik}(\boldsymbol{x}_k, \phi_{ik}) = T\left(\sum_{j=1}^{J} \lambda_{ij} x_{jk} - \mu_i \phi_{ik} \sum_{j=1}^{J} m_j x_{jk}\right) + \sum_{j=1}^{J} q_{ij} x_{jk} = T\left(\boldsymbol{\lambda}_i^{\mathsf{T}} - \mu_i \phi_{ik} \boldsymbol{m}^{\mathsf{T}} + \frac{1}{T}\boldsymbol{q}_i^{\mathsf{T}}\right)\boldsymbol{x}_k.$$

Assume that $A_i y_{ik}^2(\boldsymbol{x}_k, \phi_{ik})$ and $\boldsymbol{\lambda}_i^{\mathsf{T}}\boldsymbol{x}_k$ are differentiable functions on $x_{jk} \in [0,1]$. Then, holding all other coordinates fixed, $\tilde{g}_i(z_k, \boldsymbol{x}_k, \phi_{ik})$ is non-decreasing in x_{jk} if the following is non-negative:

$$\frac{\frac{\partial}{\partial x_{jk}}\left(A_i y_{ik}^2(\boldsymbol{x}_k, \phi_{ik})\right)}{\frac{\partial}{\partial x_{jk}}\left(\boldsymbol{\lambda}_i^{\mathsf{T}}\boldsymbol{x}_k\right)} \geq 0.$$

Taking derivatives in the above equation, and because we constrain $y_{ik}(\boldsymbol{x}_k, \phi_{ik}) \geq 0$, we get that

$$\frac{2A_i T^2 y_{ik}(\boldsymbol{x}_k, \phi_{ik})\left(\lambda_{ij} - \mu_i \phi_{ik} m_j + \frac{1}{T} q_{ij}\right)}{\lambda_{ij}} \geq 0.$$

Since the sum of non-decreasing functions on $x_{jk} \in [0,1]$ is non-decreasing, it is also non-decreasing when restricting $x_{jk} \in \{0,1\}$. Further, to demonstrate the element-wise convexity of the objective in x_{jk}, we take the second derivative of $\tilde{g}_i(z_k, \boldsymbol{x}_k, \phi_{ik})$ with respect to x_{jk} assuming that $z_k = 1$ (otherwise, element-wise convexity trivially holds). This gives the following relation:

$$\frac{2\gamma_i\left(\lambda_{ij}\left(T\mu_i\phi_{ik}\sum_{j=1}^{J} m_j x_{jk} - \sum_{j=1}^{J} q_{ij}x_{jk}\right) + \sum_{j=1}^{J}\lambda_{ij}x_{jk}\left(q_{ij} - T\mu_i m_j \phi_{ik}\right)\right)^2}{\left(\sum_{j=1}^{J}\lambda_{ij}x_{jk}\right)^3} \geq 0.$$

The non-negativity implies that the objective function is convex in each dimension x_{jk}.

For the second property, the first derivative of the objective function with respect to ϕ_{ik} gives

$$-\frac{2A_i T\,\mu_i\,\boldsymbol{m}^{\mathsf{T}}\boldsymbol{x}_k\,y_{ik}(\boldsymbol{x}_k, \phi_{ik})}{1 - z_k + \boldsymbol{\lambda}_i^{\mathsf{T}}\boldsymbol{x}_k} \leq 0.$$

This implies that the objective function is non-increasing in ϕ_{ik}. Furthermore, we can also take the second derivative of each term in the objective function with respect to ϕ_{ik}, noting that cross terms will be zero. The result is non-negative which implies element-wise convexity in ϕ_{ik}:

$$\frac{\partial^2}{\partial \phi_{ik}^2}\left(\frac{A_i y_{ik}^2(\boldsymbol{x}_k, \phi_{ik})}{1 - z_k + \boldsymbol{\lambda}_i^{\mathsf{T}}\boldsymbol{x}_k}\right) = \frac{2T\gamma_i\mu_i^2\left(\boldsymbol{m}_j^{\mathsf{T}}\boldsymbol{x}_k\right)^2}{1 - z_k + \boldsymbol{\lambda}_i^{\mathsf{T}}\boldsymbol{x}_k} \geq 0.$$

We can now construct the Hessian which has non-negative diagonal entries and off-diagonal elements equal to zero. The resulting matrix is positive semi-definite implying joint convexity. □

EC.4. Additional Fluid Model Results: Cost and Discretization

We first examine the impact of discretization on the optimal cost by comparing the direct optimization approach (referred to as *dir*) with Algorithm EC.1 (referred to as *alg*). As shown in Table EC.5, the relative cost gap between the two methods is minimal, indicating comparable performance.

We next examine how the number of discretized time steps N affects policy performance, using the same fluid model setup as in Section 6.1 by plotting the optimal allocations and fluid levels over time.

Cost	*dir*	*alg*	Gap
sqrt	13128.55	13132.52	0.03%
quad	3370025.23	3383236.61	0.390%
r-logit	4970.74	4975.36	0.093%

Table EC.5 Comparison between the objective values under direct optimization (*dir*) and capacity allocation algorithm (*alg*) for three of the cost function. The values are computed using 20 independent replications. The cost gap is calculated as $(\text{cost}_{\text{alg}} - \text{cost}_{\text{dir}})/\text{cost}_{\text{dir}}$.

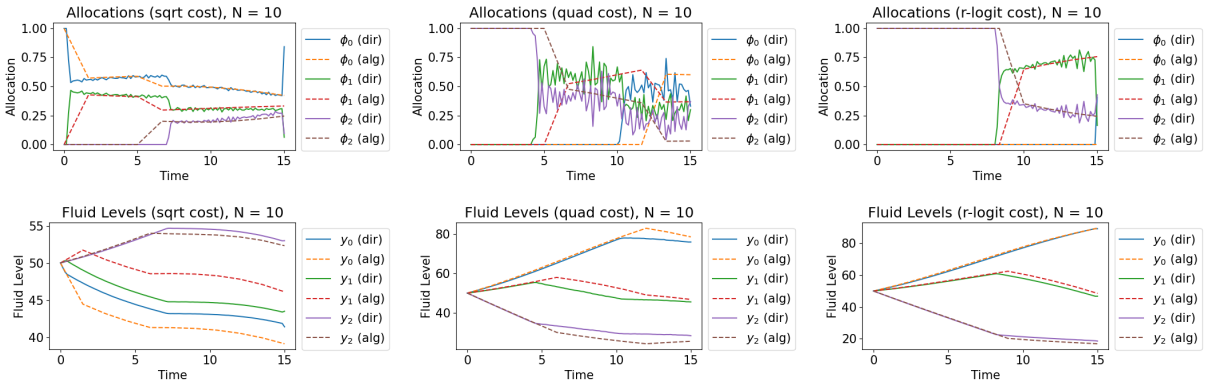


Figure EC.1 Comparison of FL-S (sqrt), FL-R (quad), and FL-Log (r-logit) under $N = 10$.

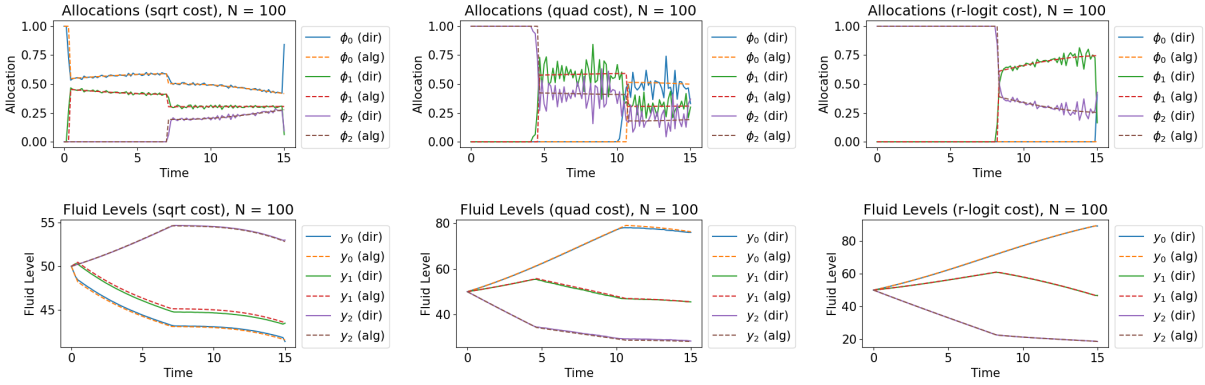


Figure EC.2 Comparison of FL-S (sqrt), FL-R (quad), and FL-Log (r-logit) under $N = 100$.

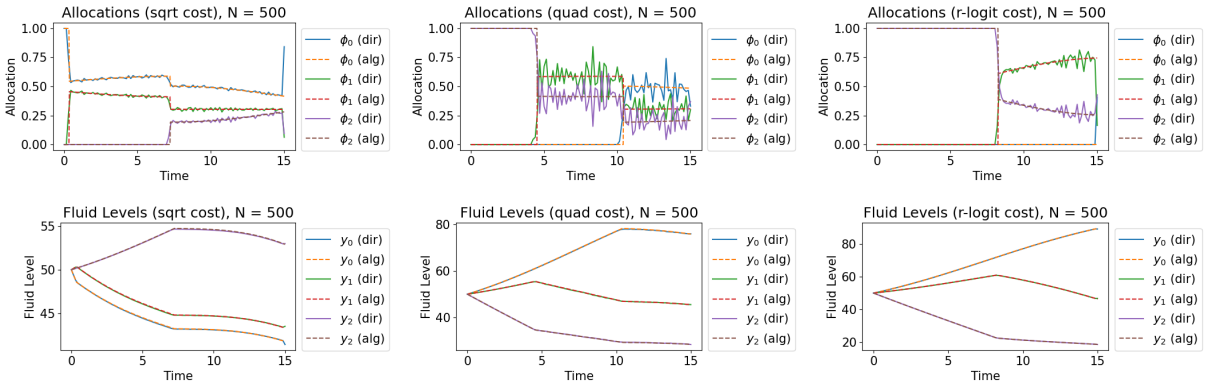


Figure EC.3 Comparison of FL-S (sqrt), FL-R (quad), and FL-Log (r-logit) under $N = 500$.

EC.5. Additional Clustering Experiments: Equity and Efficiency

We first compare the fraction of patients served across clustering solutions. We see that as the number of clusters increases, the equity benefits of pooling diminish. That is, disparities in the fraction of served patients are more pronounced in single-region clusters. In contrast, when the system is partitioned into fewer clusters, pooled regions exhibit more balanced service across patient classes, indicating that regional pooling promotes a more equitable distribution of capacity amongst patients of the same class.

Figure EC.4 illustrates this effect using a linear cost function. We report the fraction of patients served in each pool for the four clustering solutions in Table 3, where we vary the weight placed on the travel disutility cost. In the bottom-right subplot, each region operates as a separate pool and allocation outcomes are highly imbalanced across urgency classes. For example, in the CE region, more than 90% of urgent patients are served, compared with fewer than 40% of non-urgent patients. In contrast, the top-left subplot shows CE clustered with "C, TC, MH". The fraction of served non-urgent patients increases to nearly 50%.

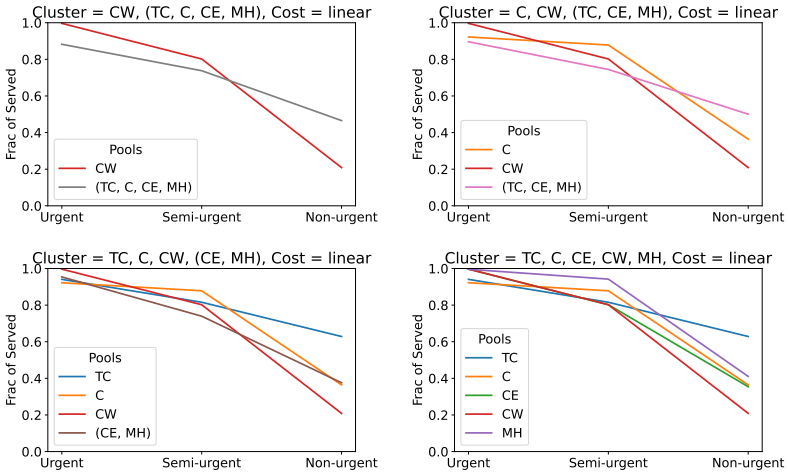


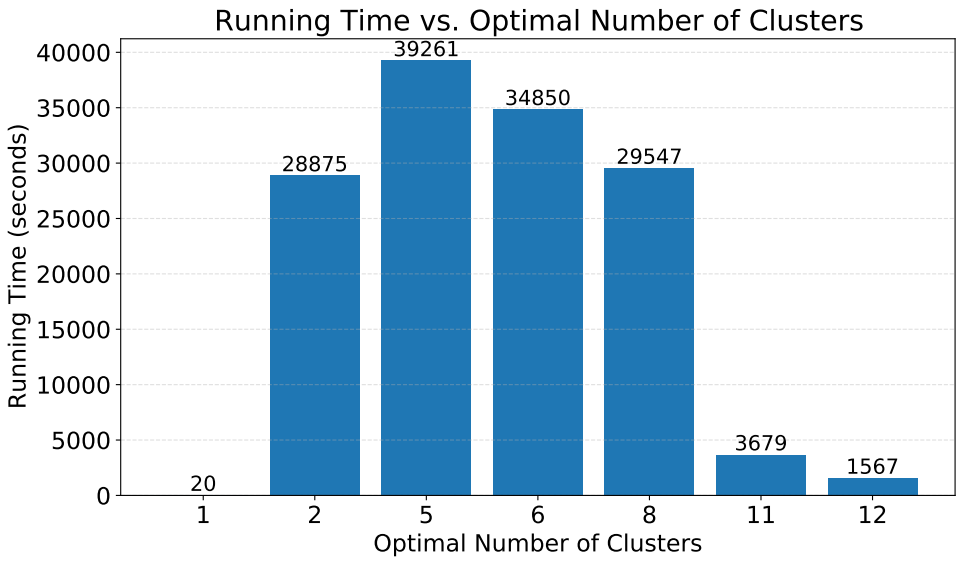
Figure EC.4 **Equity comparison of capacity allocation across clusters. Each plot shows the fraction of patients served in one of the optimal clusters reported as we vary the emphasis (ε) on travel cost in Table 3. For example, in the bottom-left plot, the optimal cluster configuration "TC, C, CW, (CE, MH)" yields four hospital pools: TC, C, CW, and (CE, MH). The first three pools consist of single regions (TC, C, and CW), while the last pool combines two regions, CE and MH.**

We next assess the computational efficiency of the pooling algorithm by applying it to a larger-scale instance from our case study. Specifically, we consider 15 Ontario hospitals with MRI machines located outside the Greater Toronto Area, namely, in the North (North Bay, Sudbury), North East (Kingston), and Champlain (Ottawa) LHINs. Notice that, in contrast to the main analyses, we analyze pooling at the hospital level, as these regions contain relatively few institutions with MRI services. The experiment evaluates how the optimal number of clusters, K^*, varies as a function of ε (see Figure EC.5). The non-smooth increases in

K^* arise because the disutility cost is proportional to the maximum pairwise distance, which does not scale linearly within each cluster, as hospitals are typically located in regions with higher population density.

In terms of performance, the runtime of the Benders algorithm depends strongly on the resulting number of clusters (or emphasis on ε). For small K^* values (e.g., $K^* = 1$), the algorithm terminates almost immediately (under 30 seconds). Instances with intermediate-sized clusters (i.e., $K^* \in \{5, 6, 8\}$) take the longest to solve. While the optimal pooling configuration is found within one minute on average across all instances, in these cases, optimality is proven after approximately 10 hours. These runtime peaks correspond to regions in which the clustering landscape is most complex and small changes in ε can shift hospitals across clusters. For larger K^* values, the algorithm again converges quickly. Overall, despite these fluctuations, the experiments show that the Benders algorithm remains tractable even for fairly large problem instances.

Figure EC.5 The runtime as a function of optimal number of clusters in seconds.



EC.6. Alternative Clustering Approach

One solution approach is to reformulate POOL using an epigraphical representation of the clustering cost to admit a second-order conic (SOC) representation (e.g., Lejeune and Margot, 2025). This produces a mixed-integer, SOC programming model which can be directly computed by optimization solvers.

PROPOSITION EC.1. *Let $\boldsymbol{u}_{ik}$ be a J-dimensional vector with non-negative elements u_{ijk} and $v_{ik} \geq 0$. Then, problem POOL can be reformulated as a second-order conic program*

$$\min_{u_{ijk}, v_{ik}, w_{jj'k}, z_k, x_{jk}, \phi_{ik}, \rho_k} \sum_{k=1}^{J} \sum_{i=1}^{I} v_{ik} + \varepsilon \sum_{k=1}^{J} \rho_k \quad \textit{s.t.} \tag{SOCP}$$

$$\left\| \begin{matrix} v_{ik} - \left(1 - z_k + \boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k\right) \\ 2\sqrt{\gamma_i} T \left(\boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k - \mu_i \boldsymbol{m}^\mathsf{T} \boldsymbol{u}_{ik} + \frac{1}{T} \boldsymbol{q}_i^\mathsf{T} \boldsymbol{x}_k\right) \end{matrix} \right\|_2 \leq v_{ik} + \left(1 - z_k + \boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k\right) \qquad \forall i, k, \tag{EC.3}$$

$$u_{ijk} \leq x_{jk} \qquad \forall i, j, k \tag{EC.4}$$

$$u_{ijk} \leq \phi_{ik} \qquad \forall i, j, k \tag{EC.5}$$

$$u_{ijk} \geq \phi_{ik} - (1 - x_{jk}) \qquad \forall i, j, k \tag{EC.6}$$

$$(2) - (9)$$

$$v_{ik}, w_{jj'k}, u_{ijk}, \phi_{ik}, \rho_k, y_{ik} \geq 0, \qquad \forall i, j, k,$$

$$x_{jk}, z_k, \in \{0, 1\}, \qquad \forall j, k.$$

Constraint (EC.3) is a second-order conic constraint, while (EC.4)-(EC.6) linearize the product of $\phi_{ik} \in [0, 1]$ and $x_{jk} \in \{0, 1\}$. Notice that SOCP introduces $O(I \times J \times K)$ variables as compared to MST. However, unlike the Benders approach which iteratively refines the solution by solving a sequence of master and subproblems, SOCP is solved in a single iteration as a second-order conic, mixed-integer program (SOCP-MIP). While this direct approach may often lead to improved computational performance (Ahmadi-Javid and Hoseinpour, 2022), our numerical experiments indicate that this is not the case for this problem.

Proof of Proposition EC.1:

We start by writing the problem in its epigraphical representation by introducing the variable $v_{ik} \geq 0$ for all i and k. Utilizing vector notation, we get that POOL admits the following constraint

$$v_{ik} \geq \frac{\gamma_i T^2 \left(\boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k - \mu_i \phi_{ik} \boldsymbol{m}^\mathsf{T} \boldsymbol{x}_k + \frac{1}{T} \boldsymbol{q}_i^\mathsf{T} \boldsymbol{x}_k\right)^2}{1 - z_k + \boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k} \qquad \forall i, k.$$

Since the denominator is always positive, we can multiply both sides by $1 - z_k + \boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k$ to get rid of the fraction. Then, the corresponding constraint

$$v_{ik} \left(1 - z_k + \boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k\right) \geq T^2 \gamma_i \left(\boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k - \mu_i \phi_{ik} \boldsymbol{m}^\mathsf{T} \boldsymbol{x}_k + \frac{1}{T} \boldsymbol{q}_i^\mathsf{T} \boldsymbol{x}_k\right)^2 \quad \forall i, k,$$

$$\left(v_{ik} + \left(1 - z_k + \boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k\right)\right)^2 - \left(v_{ik} - \left(1 - z_k + \boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k\right)\right)^2 \geq 4T^2 \gamma_i \left(\boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k - \mu_i \phi_{ik} \boldsymbol{m}^\mathsf{T} \boldsymbol{x}_k + \frac{1}{T} \boldsymbol{q}_i^\mathsf{T} \boldsymbol{x}_k\right)^2 \quad \forall i, k.$$

Simplifying and re-arranging gives

$$\left\| \begin{matrix} v_{ik} - \left(1 - z_k + \boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k\right) \\ 2T\sqrt{\gamma_i} \left(\boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k - \mu_i \boldsymbol{m}^\mathsf{T} \phi_{ik} \boldsymbol{x}_k + \frac{1}{T} \boldsymbol{q}_i^\mathsf{T} \boldsymbol{x}_k\right) \end{matrix} \right\|_2 \leq v_{ik} + \left(1 - z_k + \boldsymbol{\lambda}_i^\mathsf{T} \boldsymbol{x}_k\right) \qquad \forall i, k.$$

Since $\phi_{ik} \in [0, 1]$, and $\boldsymbol{x}_k$ is a vector of binary variables, we introduce the transformed variable $u_{ijk} = \phi_{ik} x_{jk}$ to linearize the product of a binary and a continuous (but bounded) variable. To ensure this transformation has the desired behavior, constraints (EC.4)-(EC.6) are introduced. □