An Operational View on Managing Mass Trauma Events

Noa Zychlinski
Faculty of Data and Decision Sciences,
Technion – Israel Institute of Technology, Haifa 3200003, Israel

Problem definition: Mass Trauma Events (MTEs)—such as wars, natural disasters, and terror attacks—present significant operational challenges. Affected populations require both immediate and prolonged mental health support, complicating response efforts. Post-traumatic Stress Disorder (PTSD) can have lasting effects and imposes a substantial economic burden, making early intervention critical to improving outcomes. The October 7, 2023, terror attack in southern Israel caused widespread trauma. Survivors, responders, and many others were exposed to extreme atrocities, placing an estimated 5.3% of the population (over 520,000 individuals) at risk of developing PTSD and related conditions. This crisis underscores the urgent need for practical policies to deliver timely mental health care amid a surge in demand on an already strained system.

Methodology/results: We study the coordination of group and individual therapy channels in a multi-server queueing setting. Group therapy can alleviate immediate workload but may lead to increased follow-up demand for individual treatment. Our model captures this trade-off and the interdependence between therapy channels, while accounting for key mental health system features such as patient no-shows and dropouts. Using a fluid approximation, we derive index-based policies tailored to the Surge, Recovery, and Long-Term Phases of MTEs, integrating time-varying, transient, and steady-state dynamics. Drawing on data from the October 7 attack and prior MTEs, we show that our policies can shorten the recovery phase by approximately six months, reduce queue lengths by 31%, and increase total cost savings by 52%, relative to a commonly accepted benchmark policy which we adapted to incorporate group therapy, no-shows, and dropouts. These improvements result from the embedded channel coordination in our policies.

Managerial implications: Our results highlight the value of channel-specific coordination in mental health scheduling policies for traumatized populations. The index-based rules we propose are simple to implement and offer actionable guidance for practitioners and policymakers managing care delivery after MTEs. Applying these policies can enhance support for at-risk populations, reduce system strain, and strengthen community recovery and resilience.

Key words: Stochastic modeling, healthcare operations management, fluid models, scheduling queues, mass casualty events

1. Introduction

Mass trauma events (MTEs)—such as wars, natural disasters, and terror attacks—create widespread psychological distress and place acute pressure on mental health systems.

Unlike mass casualty events (MCEs), where injuries are physical and immediate, the psychological harms associated with MTEs often emerge gradually and persist long after the event (Chriman and Dougherty 2014, Hirschberger 2018, Makari and Friedman 2024). Without timely intervention, these mental health consequences—most notably Post-Traumatic Stress Disorder (PTSD)—impose severe human, public health, and economic burdens (von der Warth et al. 2020, Davis et al. 2022).

MTEs of various scales continue to occur globally each year. From the ongoing Russia-Ukraine war to earthquakes in Myanmar, Afghanistan, Morocco, and Japan, the Southern California wildfires, and terror attacks such as the Crocus City Hall shooting and the October 7 Hamas attack and subsequent Hamas-Israel war, recent events underscore the growing need for scalable mental health responses.

The October 7, 2023, terror attack in Israel further illustrates the magnitude of this challenge. Mental health services reported a 200% surge in patients, a 25% rise in psychiatric medication use, and a 52% increase in anxiety-related cases (The Jerusalem Post, 2024). Projections estimate that over 520,000 individuals may develop PTSD (Katsoty et al. 2024).

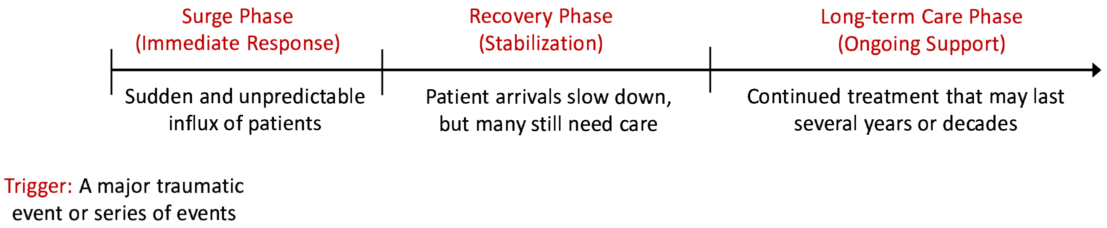
Experience from prior MTEs underscores the long-term nature of these mental health needs. More than two decades after the 9/11 attacks, Mount Sinai's World Trade Center Mental Health Program continues to treat approximately 700 patients, with new admissions ongoing (Mount Sinai Blog, 2021). This highlights the enduring demand placed on mental health systems and the importance of operational models that account for long-term effects.

Mental health systems responding to MTEs must manage large, time-varying surges in demand. These systems typically progress through three distinct operational phases, as illustrated in Figure 1, each with unique dynamics, objectives, and constraints. Despite the growing frequency of MTEs and the operational challenges they pose, the OR/OM literature offers limited guidance on how to allocate resources dynamically during such crises.

To address this gap, we develop an analytical queueing model that supports dynamic scheduling and capacity allocation decisions across the three response phases introduced above.

The *Surge Phase* marks the onset of an MTE response, during which the mental health system experiences a sharp and overwhelming increase in demand, with highly variable

Figure 1 **Three response phases in the aftermath of MTEs.**


Surge Phase
(Immediate Response)
Recovery Phase
(Stabilization)
Long-term Care Phase
(Ongoing Support)
Sudden and unpredictable influx of patients
Patient arrivals slow down, but many still need care
Continued treatment that may last several years or decades
Trigger: A major traumatic event or series of events

arrival rates. The duration of this phase depends on the event's scale and nature. For example, following the 9/11 and October 7 attacks, elevated demand persisted for several months (Herman et al. 2002).

In the second phase, the *Recovery Phase*, arrival rates begin to stabilize, but a substantial backlog of untreated patients remains. The goal is to maximize treatment throughput and cost-effectively return the system to a manageable state (Gibbs and Skyler 2004).

The final *Long-Term Care Phase* involves extended therapy for patients with enduring needs. This phase may last years or even decades (Bowler et al. 2016). Across all phases, a central operational challenge is how to cost-effectively schedule and prioritize access to care.

While PTSD has no cure, short-term psychological treatments can reduce symptoms and improve quality of life. Cognitive-behavioral therapy (CBT) is a trauma-focused approach shown to be both cost-effective and effective in the short and long term (von der Warth et al. 2020). Other treatments include prolonged exposure, cognitive-processing therapy, and EMDR (eye movement desensitization and reprocessing).

Group therapy is a common approach in mental health care that helps patients share experiences, reduce isolation, and build coping skills in a supportive setting (Sloan et al. 2012, Yalom and Leszcz 2020). During MTEs, it increases treatment capacity and reduces wait times. Patients with persistent symptoms may be referred to individual therapy for more intensive care. This tiered structure balances efficiency with personalized treatment.

Coordinating group and individual therapy is crucial for effective and efficient care delivery. While group sessions improve throughput and alleviate immediate resource constraints, they may reduce individual attention and therapeutic effectiveness. Patients who do not respond well may require follow-up individual care, further straining system capacity. Clinical outcomes are also sensitive to group size (Daley et al. 1983, Dueweke et al. 2022), but the optimal size that balances efficiency and effectiveness remains unclear.

The paper's contributions are as follows:

- *Integrated Modeling of Group and Individual Therapy.* We develop a multi-server dynamic queueing model where therapists provide series of sessions aimed at trauma processing and recovery. Patients begin with group sessions and, if needed, proceed to individual therapy. Our model captures the dependency and trade-off between group and individual therapy in addressing mental health needs following MTEs. The model explicitly incorporates patient no-shows and dropouts, which are particularly prevalent in mental health settings, especially PTSD.

- *Phase-Specific Index Policies for Optimal Scheduling.* Using a fluid approximation, we derive optimal scheduling policies tailored to the Surge, Recovery, and Long-Term Care Phases – each with distinct dynamics, objectives, and constraints – integrating time-varying, transient, and steady-state elements. We also demonstrate how our model can help determine the optimal group size that balances efficiency and effectiveness in this setting.

Using data from the October 7 terror attack and previous MTEs, we demonstrate that our policies significantly improve system performance – shortening the recovery phase by nearly six months and increasing overall cost savings by more than half – compared to an adapted version of the $c\mu/\theta$ rule that accounts for group therapy, no-shows, and dropouts.

- *Coordinated Resource Allocation Across Channels.* We demonstrate the critical role of coordinated resource allocation between group and individual therapy channels. Since group therapy serves as a feeder channel to individual therapy, effective scheduling and resource allocation require coordination across both channels. This coordination leads to significantly improved system outcomes compared to policies that treat the channels independently.

The remainder of the paper is organized as follows: Section 2 reviews related literature. Section 3 introduces the model components, assumptions and formulations. Sections 4, 5, and 6 analyze the Long-Term Care, Recovery, and Surge Phases, respectively. Section 7 presents a case study using data from the October 7 MTE. Section 8 concludes the paper and offers future research directions. The appendix includes a multi-class extension, additional numerical results and technical proofs.

2. Literature Review

Due to the limited operational research on MTEs, our review focuses on two related streams of literature: (i) OR/OM studies on managing MCEs and (ii) scheduling and resource allocation in multi-server queues. Although these areas may appear to overlap, they differ fundamentally in their analytical focus. Traditional queueing models emphasize steady-state optimization, whereas MCE-related research must capture transient dynamics driven by sudden and overwhelming surges in demand.

Related to our focus on group-based service delivery, the healthcare OM literature has examined shared medical appointments (SMAs), where patients are treated in groups rather than individually. Empirical studies in glaucoma care have shown SMAs improve patient engagement, satisfaction, and compliance (Sönmez et al. 2023, Buell et al. 2024). While these works demonstrate the practical value of group service delivery, our contribution lies in developing an analytical model to optimize such systems.

2.1. OR/OM Literature on Managing MCEs

A substantial body of OR/OM work has focused on patient prioritization during MCEs. Jacobson et al. (2012) proposed heuristic policies based on patients' wait tolerance, service time, and rewards, showing that simple state-dependent rules can perform well. Mills et al. (2018) developed a fluid model for patient transfers during MCEs, demonstrating improvements over the START triage protocol by incorporating survival probabilities and resource constraints. Sun et al. (2018) identified optimal strategies for balancing triage and service times, introducing a "switching curve" that guides when triage is beneficial. Li et al. (2020) applied index policies based on Whittle's restless bandits to prioritize patient classes under finite and uncertain horizons. Rezapour et al. (2022) derived optimal casualty treatment strategies for large-scale incidents, showing the advantage of dynamic resource allocation. Shi et al. (2023) extended this line of work by incorporating victim deterioration trajectories and resource availability into emergency planning.

Resource allocation under MCEs has also received attention. Cohen et al. (2014) optimized surgeon allocation using two policies—static prioritization and dynamic switching—offering actionable strategies for both planning and real-time response. Lodree et al. (2019) proposed heuristics for queue management that prioritize critical patient classes with buffer capacity for future arrivals. For a broad overview, see the review by Farahani et al. (2020).

A related setting arises during pandemics, where demand exceeds capacity. Motivated by COVID-19, Chan et al. (2021) used a fluid model to study inter-facility transfers, aiming to alleviate hospital congestion and improve equity in patient distribution.

Most MCE studies adopt a clearing system model, where a sudden influx of patients must be treated and the system eventually empties. In contrast, MTEs unfold over three operational phases: a Surge Phase with highly variable arrivals, a Recovery Phase with stabilized arrivals and a backlog, and a Long-Term Phase with prolonged care needs. Full clearance is rarely attainable in MTEs, even over extended horizons.

Our work contributes by modeling each phase as a distinct operational problem, incorporating time-varying, transient, and steady-state dynamics. Unlike most MCE-focused models, we also account for key features relevant to mental health systems—such as group therapy, follow-up individual care, no-shows, and dropouts—within a tractable modeling framework.

2.2. Scheduling and Resource Allocation in Multi-Server Queues

Our work is closely related to the literature on scheduling in multi-class queueing systems. Numerous generalizations of the classical $c\mu$ rule (Cox and Smith 1961)—which is optimal for single-server queues with linear holding costs—have been developed, with many establishing asymptotic optimality via fluid models (Van Mieghem 1995, Mandelbaum and Stolyar 2004, Huang et al. 2015). In the many-server setting, Atar et al. (2010) established the asymptotic optimality of the $c\mu/\theta$ rule under customer abandonment. Subsequent work has extended this framework to more general cost structures (Long et al. 2020), proactive service policies (Hu et al. 2022), heterogeneous server pools (Long et al. 2024), and the use of AI to reduce diagnostic errors (Cai and Zychlinski 2025).

Our work also connects to models with re-work or readmissions, common in healthcare. Dai and Weiss (1996) analyzed fluid models for re-entrant lines, while Yom-Tov and Mandelbaum (2014) developed the Erlang-R model to support staffing under time-varying arrivals. More recently, Chan et al. (2024) studied policies to reduce return probabilities in systems where customers may re-enter after discharge.

In contrast to prior work, we develop dynamic scheduling policies tailored to distinct operational phases following mass trauma events. Our model incorporates several features specific to mental health systems, including the interaction between group and individual therapy, patient no-shows, and dropouts.

We contribute to this literature by modeling the coordination between group and individual treatment channels, where group sessions serve as feeders for individual therapy. This interdependence—spanning time-varying, transient, and steady-state regimes—introduces analytical challenges. Our structural results and numerical experiments demonstrate the robustness and effectiveness of the proposed policies.

3. The Stochastic Model

We structure our analysis around the three response phases that follow an MTE, as illustrated in Figure 1. Across these phases, we address the central question of how to cost-effectively schedule individuals seeking mental health support and allocate resources across both treatment channels. Before analyzing each phase in detail, we first present the foundational elements of our coordinated model for group and individual therapy services.

We consider a Markovian multi-server queueing model with N servers. These servers are shared between two service channels: one for group therapy, denoted by $\mathbf{m}$, and the other for individual therapy, denoted by $\mathbf{s}$, as illustrated in Figure 2.

As we explain shortly, the service in both channels includes a sequence of sessions aimed at processing traumatic events and providing coping strategies. The number of sessions is determined at the outset of treatment, when all sequential appointments are allocated.

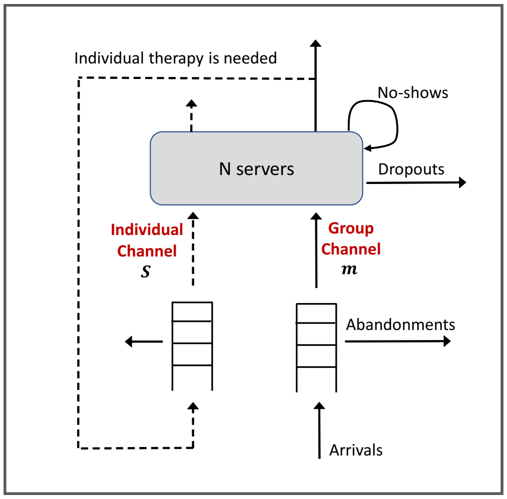
Let $\mathcal{J} = \{\mathbf{m}, \mathbf{s}\}$ denote the set of treatment channels. We use a subscript j, where $j \in \mathcal{J}$, to distinguish parameters specific to each channel. In Appendix C, we extend the model to a multi-class setting, where each class may receive treatment through both channels.

Arriving patients first enter the group therapy channel, where each patient requires $m_{\mathbf{m}}$ servers (with $1/m_{\mathbf{m}}$ representing the group size). We focus on open group therapy, which allows members to join or leave as they complete their assigned sessions. This structure is well-suited for MTEs, enabling immediate, scalable support without the delays of closed group formation. It maximizes access, reduces wait times, and accommodates diverse trauma responses by offering flexible, individualized care.

Group sessions may be less effective for some patients. Therefore, upon completion of the group session series, some patients require additional individual therapy with probability $p(m_{\mathbf{m}})$, while with probability $1 - p(m_{\mathbf{m}})$, they exit the system. Although the probability of requiring additional individual therapy depends on the group size, for simplicity, we will use p in place of $p(m_{\mathbf{m}})$. In Section 4.1, we analyze the effects of different patterns of $p(m_{\mathbf{m}})$.

To keep our model general and facilitate the generalization to multiple classes, we assume that each patient in the individual channel requires $m_{\mathbf{s}}$ servers. Note that $m_{\mathbf{m}}$ and $m_{\mathbf{s}}$ can take any positive real value. In practice, $m_{\mathbf{m}}$ is typically a rational number, determined by the therapist-to-patient ratio. For instance, $m_{\mathbf{m}} = 1/5$ means that each patient requires one-fifth of a server (equivalently, a group consists of five participants) and $m_{\mathbf{s}} = 1$. By adjusting the unit of measurement, one-fifth of a server can be redefined as a unit of service capacity, in which case we set $m_{\mathbf{m}} = 1$ and $m_{\mathbf{s}} = 5$. In our simulation experiments, we use this integer adjustment.

Figure 2 **Model illustration of group and individual service channels with no-shows, dropouts and abandonments.**



Next, we introduce the stochastic components of the model. Let $X_j(t)$ denote the total number patients in Channel j, $j \in \mathcal{J}$ at time t, $t \geq 0$. Similarly, let $Q_j(t)$ represent the number of patients waiting in queue for Channel j at time t. We define $X(t) = (X_j(t),\ j \in \mathcal{J})$ and $Q(t) = (Q_j(t),\ j \in \mathcal{J})$, so that the system state at time t is described by $(X(t), Q(t))$. We denote by $Z_j(t)$ the number of servers assigned to patients in Channel j at time t. The decision variables $Z(t) = (Z_j(t),\ j \in \mathcal{J})$ must satisfy the following conditions for every $t \geq 0$ and $j \in \mathcal{J}$:

$$\sum_{j \in \mathcal{J}} Z_j(t) \leq N, \qquad X_j(t) - \frac{Z_j(t)}{m_j} = Q_j(t) \geq 0, \qquad \frac{Z_j(t)}{m_j} \in \mathbb{N}^{+}.$$

Note that the number of servers in our model represents full-time equivalent (FTE) resources, reflecting the effective availability for MTE-related treatment after accounting for existing commitments, such as time allocated to pre-MTE mental health needs.

Service, abandonment and associated costs. The treatment for patients at risk of developing PTSD consists of a series of 12 to 24 psychological sessions, scheduled once or twice a week; the number of sessions and their frequency is set depending on the patient's condition and type of exposure. Since the number of sessions is determined at the outset of treatment, when all sequential appointments are allocated with *the same* therapist, we model the entire appointment sequence as a single service time. If the group sessions prove ineffective, the patient begins an individual series of sessions, with a fixed duration set at the outset of treatment.

While awaiting the first (group/individual) session, patients may leave the queue in search of alternative support. Treatment and patience times for each class of patients follow exponential distributions with rates μ_j^b and θ_j, respectively, for Channel $j \in \mathcal{J}$.

Let $\Gamma_j(t)$ and $D_j(t)$ denote the cumulative number of abandonments and treatment completions, respectively, in Channel j, up to time $t \geq 0$. Let h_j denote the holding cost per time unit for each patient in Channel j, and b_j denote the cost savings from treatment completion of each Channel j patient[1]. Finally, an abandonment cost, α_j, is incurred for each patient who abandons while waiting for treatment in Channel j.

No-shows and dropouts. No-shows and dropouts are common in mental health treatments, particularly for PTSD patients (Milicevic et al. 2020, Xaba et al. 2024, Fenger et al. 2011). No-shows occur when a patient misses an appointment without prior notice. Beyond the negative impact on the patient's well-being and treatment outcome, no-shows result in unoccupied and wasted treatment slots, as each no-show patient rejoins the system for an additional appointment. Moreover, no-shows can occur multiple times during a session series (i.e., patients may miss more than one session), further disrupting the scheduling process. To capture this capacity loss, we define β_j as the probability of a patient being served in Channel j to show up for a scheduled appointment, where $1-\beta_j$ represents the probability of a no-show.

[1] Incorporating different cost savings for patients who complete their service in the group channel but require additional individual service can be easily achieved by partitioning $D_{\mathbf{m}}(T)$ into two groups: those requiring additional individual service and those who do not. Each group can then be assigned its corresponding cost savings. For simplicity in the formulation, we leave them combined.

To model the total service requirement accounting for no-shows, recall that μ_j^b denotes the mean required service time, assuming patients attend all scheduled appointments. The number of attended appointments required to complete treatment can be viewed as a binomial random variable with n attempts and a success probability of β_j. The expected number of attended appointments is therefore $n\beta_j$, which equals $1/\mu_j^b$, the required number of completed appointments. Consequently, the total number of required appointment attempts – accounting for no-shows – corresponds to $1/(\beta_j\mu_j^b) = 1/\mu_j$, where we define $\mu_j := \beta_j\mu_j^b$ as the *effective* service rate. The fact that the effective service rate is lower than the basic service rate (i.e., effective service times are longer) contributes to greater system overload and cost.

This modeling approach is similar to Huang et al. (2015), who studied the scheduling of in-process (IP) patients in the emergency department (ED) who occasionally return for follow-up checks. However, in Huang et al. (2015), patient returns occur upon treatment completion, whereas in our case, the service consists of a series of sessions, during which a patient may miss and reschedule multiple appointments due to no-shows.

While no-shows allow the patient to remain in the system and attend subsequent sessions after missing one, dropouts occur when a patient completely quits the program. In this case, the patient permanently leaves the system, and their future appointments can be reassigned to other patients. We define γ_j as the dropout rate from Channel j, and h_j^d as the cost associated with each such dropout. We denote by $\Gamma_j^d(t)$ the cumulative number of dropouts, respectively, from Channel j, up to time $t \geq 0$.

3.1. Scheduling Policy and Overall Cost Savings

A scheduling policy $\pi \in \Omega$ determines the allocation of servers channels, where Ω denotes the set of admissible controls, namely, all non-anticipating scheduling policies. That is, server allocations are made based on the current state (X, Q) only. Under such scheduling policies, $\{(X(t), Q(t)) : t \geq 0\}$ is a Markov process.

Since the process $\{(X(t), Q(t)) : t \geq 0\}$ depends on the scheduling policy π, we can explicitly indicate this dependence by expressing the stochastic process as $\{(X^\pi(t), Q^\pi(t)) : t \geq 0\}$, along with $D_j^\pi(t)$, $\Gamma_j^\pi(t)$, and $\Gamma_j^{\pi,d}(t)$. For simplicity, we will omit the subscript π when the context makes the dependence on the policy clear.

The overall cost savings for both channels over $[0,T]$ is, therefore,

$$\mathbb{E}\left[\sum_{j\in\mathcal{J}}\left[b_jD_j(T)-\alpha_j\Gamma_j(T)-h_j^d\Gamma_j^d(T)\right]-\int_0^T\sum_{j\in\mathcal{J}}h_jQ_j(t)\mathrm{d}t\right].$$

The Markovian modeling assumption implies that for any $j\in\mathcal{J}$,

$$\mathbb{E}\left[D_j(T)\right]=\frac{\mu_j}{m_j}\mathbb{E}\left[\int_0^T Z_j(t)\,\mathrm{d}t\right],\quad \mathbb{E}\left[\Gamma_j^d(T)\right]=\frac{\gamma_j}{m_j}\mathbb{E}\left[\int_0^T Z_j(t)\,\mathrm{d}t\right],$$

and

$$\mathbb{E}\left[\Gamma_j(T)\right]=\theta_j\mathbb{E}\left[\int_0^T Q_j(t)\,\mathrm{d}t\right].$$

Therefore, the overall cost savings can be rewritten as

$$\mathbb{E}\left[\int_0^T\sum_{j\in\mathcal{J}}\left[\frac{1}{m_j}\left(b_j\mu_j-h_j^d\gamma_j\right)Z_j(t)-\left(h_j+\alpha_j\theta_j\right)Q_j(t)\right]\,\mathrm{d}t\right].$$

By defining the generalized cost savings and the generalized holding cost as follows:

$$r_j:=\frac{1}{m_j}\left(b_j\mu_j-h_j^d\gamma_j\right),\quad c_j:=h_j+\alpha_j\theta_j, \tag{1}$$

the overall cost savings becomes:

$$\mathbb{E}\left[\int_0^T\sum_{j\in\mathcal{J}}\left[r_jZ_j(t)-c_jQ_j(t)\right]\,\mathrm{d}t\right].$$

Each of the three response phases we analyze is associated with a different cost savings objective: transient cost savings maximization during and immediately after the event, and long-run average cost savings maximization in the later stage. In all cases, the problem is an MDP. The curse of dimensionality (Papadimitriou and Tsitsiklis 1999) – a large (infinite) state-space and policy-space – makes it prohibitively hard to solve and characterize the optimal scheduling policy. To gain structural insights into the optimal scheduling policy, we take a deterministic fluid approach. Fluid models are known to provide good approximation of the first-order mean dynamics of stochastic systems, and are thus useful for a variety of applications related to service operations management (Whitt 2002).

Moreover, analyzing the stochastic system in this setting is challenging because both channels are interdependent, and the group channel involves a large number of possible server-to-patient permutations (Armony and Bambos 2003, Zychlinski et al. 2023, Grosof and Harchol-Balter 2023). The deterministic continuous fluid approach we adopt enables a more tractable analysis and yields structural insights. Given the overloaded nature of systems during MTEs, fluid models provide a natural and effective framework for deriving operational insights into the optimal solution.

3.2. The Fluid Model

For the fluid model, we remove the integer constraints and replace the discrete stochastic processes (i.e., arrivals, departures, abandonments, no-shows, dropouts, and referrals from the group channel to the individual one) by their corresponding deterministic flow rates. Denote by $\lambda(t)$ the arrival rate of Class i customers to the group channel at time t.

In the fluid model, we use lowercase $x_j(t)$ and $q_j(t)$, $j \in \mathcal{J}$, to denote the fluid content in the system and in the queue, respectively. We use $z_j(t)$ to denote the allocation of service capacity to Channel j, such that $\sum_{j\in\mathcal{J}} z_j(t) \le N$, for all $t \ge 0$. Similarly to the stochastic system, we define $x(t) = (x_j(t),\ j \in \mathcal{J})$, $q(t) = (q_j(t),\ j \in \mathcal{J})$, and $z(t) = (z_j(t),\ j \in \mathcal{J})$.

The fluid scheduling policy π determines how to prioritize and allocate resources across both channels. Specifically, the fluid dynamics is characterized by the following set of differential equations (DEs):

$$\dot{q}_{\mathbf{m}}(t) = \lambda(t) - \theta_{\mathbf{m}} q_{\mathbf{m}}(t) - (\mu_{\mathbf{m}} + \gamma_{\mathbf{m}})\, z_{\mathbf{m}}(t)/m_{\mathbf{m}};$$
$$\dot{q}_{\mathbf{s}}(t) = p\mu_{\mathbf{m}} z_{\mathbf{m}}(t)/m_{\mathbf{m}} - \theta_{\mathbf{s}} q_{\mathbf{s}}(t) - (\mu_{\mathbf{s}} + \gamma_{\mathbf{s}})\, z_{\mathbf{m}}(t)/m_{\mathbf{s}}.$$

The first DE applies to the group channel and accounts for abandonments while waiting, treatment completion rate, and dropout rate from treatment. Note that no-shows are factored into the system's dynamics through the effective service rate μ_j. The second DE applies to the individual channel. Here, the arrival rate is the treatment completion rate from the group channel multiplied by the probability of requiring an additional individual treatment. Note that on average in Channel j, each unit of service capacity allocated enables the treatment of μ_j/m_j patients over one unit of time.

Definition 1 (Admissible Fluid Scheduling Policy). A scheduling policy is said to be admissible if it is Markovian and satisfies the following constraints for all $t \ge 0$:

$$\sum_{j\in\mathcal{J}} z_j(t) \le N, \quad z_j(t) \ge 0, \qquad \dot{q}_j(t) \ge 0 \ \text{ whenever } q_j(t) = 0, \quad j \in \mathcal{J},$$

where the third constraint ensures that the queue length processes remain non-negative, preventing the queues from becoming negative when they are empty.

In the following sections, we analyze the three response phases following an MTE and define the transition points between them. Rather than proceeding chronologically, we present the phases in order of analytical complexity. This structure reflects our solution approach: the Long-Term Care Phase provides the foundation for the Recovery Phase, which in turn informs the Surge Phase.

We first solve the Long-Term Care and Recovery Phases independently, each with its own dynamics, constraints, and objectives. The Surge Phase is then addressed sequentially, incorporating the Recovery Phase's value function into its objective. In this framework, the Long-Term Care Phase yields the $\mathcal{P}$ rule, which determines optimal prioritization between group and individual channels. This rule underpins the more advanced scheduling policies used in the Recovery and Surge Phases.

4. Long-Term Care Phase

The impact of MTEs on patients' mental health can persist for years and even decades (Bowler et al. 2016). This section focuses on developing a scheduling policy and long-term resource allocation that maximizes the long-run average cost savings. We assume that arrivals to the group channel follow a time-homogeneous Poisson process with rate λ.

The long-run cost savings maximization problem for the fluid model is formulated as the following infinite-dimensional linear program:

$$\begin{aligned} \max_{z,q} \quad & \liminf_{T\to\infty} \frac{1}{T}\int_0^T \sum_{j\in\mathcal{J}} \left[r_j z_j(t) - c_j q_j(t)\right] \mathrm{d}t \\ \text{s.t.} \quad & \dot{q}_{\mathbf{m}}(t) = \lambda - \theta_{\mathbf{m}} q_{\mathbf{m}}(t) - (\mu_{\mathbf{m}} + \gamma_{\mathbf{m}})\, z_{\mathbf{m}}(t)/m_{\mathbf{m}}, \quad t \geq 0 \\ & \dot{q}_{\mathbf{s}}(t) = p\mu_{\mathbf{m}} z_{\mathbf{m}}(t)/m_{\mathbf{m}} - \theta_{\mathbf{s}} q_{\mathbf{s}}(t) - (\mu_{\mathbf{s}} + \gamma_{\mathbf{s}})\, z_{\mathbf{s}}(t)/m_{\mathbf{s}}, \quad t \geq 0 \\ & \sum_{j\in\mathcal{J}} z_j(t) \leq N, \quad z_j(t) \geq 0, \quad q_j(t) \geq 0, \quad j \in \mathcal{J}, \quad t \geq 0. \end{aligned} \tag{2}$$

If the fluid dynamics converge to an equilibrium point $(\bar{q}, \bar{z})$ as $t \to \infty$, then maximizing the long-run average cost savings reduces to a *finite*-dimensional linear program. This observation, formalized in Theorem 1, leads to the following problem formulation:

$$\begin{aligned} \max_{\bar{q},\bar{z}} \quad & \sum_{j\in\mathcal{J}} \left[r_j \bar{z}_j - c_j \bar{q}_j\right] \\ \text{s.t.} \quad & \bar{q}_{\mathbf{m}} = \tfrac{1}{\theta_{\mathbf{m}}} \left(\lambda - (\mu_{\mathbf{m}} + \gamma_{\mathbf{m}})\, \tfrac{\bar{z}_{\mathbf{m}}}{m_{\mathbf{m}}}\right), \\ & \bar{q}_{\mathbf{s}} = \tfrac{1}{\theta_{\mathbf{s}}} \left(p\mu_{\mathbf{m}} \tfrac{\bar{z}_{\mathbf{m}}}{m_{\mathbf{m}}} - (\mu_{\mathbf{s}} + \gamma_{\mathbf{s}})\, \tfrac{\bar{z}_{\mathbf{s}}}{m_{\mathbf{s}}}\right), \\ & \sum_{j\in\mathcal{J}} \bar{z}_j \leq N, \quad \bar{z}_j \geq 0, \quad \bar{q}_j \geq 0, \quad j \in \mathcal{J}. \end{aligned} \tag{3}$$

Rearranging (3) and removing the constants that do not affect the optimization yield the following equivalent problem:

$$\begin{aligned}
\max_{\bar{z}} \quad & \mathcal{P}_{\mathbf{m}} \bar{z}_{\mathbf{m}} + \mathcal{P}_{\mathbf{s}} \bar{z}_{\mathbf{s}} \\
\text{s.t.} \quad & 0 \leq \bar{z}_{\mathbf{m}} \leq \frac{\lambda m_{\mathbf{m}}}{(\mu_{\mathbf{m}} + \gamma_{\mathbf{m}})}, \\
& 0 \leq \bar{z}_{\mathbf{s}} \leq \frac{p \mu_{\mathbf{m}} m_{\mathbf{s}}}{(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}})\, m_{\mathbf{m}}} \ \bar{z}_{\mathbf{m}}, \\
& \bar{z}_{\mathbf{m}} + \bar{z}_{\mathbf{s}} \leq N,
\end{aligned} \tag{4}$$

where the $\mathcal{P}$ indexes are defined as:

$$\begin{aligned}
\mathcal{P}_{\mathbf{m}} &:= r_{\mathbf{m}} + \frac{c_{\mathbf{m}}}{\theta_{\mathbf{m}} m_{\mathbf{m}}} (\gamma_{\mathbf{m}} + \mu_{\mathbf{m}}) - \frac{p c_{\mathbf{s}}}{\theta_{\mathbf{s}} m_{\mathbf{m}}} \mu_{\mathbf{m}}, \\
\mathcal{P}_{\mathbf{s}} &:= r_{\mathbf{s}} + \frac{c_{\mathbf{s}}}{\theta_{\mathbf{s}} m_{\mathbf{s}}} (\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}).
\end{aligned} \tag{5}$$

Recall that r_j and c_j, $j \in \mathcal{J}$, as defined in (1), depend on rates and costs of service, no-shows, dropouts, and abandonments.

Since problem (4) is a linear program, the optimal solution tends to assign a larger value to $\bar{z}_i$ when it has a larger $\mathcal{P}_i$ coefficient. Furthermore, we observe that the second constraint in (4) defines a dependency between $\bar{z}_m$ and $\bar{z}_s$ – the group channel serves as the feeding source for the individual channel. Therefore, the long-run resource allocation must also be interconnected.

To better understand this dependency, we first define the following index policy.

Definition 2. (The $\mathcal{P}$ index rule)

- When $\mathcal{P}_{\mathbf{m}} > P_{\mathbf{s}}$: Prioritize group patients over the individual ones.
- When $\mathcal{P}_{\mathbf{m}} < P_{\mathbf{s}}$: Prioritize individual patients over the group ones.

Theorem 1 establishes the optimality of the $\mathcal{P}$ rule scheduling policy and characterizes the long-run resource allocation to the group channel and subsequent individual channel. For the sake of simplicity, we assume that the indices are unique; otherwise, multiple optimal scheduling policies may prevail, which can complicate the analysis.

Theorem 1 **(optimality of the $\mathcal{P}$ index rule)**. *For the long-run cost savings maximization problem (2), with $\theta_j > 0$, $j \in J$, and any initial condition, the $\mathcal{P}$ rule in Definition 2 is optimal, and*

- ***If*** $\boldsymbol{\mathcal{P}_{\mathbf{m}} > \mathcal{P}_{\mathbf{s}}}$ *(group patients are prioritized): The long-run average resource allocation is:*

$$\bar{z}_{\mathbf{m}} = N \wedge \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}, \quad \bar{z}_{\mathbf{s}} = (N - \bar{z}_{\mathbf{m}}) \wedge \left(\frac{p \mu_{\mathbf{m}} m_{\mathbf{s}}}{(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}})\, m_{\mathbf{m}}} \bar{z}_{\mathbf{m}} \right), \tag{6}$$

where $x \wedge y = \min(x, y)$*.*

- ***If*** $\boldsymbol{\mathcal{P}_{\mathbf{m}} < \mathcal{P}_{\mathbf{s}}}$ *(individual patients are prioritized): The long-run average resource allocation is:*

$$\bar{z}_{\mathbf{m}} = \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}} \wedge \frac{(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}})\, m_{\mathbf{m}} N}{m_{\mathbf{m}} (\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}) + p \mu_{\mathbf{m}} m_{\mathbf{s}}}, \quad \bar{z}_{\mathbf{s}} = \frac{p \mu_{\mathbf{m}} m_{\mathbf{s}}}{(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}})\, m_{\mathbf{m}}} \bar{z}_{\mathbf{m}}. \tag{7}$$

When the group channel is prioritized over the individual channel ($\mathcal{P}_{\mathbf{m}} > \mathcal{P}_{\mathbf{s}}$), resources are allocated to both service channels separately: first to the group channel, then to the individual channel. In contrast, when the individual channel is prioritized ($\mathcal{P}_{\mathbf{m}} < \mathcal{P}_{\mathbf{s}}$), resource allocation must be coordinated to ensure that the group channel—serving as a feeder to the individual channel—receives sufficient capacity. In this case, resources must be allocated jointly across both channels to ensure that the allocation to the individual channel satisfies the upper bound constraint in (4). In Appendix C.1.2 we discuss how to manage idleness in small systems.

4.1. Numerical Experiments – Setting the Optimal Group Size

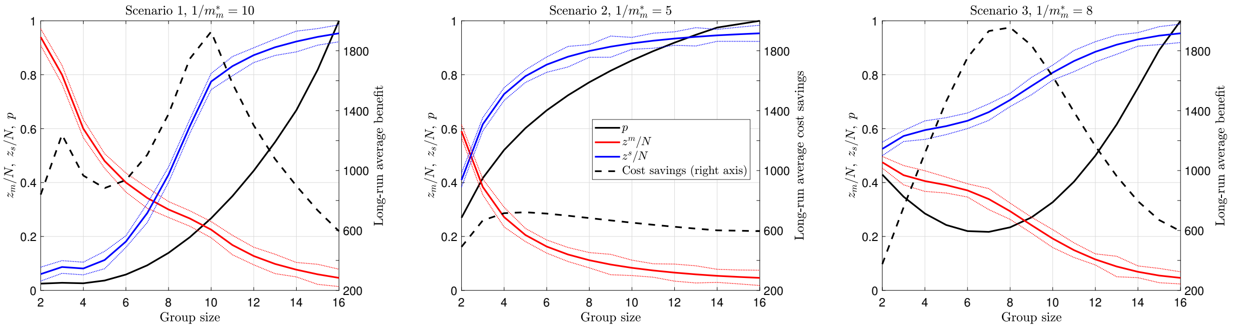
Empirical evidence indicates that clinical outcomes are influenced by group size. For instance, effectiveness may follow a quadratic relationship with group size, or fall within specific ranges depending on the diagnosis and treatment context (Daley et al. 1983, Yalom and Leszcz 2020, Dueweke et al. 2022). To capture these variations, our model assumes a general relationship between group size and effectiveness.

This section serves two purposes. The first is to demonstrate the optimal group size under different group-size effect assumptions, and the second is to evaluate the effectiveness of the proposed $\mathcal{P}$ rule in the corresponding stochastic system via simulation.

Specifically, in addition to determining the optimal scheduling and long-run average resource allocation for each channel, the formulation in (4)–(5) can be used to identify the optimal group size, $1/m_{\mathbf{m}}$. The effectiveness of each group size is captured through the probability of requiring additional individual therapy following group therapy, which depends on the structure of $p(m_{\mathbf{m}})$. Given this relationship, the optimal group size for each class can be found by solving the small-dimensional optimization problem in (4) for a range of reasonable values (typically between 3 and 15).

Figure 3 illustrates three group-size effects reported in the literature on optimal resource allocation between group and individual therapy channels. In Scenarios 1 and 2, $p(m_{\mathbf{m}})$ is increasing convex or concave, respectively, indicating that as group size increases, groups become less effective. In Scenario 3, $p(m_{\mathbf{m}})$ follows a quadratic pattern, where both small and large groups are less effective.

Figure 3 **Long-run average resource allocation for different group size effects. The solid lines represent the fluid solution, while the dashed lines represent the 95% confidence intervals based on the simulation results. The parameters are** $\lambda = 120$, $N = 50$, $\mu_{\mathbf{m}} = 0.9$, $\mu_{\mathbf{s}} = 0.6$, $\gamma_{\mathbf{m}} = 0.1$, $\gamma_{\mathbf{s}} = 0.1$, $\theta_{\mathbf{m}} = 0.1$, $\theta_{\mathbf{s}} = 0.1$, $m_{\mathbf{s}} = 1$, $r_{\mathbf{m}} = 26$, $r_{\mathbf{s}} = 46$, $c_{\mathbf{m}} = 2$, $c_{\mathbf{s}} = 16$.



Each plot shows the optimal long-run resource allocation for each channel – $z_{\mathbf{m}}$ and $z_{\mathbf{s}}$ – as a function of group size, based on both the optimal fluid solution and a stochastic simulation model. The fluid solution is obtained by solving the linear program in (4)–(5), while the stochastic simulation solution is obtained by following the $\mathcal{P}$ rule for channel prioritization in the simulation and calculating the long-run resource allocation and cost savings. In the latter, for each group size and effect, we computed the average and the 95% confidence interval over 100 replications, each with $T = 10,000$.

The long-run average cost savings, shown as a dashed line corresponding to the right y-axis, varies in shape across scenarios and affects the optimal group size. The optimal group size is 10 in the first scenario, 5 in the second scenario, and 8 in the third scenario. These findings demonstrate that there is no one-size-fits-all group size. Therefore, decision-makers must incorporate the group size effect for each patient class when designing such systems. Additionally, resource allocation to both channels must be done in coordination since the channels are interdependent. Lastly, these findings suggest that the fluid-based $\mathcal{P}$-rule policy is accurate and effective when implemented in the original stochastic system.

5. Recovery Phase

Some time after the initial Surge Phase, arrival rates begin to stabilize; however, a significant backlog of patients requiring mental health support remains. The goal is to determine an optimal scheduling policy that efficiently addresses the backlog and transitions the system toward a more sustainable operational state[2].

In Section 6, we analyze the Surge Phase, characterized by highly variable arrival rates and critically insufficient resources to meet demand. In contrast, during the Recovery Phase, we assume that the system has sufficient long-run capacity to serve all incoming patients. Specifically, we impose the condition:

$$\rho := \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}} + \frac{\lambda p m_{\mathbf{s}}}{\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}} < N, \tag{8}$$

which guarantees that the system is capable of eventually clearing the backlog.

Accordingly, we define the transition from the Surge Phase to the Recovery Phase to occur at the first time when two conditions are met: (i) the arrival rate stabilizes—that is, the average arrival rate λ remains approximately constant over time—and (ii) this stabilized λ satisfies inequality (8). While detecting a constant arrival rate is inherently data-driven, a practical implementation may rely on observing sustained periods without significant changes. In Section 7, we demonstrate how to identify this transition using real arrival rate data.

We define the stopping time $\tau := \inf\{t \geq 0 : q_{\mathbf{m}}(t) + q_{\mathbf{s}}(t) = 0\}$, as the transition point from the Recovery Phase to the Long-Term Care Phase, that is, the first time at which both queues are fully depleted. Based on the proof of Theorem 1, and under the assumption that the system has sufficient capacity in this phase (i.e., condition (8) holds), there exists a scheduling policy such that $\tau < \infty$ for any initial condition.

[2] At the end of this phase when transition to the Long-Term Phase, the number of therapists which was expanded through therapists recruitment from the private sector or reserves and the extension of working hours of part-time therapists for the Surge and Recovery Phase can be reduced.

We focus on the following transient optimization problem

$$\begin{aligned} \max_{z,q} \quad & \int_0^{\tau} \sum_{j\in\mathcal{J}} \left[r_j z_j(t) - c_j q_j(t)\right] \,\mathrm{d}t \\ \text{s.t.} \quad & \dot{q}_{\mathbf{m}}(t) = \lambda - \theta_{\mathbf{m}} q_{\mathbf{m}}(t) - \left(\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}\right) z_{\mathbf{m}}(t)/m_{\mathbf{m}}, \\ & \dot{q}_s(t) = p\mu_{\mathbf{m}} z_{\mathbf{m}}(t)/m_{\mathbf{m}} - \theta_{\mathbf{s}} q_{\mathbf{s}}(t) - \left(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}\right) z_{\mathbf{s}}(t)/m_{\mathbf{s}}, \\ & \sum_{j\in\mathcal{J}} z_j(t) \le N, \quad z_j(t) \ge 0, \quad q_j(t) \ge 0, \quad j \in \mathcal{J}. \end{aligned} \tag{9}$$

Let τ^* denote the time required to empty the queue under the optimal policy. Problem (9) falls under the category of optimal control with state constraints. These problems are typically difficult to solve due to the boundary conditions they impose (Hartl et al. 1995).

We begin by establishing that the exclusion of irregular boundary behaviors (i.e., the non-negativity of queue lengths) holds strictly and proving that it suffices to focus on trajectories that avoid both chattering points and chattering intervals. These results are formally developed in Appendix A.

Next, we note that the transient system dynamics are complex due to specific features of our model—particularly the asymmetric dependency between both service channels, with one serving as the source for the other. To gain insight into the solution structure, we solve the discrete-time version of (9), formulated as the following finite-dimensional linear program:

$$\begin{aligned} \max_{z,q} & \sum_{t=0}^{\tau} \sum_{j\in\mathcal{J}} \Delta t \left[r_j z_j(t) - c_j q_j(t)\right] \\ \text{s.t.} \quad & q_m(t+\Delta t) = q_m(t) + \Delta t \left[\lambda - \theta_{\mathbf{m}} q_{\mathbf{m}}(t) - \frac{(\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}) z_{\mathbf{m}}(t)}{m_{\mathbf{m}}}\right], \quad t \in [0, \tau - \Delta t], \\ & q_{\mathbf{s}}(t+\Delta t) = q_{\mathbf{s}}(t) + \Delta t \left[\frac{p\mu_{\mathbf{m}} z_{\mathbf{m}}(t)}{m_{\mathbf{m}}} - \theta_{\mathbf{s}} q_{\mathbf{s}}(t) - \frac{(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}) z_{\mathbf{s}}(t)}{m_{\mathbf{s}}}\right], \quad t \in [0, \tau - \Delta t], \\ & \sum_{j\in\mathcal{J}} z_j(t) \le N, \quad z_j(t) \ge 0, \quad q_j(t) \ge 0, \quad j \in \mathcal{J}, \quad t \in [0, \tau - \Delta t], \end{aligned} \tag{10}$$

where Δt is the discretization step size.

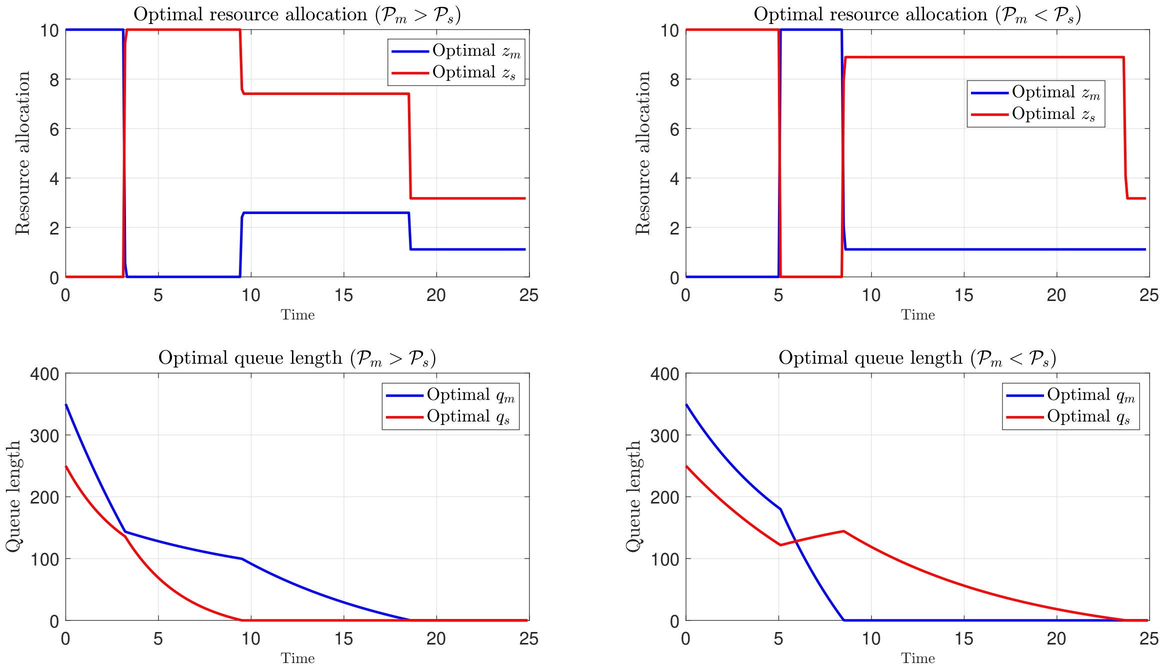
Our extensive numerical experiments show that in some cases, the optimal policy in the transient setting is to follow the $\mathcal{P}$ index rule throughout the entire horizon. In other cases – particularly when initial queue lengths are large – the optimal policy initially applies strict priority according to the $\mathcal{P}$ rule, and then at some point the priority is switched. Similar switching behavior has been observed in other transient queueing systems (Cohen et al. 2014, Hu et al. 2022), though those models do not involve re-entrant flows or interdependent service channels, as in our setting.

Figure 4 presents two such scenarios, each with $N = 10$ servers and initial queue lengths of $q_{\mathbf{m}}(0) = 350$ and $q_{\mathbf{s}}(0) = 250$. The top plots illustrate the optimal resource allocation to both channels over time, while the bottom plots depict the evolution of the corresponding queue lengths. In the left scenario, $\mathcal{P}_{\mathbf{m}} > \mathcal{P}_{\mathbf{s}}$. Initially, all resources are allocated to the group channel, in line with the $\mathcal{P}$ index rule. We note that the decrease in q_s is not linear due to abandonments. Around time $t \approx 4$, the priority shifts, and all resources are allocated to the individual channel. This continues until the individual queue is depleted. At this point, resources are allocated in coordination across both channels, ensuring that patients completing the group channel who require the individual channel do not form a queue. Once both queues are empty, the server allocation ensures that they remain empty, as there are now sufficient resources to treat all arriving patients.

In the right scenario, the situation is reversed, with $\mathcal{P}_{\mathbf{m}} < \mathcal{P}_{\mathbf{s}}$. Initially, all resources are allocated to the individual channel, again consistent with the $\mathcal{P}$ index rule. At time $t = 5$, the priority switches, and all resources are shifted to the group channel. This continues until the group queue is emptied. Afterward, the group channel receives the necessary resources to maintain a zero queue, while the remaining servers are allocated to the individual channel until its queue is cleared. Once both queues are empty, the server allocation ensures that they remain empty, as there are now sufficient resources to treat all arriving patients.

To gain a deeper understanding of the switching phenomenon, we analyze the optimal trajectories of the queue lengths in both channels, starting from various initial conditions as indicated by the top-right starting point of each line. Figure 5 presents four representative examples. The top plots use the same parameters as in Figure 4, but with different initial queue lengths. At the beginning, when the queue lengths are very large following the Surge Phase, the optimal prioritization adheres to the $\mathcal{P}$ rule (Channel $\mathbf{m}$ on the left plot and Channel $\mathbf{s}$ on the right plot). As the queue lengths decrease, the priority switches to the other channel. The dashed lines in these figures connect the points where the priority switches. From that point onward, the same priority is maintained until both queues are empty. Notably, in these two cases, as well as in all other cases we examined, the switching line appears to follow a nearly linear pattern.

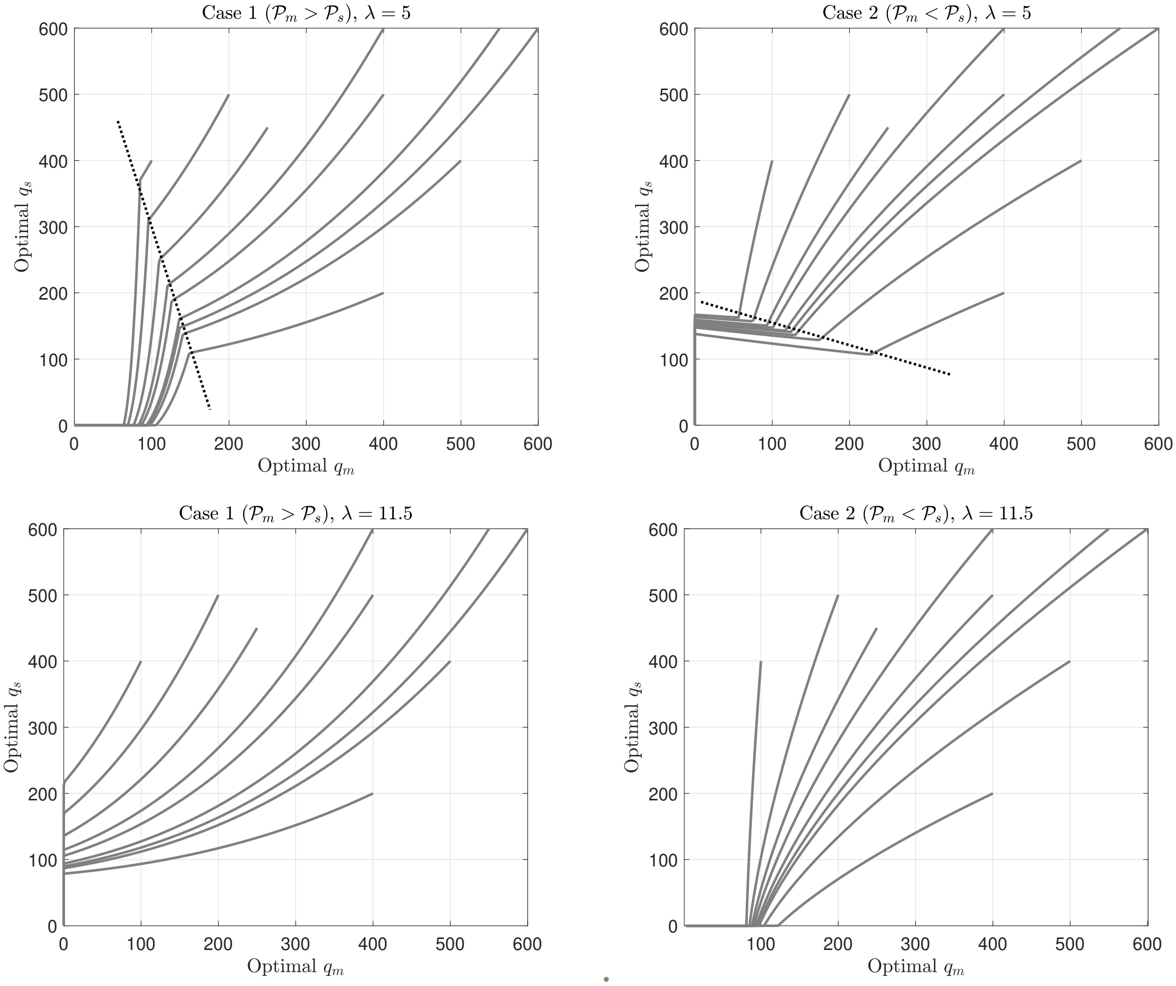
Figure 4 **Optimal transient resource allocation and queue length for each channel. In the left plots:** $N=10$, $\lambda=5$, $\theta_{\mathbf{m}}=0.1$, $\theta_{\mathbf{s}}=0.3$, $\mu_{\mathbf{m}}=0.8$, $\mu_{\mathbf{s}}=0.6$, $\gamma_{\mathbf{m}}=0.1$, $\gamma_{\mathbf{s}}=0.1$, $p=0.5$, $m_{\mathbf{m}}=1/5$, $m_{\mathbf{s}}=1$, $r_{\mathbf{m}}=100\mu_{\mathbf{m}}/m_{\mathbf{m}}$, $r_{\mathbf{s}}=400\mu_{\mathbf{s}}/m_{\mathbf{s}}$, $c_{\mathbf{m}}=2$, $c_{\mathbf{s}}=4$, $q_{\mathbf{m}}(0)=350$, $q_{\mathbf{s}}(0)=250$. **In the right plots:** $N=10$, $\lambda=5$, $\theta_{\mathbf{m}}=0.15$, $\theta_{\mathbf{s}}=0.1$, $\mu_m=0.8$, $\mu_{\mathbf{s}}=0.6$, $\gamma_{\mathbf{m}}=0.1$, $\gamma_{\mathbf{s}}=0.1$, $p=0.5$, $m_{\mathbf{m}}=1/5$, $m_{\mathbf{s}}=1$, $r_{\mathbf{m}}=50\mu_{\mathbf{m}}/m_{\mathbf{m}}$, $r_{\mathbf{s}}=400\mu_{\mathbf{s}}/m_{\mathbf{s}}$, $c_{\mathbf{m}}=2$, $c_{\mathbf{s}}=3$, $q_{\mathbf{m}}(0)=350$, $q_{\mathbf{s}}(0)=250$.



To determine this curve, we solve the discrete-time version (10) for a range of representative initial queue lengths, $(q_{\mathbf{m}}(0), q_{\mathbf{s}}(0))$, that sufficiently span the relevant state space. These initial conditions generate a set of optimal trajectories from which we identify the boundary between the two priority regimes. We approximate this boundary by fitting a linear function of the form $q_s = \kappa q_m + d$, where κ is the slope and d is the intercept. Specifically, we apply simple linear regression to the set of switching points (q_m^i, q_s^i), where $i = 1, 2, \ldots, I$ indexes the selected representative points. This approach provides flexibility in capturing the approximate slope of the boundary curve. In practice, the switching curve appears nearly linear, so this approximation is both simple and effective for visualization and interpretation.

Therefore, for the Recovery Phase, we propose the *recovery-based* $\mathcal{P}$ *rule*, which utilizes this switching curve: as long as $q_{\mathbf{m}} + q_{\mathbf{s}}$ exceeds the curve, the system follows the $\mathcal{P}$ rule; once $q_{\mathbf{m}} + q_{\mathbf{s}}$ falls below the curve, the system switches priority to the inverse $\mathcal{P}$ rule.

Figure 5 **Optimal trajectories of queue lengths for different initial conditions. In the left plots (Case 1):** $T = 25$, $N = 10$, $\theta_{\mathbf{m}} = 0.1$, $\theta_{\mathbf{s}} = 0.3$, $\mu_{\mathbf{m}} = 0.8$, $\mu_{\mathbf{s}} = 0.6$, $\gamma_{\mathbf{m}} = 0.1$, $\gamma_{\mathbf{s}} = 0.1$, $p = 0.5$, $m_{\mathbf{m}} = 1/5$, $m_{\mathbf{s}} = 1$, $r_{\mathbf{m}} = 100\mu_{\mathbf{m}}/m_{\mathbf{m}}$, $r_{\mathbf{s}} = 400\mu_{\mathbf{s}}/m_{\mathbf{s}}$, $c_{\mathbf{m}} = 2$, $c_{\mathbf{s}} = 4$. **In the right plots (Case 2):** $T = 50$, $N = 10$, $\theta_{\mathbf{m}} = 0.15$, $\theta_{\mathbf{s}} = 0.1$, $\mu_{\mathbf{m}} = 0.8$, $\mu_{\mathbf{s}} = 0.6$, $\gamma_{\mathbf{m}} = 0.1$, $\gamma_{\mathbf{s}} = 0.1$, $p = 0.5$, $m_{\mathbf{m}} = 1/5$, $m_{\mathbf{s}} = 1$, $r_{\mathbf{m}} = 50\mu_{\mathbf{m}}/m_{\mathbf{m}}$, $r_{\mathbf{s}} = 400\mu_{\mathbf{s}}/m_{\mathbf{s}}$, $c_{\mathbf{m}} = 2$, $c_{\mathbf{s}} = 3$.



Note that the bottom plots of Figure 5 use the same parameters as the top ones, except for a higher arrival rate λ, resulting in a heavier system load (with $\rho = 0.45$ in the top plots and $\rho = 1$ in the bottom plots). When the system is highly loaded – as is often the case in MTEs, which are the focus of this study – the optimal policy maintains a consistent priority throughout the horizon, adhering to the standard $\mathcal{P}$ rule. In such cases, the recovery-based $\mathcal{P}$ rule coincides with the standard $\mathcal{P}$ rule.

In Appendix B.1, we assess the performance of the proposed policy through stochastic simulation. The results show that the difference between the optimal fluid objective and the average transient cost savings in the stochastic model is very small across all cases. This difference is particularly small when the traffic intensity is high and no priority switching occurs, similar to the bottom plots in Figure 5.

Remark 1 (Setting the optimal group size). While the analysis of optimal group size was demonstrated explicitly for the Long-Term Care Phase in Section 4.1, a similar procedure could, in principle, be applied to the surge and recovery phases. Specifically, one could evaluate a range of candidate group sizes, apply the proposed scheduling policy for each, derive the corresponding switching curves (which would differ across group sizes), and select the group size that maximizes cost savings. Such a procedure would need to be conducted offline, based on a forecasted arrival pattern. This type of analysis could also serve as part of preparedness planning for future MTEs.

6. Surge Phase

In typical MCEs, the Surge Phase – where arrival rates fluctuate significantly over time and resources are critically insufficient to meet demand – is relatively short, as most casualties receive treatment within hours of the event (Arnold et al. 2003). However, in the case of MTEs, the surge phase can be considerably longer, as patients may take time to recognize the need for mental health support. Moreover, in prolonged MTEs such as wars, series of terror attacks, or natural disasters, the Surge Phase itself can be extended.

We assume a finite duration T for the Surge Phase, during which the patient arrival rate is given by $\lambda(t)$ for $t \in [0,T]$. The transition to the Recovery Phase occurs at time T, which we define as the first time when the arrival rate stabilizes (i.e., becomes approximately constant) and satisfies the capacity condition in (8). This ensures that the system is capable of eventually clearing the backlog. Following Hu et al. (2022), we define the time required to clear the remaining backlog as $\tau := \inf\{t \geq T : q_{\mathbf{m}}(t) + q_{\mathbf{s}}(t) = 0\} - T$, and denote by τ^* the duration of this clearance period under the optimal policy. The endpoint of the Recovery Phase (i.e., time $T + \tau^*$) marks the transition to the Long-Term Care Phase.

The optimal optimization problem for the Surge Phase is, therefore,

$$\begin{aligned} \max_{z,q} \quad & \int_0^T \sum_{j\in\mathcal{J}} \left[r_j z_j(t) - c_j q_j(t)\right] \mathrm{d}t + F\left(q_{\mathbf{m}}(T), q_{\mathbf{s}}(T)\right) \\ \text{s.t.} \quad & \dot{q}_{\mathbf{m}}(t) = \lambda(t) - \theta_{\mathbf{m}} q_{\mathbf{m}}(t) - \left(\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}\right) z_{\mathbf{m}}(t)/m_{\mathbf{m}}, \quad t \geq 0 \\ & \dot{q}_s(t) = p\mu_{\mathbf{m}} z_{\mathbf{m}}(t)/m_{\mathbf{m}} - \theta_{\mathbf{s}} q_{\mathbf{s}}(t) - \left(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}\right) z_{\mathbf{s}}(t)/m_{\mathbf{s}}, \quad t \geq 0 \\ & \sum_{j\in\mathcal{J}} z_j(t) \leq N, \quad z_j(t) \geq 0, \quad q_j(t) \geq 0, \quad j \in \mathcal{J}, \quad t \geq 0, \end{aligned} \tag{11}$$

where $F\left(q_{\mathbf{m}}(T), q_{\mathbf{s}}(T)\right)$ is the optimal objective value for the Recovery Phase problem (9) studied in Section 5, only with a time shift from $[0,\tau]$ to $[T, T+\tau]$ and initial conditions $(q_{\mathbf{m}}(T), q_{\mathbf{s}}(T))$. As a result of this relationship, we can leverage some of the insights derived in Section 5 to the Surge Phase.

To illustrate these insights, we use the time-varying arrival rates of patients needing mental health support, as shown in Figure 6. Scenario 1, with multiple peaks and fluctuations, could represent a large-scale disaster with recurring stressors, such as an earthquake with repeated tremors, where anxiety and trauma responses resurface with each event. Scenario 2, characterized by a single sharp peak followed by a steady decline, suggests a sudden, intense incident in which the initial shock triggers a surge in mental health support needs that gradually stabilizes, as might occur following a terror attack. Scenario 3, with multiple peaks of diminishing intensity, might reflect a prolonged crisis, such as an ongoing war or sequential traumatic events, leading to periodic surges in mental health support as individuals process each new trauma.

Figure 6 **Scenarios for time-varying arrival rates during the Surge Phase.**

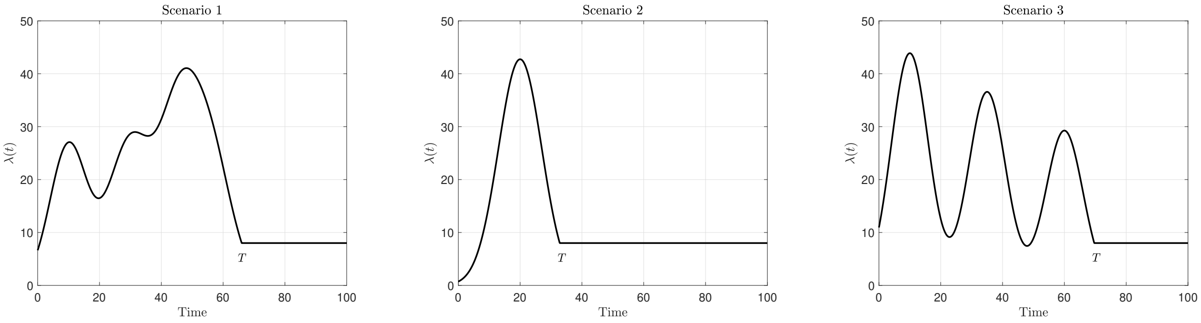
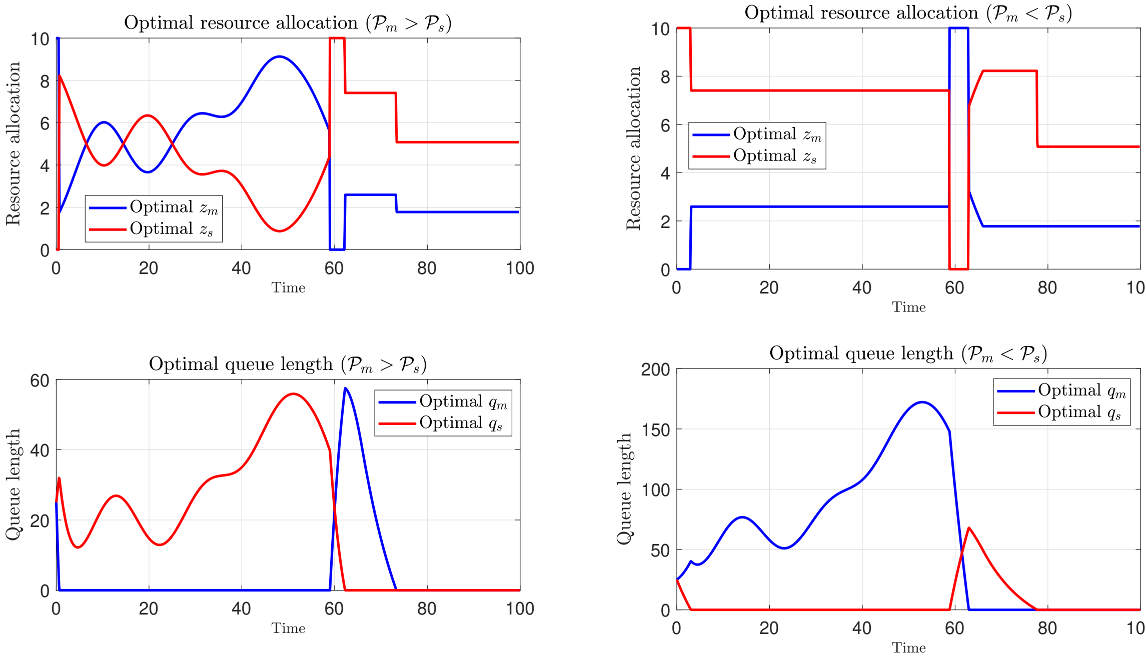


Figure 7 presents the same two scenarios from Figure 4, but with the the time-varying arrival rate illustrated in Figure 6, Scenario 1.

In the left plot, where $\mathcal{P}_{\mathbf{m}} > \mathcal{P}_{\mathbf{s}}$, we observe that at the onset of the surge, the optimal policy follows the $\mathcal{P}$ rule, prioritizing the group channel by allocating the necessary resources to it and distributing the remaining resources to the individual channel. Around $t \approx 60$, the priority shifts to the individual channel, and the system continues this allocation until both queues are depleted. In the right plot, where $\mathcal{P}_{\mathbf{m}} < \mathcal{P}_{\mathbf{s}}$, the situation is reversed. Initially, the optimal policy prioritizes the individual channel, in accordance with the $\mathcal{P}$ rule, allocating resources in such a way that the individual channel maintains a zero queue.

Around $t \approx 60$, the priority switches to the group channel, and the system continues this allocation until both queues are cleared.

Figure 7 Optimal time-varying resource allocation and queue length for each channel, using the same two scenarios as in Figure 4 under the arrival pattern from Figure 6, Scenario 1.



The switching point during the Surge Phase is complex and highly sensitive to both the initial state and the phase duration. Therefore, we propose the *surge-based* $\mathcal{P}$ rule, which is informed by the analysis of the Recovery Phase in Section 5 and our numerical findings.

The surge-based $\mathcal{P}$ rule utilizes the switching curve derived in Section 5. During the Surge Phase, for $t \in [0, T)$, the system follows the $\mathcal{P}$ rule. Afterwards, if $q_{\mathbf{m}} + q_{\mathbf{s}}$ exceeds the switching curve, the $\mathcal{P}$ rule is applied; otherwise, the switched $\mathcal{P}$ rule is followed when the system is below the curve. Note that the suggested surge-based $\mathcal{P}$ is applied dynamically without requiring T to be determined in advance. Once arrival rates begin to stabilize (assuming no recurring events), the Recovery-based policy can be implemented.

In Appendix B.2, we present additional numerical results comparing different versions of the $\mathcal{P}$ rule with the optimal fluid value function across the joint Surge and Recovery Phases. The results show that while the recovery-based $\mathcal{P}$ rule performs reasonably well during the Surge Phase, the surge-based $\mathcal{P}$ rule consistently outperforms the other policies across all scenarios and achieves near-optimal performance.

7. Case Study: The October 7, 2023 Terror Attack

In this section, we implement the proposed policies and evaluate their effectiveness in the context of the October 7 terror attack and the ongoing war, which triggered unprecedented demand for mental health support. For this case study, we apply the extended multi-class framework developed in Appendix C, where each class may receive treatment through both channels.

Following Katsoty et al. (2024), we consider six patient classes summarized in Table 1, which differ in exposure intensity, context, and type. Class sizes are based on official national databases, and PTSD prevalence rates are drawn from relevant literature corresponding to each exposure category.

Table 1 Exposure-based classes to the October 7 terror attacks (source: Katsoty et al. 2024).

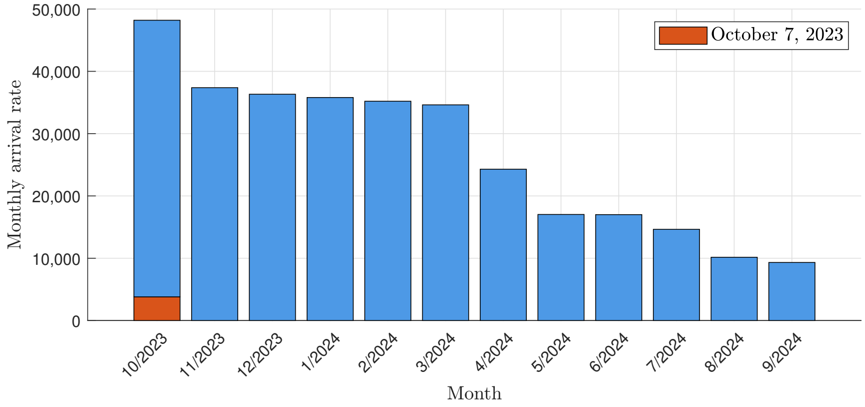
Class	Exposure type	Class size	Estimated PTSD prevalence	Expected number of PTSD cases
1	Direct exposure to the terror attack	39,664	0.31	12,366
2	Close proximity to the terror attack	121,061	0.1	12,372
3	War-involved soldiers	144,227	0.08	11,021
4	Civilians under intense exposure to rocket attacks (living up to 25 miles from the Gaza Strip)	1,069,011	0.1	109,249
5	Civilians under moderate exposure to rocket attacks (living up to 25–50 miles from the Gaza Strip)	4,960,469	0.06	304,556
6	Indirectly affected communities (living more than 50 miles from the Gaza Strip)	3,433,286	0.02	70,714
Total				520,278

The arrival rate data are drawn from the Home Front Command's Information and Knowledge Center and two major Trauma and Resiliency Centers. Figure 8 shows the monthly arrival rate from October 7, 2023, to September 2024. Notably, the rate on October 7 was 500% higher than the pre-event average.

Due to the intensity of the event and the ongoing war, the Surge Phase in this case spans the first year, during which arrival rates remain extremely high and variable. Stabilization occurs toward the last two months, marking the start of the Recovery Phase at the beginning of the second year. This phase continues until all queues are cleared, after which the Long-Term Care Phase begins.

The service consists of a series of therapy sessions aimed at trauma processing and coping. Session length is determined at treatment initiation, typically ranging from 12 to

Figure 8 **Monthly arrival rate from October 2023 to September 2024.**



24 sessions depending on exposure severity (Kar 2011). For instance, PTSD from terrorist attacks typically requires 20–24 sessions, while evacuees facing complex trauma may need 16–20 sessions (Paunovic and Öst 2001). Accordingly, we use a 24-session series for Class 1, 20 sessions for Classes 2 and 3, 16 for Class 4, and 12 for Classes 5 and 6.

Group sizes follow standard clinical practice: 5 participants for Classes 1–3 and 8 for Classes 4–6 (Yalom and Leszcz 2020, Dueweke et al. 2022). The probability of requiring follow-up individual therapy after group sessions is set to 50% for Classes 1–3 and 20% for Classes 4–6, based on empirical studies of group therapy effectiveness (Daley et al. 1983, Yalom and Leszcz 2020, Dueweke et al. 2022).

Treatment cost savings and abandonment costs were estimated from psychiatric and economic literature on PTSD. This literature distinguishes between civilian and military populations (Davis et al. 2022), incorporates socio-demographic factors (Priebe et al. 2009), and includes economic evaluations of PTSD treatments (von der Warth et al. 2020, Watkins et al. 2018). For example, the total excess economic burden of PTSD in the U.S. is estimated at $18,640 for civilians and $25,684 for military personnel (Davis et al. 2022). Based on this, we assign values of $26,000 for Classes 1 and 3, and $24,000, $22,000, $20,000, and $18,000 for Classes 2, 4, 5, and 6, respectively.

Finally, estimates for no-show and dropout rates, and their associated costs, are drawn from prior empirical studies. The Veterans Health Administration reports an average no-show rate of 18% in mental health clinics (Milicevic et al. 2020), while the premature dropout rate is estimated at 11.7% (Xaba et al. 2024, Fenger et al. 2011).

Using these parameters as well as a daily arrival rate over a five-year period beginning on October 7, 2023, we evaluate the system's performance across both the surge and recovery phases. Specifically, we implemented the multi-class extension policies developed in Appendix C in a simulation model over both phases by dynamically prioritizing the classes and channels according to their respective indices. For each policy, we computed the evolution of queue lengths over time for each class, identified when the recovery phase ends, and evaluated the long-run average cost savings.

Figure 9 illustrates the queue lengths of each class under the $\mathcal{P}$ rule as well as under an adjusted version of the $c\mu/\theta$ rule (Atar et al. 2004). To ensure a fair comparison, we adapted the standard $c\mu/\theta$ rule to account for the unique features of our model, such as group therapy and dropouts. In particular, the adjusted $c\mu/\theta$ rule prioritizes the classes and channels according to the following index for Class i and Channel j:

$$r_{i,j} + \frac{c_{i,j}}{\theta_{i,j} m_{i,j}} \left(\gamma_{i,j} + \mu_{i,j}\right),$$

where $r_{i,j}$ and $c_{i,j}$ are the generalized cost savings and holding costs, as defined in (C.3).

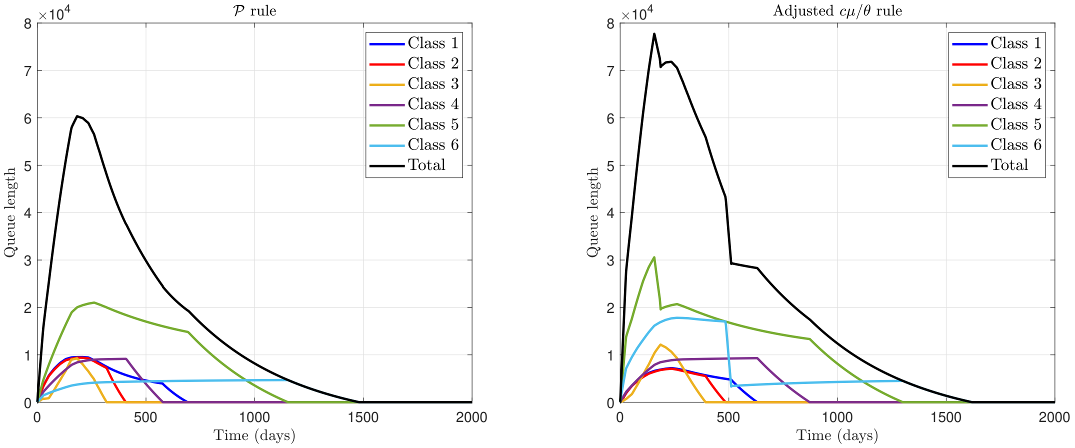
Note that both scheduling policies are implemented dynamically: servers are assigned to channels and customers only when there is demand, and available servers prioritize across channels and customer classes based on their indices.

The key difference between this policy and ours lies in the coordination between channels embedded in the $\mathcal{P}$ rule. Specifically, for Classes 1–3, the $\mathcal{P}$ rule prioritizes the individual channel over the group channel (i.e., $\mathcal{P}_{\mathbf{s}} > \mathcal{P}_{\mathbf{m}}$). As a result, the indices for these classes—and the corresponding prioritization and capacity allocations—are determined jointly across both channels to ensure that patients requiring follow-up individual care are accommodated. In contrast, the adapted $c\mu/\theta$ rule treats each channel separately, prioritizing patients within each one independently. This coordination in the $\mathcal{P}$ rule leads to significant performance differences. With the $\mathcal{P}$ rule, the recovery phase concludes approximately six months earlier than with the adjusted $c\mu/\theta$ rule, the total queue length over the five-year period is reduced by 31%, and the total cost savings are 52% higher.

8. Concluding Remarks

This study addresses the operational challenges posed by MTEs, highlighting the need for timely and effective mental health interventions to mitigate PTSD and related conditions that affect large populations.

Figure 9 Queue length for each class under the $\mathcal{P}$ rule and the adjusted $c\mu/\theta$ rule over a five-year period.



We analyze the coordination between group and individual therapy across the Surge, Recovery, and Long-Term Phases of MTEs and develop cost-effective policies for prioritizing and allocating resources. Our model captures key real-world complexities, including no-shows, dropouts, and referrals from group to individual care.

Through a case study, we illustrate how the proposed framework can inform concrete operational decisions in realistic settings. In particular, we show that policies guided by the $\mathcal{P}$ rule lead to faster system recovery, shorter wait times, and substantial cost savings relative to existing benchmarks. These improvements go beyond theoretical value—they demonstrate how operational models can enhance the efficiency and equity of mental health systems during prolonged crises.

Overall, our findings emphasize the critical role of planning and coordination—especially the dynamic balancing of both group and individual therapy—in building system resilience and improving patient outcomes following mass trauma. These insights can help policymakers and practitioners design more scalable and responsive mental health services in preparation for future MTEs.

Further research is warranted. A promising direction involves low-intensity "waitlist treatment" delivered by peer supporters or lay counselors, which can offer accessible, scalable PTSD care and reduce the need for intensive services (Forneris et al. 2013, Levin et al. 2022). In the near future, such support may be delivered by virtual therapists or AI tools (e.g., chatbots) to provide immediate psychological care, bridge accessibility gaps, and assist in crisis intervention (Omarov et al. 2023). While these approaches require initial investment in training and coordination, they offer potential to extend limited resources and improve access for at-risk populations.

References

Armony M, Bambos N (2003) Queueing dynamics and maximal throughput scheduling in switched processing systems. *Queueing systems* 44:209–252.

Arnold J, Tsai MC, Halpern P, Smithline H, Stok E, Ersoy G (2003) Mass-casualty, terrorist bombings: Epidemiological outcomes, resource utilization, and time course of emergency needs (part i). *Prehospital and Disaster Medicine* 18(3):220–234.

Atar R, Giat C, Shimkin N (2010) The $c\mu/\theta$ rule for many-server queues with abandonment. *Operations Research* 58(5):1427–1439.

Atar R, Mandelbaum A, Reiman M (2004) Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability* 14(3):1084–1134.

Baron O, Berman O, Krass D, Wang J (2014) Using strategic idleness to improve customer service experience in service networks. *Operations Research* 62(1):123–140.

Bowler R, Kornblith E, Li J, Adams S, Gocheva V, Schwarzer R, Cone J (2016) Police officers who responded to 9/11: Comorbidity of ptsd, depression, and anxiety 10–11 years later. *American Journal of Industrial Medicine* 59(6):425–436.

Buell R, Ramdas K, Sönmez N, Srinivasan K, Venkatesh R (2024) Shared service delivery can increase client engagement: A study of shared medical appointments. *Manufacturing & Service Operations Management* 26(1):154–166.

Cai J, Zychlinski N (2025) When ai is not enough: Reducing diagnostic errors with radiologist oversight. *Service Science* .

Chan T, Huang S, Sarhangian V (2024) Dynamic control of service systems with returns: Application to design of postdischarge hospital readmission prevention programs. *Operations Research* .

Chan T, Pogacar F, Sarhangian V, Hellsten E, Razak F, Verma A (2021) Optimizing inter-hospital patient transfer decisions during a pandemic: A queueing network approach. *working paper* .

Chriman A, Dougherty J (2014) *Mass Trauma: Disasters, Terrorism, and War* (DigitalCommons@University of Nebraska - Lincoln: Uniformed Services University of the Health Sciences, U.S. Department of Defense), https://digitalcommons.unl.edu/.

Cohen I, Mandelbaum A, Zychlinski N (2014) Minimizing mortality in a mass casualty event: Fluid networks in support of modeling and staffing. *IIE Transactions* 46(7):728–741.

Cox D, Smith W (1961) Queues. *Methuen, London* .

Dai J, Weiss G (1996) Stability and instability of fluid models for reentrant lines. *Mathematics of Operations Research* 21(1):115–134.

Daley P, Bloom L, Deffenbacher J, Stewart R (1983) Treatment effectiveness of anxiety management training in small and large group formats. *Journal of Counseling Psychology* 30(1):104.

Davis L, Schein J, Cloutier M, Gagnon-Sanschagrin P, Maitland J, Urganus A, Guerin A, Lefebvre P, Houle C (2022) The economic burden of posttraumatic stress disorder in the United States from a societal perspective. *The Journal of Clinical Psychiatry* 83(3):40672.

Dueweke A, Higuera D, Zielinski M, Karlsson M, Bridges A (2022) Does group size matter? Group size and symptom reduction among incarcerated women receiving psychotherapy following sexual violence victimization. *International Journal of Group Psychotherapy* 72(1):1–33.

Farahani R, Lotfi M, Baghaian A, Ruiz R, Rezapour S (2020) Mass casualty management in disaster scene: A systematic review of OR&MS research in humanitarian operations. *European Journal of Operational Research* 287(3):787–819.

Fenger M, Mortensen E, Poulsen S, Lau M (2011) No-shows, drop-outs and completers in psychotherapeutic treatment: Demographic and clinical predictors in a large sample of non-psychotic patients. *Nordic Journal of Psychiatry* 65(3):183–191.

Forneris C, Gartlehner G, Brownley K, Gaynes B, Sonis J, Coker-Schwimmer E, Jonas D, Greenblatt A, Wilkins T, Woodell C (2013) Interventions to prevent post-traumatic stress disorder: A systematic review. *American Journal of Preventive Medicine* 44(6):635–650.

Gibbs L, Skyler E (2004) Addressing the health impacts of 9/11: Report and recommendations to mayor Michael R. Bloomberg. Technical report, World Trade Center Health Panel, New York, NY, report to Mayor Michael R. Bloomberg.

Grasser LR, Javanbakht A (2019) Treatments of posttraumatic stress disorder in civilian populations. *Current Psychiatry Reports* 21:1–19.

Grosof I, Harchol-Balter M (2023) Serverfilling: A better approach to packing multiserver jobs. *Proceedings of the 5th workshop on Advanced tools, programming languages, and PLatforms for Implementing and Evaluating algorithms for Distributed systems*, 1–5.

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

Herman D, Felton C, Susser E (2002) Mental health needs in New York State following the September 11th attacks. *Journal of Urban Health* 79:322–331.

Hirschberger G (2018) Collective trauma and the social construction of meaning. *Frontiers in Psychology* 9:1441.

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

Huang J, Carmeli B, Mandelbaum A (2015) Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research* 63(4):892–908.

Jacobson E, Argon N, Ziya S (2012) Priority assignment in emergency response. *Operations Research* 60(4):813–832.

Kar N (2011) Cognitive behavioral therapy for the treatment of post-traumatic stress disorder: A review. *Neuropsychiatric Disease and Treatment* 167–181.

Katsoty D, Greidinger M, Neria Y, Segev A, Lurie I (2024) A prediction model of PTSD in the Israeli population in the aftermath of October 7th, 2023, terrorist attack and the Israel-Hamas war. *Israel Journal of Health Policy Research* 13(1):63.

Levin M, Hicks E, Krafft J (2022) Pilot evaluation of the stop, breathe & think mindfulness app for student clients on a college counseling center waitlist. *Journal of American College Health* 70(1):165–173.

Li D, Ding L, Connor S (2020) When to switch? Index policies for resource scheduling in emergency response. *Production and Operations Management* 29(2):241–262.

Lodree E, Altay N, Cook R (2019) Staff assignment policies for a mass casualty event queuing network. *Annals of Operations Research* 283(1):411–442.

Long Z, Shimkin N, Zhang H, Zhang J (2020) Dynamic scheduling of multiclass many-server queues with abandonment: The generalized cμ/h rule. *Operations Research* 68(4):1218–1230.

Long Z, Zhang H, Zhang J, Zhang Z (2024) The generalized c/μ rule for queues with heterogeneous server pools. *Operations Research* 72(6):2488–2506.

Makari G, Friedman R (2024) Collective trauma and commemoration–A moment of silence, please. *New England Journal of Medicine* 391(6):487–489.

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

Milicevic AS, Mitsantisuk K, Tjader A, Vargas D, Hubert T, Scott B (2020) Modeling patient no-show history and predicting future appointment behavior at the veterans administration's outpatient mental health clinics: NIRMO-2. *Military Medicine* 185(7-8):e988–e994.

Mills A, Argon N, Ziya S (2018) Dynamic distribution of patients to medical facilities in the aftermath of a disaster. *Operations Research* 66(3):716–732.

Omarov B, Zhumanov Z, Gumar A, Kuntunova L (2023) Artificial intelligence enabled mobile chatbot psychologist using AIML and cognitive behavioral therapy. *International Journal of Advanced Computer Science and Applications* 14(6).

Papadimitriou C, Tsitsiklis J (1999) The complexity of optimal queuing network control. *Mathematics of Operations Research* 24(2):293–305.

Paunovic N, Öst L (2001) Cognitive-behavior therapy vs exposure therapy in the treatment of PTSD in refugees. *Behaviour Research and Therapy* 39(10):1183–1197.

Priebe S, Matanov A, Janković Gavrilović J, McCrone P, Ljubotina D, Knežević G, Kučukalić A, Frančišković T, Schützwoh M (2009) Consequences of untreated posttraumatic stress disorder following war in former Yugoslavia: Morbidity, subjective quality of life, and care costs. *Croatian medical journal* 50(5):465–475.

Rezapour S, Baghaian A, Naderi N, Farzaneh M (2022) Dynamic on-site treatment strategy for large-scale mass casualty incidents with rescue operation. *Computers & Industrial Engineering* 163:107796.

Shi Y, Liu N, Wan G (2023) Treatment planning for victims with heterogeneous time sensitivities in mass casualty incidents. *Operations Research* .

Sloan D, Bovin M, Schnurr P (2012) Review of group treatment for PTSD. *Journal of Rehabilitation Research & Development* 49(5).

Sönmez N, Srinivasan K, Venkatesh R, Buell R, Ramdas K (2023) Evidence from the first shared medical appointments (smas) randomised controlled trial in india: Smas increase the satisfaction, knowledge, and medication compliance of patients with glaucoma. *PLOS Global Public Health* 3(7):e0001648.

Sun Z, Argon N, Ziya S (2018) Patient triage and prioritization under austere conditions. *Management Science* 64(10):4471–4489.

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

von der Warth R, Dams J, Grochtdreis T, König HH (2020) Economic evaluations and cost analyses in post-traumatic stress disorder: A systematic review. *European Journal of Psychotraumatology* 11(1):1753940.

Watkins L, Sprang K, Rothbaum B (2018) Treating PTSD: A review of evidence-based psychotherapy interventions. *Frontiers in Behavioral Neuroscience* 12:258.

Whitt W (2002) Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and their Application to Queues. *Space* 500:391–426.

Xaba F, Lowane M, Chelule P, Shilubane H (2024) Failure to keep psychiatric appointments at primary healthcare facilities: Mental health care users missed ongoing clinical visits in ekurhuleni district in gauteng province, South Africa. *International Journal of Innovative Research and Scientific Studies* 7(2):645–652.

Yalom I, Leszcz M (2020) *The Theory and Practice of Group Psychotherapy* (Basic Books).

Yom-Tov G, Mandelbaum A (2014) Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management* 16(2):283–299.

Zychlinski N, Chan C, Dong J (2023) Managing queues with different resource requirements. *Operations Research* 71(4):1387–1413.

Appendices

Appendix A: Excluding Irregular Boundary Behaviors in Section 5

Recall the transient optimization problem in (9). We begin by excluding irregular boundary behaviors where state constraints (i.e., non-negativity of queue lengths) hold tightly. One such challenging boundary behavior is known as *chattering*, where the trajectory $q_j(t)$, for $j \in \mathcal{J}$, oscillates infinitely fast between zero and positive values (Hu et al. 2022). Specifically, a time point $\tilde{t}$ is called a *chattering point* of the state trajectory q_j if $q_j(\tilde{t}) = 0$ and for any $\delta > 0$, there exist times s' and $s'' \in (\tilde{t} - \delta, \tilde{t} + \delta)$ such that $q_j(s') > 0$ and $q_j(s'') = 0$. An interval is referred to as a *chattering interval* if any of its sub-intervals contains at least one chattering point.

We next prove that it suffices to focus on trajectories that avoid both chattering points and chattering intervals.

Lemma A.1 **(excluding irregular boundary behaviors)**. *For the transient optimal control problem (9), we can assume that the state trajectories are free of chattering behavior without compromising optimality.*

Appendix B: Additional Numerical Results

B.1. Additional Numerical Results for Section 5

In this section we use stochastic simulation to assess the performance of the recovery-based $\mathcal{P}$ rule. Table 2 provides a comparison between the fluid solution, obtained by solving (9), and the average simulation results, conducted with 1500 replications for each initial condition.

The results indicate that the difference between the optimal fluid objective and the average transient cost savings in the stochastic model is very small across all cases and is especially small when the traffic intensity is high and no priority switching occurs.

Table 2 Comparison between fluid solution and simulation results for different initial conditions. The parameters are as in Figure 5.

ρ	Initial conditions	$P_m > P_s$			$P_m < P_s$		
	$(q_{\mathbf{m}}(0), q_{\mathbf{s}}(0))$	Fluid Solution	Simulation	Difference	Fluid Solution	Simulation	Difference
0.45	(100, 400)	467,590	467,096	0.11%	725,340	723,699	0.23%
	(250, 450)	600,860	597,916	0.49%	503,250	500,784	1.34%
	(400, 500)	613,420	599,791	2.22%	526,840	515,135	0.36%
	(600, 600)	614,840	611,000	0.62%	541,440	538,058	1.26%
	(400, 200)	646,420	638,215	1.27%	554,790	547,748	2.50%
1	(100,400)	672,500	671,962	0.08%	1,045,400	1,043,936	0.14%
	(250,450)	682,770	680,927	0.27%	1,012,000	1,002,993	0.89%
	(400,500)	682,260	674,209	1.18%	978,450	976,395	0.21%
	(600,600)	667,060	665,059	0.30%	924,380	917,447	0.75%
	(400,200)	722,240	715,451	0.94%	1,056,600	1,039,060	1.66%

B.2. Additional Numerical Results for Section 6

Table 3 presents a comparison of the performance of four different policies against the optimal policy, which is derived by solving (11) for various scenarios. The four policies include: the surge-based $\mathcal{P}$ rule, the recovery-based $\mathcal{P}$ rule developed in Section 5 (where the $\mathcal{P}$ rule is applied above the switching curve and the inverse $\mathcal{P}$ rule below it) as well as both the $\mathcal{P}$ rule and its inverse (where priorities are switched).

Table 3 cost savings comparison (in percentages) of the $\mathcal{P}$ rule against the fluid value function for the joint Surge and Recovery phases. The parameters for the classes are as in Figure 5.

Arrival rate (Figure 6)	Case	$\mathcal{P}$ rule	Inverse $\mathcal{P}$ rule	Recovery-based policy	Surge-based policy
Scenario 1	$\mathcal{P}_m > \mathcal{P}_s$	4.04%	12.84%	1.22%	0.21%
	$\mathcal{P}_m < \mathcal{P}_s$	2.28%	6.45%	3.64%	0.14%
Scenario 2	$\mathcal{P}_m > \mathcal{P}_s$	4.73%	8.48%	2.05%	0.52%
	$\mathcal{P}_m < \mathcal{P}_s$	2.35%	3.87%	1.57%	0.28%
Scenario 3	$\mathcal{P}_m > \mathcal{P}_s$	3.92%	15.44%	1.02%	0.35%
	$\mathcal{P}_m < \mathcal{P}_s$	1.88%	5.68%	3.85%	0.40%

The results show that, while the recovery-based $\mathcal{P}$ rule performs reasonably well during the Surge Phase, the surge-based $\mathcal{P}$ rule consistently outperforms the other policies across all scenarios and achieves near-optimal performance.

Appendix C: The Multiple-Class Extension

In some cases of MTEs, patients may experience varying levels of exposure or intensity. Consequently, the exposed population can be categorized into different classes based on the intensity, context, and type of traumatic exposure they endured. Treatment cost savings, for example, may vary across classes (Grasser and Javanbakht 2019), depending on factors such as the type of traumatic exposure.

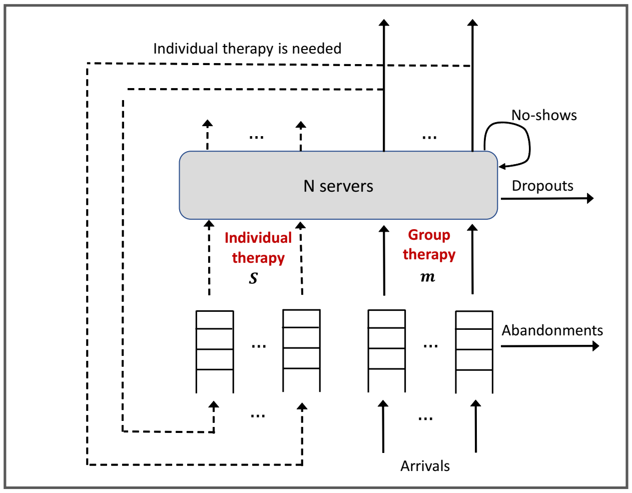
In this section, we extend the policies proposed across the three response stages to accommodate I classes of patients, denoted by $\mathcal{I} = \{1, \dots, I\}$. Each class is processed through a group channel, the size of which depends on the class, followed by individual channels, as illustrated in Figure C.1.

To incorporate distinct parameters for each class and channel, we introduce a subscript i for all parameters. For example, $\mu_{i,j}$ denotes the service rate of Class i patients in Channel j, and $m_{i,\mathbf{m}}$ represents the required number of servers for Class i in the group channel (i.e., the group size of Class i is $1/m_{i,\mathbf{m}}$).

Recall that both $m_{i,\mathbf{m}}$ and $m_{i,\mathbf{s}}$ can take any positive real value. In practice, these are typically rational numbers, determined by the therapist-to-patient ratio. For instance, $m_{i,\mathbf{m}} = 1/5$ and $m_{j,\mathbf{m}} = 1/10$ indicate that each Class i patient requires one-fifth of a server, while each Class j patient requires one-tenth of a server. Equivalently, a Class i group consists of five participants, whereas a Class j group consists of ten participants.

To ensure that groups remain homogeneous within their respective classes, we can adjust the unit of measurement. For example, one-tenth of a server can be redefined as a unit of service capacity. In this case, we set $m_{i,\mathbf{m}} = 2$, $m_{j,\mathbf{m}} = 1$, and $m_{i,\mathbf{s}} = m_{j,\mathbf{s}} = 10$.

Figure C.1 **Model illustration with multiple classes of patients.**



C.1. Long-Term Care Phase

In the Long-Term Phase, we assume that each customer class arrives according to a homogeneous Poisson process with rate λ_i, $i \in \mathcal{I}$. Additionally, we retrieve the class index, i to the subscript of each class-related parameter. The extension of (4) to $\mathcal{I}$ classes of customers yields the following linear program:

$$\begin{aligned} \max_{\bar{z}} \quad & \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \mathcal{P}_{i,j}\bar{z}_{i,j} \\ \text{s.t.} \quad & 0 \le \bar{z}_{i,m} \le \frac{\lambda_i}{m_i\mu_{i,m}+\gamma_{i,m}}, \quad i\in\mathcal{I}, \\ & 0 \le \bar{z}_{i,s} \le \frac{m_i\mu_{i,m}p_i(m_i)}{\mu_{i,s}+\gamma_{i,s}}\bar{z}_{i,m}, \quad i\in\mathcal{I}, \\ & \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \bar{z}_{i,j} \le N, \end{aligned} \tag{C.1}$$

where

$$\begin{aligned} \mathcal{P}_{i,m} &= r_{i,m} + \frac{c_{i,\mathbf{m}}}{\theta_{i,\mathbf{m}}m_{i,\mathbf{m}}}\left(\gamma_{i,\mathbf{m}}+\mu_{i,\mathbf{m}}\right) - \frac{p_i c_{i,\mathbf{s}}}{\theta_{i,\mathbf{s}}m_{i,\mathbf{m}}}\mu_{i,\mathbf{m}}, \\ \mathcal{P}_{i,\mathbf{s}} &:= r_{i,\mathbf{s}} + \frac{c_{i,\mathbf{s}}}{\theta_{i,\mathbf{s}}m_{i,\mathbf{s}}}\left(\mu_{i,\mathbf{s}}+\gamma_{i,\mathbf{s}}\right). \end{aligned} \tag{C.2}$$

and the generalized cost savings and holding cost are

$$r_{i,j} := \frac{1}{m_{i,j}}\left(b_{i,j}\mu_{i,j} - h^d_{i,j}\gamma_{i,j}\right), \quad c_{i,j} := h_{i,j} + \alpha_{i,j}\theta_{i,j}. \tag{C.3}$$

The complexity of scheduling multiple classes in this setting arises from the dependency in resource allocation between both channels, as the group channel feeds the individual channel. Specifically, when $\mathcal{P}_i > \mathcal{P}_{i,m}$ for some Class i, the resource allocation must be coordinated between both channels: to provide the individual channel with its maximal possible allocation, the group channel needs sufficient resources. Therefore, optimal scheduling requires jointly considering both channels of each class. Conversely, when $\mathcal{P}_i < \mathcal{P}_{i,m}$, both channels can be scheduled separately.

Proving the scheduling optimality in the multiple-class case requires following the analysis used to prove Theorem 1. This involves first proving that under the suggested scheduling rule, the fluid approximation converges to a globally asymptotically stable equilibrium point, and then proving that the optimal solution to the long-run cost savings maximization problem in the multi-class case is the globally asymptotically stable equilibrium. This approach, however, becomes prohibitively tedious with too many scenarios to consider. Therefore, we provide an algorithm that extends the essence of the single-class rule for setting the prioritization among classes and channels. Note that for a given set of parameters, this algorithm needs to be run *once*.

First, we introduce the weighted average index for each class, which is necessary for setting the scheduling policy (see Section C.1.1 for an explanation regarding this term):

$$\bar{\mathcal{P}}_i = \frac{\mu_{i,\mathbf{s}}+\gamma_{i,\mathbf{s}}}{\mu_{i,\mathbf{s}}+\gamma_{i,\mathbf{s}}+p_i\mu_{i,\mathbf{m}}m_{is}}\mathcal{P}_{i,\mathbf{m}} + \frac{p_i\mu_{i,\mathbf{m}}m_{i,\mathbf{s}}}{\beta_{i,\mathbf{s}}\mu_{i,\mathbf{s}}+\gamma_{i,\mathbf{s}}+p_i\mu_{i,\mathbf{m}}m_{i,\mathbf{s}}}\mathcal{P}_{i,\mathbf{s}}. \tag{C.4}$$

Note that if multiple classes share the same $\mathcal{P}$ index, multiple optimal scheduling policies may prevail; this can complicate the analysis. For the sake of simplicity, we assume that all indices are unique.

The following algorithm uses the sorted set $\mathcal{S}$, which includes the indexes of each class to determine the priority among classes.

Algorithm 1 (Multi-class scheduling with group and individual channels)

```
1. Set $\mathcal{S}=\varnothing$
2. For each Class $i$, $i\in\mathcal{I}$:
    (a) If $\bar{\mathcal{P}}_i<\mathcal{P}_{i,m}$, then $\mathcal{S}\leftarrow\mathcal{S}\cup\{\mathcal{P}_{i,m},\mathcal{P}_i\}$
    (b) Otherwise, $\mathcal{S}\leftarrow\mathcal{S}\cup\{\bar{\mathcal{P}}_i\}$
3. Sort the set $\mathcal{S}$ in a decreasing order
4. Replace the $\bar{\mathcal{P}}_i$'s in $\mathcal{S}$ with $\{\mathcal{P}_i,\mathcal{P}_{i,m}\}$
5. Return $\mathcal{S}$
```

The prioritization of classes will be done according to their order in the sorted set $\mathcal{S}$. The algorithm shares the same principles as the single-class rule: if for a specific class the group channel's index is higher than the individual channel's index, then the prioritization is set separately to the group/individual channels according to their $\mathcal{P}_m,\mathcal{P}_s$ indexes. If, however, the group channel's index is smaller than the individual channel's index, both channels are first prioritized jointly according to their integrated $\bar{\mathcal{P}}_i$ index, and their allocated resources are set jointly as in the second case of Theorem 1.

C.1.1. The Weighted Average $\bar{\mathcal{P}}$.

The joint index $\bar{\mathcal{P}}$ in (C.4) incorporates the group and individual channels. Intuitively, when $\mathcal{P}_{k,\mathbf{s}}>\mathcal{P}_{k,\mathbf{m}}$ for some Class k, we would tend to prioritize the individual channel over the group one. Since the latter is the feeding source of the former, enough resources need to be allocated to the group channel to assure that

$$\bar{z}_{k,\mathbf{s}} = \frac{p_k\mu_{k,\mathbf{m}}m_{k,\mathbf{s}}}{(\mu_{k,\mathbf{s}}+\gamma_{k,\mathbf{s}})\,m_{k,\mathbf{m}}}\bar{z}_{k,\mathbf{m}}.$$

By substituting $\bar{z}_{k,\mathbf{s}}$ in the objection function of (C.1), we get

$$\max_{\bar{z}} \sum_{\substack{i\in\mathcal{I}\\ i\neq k}}\left[\mathcal{P}_{i,\mathbf{m}}\bar{z}_{i,\mathbf{m}}+\mathcal{P}_{i,\mathbf{s}}\bar{z}_{i,\mathbf{s}}\right]+\left(\mathcal{P}_{k,\mathbf{m}}+\frac{p_k\mu_{k,\mathbf{m}}m_{k,\mathbf{s}}}{(\mu_{k,\mathbf{s}}+\gamma_{k,\mathbf{s}})\,m_{k,\mathbf{m}}}\mathcal{P}_{k,\mathbf{s}}\right)\bar{z}_{k,\mathbf{m}}. \tag{C.5}$$

When the resource constraint is active (i.e., $\sum_{i\in\mathcal{I}}[\bar{z}_{i,\mathbf{m}}+\bar{z}_{i,\mathbf{s}}]=N$), we have

$$\bar{z}_{k,\mathbf{m}} = \frac{\mu_{k,\mathbf{s}}+\gamma_{k,\mathbf{s}}}{\mu_{k,\mathbf{s}}+\gamma_{k,\mathbf{s}}+p_k\mu_{k,\mathbf{m}}m_{k,\mathbf{s}}}\left(N-\sum_{\substack{i\in\mathcal{I}\\ i\neq j}}[\bar{z}_{i,\mathbf{m}}+\bar{z}_{i,\mathbf{s}}]\right),$$

which in turn is plugged in back into (C.5) to give the following objective function

$$\max_{\bar{z}}\sum_{\substack{i\in\mathcal{I}\\ i\neq k}}[\mathcal{P}_{i,\mathbf{m}}\bar{z}_{i,\mathbf{m}}+\mathcal{P}_{i,\mathbf{s}}\bar{z}_{i,\mathbf{s}}]+\bar{\mathcal{P}}_k\left(N-\sum_{\substack{i\in\mathcal{I}\\ i\neq k}}[\bar{z}_{i,\mathbf{m}}+\bar{z}_{i,\mathbf{s}}]\right).$$

Now it is evident that resources need to be allocated jointly for both Class j's group and individual channels according to their weighted average index, $\bar{\mathcal{P}}_j$.

C.1.2. Managing Idleness. When implementing the $\mathcal{P}$ rule in stochastic systems, idleness may occur if there are not enough patients to meet the required group size. For instance, prioritizing a class with a group size of five when only three such patients are available would leave two-fifths of the server's capacity idle. This idleness, especially in small systems, can lead to sub-optimal performance (Baron et al. 2014, Zychlinski et al. 2023, Grosof and Harchol-Balter 2023). The effect is more pronounced in small systems, since, for example, one idle server out of three is significant, while one idle server out of 3000 is negligible.

It is, therefore, crucial to properly manage policy-induced idleness; that is, admitting a different class group or an individual patient, while anticipating the arrival of two more patients from the first class, could increase system throughput and overall performance. A natural way to mitigate policy-induced idleness is to add a penalty term for incurred idleness when evaluating a scheduling rule. Specifically, we introduce a tuning parameter $\Gamma\geq 0$ to penalize priority-induced idleness, resulting in the same optimization problem as (C.1) with the following adjusted objective function:

$$\max_{\bar{z},m}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\mathcal{P}_{i,j}\bar{z}_{i,j}+\Gamma\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\bar{z}_{i,j}=\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\tilde{\mathcal{P}}_{i,j}\bar{z}_{i,j},$$

where the Γ-adjusted $\mathcal{P}$ indices are $\tilde{\mathcal{P}}_{i,j}=\mathcal{P}_{i,j}+\Gamma$.

The additional term in the objective function increases the number of utilized servers (or equivalently, minimizes server idleness). The two extreme cases are: (i) when $\Gamma=0$, we prioritize according to the original $\mathcal{P}$ rule, and (ii) when Γ is large enough ($\Gamma>\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\mathcal{P}_{i,j}$), the primary objective becomes maximizing server utilization. That is, among all policies that minimize idleness, we choose the one that maximizes the $\mathcal{P}$ index. For intermediate values of Γ, different levels of idleness may occur.

The Γ parameter in the Γ-adjusted $\mathcal{P}$ index rule provides flexibility in balancing the immediate cost savings-increase rate with the priority-induced idleness. The general intuition for setting Γ is that in large or lightly loaded systems, more weight should be placed on the $\mathcal{P}$ indices (through a smaller or even zero Γ). Conversely, in small or critically loaded systems, greater emphasis should be placed on minimizing idleness (through a larger Γ). The solution approach incorporating idleness is the same as the $\mathcal{P}$ index rule, but with the Γ-adjusted $\mathcal{P}$ indices replacing the original $\mathcal{P}$ indices.

C.2. Recovery and Surge Phase

The transient optimization problem in the multiple-class case is the following linear programming:

$$\begin{aligned}
\max_{z,q}\quad & \int_0^{\tau} \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \left[r_{i,j} z_{i,j}(t) - c_{i,j} q_{i,j}(t)\right] \mathrm{d}t \\
\text{s.t.}\quad & \dot{q}_{i,\mathbf{m}}(t) = \lambda_i - \theta_{i,\mathbf{m}} q_{i,\mathbf{m}}(t) - (\mu_{i,\mathbf{m}} + \gamma_{i,\mathbf{m}})\, z_{i,\mathbf{m}}(t)/m_{i,\mathbf{m}}, \quad t \geq 0 \\
& \dot{q}_{i,\mathbf{s}}(t) = p\mu_{i,\mathbf{m}} z_{i,\mathbf{m}}(t)/m_{i,\mathbf{m}} - \theta_{i,\mathbf{s}} q_{i,\mathbf{s}}(t) - (\mu_{i,\mathbf{s}} + \gamma_{i,\mathbf{s}})\, z_{i,\mathbf{s}}(t)/m_{i,\mathbf{s}}, \quad t \geq 0 \\
& \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} z_{i,j}(t) \leq N, \quad z_{i,j}(t) \geq 0, \quad q_{i,j}(t) \geq 0, \quad i \in \mathcal{I},\ j \in \mathcal{J},\ t \geq 0.
\end{aligned} \tag{C.6}$$

By numerically solving (C.6), we observe that, similar to the single-class case, the $\mathcal{P}$-rule is optimal when the total queue lengths are large. Unlike the single-class case, however, where a single priority switch occurs when the total queue lengths become small, allowing us to characterize the switching curve, in the multi-class case, each class switches priority at different times. These switching times depend on specific relationships between the class parameters and initial conditions. Establishing these properties in a rigorous way for the multi-class case requires complex and non-trivial derivations.

Fortunately, we find that, analogous to the single-class case, when the system is highly loaded, as is typical during MTEs — the main focus of this research — there is no priority switching. The optimal solution adheres to the $\mathcal{P}$-rule throughout the entire horizon. Moreover, even when the system is moderately loaded, following the $\mathcal{P}$-rule throughout the horizon results in performance that is reasonably close to optimal. Therefore, for multi-class scenarios, we recommend adhering to the $\mathcal{P}$-rule throughout the recovery and surge phases.

Table 4 presents a simulation-based performance comparison in a two-class setting across various scenarios, including the Recovery Phase and the combined Recovery and Surge Phases. The percentages in parentheses indicate the optimality gap relative to the optimal fluid value function. For the fluid policy, we implemented in the simulation the optimal solution obtained by solving (C.6). While this policy performs very close to optimal, it is more complex to implement, as it requires tracking the solution of (C.6), which may be challenging – particularly when multiple classes and switches are involved. In contrast, following the $\mathcal{P}$ rule is straightforward and easy to implement. Moreover, the results show that the deviation of the $\mathcal{P}$ rule from the optimal policy is small—within 3.5%.

Appendix D: Proofs of Analytical Results

Proof of Theorem 1:

We begin the proof by the following definition. Considering an autonomous differential equation:

$$\dot{q}(t) = f(q(t)) \quad \text{with} \quad q(0) = q_0. \tag{D.7}$$

Suppose there exists an equilibrium point $\bar{q}$ so that $f(\bar{q}) = 0$. Then, $\bar{q}$ is *globally asymptotically stable* if for any initial condition q_0, $\lim_{t\to\infty} ||q(t) - \bar{q}|| = 0$, where $||\cdot||$ is the Euclidean norm.

Table 4 Cost savings comparison (in percentages) relative to the optimal fluid value function during the joint Surge and Recovery phases. The parameters for the classes are as described in Figure 5, with Class 1 having $\mathcal{P}_{\mathbf{m}} > \mathcal{P}_{\mathbf{s}}$ and Class 2 having $\mathcal{P}_{\mathbf{m}} < \mathcal{P}_{\mathbf{s}}$.

Recovery Phase			Surge and Recovery Phases		
$(q_{1,m}(0), q_{2,m}(0))$ $(q_{1,s}(0), q_{2,s}(0))$	Fluid policy	$\mathcal{P}$ rule	Arrival rate (Figure 6)	Fluid policy	$\mathcal{P}$ rule
(100,400) (400,600)	835,808 (0.55%)	818,999 (2.55%)	$\lambda_1(t)$ Scenario 1 $\lambda_2(t)$ Scenario 2	1,361,235 (0.48%)	1,352,070 (1.15%)
(400,600) (100,400)	778,543 (0.62%)	761,230 (2.83%)	$\lambda_1(t)$ Scenario 2 $\lambda_2(t)$ Scenario 1	1,832,624 (0.66%)	1,784,844 (3.25%)
(350,300) (300,400)	749,678 (0.81%)	737,585 (2.41%)	$\lambda_1(t)$ Scenario 1 $\lambda_2(t)$ Scenario 1	1,853,810 (0.52%)	1,810,949 (2.82%)
(300,400) (350,300)	719,809 (0.97%)	703,382 (3.23%)	$\lambda_1(t)$ Scenario 2 $\lambda_2(t)$ Scenario 2	1,213,904 (0.76%)	1,202,650 (1.68%)

For the purpose of the proof, it is helpful to explicitly rewrite (2) as follows:

$$
\begin{aligned}
\max_{q,z} \quad & \liminf_{T\to\infty} \frac{1}{T}\int_0^T \left[r_{\mathbf{m}} z_{\mathbf{m}}(t) - c_{\mathbf{m}} q_{\mathbf{m}}(t) + r_{\mathbf{s}} z_{\mathbf{s}}(t) - c_{\mathbf{s}} q_{\mathbf{s}}(t)\right] \mathrm{d}t \\
\text{s.t.} \quad & \dot q_{\mathbf{m}}(t) = \lambda - \theta_{\mathbf{m}} q_{\mathbf{m}}(t) - (\mu_{\mathbf{m}} + \gamma_{\mathbf{m}})\, z_{\mathbf{m}}(t)/m_{\mathbf{m}}, \quad t \ge 0 \\
& \dot q_s(t) = p\mu_{\mathbf{m}} z_{\mathbf{m}}(t)/m_{\mathbf{m}} - \theta_{\mathbf{s}} q_{\mathbf{s}}(t) - (\mu_{\mathbf{s}} + \gamma_{\mathbf{s}})\, z_{\mathbf{s}}(t)/m_{\mathbf{s}}, \quad t \ge 0 \\
& z_{\mathbf{m}}(t) + z_{\mathbf{s}}(t) \le 1, \quad t \ge 0; \\
& q_{\mathbf{m}}(t), q_{\mathbf{s}}(t), z_{\mathbf{m}}(t), z_{\mathbf{s}}(t) \ge 0, \quad t \ge 0.
\end{aligned}
\tag{D.8}
$$

In the first part of the proof, we establish that by following the suggested $\mathcal{P}$ rule for the system dynamics in (D.8) from any initial condition, and for $\theta_{\mathbf{m}}, \theta_{\mathbf{s}} > 0$, the globally asymptotically stable equilibria, $\bar z = (\bar z_{\mathbf{m}}, \bar z_{\mathbf{s}})$ are as in Equations (6) and (7), and the equilibrium queue lengths $\bar q = (\bar q_{\mathbf{m}}, \bar q_{\mathbf{s}})$, are given by

$$
\bar q_{\mathbf{m}} = \tfrac{1}{\theta_{\mathbf{m}}}\left(\lambda - (\mu_{\mathbf{m}} + \gamma_{\mathbf{m}})\,\tfrac{\bar z_{\mathbf{m}}}{m_{\mathbf{m}}}\right), \quad \text{and} \quad \bar q_{\mathbf{s}} = \tfrac{1}{\theta_{\mathbf{s}}}\left(p\mu_{\mathbf{m}}\tfrac{\bar z_{\mathbf{m}}}{m_{\mathbf{m}}} - (\mu_{\mathbf{s}} + \gamma_{\mathbf{s}})\,\tfrac{\bar z_{\mathbf{s}}}{m_{\mathbf{s}}}\right). \tag{D.9}
$$

In the second part of the proof, we show that the solution of (4), $\bar z^*$, and the corresponding $\bar q^*$ constitute the globally asymptotically stable equilibrium established in the first part of the proof.

Part 1. This part is based on the construction of Lyapunov functions.

- **If $\boldsymbol{\mathcal{P}_{\mathbf{m}} > \mathcal{P}_{\mathbf{s}}}$**: We consider the three sub-cases summarized in Table 5. For each sub-case we prove that the globally asymptotically stable equilibrium $\bar q = (\bar q_{\mathbf{m}}, \bar q_{\mathbf{s}})$ is. In this case, the $\mathcal{P}$ rule gives strict priority to the group channel over the individual one. In this case, there could be three options. It what follows, we analyze each such option.

—**Sub-case I: $\boldsymbol{\frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}} + \frac{p\mu_{\mathbf{m}} m_{\mathbf{s}}}{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}}) m_{\mathbf{m}}} N \le N}$. $\boldsymbol{\bar q = (0,0)}$.**

We consider the Lyapunov function

$$
V(q) = \frac{m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}|q_{\mathbf{m}} - \bar q_{\mathbf{m}}| + \frac{m_{\mathbf{s}}}{\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}}|q_{\mathbf{s}} - \bar q_{\mathbf{s}}|,
$$

where the equilibrium point $\bar q = (0,0)$, and show its asymptotic stability. To this end, we first verify that $V(\bar q) = 0$ and $V(q) \to \infty$ as $\|q\| \to \infty$. Then, we show that $\nabla_q V(q)_{\mathrm{s}} T f(q) < 0$ for $q \ne \bar q$, where $\dot q(t) = f(q(t))$, as defined in (D.7).

$*$ **When $\boldsymbol{q_{\mathbf{m}}(t) > 0}$**, all resources are allocated to the group channel. Specifically, the system dynamics in (D.8) are as follows:

$$\begin{cases} \dot{q}_{\mathbf{m}}(t) = \lambda - \frac{1}{m_{\mathbf{m}}}\left(\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}\right) N - \theta_{\mathbf{m}} q_{\mathbf{m}}(t); \\ \dot{q}_{\mathbf{s}}(t) = \frac{1}{m_{\mathbf{m}}} p \mu_{\mathbf{m}} N - \theta_{\mathbf{s}} q_{\mathbf{s}}(t). \end{cases}$$

We have

$$\begin{aligned} \nabla_q V(q)^T f(q) &= \frac{m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}\left(\lambda - \tfrac{1}{m_{\mathbf{m}}}\left(\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}\right) N - \theta_{\mathbf{m}} q_{\mathbf{m}}(t)\right) + \frac{m_{\mathbf{s}}}{\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}}\left(\tfrac{1}{m_{\mathbf{m}}} p \mu_{\mathbf{m}} N - \theta_{\mathbf{s}} q_{\mathbf{s}}(t)\right) \\ &= \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}} - N + \frac{p \mu_{\mathbf{m}} m_{\mathbf{s}}}{\left(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}\right) m_{\mathbf{m}}} N - \frac{\theta_{\mathbf{m}} q_{\mathbf{m}}(t) m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}} - \frac{\theta_{\mathbf{s}} q_{\mathbf{s}}(t) m_{\mathbf{s}}}{\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}} < 0, \end{aligned}$$

where the inequality follows from the sub-case's condition and the assumption that $\theta > 0$.

$*$ **When $\boldsymbol{q_{\mathbf{m}}(t) = 0}$ and $\boldsymbol{q_{\mathbf{s}}(t) > 0}$**, the required resources to the group channel are allocated, and any leftover resources are allocated to the individual channel. The system dynamics are, therefore,

$$\begin{cases} \dot{q}_{\mathbf{m}}(t) = \lambda - \frac{1}{m_{\mathbf{m}}}\left(\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}\right) \tilde{z}_{\mathbf{m}}; \\ \dot{q}_{\mathbf{s}}(t) = \frac{1}{m_{\mathbf{m}}} p \mu_{\mathbf{m}} \tilde{z}_{\mathbf{m}} - \frac{1}{m_{\mathbf{s}}}\left(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}\right)\left(N - \tilde{z}_{\mathbf{m}}\right) - \theta_{\mathbf{s}} q_{\mathbf{s}}(t), \end{cases}$$

where $\tilde{z}_{\mathbf{m}} = \left(\frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}} \wedge N\right)$.

We have,

$$\begin{aligned} \nabla_q V(q)^T f(q) &= \frac{m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}\left(\lambda - \tfrac{1}{m_{\mathbf{m}}}\left(\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}\right) \tilde{z}_{\mathbf{m}}\right) + \frac{m_{\mathbf{s}}}{\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}}\left(\tfrac{1}{m_{\mathbf{m}}} p \mu_{\mathbf{m}} \tilde{z}_{\mathbf{m}} - \tfrac{1}{m_{\mathbf{s}}}\left(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}\right)\left(N - \tilde{z}_{\mathbf{m}}\right) - \theta_{\mathbf{s}} q_{\mathbf{s}}(t)\right) \\ &= \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}} - N + \frac{p \mu_{\mathbf{m}} m_{\mathbf{s}}}{\left(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}\right) m_{\mathbf{m}}} \tilde{z}_{\mathbf{m}} - \frac{\theta_{\mathbf{s}} q_{\mathbf{s}}(t) m_{\mathbf{s}}}{\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}} < 0, \end{aligned}$$

where the first inequality follows from the fact that $\tilde{z}_{\mathbf{m}} \leq N$; the last inequality follows from the sub-case's condition and the assumption that $\theta > 0$.

—**Sub-case II: $\boldsymbol{\frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}} \leq N < \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}} + \frac{p \lambda m_{\mathbf{s}}}{(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}) m_{\mathbf{m}}}}$. $\boldsymbol{\bar{q} = \left(0, \frac{1}{\theta_{\mathbf{s}}}\left(\lambda p - \frac{(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}})}{m_{\mathbf{s}}}\left(N - \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}\right)\right)\right)}$.**

We consider the Lyapunov function

$$V(q) = |q_{\mathbf{m}} - \bar{q}_{\mathbf{m}}| + |q_{\mathbf{s}} - \bar{q}_{\mathbf{s}}|,$$

where the equilibrium point $\bar{q}$, and show its asymptotic stability.

$*$ **When $\boldsymbol{q_{\mathbf{m}}(t) \geq \bar{q}_{\mathbf{m}},\ q_{\mathbf{s}}(t) \geq \bar{q}_{\mathbf{s}}}$, and $\boldsymbol{q(t) \neq \bar{q}}$**, the system dynamics are:

$$\begin{cases} \dot{q}_{\mathbf{m}}(t) = \lambda - \frac{1}{m_{\mathbf{m}}}\left(\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}\right) \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}} - \theta_{\mathbf{m}} q_{\mathbf{m}}(t) = -\theta_{\mathbf{m}} q_{\mathbf{m}}(t); \\ \dot{q}_{\mathbf{s}}(t) = \lambda p - \frac{1}{m_{\mathbf{s}}}\left(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}\right)\left(N - \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}\right) - \theta_{\mathbf{s}} q_{\mathbf{s}}(t). \end{cases}$$

We, therefore, have

$$\begin{aligned} \nabla_q V(q)^T f(q) &= \lambda p - \tfrac{1}{m_{\mathbf{s}}}\left(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}\right)\left(N - \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}\right) - \theta_{\mathbf{m}} q_{\mathbf{m}}(t) - \theta_{\mathbf{s}} q_{\mathbf{s}}(t) \\ &< \lambda p - \tfrac{1}{m_{\mathbf{s}}}\left(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}\right)\left(N - \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}\right) - \theta_{\mathbf{m}} \bar{q}_{\mathbf{m}} - \theta_{\mathbf{s}} \bar{q}_{\mathbf{s}} = 0, \end{aligned}$$

where the inequality follows from the fact that $q_{\mathbf{m}}(t) \geq \bar{q}_{\mathbf{m}},\ q_{\mathbf{s}}(t) \geq \bar{q}_{\mathbf{s}}$, and $q(t) \neq \bar{q}$, while the equality arises from the conditions of this sub-case.

∗ **When $\boldsymbol{q_{\mathrm{m}}(t) < \bar{q}_{\mathrm{m}},\ q_{\mathrm{s}}(t) < \bar{q}_{\mathrm{s}}}$, and $\boldsymbol{q(t) \neq \bar{q}}$**, we get the same $\nabla_q V(q)^T f(q)$ as in the previous case with a negative sign; namely,

$$\begin{aligned}\nabla_q V(q)^T f(q) &= -\lambda p + \tfrac{1}{m_{\mathbf{s}}}\left(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}\right)\left(N - \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}\right) + \theta_{\mathbf{m}} q_{\mathbf{m}}(t) + \theta_{\mathbf{s}} q_{\mathbf{s}}(t)\\ &< -\lambda p + \tfrac{1}{m_{\mathbf{s}}}\left(\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}\right)\left(N - \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}\right) + \theta_{\mathbf{m}} \bar{q}_{\mathbf{m}} + \theta_{\mathbf{s}} \bar{q}_{\mathbf{s}} = 0,\end{aligned}$$

where the inequality follows from the fact that $q_{\mathbf{m}}(t) \geq \bar{q}_{\mathbf{m}},\ q_{\mathbf{s}}(t) \geq \bar{q}_{\mathbf{s}}$, and $q(t) \neq \bar{q}$. The last equality follows from the conditions of this sub-case.

The other two cases for the different relations between $q_{\mathbf{s},i}(t)$ and $\bar{q}_{\mathbf{s},i}$ are handled in exactly the same way and are therefore omitted.

—**Sub-case III. $\boldsymbol{N < \frac{\lambda m_{\mathrm{m}}}{\mu_{\mathrm{m}} + \gamma_{\mathrm{m}}}}$. $\boldsymbol{\bar{q} = \left(\frac{\lambda - \frac{(\mu_{\mathrm{m}} + \gamma_{\mathrm{m}})}{m_{\mathrm{m}}} N}{\theta_{\mathrm{m}}}, \frac{p\mu_{\mathrm{m}}}{\theta_{\mathrm{s}} m_{\mathrm{m}}} N\right)}$**. We consider the Lyapunov function

$$V(q) = |q_{\mathbf{m}} - \bar{q}_{\mathbf{m}}| + |q_{\mathbf{s}} - \bar{q}_{\mathbf{s}}|,$$

where the equilibrium point $\bar{q}$, and show its asymptotic stability. Since the conditions $V(\bar{q}) = 0$ and $V(q) \to \infty$ as $\|q\| \to \infty$ can easily be verified, we focus on showing that $\nabla_q V(q)_{\mathbf{s}} T f(q) < 0$ for $q \neq \bar{q}$.

∗ **When $\boldsymbol{q_{\mathrm{m}}(t) \geq \bar{q}_{\mathrm{m}},\ q_{\mathrm{s}}(t) \geq \bar{q}_{\mathrm{s}}}$, and $\boldsymbol{q(t) \neq \bar{q}}$**, all of the resources are allocated to the group channel. The system dynamics are therefore,

$$\begin{cases} \dot{q}_{\mathbf{m}}(t) = \lambda - \frac{(\mu_{\mathbf{m}} + \gamma_m)}{m_{\mathbf{m}}} N - \theta_{\mathbf{m}} q_{\mathbf{m}}(t); \\ \dot{q}_{\mathbf{s}}(t) = \frac{p\mu_{\mathbf{m}}}{m_{\mathbf{m}}} N - \theta_{\mathbf{s}} q_{\mathbf{s}}(t); \end{cases}$$

We have

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \lambda - \frac{(\mu_{\mathbf{m}} + \gamma_m)}{m_{\mathbf{m}}} N - \theta_{\mathbf{m}} q_{\mathbf{m}}(t) + \frac{p\mu_{\mathbf{m}}}{m_{\mathbf{m}}} N - \theta_{\mathbf{s}} q_{\mathbf{s}}(t)\\ &< \lambda - \frac{(\mu_{\mathbf{m}} + \gamma_m)}{m_{\mathbf{m}}} N + \frac{p\mu_{\mathbf{m}}}{m_{\mathbf{m}}} N - \theta_{\mathbf{m}} \bar{q}_{\mathbf{m}} - \theta_{\mathbf{s}} \bar{q}_{\mathbf{s}} = 0,\end{aligned}$$

where the inequality follows from the fact that $q_{\mathbf{m}}(t) \geq \bar{q}_{\mathbf{m}},\ q_{\mathbf{s}}(t) \geq \bar{q}_{\mathbf{s}}$, and $q(t) \neq \bar{q}$. The last equality follows from the conditions of this sub-case.

∗ **When $\boldsymbol{q_{\mathrm{m}}(t) < \bar{q}_{\mathrm{m}},\ q_{\mathrm{s}}(t) < \bar{q}_{\mathrm{s}}}$, and $\boldsymbol{q(t) \neq \bar{q}}$**, we have,

$$\begin{aligned}\nabla_q V(q)^T f(q) &= -\lambda + \frac{(\mu_{\mathbf{m}} + \gamma_m)}{m_{\mathbf{m}}} N + \theta_{\mathbf{m}} q_{\mathbf{m}}(t) - \frac{p\mu_{\mathbf{m}}}{m_{\mathbf{m}}} N + \theta_{\mathbf{s}} q_{\mathbf{s}}(t)\\ &< -\lambda + \frac{(\mu_{\mathbf{m}} + \gamma_m)}{m_{\mathbf{m}}} N - \frac{p\mu_{\mathbf{m}}}{m_{\mathbf{m}}} N + \theta_{\mathbf{m}} \bar{q}_{\mathbf{m}} + \theta_{\mathbf{s}} \bar{q}_{\mathbf{s}} = 0,\end{aligned}$$

where the inequality follows from the fact that $q_{\mathbf{m}}(t) < \bar{q}_{\mathbf{m}},\ q_{\mathbf{s}}(t) < \bar{q}_{\mathbf{s}}$, and $q(t) \neq \bar{q}$. The last equality follows from the conditions of this sub-case.

The other two cases for the different relations between $q_{\mathbf{s},i}(t)$ and $\bar{q}_{\mathbf{s},i}$ are handled in exactly the same way and are therefore omitted.

- **If $\boldsymbol{\mathcal{P}_{\mathrm{m}} < \mathcal{P}_{\mathrm{s}}}$**: We consider the two sub-cases described in Table 6. For each sub-case we prove what the globally asymptotically stable equilibrium $\bar{q} = (\bar{q}_{\mathbf{m}}, \bar{q}_{\mathbf{s}})$ is. To this end, we construct a Lyapunov function for each case and demonstrate the global asymptotic stability of the equilibrium point.

—**Sub-case I.** $\boldsymbol{\frac{\lambda m_{\mathrm{m}}}{\mu_{\mathrm{m}}+\gamma_{\mathrm{m}}}+\frac{\lambda p m_{\mathrm{s}}}{(\mu_{\mathrm{s}}+\gamma_{\mathrm{s}})m_{\mathrm{m}}} \leq N}$. $\boldsymbol{\bar{q}=(0,0)}$. We consider the Lyapunov function

$$V(q)=\frac{p\mu_{\mathbf{m}}m+\mu_{\mathbf{s}}}{\mu_{\mathbf{m}}m\mu_{\mathbf{s}}}|q_{\mathbf{m}}-\bar{q}_{\mathbf{m}}|+\frac{m_{\mathbf{s}}}{\mu_{\mathbf{s}}+\gamma_{\mathbf{s}}}|q_{\mathbf{s}}-\bar{q}_{\mathbf{s}}|,$$

where the equilibrium point $\bar{q}=(0,0)$, and show its asymptotic stability.

$*$ **When $\boldsymbol{q_{\mathrm{m}}(t)>0}$ and $\boldsymbol{q_{\mathrm{s}}(t)=0}$**, all of the resources are allocated to the group and individual channels. Specifically, the system dynamics in (D.8) are as follows:

$$\begin{cases} \dot{q}_{\mathbf{m}}(t)=\lambda-\frac{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}{m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})+p\mu_{\mathbf{m}}m_{\mathbf{s}}}N-\theta_{\mathbf{m}}q_{\mathbf{m}}(t); \\ \dot{q}_{\mathbf{s}}(t)=\frac{p(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})\mu_{\mathbf{m}}}{m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})+p\mu_{\mathbf{m}}m_{\mathbf{s}}}N-\frac{p\mu_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}{m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})+p\mu_{\mathbf{m}}m_{\mathbf{s}}}N=0. \end{cases}$$

We have

$$\begin{aligned} \nabla_q V(q)^T f(q) &= \frac{m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})+p\mu_{\mathbf{m}}m_{\mathbf{s}}}{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}\left(\lambda-\frac{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}{m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})+p\mu_{\mathbf{m}}m_{\mathbf{s}}}N-\theta_{\mathbf{m}}q_{\mathbf{m}}(t)\right) \\ &= \frac{\lambda m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})+\lambda p\mu_{\mathbf{m}}m_{\mathbf{s}}}{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}-N-\frac{m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})+p\mu_{\mathbf{m}}m_{\mathbf{s}}}{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}\theta_{\mathbf{m}}q_{\mathbf{m}}(t) \\ &= \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}}+\frac{\lambda p\mu_{\mathbf{m}}m_{\mathbf{s}}}{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}-N-\frac{m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})+p\mu_{\mathbf{m}}m_{\mathbf{s}}}{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}\theta_{\mathbf{m}}q_{\mathbf{m}}(t) \\ &\leq \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}}+\frac{\lambda p m_{\mathbf{s}}}{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}-N-\frac{m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})+p\mu_{\mathbf{m}}m_{\mathbf{s}}}{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}\theta_{\mathbf{m}}q_{\mathbf{m}}(t)<0, \end{aligned}$$

where the first inequality follows from the fact that $\mu_{\mathbf{m}}/(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})<1$, and the second inequality follows from the sub-case's condition and the assumption that $\theta>0$.

—**Sub-case II.** $\boldsymbol{\frac{\lambda m_{\mathrm{m}}}{\mu_{\mathrm{m}}+\gamma_{\mathrm{m}}}+\frac{\lambda p m_{\mathrm{s}}}{(\mu_{\mathrm{s}}+\gamma_{\mathrm{s}})m_{\mathrm{m}}} > N}$. $\boldsymbol{\bar{q}=\left(\frac{1}{\theta_{\mathrm{m}}}\left(\lambda-\frac{(\mu_{\mathrm{m}}+\gamma_{\mathrm{m}})(\mu_{\mathrm{s}}+\gamma_{\mathrm{s}})}{p\mu_{\mathrm{m}}m_{\mathrm{s}}+m_{\mathrm{m}}(\mu_{\mathrm{s}}+\gamma_{\mathrm{s}})}N\right),0)\right)}$. The Lyapunov functions we use is

$$V(q)=|q_{\mathbf{m}}-\bar{q}_{\mathbf{m}}|+|q_{\mathbf{s}}-\bar{q}_{\mathbf{s}}|,$$

where the equilibrium point $\bar{q}$, and show its asymptotic stability.

$*$ **When $\boldsymbol{q_{\mathrm{m}}(t)\geq\bar{q}_{\mathrm{m}},\ q_{\mathrm{s}}(t)\geq\bar{q}_{\mathrm{s}}}$, and $\boldsymbol{q(t)\neq\bar{q}}$**, all of the resources are allocated to the group channel. The system dynamics are therefore,

$$\begin{cases} \dot{q}_{\mathbf{m}}(t)=\lambda-\frac{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}{p\mu_{\mathbf{m}}m_{\mathbf{s}}+m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}N-\theta_{\mathbf{m}}q_{\mathbf{m}}(t); \\ \dot{q}_{\mathbf{s}}(t)=\frac{p\mu_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}{p\mu_{\mathbf{m}}m_{\mathbf{s}}+m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}N-\frac{p\mu_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}{p\mu_{\mathbf{m}}m_{\mathbf{s}}+m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}N-\theta_{\mathbf{s}}q_{\mathbf{s}}(t)=-\theta_{\mathbf{s}}q_{\mathbf{s}}(t); \end{cases}$$

We have

$$\begin{aligned} \nabla_q V(q)^T f(q) &= \left(\lambda-\frac{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}{p\mu_{\mathbf{m}}m_{\mathbf{s}}+m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}N\right)-\theta_{\mathbf{m}}q_{\mathbf{m}}(t)-\theta_{\mathbf{s}}q_{\mathbf{s}}(t) \\ &< \left(\lambda-\frac{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}{p\mu_{\mathbf{m}}m_{\mathbf{s}}+m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}N\right)-\theta_{\mathbf{m}}\bar{q}_{\mathbf{m}}-\theta_{\mathbf{s}}\bar{q}_{\mathbf{s}}=0, \end{aligned}$$

where the inequality follows from the fact that $q_{\mathbf{m}}(t)\geq\bar{q}_{\mathbf{m}},\ q_{\mathbf{s}}(t)\geq\bar{q}_{\mathbf{s}}$, and $q(t)\neq\bar{q}$. The last equality follows from the sub-case conditions.

$*$ **When $\boldsymbol{q_{\mathrm{m}}(t)<\bar{q}_{\mathrm{m}},\ q_{\mathrm{s}}(t)<\bar{q}_{\mathrm{s}}}$, and $\boldsymbol{q(t)\neq\bar{q}}$**, we have,

$$\begin{aligned} \nabla_q V(q)^T f(q) &= -\left(\lambda-\frac{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}{p\mu_{\mathbf{m}}m_{\mathbf{s}}+m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}N\right)+\theta_{\mathbf{m}}q_{\mathbf{m}}(t)+\theta_{\mathbf{s}}q_{\mathbf{s}}(t) \\ &< -\left(\lambda-\frac{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}{p\mu_{\mathbf{m}}m_{\mathbf{s}}+m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}N\right)+\theta_{\mathbf{m}}\bar{q}_{\mathbf{m}}+\theta_{\mathbf{s}}\bar{q}_{\mathbf{s}}=0, \end{aligned}$$

where the inequality follows from the fact that $q_{\mathbf{m}}(t)<\bar{q}_{\mathbf{m}},\ q_{\mathbf{s}}(t)<\bar{q}_{\mathbf{s}}$, and $q(t)\neq\bar{q}$, and the last equality follows from the sub-case conditions.

Part 2. Recall the long-run profit maximization problem (4). Let $\bar{z}^* = (\bar{z}^*_{\mathbf{m}}, \bar{z}^*_{\mathbf{s}})$ and $\bar{q}^* = (\bar{q}^*_{\mathbf{m}}, \bar{q}^*_{\mathbf{s}})$ denote its solution (i.e., long-run average resource allocation and corresponding queue length for each channel). To prove the optimality of the $\mathcal{P}$ rule, it suffices to show that $\bar{z}^*$ and $\bar{q}^*$ constitute the globally asymptotically stable equilibrium established in Part 1 of this proof. We, therefore, consider the same cases as in Part 1, and present the optimal solution $\bar{z}^*$ and $\bar{q}^*$.

- **If $\boldsymbol{\mathcal{P}_{\mathbf{m}} > \mathcal{P}_{\mathbf{s}}}$:** In this case, the $\mathcal{P}$ rule gives strict priority to the group channel. We consider the three sub-cases shown in Table 5.

Table 5 Optimal solution – $\mathcal{P}_{\mathbf{m}} > \mathcal{P}_{\mathbf{s}}$.

Sub-case	$(\bar{z}^*_{\mathbf{m}}, \bar{z}^*_{\mathbf{s}})$	$(\bar{q}^*_{\mathbf{s}}, \bar{q}^*_{\mathbf{s}})$
I. $\frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}} + \frac{p\mu_{\mathbf{m}} m_{\mathbf{s}}}{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}}) m_{\mathbf{m}}} N \leq N$	$\left(\frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}}, \frac{\lambda p\mu_{\mathbf{m}} m_{\mathbf{s}}}{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})}\right)$	$(0,0)$
II. $\frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}} \leq N < \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}} + \frac{p\lambda m_{\mathbf{s}}}{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}}) m_{\mathbf{m}}}$	$\left(\frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}}, N - \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}}\right)$	$\left(0, \frac{1}{\theta_{\mathbf{s}}}\left(\lambda p - \frac{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}{m_{\mathbf{s}}}\left(N - \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}}\right)\right)\right)$
III. $N < \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}}$	$(N,0)$	$\left(\frac{\lambda - \frac{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})}{m_{\mathbf{m}}} N}{\theta_{\mathbf{m}}}, \frac{p\mu_{\mathbf{m}} N}{\theta_{\mathbf{s}} m_{\mathbf{m}}}\right)$

Except for sub-case I, where there are enough resources to serve all patients, the two other sub-cases prioritize the group channel first and then the individual channel. This is align with the $\mathcal{P}$ rule prioritization in this scenario.

- **If $\boldsymbol{\mathcal{P}_{\mathbf{m}} < \mathcal{P}_{\mathbf{s}}}$:** In this case, the $\mathcal{P}$ rule gives priority to the individual channel and allocates sufficient resources to the group channel to achieve that. We consider the two sub-cases described in Table 6.

$$\bar{z}_{\mathbf{m}} = \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}} \wedge \frac{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})\, m_{\mathbf{m}} N}{m_{\mathbf{m}}\,(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}}) + p\mu_{\mathbf{m}} m_{\mathbf{s}}}, \qquad \bar{z}_{\mathbf{s}} = \frac{p\mu_{\mathbf{m}} m_{\mathbf{s}}}{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})\, m_{\mathbf{m}}} \bar{z}_{\mathbf{m}}.$$

Table 6 Optimal solution – $\mathcal{P}_{\mathbf{m}} < \mathcal{P}_{\mathbf{s}}$.

Sub-case	$(\bar{z}^*_{\mathbf{m}}, \bar{z}^*_{\mathbf{s}})$	$(\bar{q}^*_{\mathbf{m}}, \bar{q}^*_{\mathbf{s}})$
I. $\frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}} + \frac{\lambda p m_{\mathbf{s}}}{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}}) m_{\mathbf{m}}} \leq N$	$\left(\frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}}, \frac{\lambda p\mu_{\mathbf{m}} m_{\mathbf{s}}}{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})}\right)$	$(0,0)$
II. $N < \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}} + \frac{\lambda p m_{\mathbf{s}}}{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}}) m_{\mathbf{m}}}$	$\left(\frac{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}}) m_{\mathbf{m}} N}{m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})+p\mu_{\mathbf{m}} m_{\mathbf{s}}}, \frac{p\mu_{\mathbf{m}} m_{\mathbf{s}} N}{m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}}) m_{\mathbf{m}}+p\mu_{\mathbf{m}} m_{\mathbf{s}}}\right)$	$\left(\frac{1}{\theta_{\mathbf{m}}}\left(\lambda - \frac{(\mu_{\mathbf{m}}+\gamma_{\mathbf{m}})(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}}) N}{p\mu_{\mathbf{m}} m_{\mathbf{s}}+m_{\mathbf{m}}(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})}\right),0)\right)$

In sub-case I, there are enough resources to treat all patients. In sub-case II, however, resources are allocated to both channels, ensuring the group channel receives enough resources to allow the maximum allocation to the individual channel. This is also in line with the $\mathcal{P}$ rule and concludes the proof. Q.E.D.

Proof of Lemma A.1: We prove the lemma by demonstrating that the cost difference between a trajectory with chattering and one without is negligible. This shows that any admissible control policy that results in a chattering interval can be replaced by a cost-equivalent policy that yields chattering-free state trajectories. Consequently, it is optimal to consider only state trajectories without chattering when solving the transient optimal control problem in (9).

The general structure of the proof is similar to Proposition 1 in Hu et al. (2022), however, our case incorporates the group channel, treatment cost savings, dropouts and no-shows. Additionally, due to the asymmetry of our system, we need to establish the optimality of chattering-free behavior for both classes (Part I and Part II).

Part I. Consider an interval $I_1 := [0, \epsilon]$, for some small $\epsilon > 0$. The group channel starts this interval with zero queue ($q_{\mathbf{m}}(0) = 0$) and is not allocated any resources throughout the interval. Following I_1, we define another interval $I_2 = (\epsilon, \epsilon + \epsilon')$, where the group channel is allocated all resources and is emptied by the end of I_2. Let $q_{\mathbf{s}}(0) = q_{\mathbf{s}0}$, $q_{\mathbf{s}0} \in \mathbb{R}_+$. Next, we compute the state trajectories and the associated cost over the interval $[0, \epsilon + \epsilon']$.

In the first interval I_1, where $t \in [0, \epsilon]$, the state trajectories evolve as:

$$\begin{aligned} q_{\mathbf{m}}(t) &= \lambda t + o(\epsilon), \\ q_{\mathbf{s}}(t) &= q_{\mathbf{s}0} - t\left[q_{\mathbf{s}0}\theta_{\mathbf{s}} - \frac{\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}}{m_{\mathbf{s}}}N\right] + o(\epsilon). \end{aligned}$$

At the end of I_1, the queue lengths are:

$$q_{\mathbf{m}}(\epsilon) = \lambda\epsilon + o(\epsilon), \quad q_{\mathbf{s}}(\epsilon) = q_{\mathbf{s}0} - \epsilon\left[q_{\mathbf{s}0}\theta_{\mathbf{s}} - \frac{\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}}{m_{\mathbf{s}}}N\right] + o(\epsilon).$$

Using these as initial conditions for the second interval, the state trajectories in I_2, where $t \in [\epsilon, \epsilon + \epsilon']$ evolve as:

$$\begin{aligned} q_{\mathbf{m}}(t) &= q_{\mathbf{m}}(\epsilon) + (t - \epsilon)\left[-\theta_{\mathbf{m}}q_{\mathbf{m}}(\epsilon) + \lambda - \frac{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}{m_{\mathbf{m}}}N\right] + o(\epsilon), \\ q_{\mathbf{s}}(t) &= q_{\mathbf{s}}(\epsilon) + (t - \epsilon)\left[-\theta_{\mathbf{s}}q_{\mathbf{s}}(\epsilon) + \frac{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}{m_{\mathbf{m}}}Np\right] + o(\epsilon). \end{aligned}$$

By requiring that $q_{\mathbf{m}}(\epsilon') = 0$, we get that the time the group queue empties from its initial value $q_{\mathbf{m}}(\epsilon)$ is:

$$\epsilon' = \frac{\lambda\epsilon}{-\lambda + \frac{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}{m_{\mathbf{m}}}N} + o(\epsilon).$$

The total cost savings over the two intervals is:

$$C = \int_0^{\epsilon} r_{\mathbf{s}}Nt \;\mathrm{d}t + \int_{\epsilon}^{\epsilon'} r_{\mathbf{m}}Nt \;\mathrm{d}t - \int_0^{\epsilon+\epsilon'} \left[c_{\mathbf{m}}q_{\mathbf{m}}(t) + c_{\mathbf{s}}q_{\mathbf{s}}(t)\right]\mathrm{d}t$$

In contrast, we now consider a single interval of length, $\epsilon + \epsilon'$, with the same initial conditions, only now instead of allowing $q_{\mathbf{m}}$ to increase and then decrease to zero, we assign strict priority to the group channel, maintaining $\tilde{q}_{\mathbf{m}}$ at zero throughout the interval. The remaining resources are allocated to the individual channel. Similarly, we characterize the state trajectory over this interval, where $t \in [\epsilon, \epsilon + \epsilon']$ as:

$$\begin{aligned} \tilde{q}_{\mathbf{m}}(t) &= 0, \\ \tilde{q}_{\mathbf{s}}(t) &= q_{\mathbf{s}0} + t\left[-q_{\mathbf{s}0}\theta_{\mathbf{s}} + \frac{p\lambda m_m}{\mu_m + \gamma_m} - \frac{\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}}{m_{\mathbf{s}}}\left(N - \frac{\lambda m_{\mathbf{m}}}{\mu_m + \gamma_m}\right)\right] + o(\epsilon). \end{aligned}$$

The total cost savings in this case is:

$$\tilde{C} = \int_0^{\epsilon+\epsilon'} \left[r_{\mathbf{s}} \left(N - \frac{\lambda m_{\mathbf{m}}}{\mu_m + \gamma_m} \right) t + r_{\mathbf{m}} \frac{\lambda m_{\mathbf{m}} t}{\mu_m + \gamma_m} - c_{\mathbf{m}} q_{\mathbf{m}}(t) + c_{\mathbf{s}} q_{\mathbf{s}}(t) \right] \mathrm{d}t$$

By comparing C and $\tilde{C}$, we obtain:

$$\begin{aligned} C - \tilde{C} = {} & \frac{\epsilon^2}{2(\lambda - \frac{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}}{m_{\mathbf{m}}} N)^2} \Bigg(\lambda \Big(c_{\mathbf{s}} \epsilon \lambda \Big(q_{\mathbf{s}0} \theta_{\mathbf{s}}^2 - \theta_{\mathbf{s}} \frac{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}{m_{\mathbf{m}}} Np \Big) \Big) + c_{\mathbf{s}} N \Big((1 + \epsilon(\theta_{\mathbf{s}}))\lambda - \frac{\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}}{m_{\mathbf{s}}} N \Big) \mu_{\mathbf{s}} \\ & - c_{\mathbf{m}} \left(\epsilon \theta_{\mathbf{m}} \lambda^2 + \lambda \frac{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}{m_{\mathbf{m}}} N - N^2 \left(\frac{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}{m_{\mathbf{m}}} \right)^2 \right) \Bigg) \\ & + \frac{\epsilon^2}{2} \left(N \left(r_{\mathbf{s}} - r_{\mathbf{m}} + \frac{r_{\mathbf{m}} \lambda^2}{\left(-\lambda + \frac{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}}{m_{\mathbf{m}}} N \right)^2} \right) - r_{\mathbf{s}} \left(N - \frac{\lambda m_{\mathbf{m}}}{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}} \right) \frac{(\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}) N}{\left(-\lambda + \frac{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}}{m_{\mathbf{m}}} \right) m_{\mathbf{m}}} \right) = o(\epsilon) \end{aligned}$$

Moreover, at time $\epsilon + \epsilon'$, we have $q_{\mathbf{m}}(\epsilon + \epsilon') = \tilde{q}_{\mathbf{m}}(\epsilon + \epsilon') = 0$, and $q_{\mathbf{s}}(\epsilon + \epsilon') - \tilde{q}_{\mathbf{s}}(\epsilon + \epsilon') = o(\epsilon)$.

Part II. Since the channels are asymmetric, we now verify the same conclusion hold when the queue of the individual channel first increases and then empties. Specifically, consider an interval $I_1 := [0, \epsilon]$, for some small $\epsilon > 0$. The individual channel starts this interval with zero queue ($q_{\mathbf{s}}(0) = 0$) and is not allocated any resources throughout the interval. Following I_1, we define another interval $I_2 = (\epsilon, \epsilon + \epsilon')$, where the individual channel is allocated all resources and is emptied by the end of I_2. Let $q_{\mathbf{m}}(0) = q_{\mathbf{m}0}$, $q_{\mathbf{m}0} \in \mathbb{R}_+$. Next, we compute the state trajectories and the associated cost over the interval $[0, \epsilon + \epsilon']$.

In the first interval I_1, where $t \in [0, \epsilon]$, the state trajectories evolve as:

$$\begin{aligned} q_{\mathbf{m}}(t) &= q_{\mathbf{m}0} + \left[-\theta_{\mathbf{m}} q_{\mathbf{s}0} + \lambda - \frac{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}{m_{\mathbf{m}}} N \right] t + o(\epsilon), \\ q_{\mathbf{s}}(t) &= \left[\frac{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}{m_{\mathbf{m}}} Np \right] t + o(\epsilon). \end{aligned}$$

At the end of I_1, the queue lengths are:

$$q_{\mathbf{m}}(\epsilon) = q_{\mathbf{m}0} + \left[-\theta_{\mathbf{m}} q_{\mathbf{s}0} + \lambda - \frac{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}{m_{\mathbf{m}}} N \right] \epsilon + o(\epsilon), \quad q_{\mathbf{s}}(\epsilon) = \left[\frac{\mu_{\mathbf{m}} + \gamma_{\mathbf{m}}}{m_{\mathbf{m}}} Np \right] \epsilon + o(\epsilon).$$

Using these as initial conditions for the second interval, the state trajectories in I_2, where $t \in [\epsilon, \epsilon + \epsilon']$ evolve as:

$$\begin{aligned} q_{\mathbf{m}}(t) &= q_{\mathbf{m}}(\epsilon) + (t - \epsilon) \left[-\theta_{\mathbf{m}} q_{\mathbf{m}}(\epsilon) + \lambda \right] + o(\epsilon), \\ q_{\mathbf{s}}(t) &= q_{\mathbf{s}}(\epsilon) + (t - \epsilon) \left[-\theta_{\mathbf{s}} q_{\mathbf{s}}(\epsilon) - \frac{\mu_{\mathbf{s}} + \gamma_{\mathbf{s}}}{m_{\mathbf{s}}} N \right] + o(\epsilon). \end{aligned}$$

By requiring that $q_{\mathbf{s}}(\epsilon') = 0$, we get that the time the group queue empties from its initial value $q_{\mathbf{s}}(\epsilon)$ is:

$$\epsilon' = \frac{\frac{\mu_{\mathbf{m}}+\gamma_{\mathbf{m}}}{m_{\mathbf{m}}} p\epsilon}{\frac{\mu_{\mathbf{s}}+\gamma_{\mathbf{s}}}{m_{\mathbf{s}}}} + o(\epsilon).$$

The total cost savings over the two intervals is:

$$C = \int_0^{\epsilon} r_{\mathbf{m}} Nt \ \mathrm{d}t + \int_{\epsilon}^{\epsilon'} r_{\mathbf{s}} Nt \ \mathrm{d}t - \int_0^{\epsilon+\epsilon'} \left[c_{\mathbf{m}} q_{\mathbf{m}}(t) + c_{\mathbf{s}} q_{\mathbf{s}}(t) \right] \mathrm{d}t$$

In contrast, we now consider a single interval of length, $\epsilon+\epsilon'$, with the same initial conditions, only now instead of allowing $q_{\mathbf{s}}$ to increase and then decrease to zero, we assign strict priority to the individual channel, maintaining $\tilde{q}_{\mathbf{s}}$ at zero throughout the interval. The remaining resources are allocated to the individual channel. Similarly, we characterize the corresponding state trajectory over this interval, where $t \in [\epsilon, \epsilon+\epsilon']$ as:

$$\begin{aligned} \tilde{q}_{\mathbf{m}}(t) &= q_{\mathbf{m}0} - t\left[-q_{\mathbf{m}0}\theta_{\mathbf{m}} + \lambda - \frac{(m_{\mathbf{s}}+\gamma_{\mathbf{s}})\, m_{\mathbf{m}}N}{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})\, m_{\mathbf{m}} + p\mu_{\mathbf{m}}m_{\mathbf{s}}}\right] + o(\epsilon), \\ \tilde{q}_{\mathbf{s}}(t) &= 0. \end{aligned}$$

The total cost savings in this case is:

$$\tilde{C} = \int_0^{\epsilon+\epsilon'} \left[r_{\mathbf{s}}\left(\frac{p\mu_{\mathbf{m}}m_sN}{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})\, m_{\mathbf{m}} + p\mu_{\mathbf{m}}m_{\mathbf{s}}}\right)t + r_{\mathbf{m}}\left(\frac{(m_{\mathbf{s}}+\gamma_{\mathbf{s}})\, m_{\mathbf{m}}N}{(\mu_{\mathbf{s}}+\gamma_{\mathbf{s}})\, m_{\mathbf{m}} + o\mu_{\mathbf{m}}m_{\mathbf{s}}}\right)t - c_{\mathbf{m}}q_{\mathbf{m}}(t) + c_{\mathbf{s}}q_{\mathbf{s}}(t)\right] \mathrm{d}t$$

Similar to part I, we get that $C - \tilde{C} = o(\epsilon)$ as well as $q_{\mathbf{s}}(\epsilon+\epsilon') = \tilde{q}_{\mathbf{s}}(\epsilon+\epsilon') = 0$, and $q_{\mathbf{m}}(\epsilon+\epsilon') - \tilde{q}_{\mathbf{m}}(\epsilon+\epsilon') = o(\epsilon)$.

Part III. In both parts I and II, we showed that the cost under a policy that increases and then decreases one queue, and the cost under the strict priority rule keeping the queue at zero, differ by $o(\epsilon)$. Similarly, the queue lengths at time $\epsilon+\epsilon'$ under both policies differ by $o(\epsilon)$. For any interval of length L, dividing it into $O(L/\epsilon)$ small triangular trajectories (where one queue increases for ϵ units then decreases to zero), each incurs a cost difference of $o(\epsilon)$. Thus, the total cost difference between the two policies is $o(\epsilon)O(L/\epsilon)$, which tends to zero as $\epsilon \to 0$ for fixed L.

Any chattering interval consists of infinitely many such triangular trajectories. Therefore, any control policy π that causes chattering can be replaced by a cost-equivalent policy $\tilde{\pi}$, which holds the queue at zero and coincides with π elsewhere. Q.E.D.