From Waitlists to Well-Being: An Analytical Approach to AI-Supported Mental Healthcare

Junyang Cai, Noa Fani Likvornik, Noa Zychlinski
Faculty of Data and Decision Science, Technion – Israel Institute of Technology, Haifa, 3200003, Israel,
junyang.cai@campus.technion.ac.il, noalikvornik@campus.technion.ac.il, noazy@technion.ac.il

The growing burden of mental health disorders is straining healthcare systems, leading to long waitlists and delayed access to therapy. Such delays can worsen symptoms, increase dropout rates, and impose substantial societal costs. Advances in Artificial Intelligence (AI)—including supervised conversational agents—offer scalable opportunities to support patients during these waiting periods.

We develop a multi-class, multi-server queueing model to jointly optimize waitlist treatment and scheduling policies. Using a fluid approximation, we derive a simple, implementable policy that incorporates fixed and variable costs while capturing key behavioral features of psychotherapy, including no-shows, dropouts, and heterogeneous treatment effects.

Numerical experiments show that the proposed policy closely matches the exact Markov Decision Process (MDP) solution in small systems, remains robust across system scales and service-time distributions, and consistently outperforms benchmark policies. We further quantify how hiring costs affects waitlist, scheduling and recruitment decisions.

A case study calibrated to a representative Veterans Health Administration (VHA) mental health unit demonstrates that supervised waitlist treatment delivers substantial performance gains and, when coordinated with scheduling decisions, can significantly reduce the need for additional therapist recruitment. Overall, our results highlight the potential of analytically guided, AI-supported policies to improve access to mental health care.

Key words: Stochastic modeling, healthcare operations management, fluid models, scheduling queues, mental treatment with AI

1. Introduction

The growing burden of mental health disorders is placing severe strain on healthcare systems, leading to long waitlists and delays in access to therapy. Such delays often worsen symptoms, increase dropout rates, and generate substantial societal costs. This crisis has been further exacerbated by the COVID-19 pandemic (Pfefferbaum and North 2020). For instance, an estimated 1.2 million people in the UK are currently waiting for NHS mental health services (Department of Health & Social Care 2023), while between 2020 and 2022, approximately 3.4 million Australians sought mental health support, yet many faced waits of six months to a year (Australian Institute of Health and Welfare 2025).

The economic consequences are equally pressing: mental illness is estimated to cost $2.5 trillion annually worldwide, a figure projected to reach $6 trillion by 2030 (Lancet editorial 2020). Rising prevalence compounds this strain, with rates of mental health disorders increasing by 48% across 204 countries over the past three decades (GBD 2019 Mental Disorders Collaborators 2022).

Delays in treatment are especially concerning in mental health care, as they are linked to worsening symptom severity (Cuijpers et al. 2021), reduced patient engagement and treatment effectiveness (Swift et al. 2012, Steinert et al. 2017, Krendl and Lorenzo-Luaces 2022), and increased no-show and dropout rates (Milicevic et al. 2020, Xaba et al. 2024). Many patients also view long waits as a barrier to care, which can heighten distress (Steinert et al. 2017) and may even lead to increased substance use (Cai and Zychlinski 2025). These challenges underscore the critical importance of addressing the waiting period itself.

Recent advances in Artificial Intelligence (AI) offer a promising opportunity. AI-based tools—such as digital monitoring platforms and conversational agents, deployed under human supervision—can provide timely, scalable support to patients on waitlists. Recent user behavior further highlights this potential: a 2025 investigation by Zao-Sanders (2025) found that therapy and companionship rank as the most frequent applications of generative AI, indicating a widespread willingness to use AI for emotional support. This trend suggests that developing structured, clinically informed AI tools for mental health care could leverage an existing inclination toward such technologies, thereby extending their benefits in a safe and effective manner. Emerging industry implementations reflect this paradigm; for example, startups such as Mentality develop AI-driven tools, including systems designed to support patients while they are waiting for mental health treatment.

Current applications of AI in mental health are expanding across diagnosis, treatment, and patient management. Machine learning algorithms enable early detection of mental health risks (D'Alfonso 2020, Lee et al. 2021), predictive models help anticipate deterioration and treatment responses (Graham et al. 2019), and virtual therapists and conversational agents provide immediate psychological support and crisis assistance (Briot et al. 2021, Omarov et al. 2023). These capabilities are directly relevant to waitlist interventions, where scalable monitoring and timely support can mitigate patient deterioration during delays. At the same time, AI is also being explored for broader clinical applications—such as enhancing teletherapy sessions through speech and mood analysis or supporting ongoing monitoring with wearable devices and sentiment detection (Gomes et al. 2023)—highlighting the technology's wide-ranging potential in mental health care.

Within this landscape, *low-intensity waitlist treatment* (henceforth, *waitlist treatment*) has emerged as a particularly relevant intervention. These programs combine AI-driven tools with structured clinical oversight, ensuring that the system's recommendations remain aligned with evidence-based practices and established treatment protocols. Mental health professionals are involved in the design

and ongoing refinement of the intervention, conduct periodic audits of anonymized interaction samples, and monitor patient outcomes to evaluate effectiveness. In addition, automated safeguards are built in to detect high-risk situations (e.g., expressions of suicidal ideation) and trigger appropriate escalation procedures. Importantly, clinical oversight focuses on protocol design and quality assurance rather than continuous monitoring of every individual interaction, thereby protecting patient privacy while upholding safety and clinical integrity. Because much of the support can be delivered asynchronously, a single supervisor can oversee multiple patients, which enhances scalability relative to traditional one-on-one therapy (Mohr et al. 2011, 2014). Moreover, waitlist treatment can increase access to care, reduce psychological distress, and prevent deterioration or abandonment during prolonged waiting periods (Levin et al. 2022, Bennett-Levy et al. 2010).

At the same time, integrating such interventions introduces trade-offs: while waitlist treatment can improve outcomes and reduce abandonment, it also entails fixed setup costs and ongoing supervision expenses. In this paper, we develop a queueing-based analytical model to evaluate these trade-offs and to design optimal waitlist and scheduling policies for mental health systems.

Our paper makes the following key contributions:

- *Modeling AI-supported waitlist interventions in mental healthcare.* We develop a multi-class, multi-server queueing model that incorporates AI-supported waitlist treatment for patients awaiting psychotherapy. The model captures essential features of mental healthcare delivery, including no-shows, dropouts, heterogeneous treatment effects, and recovery-driven abandonment when waitlist interventions improve patient well-being.
- *Optimally integrating waitlist treatment, scheduling, and therapist recruitment.* We formulate an optimization problem to maximize long-run net benefit and derive a tractable, implementable solution using a fluid approximation combined with McCormick linearization. The framework enables the joint optimization of waitlist treatment and scheduling policies, determining which patient classes should receive waitlist support, how prioritization should be managed, and how service capacity should be allocated. We further extend the framework to incorporate therapist recruitment, showing how hiring costs shape the optimal balance between professional capacity and AI-supported treatment.

We show, through extensive numerical experiments, that the resulting policy is highly effective and robust: It closely matches the exact MDP solution in small systems and performs strongly in simulation across system scales and service-time distributions. The policy also outperforms intuitive benchmark policies and reveals an interaction between hiring costs and waitlist decisions.

- *VHA case study and policy insights.* Using evidence from the Veterans Health Administration (VHA), we calibrate the model to three prevalent diagnostic groups—major depressive disorder (MDD), anxiety disorders (AD), and post-traumatic stress disorder (PTSD). The case study demonstrates that supervised waitlist treatment can substantially improve system performance without

increasing staffing levels. By contrast, achieving comparable performance without waitlist treatment would require a substantial expansion of the therapist workforce. This highlights how well-managed waitlist treatment, coordinated with scheduling decisions, can significantly reduce the number of additional therapists that need to be recruited.

Organization. The remainder of the paper is organized as follows. Section 2 reviews related literature. Section 3 introduces the stochastic model for incorporating AI-supported waitlist treatment. Section 4 presents the fluid approximation, derives the optimal waitlist and scheduling policy, and extends the framework to incorporate therapist recruitment. Section 5 provides numerical experiments that evaluate the policy across multiple scenarios and compare it with benchmark alternatives. Section 6 applies the model to a case study of a VHA outpatient mental health unit. Finally, Section 7 concludes and outlines future research directions.

2. Literature Review

This work relates to two streams of research. The first pertains to the operations research (OR), operations management (OM) literature on mental healthcare. The second concerns scheduling queues with multiple customer classes.

2.1. OR/OM in Mental Healthcare

The OR/OM literature for mental health systems is sparse, but recent efforts have begun to explore the use of operations research in this domain. One of the earliest contributions in this area is Leff et al. (1986), who developed a linear programming–based planning model for community mental health support systems. Their model assists system managers in allocating limited resources across service programs by aggregating clients into functional groups and incorporating probabilistic estimates of service requirements associated with different care packages. The framework enables planners to evaluate trade-offs between resource constraints and service coverage in community-based mental health settings. A recent review by Noorain et al. (2023) highlights the early-stage development of operational models in mental healthcare compared to broader healthcare applications, identifying gaps and opportunities for optimization-based approaches.

Previous research has examined the operational challenges of mental health service delivery in clinical settings. For example, Williams et al. (2008) studied the operations of a large urban community mental health center, focusing on psychiatric services in an adult outpatient setting. By implementing systematic changes—such as same-day intake and psychiatric evaluations, performance tracking, and administrative streamlining—their study demonstrated significant reductions in wait times and no-show rates, ultimately improving access and resource utilization. In a related study, Koizumi et al. (2005) addressed congestion in intensive psychiatric facilities resulting from the downsizing of

state mental health institutions in Philadelphia. Their queuing network model with blocking showed that facility-specific capacity shortages created bottlenecks, and that alleviating these shortages could improve patient flow and system-wide efficiency. Arntzen et al. (2024) proposed a knapsack-based routing model to improve the allocation of individuals with severe mental illness to residential facilities. Their approach enhances both fairness and efficiency in placement decisions, significantly reducing disparities in wait times for hard-to-place clients. Applied to mental health facility allocation in Amsterdam, their findings highlight the potential of optimization techniques to support decision-making in resource-constrained service environments. Recently, Zychlinski (2026) analyzed the operation of mental health systems in the aftermath of a mass trauma event, characterized by a sudden influx of individuals requiring psychological care following exposure to a traumatic incident. The study focuses on how to dynamically balance group and individual therapy across the surge, recovery, and long-term care phases of the response. Complementing these analytical approaches, simulation modeling has been widely used to evaluate mental health service operations. Several studies, including Long and Meadows (2018) and Noorain et al. (2019), provide systematic reviews of simulation-based methods for assessing and improving mental health service delivery.

We complement this stream of research by developing and optimizing an analytical model that incorporates AI-based waitlist treatments into psychotherapy scheduling. By capturing key behavioral features such as no-shows, dropouts, and heterogeneous treatment responses, the policy we suggest provide the basis for evaluating how digital interventions can be optimally integrated into scheduling decisions in mental healthcare settings.

2.2. Scheduling Multi-Class Queues

Our paper is also closely related to the literature on scheduling multiple customer classes in stochastic processing networks. The scheduling of multiple customer classes within stochastic processing networks has been extensively studied. A foundational result in this area is the $c\mu$ rule, introduced by Cox and Smith (1961), which established the optimality of a simple index-based policy for a single-server queue with linear holding costs. Subsequent research has expanded on this rule, offering various generalizations. However, the optimality of these extensions is typically proven only in asymptotic regimes (e.g., Van Mieghem 1995, Mandelbaum and Stolyar 2004, Huang et al. 2015).

In multi-server systems, the dynamic scheduling of multiple classes with customer abandonment has been explored under different operational regimes. For example, Harrison and Zeevi (2004) and Atar et al. (2004) examined this problem under a critically loaded regime. Atar et al. (2010) further advanced this work by deriving the asymptotic optimality of the $c\mu/\theta$ rule for many-server queues with abandonment in the heavy traffic regime. More recent contributions, such as Long et al. (2020), have extended these scheduling rules to accommodate general queue length cost functions and varied customer patience distributions. Other recent developments include scheduling customers with

heterogeneous resource requirements (Zychlinski et al. 2023), managing new and returning patients in hybrid healthcare systems (Zychlinski 2024), and designing AI-assisted medical diagnostics (Cai and Zychlinski 2026).

In this work, we complement the existing literature by studying the scheduling of multiple patient classes while explicitly incorporating AI-based waitlist interventions, together with key behavioral features such as no-shows and dropouts during treatment. We further examine the trade-offs between investing in AI-supported interventions and expanding therapist capacity through additional recruitment, thereby jointly studying scheduling, waitlist, and recruitment decisions in contemporary mental health systems.

3. Stochastic Model

We study a Markovian N-server queueing system with I patient classes, distinguished by their initial diagnosis and degree of psychological distress. The parameter N denotes the number of full-time-equivalent (FTE) therapists available in the system. Figure 1 provides a model illustration of patient flow in the system.

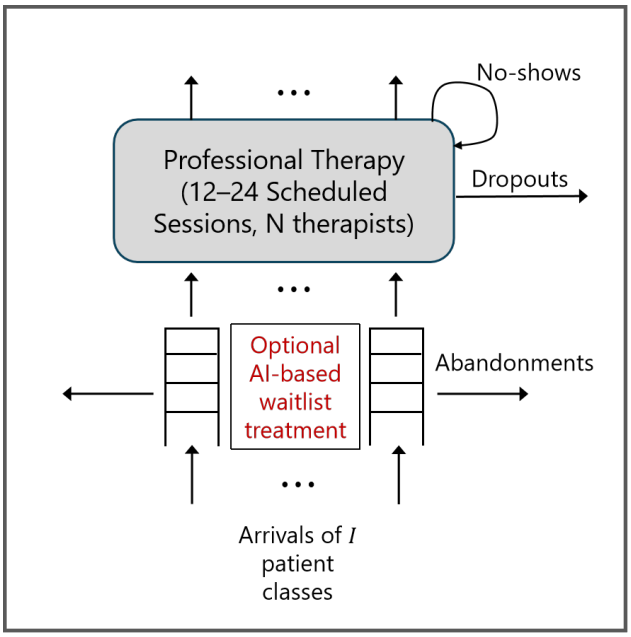

Figure 1 **Model illustration of patient flow in a mental health system with AI-supported waitlist treatment.**

Patients of Class i, $i \in \mathcal{I} := \{1, \ldots, I\}$, arrive according to a Poisson process with rate λ_i. Let $A_i(t)$ denote the cumulative number of Class i patients up to time t. Let $X_i(t)$ and $Q_i(t)$, $i \in \mathcal{I}$, denote the number of Class i customers in the system and in the queue, respectively, at time t, $t \geq 0$. Moreover, we use the notation $X(t) = (X_i(t), i \in \mathcal{I})$ and $Q(t) = (Q_i(t), i \in \mathcal{I})$. Let $Z_i(t)$ denote the number of servers assigned to Class i at time t; $Z(t) = (Z_i(t), i \in \mathcal{I})$ are the decision variables.

Service, abandonment and associated costs. The treatment for patients consists of a series of 12–24 psychological sessions, typically scheduled on a weekly basis. Since the number of sessions is determined at the *outset* of treatment and the sessions are delivered sequentially by the same therapist, we model the entire course of treatment as a single service time.

While awaiting the first 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_i^b and θ_i^n, respectively, for Class $i \in \mathcal{I}$ patients.

Let $\Gamma_i(t)$ and $D_i(t)$, $i \in \mathcal{I}$, denote the cumulative number of abandonments and treatment completions of Class i patients, respectively, up to time $t \geq 0$.

Finally, let h_i denote the holding cost per unit time for a Class i patient, b_i the benefit obtained upon treatment completion, and α_i^n the cost incurred when a Class i patient abandons the system while waiting for treatment.

No-shows and dropouts. No-shows and dropouts are common in mental health treatments (Milicevic et al. 2020, Xaba et al. 2024). 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_i as the probability of a Class i patient to show up for a scheduled appointment, where $1-\beta_i$ represents the probability of a no-show.

To model the total service requirement accounting for no-shows, we first note that $1/\mu_i^b$ represents the mean required service time assuming patients attend all scheduled appointments; equivalently, $1/\mu_i^b$ is the expected number of *attended* sessions needed to complete treatment. However, because patients may miss some of their scheduled appointments, additional appointment attempts are required. We therefore let $1/\mu_i$ denote the expected total number of appointment attempts (including those in which the patient does not show up). Since each scheduled appointment is attended with probability β_i, the expected number of attended appointments is $\beta_i(1/\mu_i)$. This quantity must equal the required number of attended sessions, $1/\mu_i^b$. Hence,

$$\beta_i \cdot \frac{1}{\mu_i} = \frac{1}{\mu_i^b} \Rightarrow \frac{1}{\mu_i} = \frac{1}{\beta_i \mu_i^b},$$

and therefore $\mu_i := \beta_i \mu_i^b$ is the resulting *effective* service rate. In the remainder of the paper, service completions occur at rate $\mu_i Z_i(t)$, so that no-shows are incorporated directly into the system dynamics through the effective service rate.

This modeling approach is conceptually similar to Huang et al. (2015), who studied the scheduling of in-process (IP) emergency-department patients who occasionally return for follow-up checks. A

key distinction, however, is that in Huang et al. (2015), such returns occur only after a treatment episode, whereas in our setting patients participate in a multi-session treatment course and may miss and reschedule multiple sessions 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_i as the dropout rate of Class i patients, and h_i^d as the cost associated with each such dropout. We denote by $\Gamma_i^d(t)$ the cumulative number of dropouts for patient Class i, up to time $t \geq 0$.

Then, by mass conservation, we have

$$X_i(t) = X_i(0) + A_i(t) - D_i(t) - \Gamma_i(t) - \Gamma_i^d(t), \quad i \in \mathcal{I}, \quad t \geq 0,$$

where $X_i(0)$, $i \in \mathcal{I}$ are all mutually independent finite random variables, and

$$Q_i(t) = X_i(t) - Z_i(t), \quad i \in \mathcal{I}, \ t \geq 0.$$

3.1. Scheduling Policy and Overall Net Benefit

A scheduling policy $\pi \in \Omega$ determines the allocation of servers to patient classes where Ω denotes the set of admissible controls. We consider Markovian policies under which the server allocations are made based on the current state (X, Q) only. In particular, the policy is non-anticipating. 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 $D_i^\pi(t)$, $\Gamma_i^\pi(t)$, and $\Gamma_i^{\pi,d}(t)$. For simplicity, we will omit the subscript π when the context makes the dependence on the policy clear.

The overall net benefit (or cost savings) over $[0, T]$ is, therefore,

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}} \left[b_i D_i(T) - \alpha_i^n \Gamma_i(T) - h_i^d \Gamma_i^d(T)\right] - \int_0^T \sum_{i\in\mathcal{I}} h_i Q_i(t) \mathrm{d}t\right].$$

The Markovian modeling assumption implies that for any $i \in \mathcal{I}$, we have

$$\mathbb{E}\left[D_i(T)\right] = \mu_i \mathbb{E}\left[\int_0^T Z_i(t)\,\mathrm{d}t\right], \ \mathbb{E}\left[\Gamma_i(T)\right] = \theta_i^n \mathbb{E}\left[\int_0^T Q_i(t)\,\mathrm{d}t\right], \ \mathbb{E}\left[\Gamma_i^d(T)\right] = \gamma_i \mathbb{E}\left[\int_0^T Z_i(t)\,\mathrm{d}t\right].$$

Therefore, the overall net benefit can be rewritten as

$$\mathbb{E}\left[\int_0^T \sum_{i\in\mathcal{I}} \left[\left(b_i\mu_i - h_i^d\gamma_i\right) Z_i(t) - \left(h_i + \alpha_i^n\theta_i^n\right) Q_i(t)\right] \mathrm{d}t\right].$$

For notational simplicity, we define the generalized benefit and generalized holding cost as follows:

$$r_i := b_i\mu_i - h_i^d\gamma_i, \quad c_i := h_i + \alpha_i^n\theta_i^n.$$

Incorporating Waitlist Treatment. In our model, AI is operationalized through supervised, AI-based waitlist treatment tools—such as conversational agents or digital monitoring platforms—that provide structured support to patients while they await therapy and thereby influence waiting-time experience, abandonment behavior, and recovery outcomes. Specifically, we denote quantities modified by the presence of waitlist treatment with a superscript w. For example, the holding-cost rate for Class i under waitlist treatment is denoted by h_i^w.

Let $W_i \in \{0,1\}$, $i \in \mathcal{I}$, be a binary decision variable, where $W_i = 1$ indicates that a waitlist treatment is implemented for Class i, and $W_i = 0$ otherwise. We denote the collection of such decisions by $W = (W_1, \ldots, W_I)$.

For classes where waitlist treatment is beneficial and alleviates patient distress, we have $h_i^w \leq h_i$, whereas if the intervention exacerbates patient conditions, the inequality is reversed. Since patients receive waitlist-based care while waiting, both the abandonment rate $\theta_i^{w,n}$ and the associated abandonment penalty $\alpha_i^{w,n}$ may differ from (and are plausibly no greater than) their baseline values θ_i^n and α_i^n; however, we do not impose this inequality in our analysis.

In addition, waitlist treatment can accelerate recovery for a subset of patients, reducing their need for subsequent professional treatment. Such patients may exit the queue voluntarily; unlike conventional abandonments, these departures are beneficial and therefore save cost. We denote the rate of recovery-driven abandonment for Class i by $\theta_i^{w,p}$. The total abandonment rate under waitlist treatment is then

$$\theta_i^w := \theta_i^{w,p} + \theta_i^{w,n}.$$

Providing waitlist treatment also entails operational costs. These include:

1. *Direct costs*, such as per-patient expenses for human supervision during waiting, denoted by a_i for Class i. The resulting generalized holding cost under waitlist treatment for Class i is

$$c_i^w := h_i^w + \alpha_i^{w,n}\theta_i^{w,n} + a_i - b_i\theta_i^{w,p},$$

where the final term captures the benefit associated with recovery-driven abandonments.

2. *Overhead costs* refer to the recurring, class-specific expenditures that arise from implementing AI-based waitlist treatment, independent of the number of patients served. These include ongoing activities such as software maintenance, integration with existing healthcare infrastructure, and supervisory training.

We denote by $f_i(t)$ the cumulative overhead cost associated with Class i up to time t, and define $\bar{f}_i < \infty$ as the long-run average overhead cost for Class i:

$$\bar{f}_i := \lim_{T\to\infty} \frac{1}{T} f_i(T).$$

The overall net benefit under a scheduling policy π and a waitlist treatment policy W over horizon $[0,T]$ is therefore

$$B_T(\pi,W) := \int_0^T \sum_{i\in\mathcal{I}} \Big[r_i Z_i(t) - \big(c_i - W_i(c_i - c_i^w)\big) Q_i(t) \Big]\, \mathrm{d}t - \sum_{i\in\mathcal{I}} f_i(T) W_i.$$

Note that the generalized holding cost is c_i^w when waitlist treatment is utilized ($W_i = 1$), and c_i otherwise.

Our goal is to jointly determine a scheduling policy π and a waitlist treatment policy W that maximize the long-run average expected net benefit:

$$\begin{aligned} \max_{\pi\in\Omega,\, W} \quad & \liminf_{T\to\infty} \frac{1}{T}\, \mathbb{E}[B_T(\pi,W)] \\ \text{s.t.} \quad & Q_i(t) = X_i(t) - Z_i(t) \geq 0, \quad i\in\mathcal{I}, \\ & \sum_{i\in\mathcal{I}} Z_i(t) \leq N, \quad Z_i(t) \geq 0, \quad i\in\mathcal{I}, \\ & W_i \in \{0,1\}, \quad i\in\mathcal{I}. \end{aligned} \tag{1}$$

The resulting optimization problem is a Markov decision process (MDP). Due to the curse of dimensionality (Papadimitriou and Tsitsiklis 1999)—arising from both the large (potentially infinite) state and policy spaces—directly solving for the exact optimal policy is intractable. To gain structural insights, we therefore adopt a deterministic fluid approximation. Fluid models are widely used to capture first-order system dynamics and provide tractable approximations in service operations management (see, e.g., Whitt 2002, Zychlinski 2023).

4. Fluid Model and Solution

To develop an implementable policy and to derive structural insights, we approximate the stochastic system introduced earlier by a deterministic fluid model, in which random fluctuations are replaced by continuous rates. Let $\bar{x}_i$, $\bar{q}_i$, and $\bar{z}_i$, $i\in\mathcal{I}$, denote the steady-state fluid analogues of the expected number of patients in the system, the queue, and the service capacity allocated to Class i, respectively. The waitlist decision variable for Class i in the fluid model is $w_i \in \{0,1\}$.

In steady state, the inflow and outflow of each class must balance. The total departure rate from Class i consists of: (i) service completions at rate $\mu_i \bar{z}_i$, (ii) treatment dropouts at rate $\gamma_i \bar{z}_i$, and (iii) abandonment at rate $\big(\theta_i^n(1-w_i) + \theta_i^w w_i\big)\bar{q}_i$. Hence, for every $i\in\mathcal{I}$,

$$\lambda_i = (\mu_i + \gamma_i)\bar{z}_i + \big[\theta_i^n(1-w_i) + \theta_i^w w_i\big]\bar{q}_i, \qquad \bar{q}_i, \bar{z}_i \geq 0,$$

which is equivalently written as

$$\bar{q}_i = \frac{\lambda_i - (\mu_i+\gamma_i)\bar{z}_i}{\theta_i^n(1-w_i) + \theta_i^w w_i} \geq 0, \qquad \bar{z}_i \geq 0.$$

Since w_i is binary, it is convenient to express $\bar{q}_i$ in the unified form

$$\bar{q}_i = \left(\frac{1-w_i}{\theta_i^n} + \frac{w_i}{\theta_i^w}\right)\left(\lambda_i - (\mu_i+\gamma_i)\bar{z}_i\right), \qquad i \in \mathcal{I}.$$

Indeed, when $w_i = 0$ the denominator is θ_i^n, and when $w_i = 1$ it becomes θ_i^w.

The fluid counterpart of the stochastic control problem in (1) is, therefore, the following optimization problem

$$\begin{aligned}
\max_{\bar{q},\bar{z},w} \quad & \sum_{i\in\mathcal{I}} \left[r_i\bar{z}_i - \left(c_i - w_i(c_i - c_i^w)\right)\bar{q}_i - \bar{f}_i w_i\right] \\
\text{s.t.} \quad & \bar{q}_i = \left(\frac{1-w_i}{\theta_i^n} + \frac{w_i}{\theta_i^w}\right)\left(\lambda_i - (\mu_i+\gamma_i)\bar{z}_i\right), \quad i \in \mathcal{I}, \\
& \sum_{i\in\mathcal{I}} \bar{z}_i \le N, \\
& \bar{q}_i, \bar{z}_i \ge 0, \quad w_i \in \{0,1\}, \quad i \in \mathcal{I}.
\end{aligned} \tag{2}$$

Problem (2) is nonlinear due to bilinear terms involving the products $\bar{q}_i w_i$ in the objective function and $\bar{z}_i w_i$ in the constraints. To obtain a tractable formulation, we reformulate it as a mixed-integer linear program (MILP) using algebraic manipulation and McCormick linearization (McCormick 1976). Specifically, we introduce the auxiliary variable $y_i := w_i\bar{z}_i$ for each $i \in \mathcal{I}$. Because w_i is binary, $y_i = 0$ when $w_i = 0$ and $y_i = \bar{z}_i$ when $w_i = 1$. The resulting MILP, which can be solved using standard solvers, is given in (3).

$$\begin{aligned}
\max_{\bar{z},w,y} \quad & \sum_{i\in\mathcal{I}} \left[\mathcal{P}_i\,\bar{z}_i + \left(\mathcal{P}_i^w - \mathcal{P}_i\right) y_i - \lambda_i\left(\frac{c_i}{\theta_i^n}(1-w_i) + \frac{c_i^w}{\theta_i^w}w_i\right) - \bar{f}_i\,w_i\right] \\
\text{s.t.} \quad & \bar{z}_i \le \frac{\lambda_i}{\mu_i+\gamma_i}, \qquad i \in \mathcal{I}, \\
& \sum_{i\in\mathcal{I}} \bar{z}_i \le N, \\
& y_i \le \frac{\lambda_i}{\mu_i+\gamma_i}\,w_i, \qquad i \in \mathcal{I}, \\
& \bar{z}_i - \frac{\lambda_i}{\mu_i+\gamma_i}(1-w_i) \le y_i \le \bar{z}_i, \qquad i \in \mathcal{I}, \\
& \bar{z}_i, y_i \ge 0, \quad w_i \in \{0,1\}, \qquad i \in \mathcal{I},
\end{aligned} \tag{3}$$

where $\mathcal{P}_i$ coefficients are defined as follows:

$$\mathcal{P}_i := \frac{r_i\theta_i^n + c_i(\mu_i+\gamma_i)}{\theta_i^n}, \qquad \mathcal{P}_i^w := \frac{r_i\theta_i^w + c_i^w(\mu_i+\gamma_i)}{\theta_i^w}, \qquad i \in \mathcal{I}. \tag{4}$$

Proposition 1 establishes the equivalence between the nonlinear formulation in (2) and its MILP reformulation in (3). The proof is provided in Appendix B.

PROPOSITION 1. *Problems (2) and (3) are equivalent: they share the same feasible objective values and therefore attain the same optimal value.*

4.1. The Dynamic $\mathcal{P}$-Index Scheduling Policy

Once the waitlist configuration w^* is determined, the fixed costs $\bar{f}_i w_i^*$ become sunk and can therefore be omitted from the remaining optimization. Substituting w^* into Problem (3) reduces the problem to the following linear program (LP):

$$\begin{aligned} \max_{\bar{z}} \quad & \sum_{i\in\mathcal{I}} \left(\mathcal{P}_i(1-w_i^*) + \mathcal{P}_i^w w_i^*\right) \bar{z}_i \\ \text{s.t.} \quad & 0 \le \bar{z}_i \le \frac{\lambda_i}{\mu_i+\gamma_i}, \qquad i\in\mathcal{I}, \\ & \sum_{i\in\mathcal{I}} \bar{z}_i \le N. \end{aligned} \tag{5}$$

This LP admits a simple and interpretable optimal solution: classes are prioritized according to their waitlist-adjusted index

$$\bar{\mathcal{P}}_i := \mathcal{P}_i(1-w_i^*) + \mathcal{P}_i^w w_i^*, \qquad i\in\mathcal{I},$$

where the $\mathcal{P}$-indices are defined in (4). Since the objective is linear in $\bar{z}_i$ and the allocation to each class is subject to an upper bound $\lambda_i/(\mu_i+\gamma_i)$, the optimal allocation fills capacity starting with the class having the largest index $\bar{\mathcal{P}}_i$, then the next largest, and so on, until the capacity constraint becomes binding or each class has been allocated its maximum feasible capacity.

Formally, after reordering the class indices such that $\bar{\mathcal{P}}_{(1)} \ge \bar{\mathcal{P}}_{(2)} \ge \cdots \ge \bar{\mathcal{P}}_{(I)}$, the optimal fluid allocation satisfies

$$\bar{z}^* = \left(\frac{\lambda_{(1)}}{\mu_{(1)}+\gamma_{(1)}}, \ldots, \frac{\lambda_{(i_0-1)}}{\mu_{(i_0-1)}+\gamma_{(i_0-1)}}, N - \sum_{j=1}^{i_0-1} \frac{\lambda_{(j)}}{\mu_{(j)}+\gamma_{(j)}}, 0, \ldots, 0 \right),$$

where

$$i_0 = \max\left\{ i \in \{1,\ldots,I+1\} : \sum_{j=1}^{i-1} \frac{\lambda_{(j)}}{\mu_{(j)}+\gamma_{(j)}} < N \right\}.$$

The LP in (5) also induces a *dynamic* scheduling policy for the stochastic system. Specifically, whenever a therapist becomes available, the system admits a customer from the class with the highest *current* waitlist-adjusted index $\bar{\mathcal{P}}_i$ among those with customers waiting. If the highest-index class has no customers waiting, the server selects the class with the next highest index, and so on. In Section 5, we evaluate the performance of this policy using a stochastic simulation model.

4.2. More Therapists, AI, or Both?

Excessive wait times for mental health services have sparked debate about the most effective way to expand access (Stringer 2024). In this section, we extend the analysis to include the decision of whether, and how many, additional therapists should be recruited, alongside the choice of waitlist and scheduling policies. Human therapists deliver personalized care but come with substantial ongoing

costs, while supervised AI-based waitlist treatments can be deployed at low marginal cost and scaled broadly, though their effectiveness may vary across patients. Our objective is to quantify these trade-offs and determine the optimal combination of waitlist, scheduling and recruitment decisions.

To this end, we introduce an additional decision variable $K \geq 0$, which increases the number of full-time-equivalent therapists from N to $N+K$. We consider a hiring cost c_k per new therapist. Problem (3) then becomes the following MILP:

$$
\begin{aligned}
\max_{\bar{z},w,K} \quad & \sum_{i\in\mathcal{I}} \left[\left(\mathcal{P}_i\left(\bar{z}_i - y_i\right) + \mathcal{P}_i^w y_i\right) - \lambda_i \left(\frac{c_i}{\theta_i^n}\left(1-w_i\right) + \frac{c_i^w}{\theta_i^w} w_i \right) \right] - \sum_{i\in\mathcal{I}} \bar{f}_i w_i - c_k K \\
\text{s.t.} \quad & 0 \leq \bar{z}_i \leq \frac{\lambda_i}{\mu_i+\gamma_i}, \quad i \in \mathcal{I}; \\
& \sum_{i\in\mathcal{I}} \bar{z}_i \leq N+K; \\
& y_i \leq \frac{\lambda_i}{\mu_i+\gamma_i} w_i, \quad i \in \mathcal{I}; \\
& \bar{z}_i - \frac{\lambda_i}{\mu_i+\gamma_i}(1-w_i) \leq y_i \leq \bar{z}_i, \quad i \in \mathcal{I}; \\
& \bar{z}_i, y_i \geq 0,\ w_i \in \{0,1\},\ i \in \mathcal{I},\ K \geq 0.
\end{aligned}
$$

Note that for any fixed parameter set, the optimization is solved *once*; after w and K are set, the induced scheduling policy is the exact index-based policy introduced in (5), only with $N+K$ therapists. In Sections 5, we numerically examine the effect of hiring costs on the optimal waitlist, scheduling and recruitment decisions.

5. Numerical Experiments

This section evaluates the performance of our proposed policy under various scenarios. We begin by comparing its performance with the exact MDP solution. Since solving the MDP becomes prohibitively expensive as the system grows, we then use stochastic simulation to evaluate the policy for larger systems, including comparisons against benchmark policies and an examination of different service-time distributions. Finally, we examine how hiring costs affect the optimal waitlist, scheduling and recruitment decisions.Throughout this section, we focus on two patient classes whose parameter values are chosen to be representative of mental health service settings and are aligned with the calibration used in the VHA case study presented in Section 6 (see Appendix C for the parameter values used).

We first compare the fluid-model solution (2) with the exact MDP solution (1); See Appendix A for details on the MDP formulation and its numerical solution via relative value iteration. Table 1 summarizes the results for a two-class system with $N = 2, 3, 4$. To ensure consistent scaling across system sizes, we proportionally increase both the arrival rates and the number of servers, while preserving the relative class proportions calibrated in the VHA case study presented in Section 6.

To assess robustness to parameter uncertainty, each scenario is evaluated over 100 independent perturbations. In each perturbation, model parameters are randomly varied within ±10% of their baseline values. For each system size, the table reports the mean performance gap between the MDP and fluid solutions, its corresponding 95% confidence interval (CI), and the maximal observed gap.

The results demonstrate an excellent fit between the fluid approximation and the exact MDP solution. The performance gap decreases as system size increases, consistent with the asymptotic accuracy of the fluid approximation. Moreover, the computational burden of solving the MDP increases rapidly with system size. For example, each perturbation with $N = 4$ requires approximately eight minutes of computation time, whereas the fluid solution is obtained within seconds for all system sizes. As the system grows larger, this disparity in computational tractability becomes even more pronounced, underscoring the practical advantage of the fluid-based approach.

Table 1 Benefit comparison between the fluid-model solution and the exact MDP solution.

N	**Average Gap (%)**	**95% CI (%)**	**Maximum Gap (%)**
2	1.62%	1.26%-2.54%	2.70%
3	0.97%	0.64%-1.50%	1.60%
4	0.85%	0.41%-1.46%	1.49%

For larger systems, where solving the MDP becomes computationally prohibitive, we evaluate the policy using a stochastic simulation model. We first compare the fluid-model predictions with simulated performance across different system sizes and service-time distributions.

To implement the policy in the simulation model, we first compute the optimal waitlist decisions w_i for all $i \in \mathcal{I}$ and the corresponding priority indices $\bar{\mathcal{P}}_i$. The policy is then implemented *dynamically*: whenever a server becomes available, a patient from the class with the highest priority index among those waiting is admitted to service.

We consider a two-class system consisting of Class A and Class B patients, and evaluate system performance across different system sizes. To ensure consistent scaling across system sizes, we proportionally increase both the arrival rates and the number of servers. Solving the MILP formulation in (3) yields the optimal waitlist decisions $w_A^* = 0$ and $w_B^* = 1$, indicating that Class B patients are assigned waitlist treatment whereas Class A patients are not. Furthermore, the resulting priority indices satisfy $\bar{\mathcal{P}}_A > \bar{\mathcal{P}}_B$, implying that Class A patients receive higher priority under the dynamic index-based scheduling policy.

Figure 2 presents the queue lengths (left panel) and server allocations (right panel) obtained from both the fluid approximation (solid lines) and stochastic simulation (dashed lines) across different system scales and under two commonly used service-time distributions: exponential and lognormal. Queue lengths and server allocations closely match the fluid model even in relatively small systems

with $N = 5$. These results—highly accurate even under lognormal service times—demonstrate the effectiveness and robustness of the suggested policy.

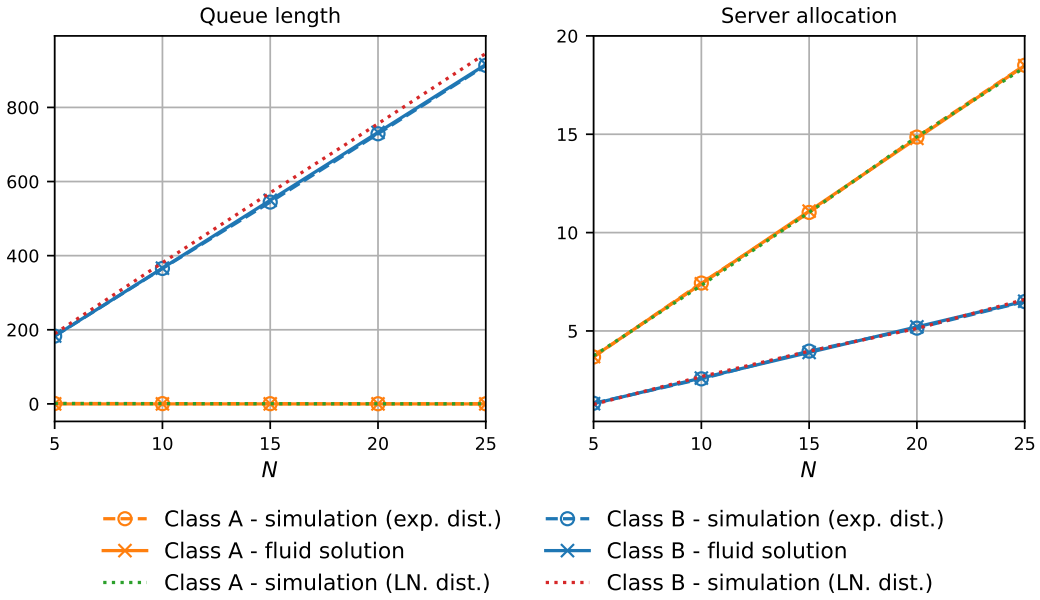

Figure 2 **Comparison of fluid model solutions and simulation results across different system scales under two service-time distributions: exponential (exp. dist.) and lognormal (LN. dist.).**

Next, we compare the performance of our suggested policy to two other benchmarks: (i) a *no-waitlist* policy, in which no class receives waitlist treatment and scheduling follows an *adjusted* $c\mu/\theta$ rule, where the $c\mu/\theta$ index from Atar et al. (2010) is augmented to incorporate treatment-completion benefits, dropouts, and no-shows; and (ii) a *uniform-waitlist* policy, in which all classes receive waitlist treatment and scheduling follows the corresponding adjusted $c^w\mu/\theta^w$ rule, reflecting the modified costs and abandonment structure under waitlist treatment. Figure 3 presents this comparison across different system scales. Under both exponential and lognormal service-time distributions, the proposed policy outperforms both benchmark policies at all system sizes. Importantly, although waitlist treatment can reduce waiting times and mitigate negative abandonment, these benefits do not necessarily outweigh the additional costs associated with implementing waitlist treatment for both classes.

Lastly, we examine how hiring costs affect system performance. Figure 4 illustrates how the optimal server allocation, waitlist-treatment decisions, and the number of newly recruited therapists K vary as a function of the hiring cost c_k. In this experiment, we set $N = 10$, so that the total service capacity equals $N + K = 10 + K$. The arrival rates are scaled proportionally to preserve the same relative demand across classes.

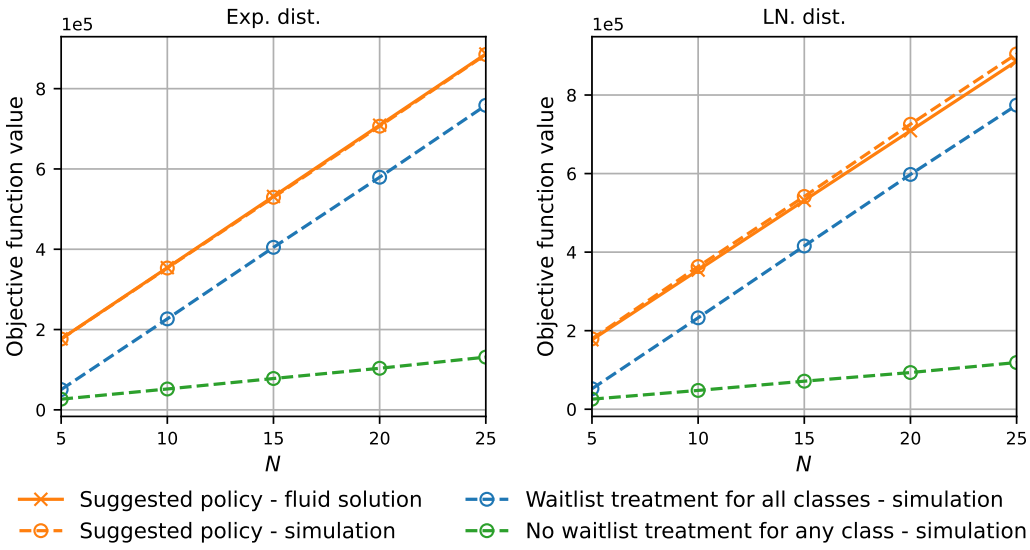

Figure 3 **Policy comparison under different system scales and service-time distributions.**

When hiring costs are low ($c_k < 1{,}010$), expanding therapist capacity is cost-effective, and the system recruits $K = 13$ additional therapists. In this regime, no class receives waitlist treatment, and Class B is prioritized, receiving most of the service capacity.

As hiring costs increase ($1{,}010 < c_k < 1{,}971$), the optimal number of new hires decreases to $K = 1$. In this range, only Class B receives waitlist treatment, which lowers its effective priority index and reverses the service allocation pattern: Class A is now allocated most of the service capacity, while Class B relies more heavily on waitlist support.

When hiring costs become sufficiently high ($c_k > 1{,}971$), it becomes more cost-effective to rely entirely on waitlist treatment rather than recruit additional therapists ($K = 0$). In this regime, waitlist treatment is implemented for all classes, and Class A receives the full service capacity.

Overall, Figure 4 demonstrates that therapist recruitment decisions can substantially reshape the optimal waitlist and scheduling policy.

6. Case Study: Representative VHA Outpatient Mental Health Unit

To evaluate the performance of the proposed wailist and scheduling policy, we calibrate a case study to a representative VHA outpatient mental health unit. The VHA is the largest integrated healthcare provider in the United States (U.S. Department of Veterans Affairs 2024c). It delivers mental health services through a nationwide network of medical centers, outpatient mental health units, and community-based outpatient clinics, collectively providing care to more than 1.5 million veterans with mental health conditions each year (U.S. Department of Veterans Affairs 2023b). Over the past two decades, the system has faced persistent challenges, including capacity shortages, excessive

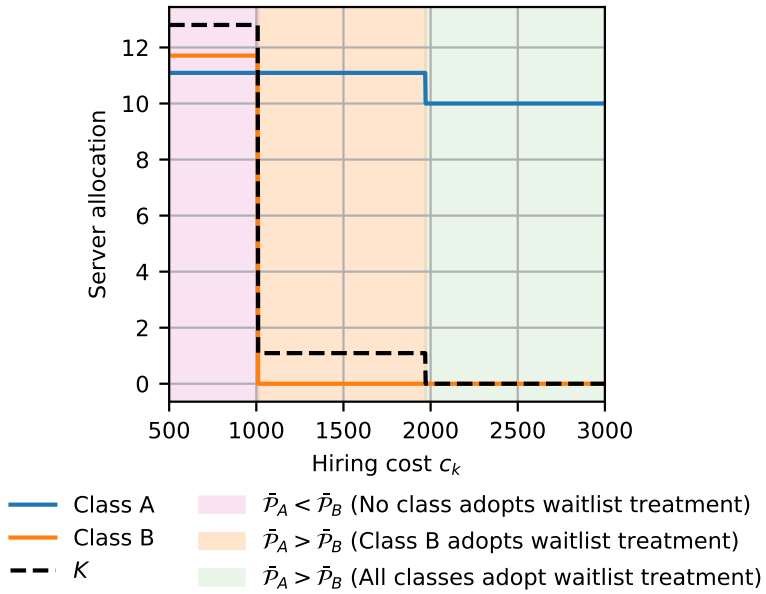

Figure 4 **Optimal server allocation and waitlist-treatment decisions as a function of hiring costs.**

wait times, and difficulty matching staffing levels to rising demand (U.S. Government Accountability Office 2021, Smith et al. 2023). These pressures culminated in the well-documented access-to-care crisis of 2014, which led to major reforms aimed at improving transparency, modernizing scheduling systems, and expanding the mental health workforce. Subsequent evaluations report that outpatient mental health demand has increased by nearly 90% since 2006, while mental health staffing within the VHA has grown by a factor of approximately 2.6 over the same period (U.S. Government Accountability Office 2021). Even with these expansions, many VHA facilities continue to struggle with balancing demand and capacity, making waitlist management an ongoing challenge.

The diagnostic classes considered in this case study reflect the conditions most commonly treated within VHA mental health services. Recent data indicate that nearly *one third* of VHA users have a confirmed mental health diagnosis, with MDD, AD, and PTSD accounting for substantial shares of this population (U.S. Department of Veterans Affairs 2023a). These conditions differ in prevalence, treatment duration, and clinical urgency. Accordingly, we focus on the class set

$$\mathcal{I} = \{\text{MDD, AD, PTSD}\},$$

which together represent the majority of VHA outpatient mental health demand and capture distinct patient trajectories relevant to our analysis.

Before turning to the detailed parameter calibration, we note that sensitivity checks in which the calibrated parameters were varied by $\pm 5\%$ showed that the resulting waitlist and scheduling decisions remained unchanged, indicating the robustness of the policy to modest perturbations in parameter estimates.

6.1. Parameter estimation

System size. We model a medium-sized outpatient mental health clinic with $N = 50$ FTE therapists. This choice aligns with VHA staffing patterns: mental health staffing increased from fewer than 7,000 to more than 18,000 FTE therapists between 2006 and 2021 (Smith et al. 2023), and has likely continued to grow given persistent demand and ongoing national hiring initiatives. The VHA operates over 1,100 community-based outpatient clinics (U.S. Department of Veterans Affairs 2024c), and assuming that not all provide mental health services yields an average staffing level consistent with our system-size estimate.

Arrival rates. Public VHA data do not report diagnosis-specific arrival rates at the clinic level. System-wide reports, however, indicate that approximately 31% of VHA users have at least one mental health diagnosis (U.S. Department of Veterans Affairs 2023a). To estimate diagnosis-specific arrival rates for a single outpatient unit, we use the following approximation:

$$\begin{aligned}\text{Annual arrival rate} \approx\ & \text{Prevalence among VHA users} \\ & \times \text{Population served} \\ & \times \text{Proportion seeking care.}\end{aligned}$$

Here, *prevalence among VHA users* denotes the fraction of veterans with a given diagnosis. Based on recent VHA reports, prevalence is estimated at 10% for MDD, 7% for AD, and 14% for PTSD (U.S. Department of Veterans Affairs 2023a). *Population served* refers to the number of enrolled veterans in the medical center's catchment area who may initiate care. We consider a representative VHA outpatient unit serving 50,000 veterans annually, which is consistent with reported facility volumes; for example, the VA Salt Lake City Health Care System serves more than 70,000 enrolled veterans per year (U.S. Department of Veterans Affairs 2024b), and the Harry S. Truman Memorial Veterans' Hospital, which serves more than 40,000 veterans annually (U.S. Department of Veterans Affairs 2024a). *Proportion seeking care* represents the fraction of diagnosed veterans who engage in outpatient mental health treatment. VHA data indicate that approximately 84% of veterans with a confirmed mental health diagnosis receive outpatient care each year (U.S. Department of Veterans Affairs 2023a).

Combining these components yields approximately 4,200, 2,940, and 5,880 new annual treatment episodes for MDD, AD, and PTSD, respectively. Converting to weekly rates gives $\lambda_{\text{MDD}} = 81$, $\lambda_{\text{AD}} = 57$, $\lambda_{\text{PTSD}} = 113$.

Service and no-show rates. Service consists of a course of psychotherapy sessions, typically scheduled once a week. Because the number of sessions is determined at intake and all sequential appointments are allocated to the same therapist, we model the entire course as a *single* service. The

baseline weekly service rate per full-time therapist equals the therapist's weekly session capacity (40) divided by the typical number of sessions in a course. For MDD, a standard course includes roughly 12–15 sessions (Alavi et al. 2023); using the midpoint (13.5) gives $\mu^b_{\text{MDD}} = 40/13.5 = 2.96$. We use the same baseline for AD (NICE 2020), i.e., $\mu^b_{\text{AD}} = 2.96$. For PTSD, prolonged exposure generally involves 8–16 sessions (Lee et al. 2022); using the midpoint yields $\mu^b_{\text{PTSD}} = 40/12 = 3.33$.

Attendance variability further increases the effective service time: missed appointments must be rescheduled and extend the total number of attempts. Following our no-show modeling in Section 3, the effective service rate is $\mu_i = \beta_i \mu^b_i$, where β_i denotes the show-up probability. Using an empirically observed no-show rate of 18% in mental health clinics (Milicevic et al. 2020), we set $\beta_i = 0.82$ and thus $\mu_i = 0.82\,\mu^b_i$ for all $i \in \mathcal{I}$.

Dropout and abandonment rates. The premature dropout rate during treatment (i.e., leaving therapy before completing the planned sequence of sessions) varies across settings and treatment modalities. Meta-analytic evidence indicates that approximately 16–20% of adult psychotherapy patients discontinue treatment prematurely (Swift and Greenberg 2012, Lewis et al. 2020). To calibrate this parameter in our model, we take a representative dropout probability of 16% over a typical six-month course of care and assume that time to dropout follows an exponential distribution with rate γ. This yields $0.16 = 1 - e^{-\gamma T}$, where $T = 0.5$ years. Solving, $\gamma = -2\ln(0.84) \approx 0.348$ per year, which corresponds to a weekly dropout rate of approximately $\gamma/52 \approx 0.0067$.

To calibrate the *baseline abandonment rate* θ^n_i, we interpret abandonment from the waitlist as failure to attend thue first scheduled clinical appointment. Operational evidence indicates that this form of disengagement is substantial: Williams et al. (2008) report that 39–50% of patients at a large community mental health center did not attend their initially scheduled psychiatric evaluation. Additional evidence from primary mental healthcare settings documents similarly high rates of disengagement during the waiting period (Xaba et al. 2024). Interpreting these proportions as exponential hazards over typical waiting intervals of six to twelve weeks yields weekly rates in the range of 0.04–0.12. As a representative conservative value within this range, we set $\theta^n_i = 0.04$.

To calibrate the *positive abandonment rate* $\theta^{w,p}_i$ at which patients meaningfully improve under waitlist treatment and no longer require psychotherapy, we draw on recent empirical evidence from AI-enabled mental health interventions. Fitzpatrick et al. (2017) report that users of the fully automated Woebot conversational agent experienced significant reductions in depressive symptoms over a two-week engagement period. Consistent with this, Inkster et al. (2023) show that users of the Wysa conversational agent exhibit significant improvements in depression and anxiety within the first weeks of use, with effects sustained over longer periods. A broader review of AI-based mental health chatbots (Boucher et al. 2021) similarly concludes that such tools consistently yield clinically

meaningful symptom reductions, often emerging within two to four weeks. Because the literature does not yet provide precise early-response rates across heterogeneous adult populations, we adopt a conservative modeling assumption and posit that approximately 10–15% of patients experience meaningful improvement over a two to three week period. Converting these proportions to exponential hazards over this period yields weekly rates in the range of 0.035–0.08. As a representative conservative value within this range, we set $\theta_i^{w,p} = 0.06$.

Finally, we estimate the *negative abandonment rate* $\theta_i^{w,n}$, which captures the small but nonzero probability that a patient disengages from the system even while receiving waitlist treatment, we draw on evidence documenting high engagement with conversational AI mental health tools. Studies of automated agents such as Woebot and Wysa report sustained multi-week use with very limited early disengagement among active users (Fitzpatrick et al. 2017, Inkster et al. 2023). Because the literature does not yet provide precise quantitative estimates for disengagement during supported waitlist periods, we adopt a conservative assumption that a small fraction of patients nonetheless abandon the system over typical waiting intervals. This corresponds to weekly exponential hazards in the range of 0.0005–0.002. We therefore set $\theta_i^{w,n} = 0.0005$, reflecting the low but nonzero likelihood of disengagement under supervised waitlist care.

Treatment Benefits. We value health benefits b_i using quality-adjusted life years (QALYs). Incremental QALYs relative to no treatment are 0.12 (MDD), 0.09 (AD), and 0.13 (PTSD) (Ontario et al. 2017, Mavranezouli et al. 2020). Using \$100,000 per QALY (Managed Healthcare Executive 2023) yields $b_{\text{MDD}} = \$12{,}000$, $b_{\text{AD}} = \$9{,}000$, and $b_{\text{PTSD}} = \$13{,}000$.

Holding and supervision costs. Weekly holding costs h_i represent the per-patient societal burden of remaining untreated, including functional impairment, productivity losses, and excess healthcare utilization. Using annual societal cost estimates (Greenberg et al. 2023) and dividing by 52 yields $h_{\text{MDD}} = \$324$, $h_{\text{AD}} = \$196$, and $h_{\text{PTSD}} = \$378$. Meta-analytic evidence indicates that low-intensity and digital psychological interventions can produce clinically meaningful symptom improvement (Cuijpers et al. 2021). Because a substantial share of the societal burden in h_i is driven by productivity losses and functional impairment—factors shown to improve appreciably with even partial reductions in symptom severity (Greenberg et al. 2023, Kavelaars et al. 2023, Davis et al. 2022)—we model the weekly burden under waitlist treatment as significantly reduced but not eliminated. To capture these differential burden–severity relationships across conditions, we apply class-specific reductions to 10% of the basic holding costs: $h^w_{\text{MDD}} = 32.4$, $h^w_{\text{AD}} = 19.6$, and $h^w_{\text{PTSD}} = 37.8$.

The parameter a_i represents the variable cost of clinician supervision required to support AI-based waitlist treatment. In guided digital interventions, therapist involvement is typically limited to brief, asynchronous feedback or monitoring. For example, Hedman et al. (2011) report that therapists

providing internet-based cognitive behavioral therapy spend on average only 5.5 minutes per week per patient, highlighting the minimal time demands of guided digital care. Community-based support models likewise demonstrate that effective oversight can be delivered with low-intensity, intermittent supervision Lewin et al. (2005). To translate these patterns into a conservative weekly supervision cost, we assume that AI-supported waitlist treatment requires approximately 1–5 minutes of clinician time per patient per week. Using a conservative fully loaded clinician wage of \$60–\$80 per hour (U.S. Bureau of Labor Statistics 2023b), the implied weekly marginal supervision cost is at most \$1–\$5 and often lower when supervision is batched across patients. To reflect a small but non-negligible variable cost, we set $a_i = 1$ for all $i \in \mathcal{I}$.

Abandonment and dropout costs. The parameter α_i^n represents the discrete harm associated with abandoning the system before receiving treatment. This harm reflects short-run consequences of disengagement—such as brief periods of unmanaged deterioration, disruption of care pathways, and delays associated with re-entering the system—which have been documented in studies of missed psychiatric appointments and early-care disengagement (Xaba et al. 2024, Forneris et al. 2013). These harms are distinct from the long-run health benefits b_i forgone when treatment is never initiated. Because these effects are short-lived relative to the full duration of treatment benefit, we model them as a small fraction of b_i. Accordingly, we set $\alpha_i^n = 0.02\, b_i$, yielding $\alpha_{\text{MDD}}^n = 240$, $\alpha_{\text{AD}}^n = 180$, and $\alpha_{\text{PTSD}}^n = 260$.

Dropout during treatment and negative abandonment under waitlist treatment are less harmful, as patients in these scenarios typically receive partial benefit before disengaging or remain monitored and supported. To reflect this mitigation while keeping penalties small relative to b_i, we set $\alpha_i^{w,n} = h_i^d = 0.01\, b_i$, producing $\alpha_{\text{MDD}}^{w,n} = h_{\text{MDD}}^d = 120$, $\alpha_{\text{AD}}^{w,n} = h_{\text{AD}}^d = 90$, and $\alpha_{\text{PTSD}}^{w,n} = h_{\text{PTSD}}^d = 130$.

Long-run average overhead cost. The fixed-cost parameter $\bar{f}_i$ represents the long-run average weekly overhead associated with implementing and maintaining the AI-based waitlist intervention for Class i, rather than a one-time expenditure. Economic evaluations of digital health interventions consistently show that such systems incur substantial fixed costs for development, iterative content updates, implementation, and technical maintenance (Hedman et al. 2011, Gomes et al. 2022, Buntrock 2024). In line with these peer-reviewed findings, industry assessments estimate that developing and deploying a secure, compliant digital mental health or AI-enabled application typically requires an initial investment between \$50,000 and \$500,000, with first-year total costs—including hosting, compliance, monitoring, and maintenance—often reaching \$250,000–\$1,200,000 (PerfectionGeeks Technologies 2024). Amortizing a representative system-level investment of \$500,000 over a five-year operational horizon yields a weekly overhead of

$$\frac{500{,}000/5}{52} \approx 1{,}923 \text{ USD/week}.$$

Allocating this amount evenly across the three diagnostic classes considered in our model gives approximately \$640 per class per week; accordingly, we set $\bar{f}_i = 640$.

6.2. Results and Insights

Under the case-study parameter estimates, the optimal policy assigns waitlist treatment to patients with MDD and AD. This reduces their effective holding costs, which in turn leads to an optimal prioritization order in which PTSD patients are served first, followed by MDD and then AD. Quantitatively, adopting this policy yields an approximate 210% increase in long-run net benefit relative to a system without waitlist treatment. This sizable gain underscores the value of targeted, diagnosis-specific waitlist interventions in resource-constrained mental health settings.

We next examine the effect of system size and analyze outcomes when therapist recruitment decisions are allowed in addition to waitlist-treatment and scheduling decisions. Figure 5 illustrates the optimal waitlist-treatment policy for each diagnostic class as a function of the number of therapists, N. When resources are scarce ($N \leq 40$), all classes receive waitlist treatment, as waiting times are excessive across the board. For a moderate number of therapists ($40 < N \leq 70$), only AD and MDD patients are assigned to waitlist treatment, since there are sufficient therapists to prioritize and directly serve PTSD patients. This range includes the baseline case with $N = 50$ discussed earlier. When therapist availability increases further ($70 < N \leq 95$), only AD patients continue to receive waitlist treatment. Finally, when $N > 95$, the number of therapists is sufficient to meet demand, and waitlist treatment is no longer required for any class.

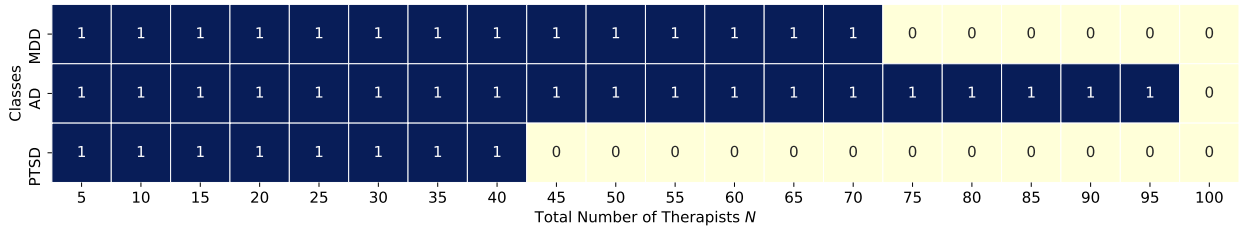

Figure 5 **Optimal waitlist-treatment policy for each diagnostic class as a function of the number of therapists. A value of 1 indicates that waitlist treatment is applied to the corresponding class, while a value of 0 indicates that it is not.**

Results with therapist recruitment. Next, we evaluate the solution developed in Section 4.2, which incorporates therapist recruitment in addition to optimal waitlist treatment and scheduling. According to U.S. Bureau of Labor Statistics (2023a), the national mean annual wage for mental health therapists is approximately \$60,000–\$65,000, corresponding to a weekly wage of about \$1,150. After accounting for payroll taxes, benefits, training, supervision, licensing, and administrative costs,

we conservatively estimate the effective weekly cost of an additional full-time-equivalent therapist to be \$2,000; accordingly, we set $c_k = 2{,}000$.

Under the case-study parameter values, we find that providing waitlist treatment eliminates the need to recruit additional therapists. In contrast, when waitlist treatment is not available, achieving the same level of benefit would require the recruitment of an *additional* 48 therapists—nearly doubling the current staffing level.

Figure 6 maps the optimal waitlist and recruitment policies as functions of the fixed cost of waitlist treatment and the recruitment cost, for different numbers of therapists. Beyond illustrating the optimal policy regions, Figure 6 reveals several important insights. Waitlist treatment and therapist recruitment are not simple substitutes: in many regimes the optimal policy employs both, indicating complementarity between the two. The role of waitlist treatment also evolves with system size, shifting from broad deployment under severe therapist shortages to diagnosis-specific targeting as staffing increases. Finally, recruitment costs shape not only whether additional therapists are hired but also how and for which classes waitlist treatment is deployed, underscoring the importance of coordinating waitlist treatment with workforce planning.

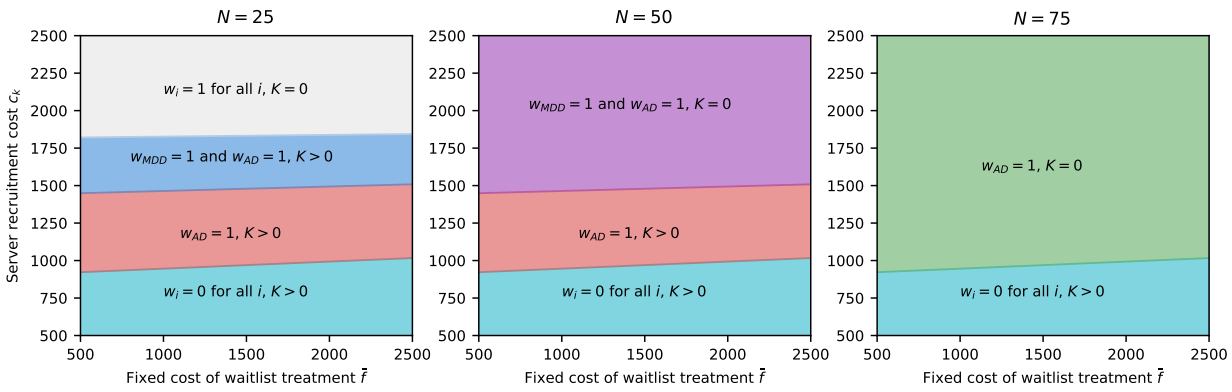

Figure 6 **Optimal waitlist-treatment and recruitment policy as a function of the fixed cost of waitlist treatment and the therapist recruitment cost.**

7. Conclusions and Future Directions

Amid rising demand for mental health services and persistent therapist shortages, supervised AI-supported waitlist treatment offers a scalable means to support patients while alleviating system pressure. We develop an analytical framework that integrates waitlist treatment, scheduling, and therapist recruitment within a multi-class, multi-server queueing model, capturing both implementation costs and key behavioral features of psychotherapy, including no-shows, dropouts, heterogeneous treatment effects, and recovery-based abandonment.

Using a fluid approximation, we reformulate the control problem as a tractable MILP whose solution yields both the optimal waitlist-treatment configuration and a simple, index-based scheduling rule. Numerical experiments show that the resulting policy closely matches the exact MDP solution in small systems and remains robust across system scales and service-time distributions. They also reveal an interaction between waitlist treatment and therapist recruitment. A case study calibrated to a representative VHA mental health unit demonstrates that supervised waitlist treatment can substantially improve system performance while reducing the need to recruit additional therapists, underscoring the value of analytically guided AI support in workforce-constrained mental health systems.

Future research may extend this framework to time-varying and transient settings, such as those arising after mass-trauma events or natural disasters, where demand surges and resource availability change rapidly. Understanding how waitlist treatment should adapt in these environments may yield insights for designing interventions that are effective in the short run and resilient over time.

Another important direction is to incorporate AI-based support for patients with substance use disorders following rehabilitation, a critical period characterized by high relapse risk and the need for continuous support. Given the scale of the crisis—approximately 600,000 drug-related deaths globally each year, with nearly 80% involving opioids (World Health Organization 2024)—AI-based support has the potential to generate substantial societal benefits. In this context, operational and dynamic models can play a key role in designing and optimizing policies under behavioral responses such as moral hazard (Cai and Zychlinski 2025), which may arise if patients rely excessively on AI-based support, potentially reducing engagement with treatment or increasing the risk of relapse.

Acknowledgments

The authors sincerely thank Editor Natarajan Gautam, the anonymous Associate Editor and two reviewers for their insightful and constructive feedback, which has greatly strengthened this work. Partial financial support was provided by the Israel Science Foundation [Grant 277/21] and the Bernard M. Gordon Center for Systems Engineering at the Technion.

Junyang Cai is a PhD student in the Faculty of Data and Decision Sciences at the Technion. He earned his Master of Management degree from Shanghai University in 2024. His research interests focus on operations research and operations management, with particular emphasis on applications in the healthcare sector.

Noa Fani Likvornik is an MSc student at the Faculty of Data and Decision Sciences at the Technion. She completed her bachelor's degree with honors at the Technion. Her interests focus on project management and the continuous improvement of operational processes and supply chains.

Noa Zychlinski is an Assistant Professor in the Faculty of Data and Decision Science at the Technion – Israel Institute of Technology. Noa completed her postdoctoral fellowship in the Division of Decision, Risk, and Operations at Columbia Business School. Her research interests focus on service and healthcare operations management, the analysis of queueing networks and their applications, the theory of stochastic process approximation, and data analysis of large service systems.

References

Alavi, N., E. Moghimi, C. Stephenson, G. Gutierrez, J. Jagayat, A. Kumar, Y. Shao, S. Miller, C.S. Yee, A. Stefatos. 2023. Comparison of online and in-person cognitive behavioral therapy in individuals diagnosed with major depressive disorder: a non-randomized controlled trial. *Frontiers in Psychiatry*, 14 1113956.

Arntzen, R.J., R. Bekker, R.D. van der Mei. 2024. Knapsack-based routing for mental health placements: A split-horizon approach. *working paper*, .

Atar, R., C. Giat, N. Shimkin. 2010. The cμ/θ rule for many-server queues with abandonment. *Operations Research*, 58 (5), 1427-1439.

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

Australian Institute of Health and Welfare. 2025. Australia's mental health system. Available at: https://www.aihw.gov.au/mental-health/overview/australias-mental-health-system.

Bennett-Levy, J., D. Richards, P. Farrand, H. Christensen, K. Griffiths, D. Kavanagh, B. Klein, M.A. Lau, J. Proudfoot, L. Ritterband. 2010. *Oxford guide to low intensity CBT interventions*. OUP Oxford.

Boucher, E.M., N.R. Harake, H.E Ward, S.E. Stoeckl, J. Vargas, J. Minkel, A.C. Parks, R. Zilca. 2021. Artificially intelligent chatbots in digital mental health interventions: A review. *Expert Review of Medical Devices*, 18 (sup1), 37-49.

Briot, K., A. Pizano, M. Bouvard, A. Amestoy. 2021. New technologies as promising tools for assessing facial emotion expressions impairments in ASD: a systematic review. *Frontiers in psychiatry*, 12 634756.

Buntrock, C. 2024. Cost-effectiveness of digital interventions for mental health: Current evidence, common misconceptions, and future directions. *Frontiers in Digital Health*, 6 1486728.

Cai, J., N. Zychlinski. 2025. Expanding naloxone accessibility: A lifesaver or a risky setback? *Production and Operations Management*, forthcoming.

Cai, J., N. Zychlinski. 2026. When AI is not enough: Reducing diagnostic errors with radiologist oversight. *Service Science, forthcoming*, .

Cox, D.R., W.L. Smith. 1961. Queues. *Methuen, London*, .

Cuijpers, P., E. Karyotaki, M. Ciharova, C. Miguel, H. Noma, T.A. Furukawa. 2021. The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: A meta-analysis. *Acta Psychiatrica Scandinavica*, 144 (3), 288-299.

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

Department of Health & Social Care. 2023. Progress in improving mental health services in England. Available at: https://www.nao.org.uk/wp-content/uploads/2023/02/Progress-in-improving-mental-health-services-CS.pdf.

D'Alfonso, S. 2020. AI in mental health. *Current opinion in psychology*, 36 112-117.

Fitzpatrick, K. Kara, A. Darcy, M. Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR mental health*, 4 (2), e7785.

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

GBD 2019 Mental Disorders Collaborators. 2022. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019. *The Lancet Psychiatry*, 9 (2), 137-150.

Gomes, M., E. Murray, J. Raftery. 2022. Economic evaluation of digital health interventions: Methodological issues and recommendations for practice. *Pharmacoeconomics*, 40 (4), 367-378.

Gomes, N., M. Pato, A. Ribeiro Lourenco, N. Datia. 2023. A survey on wearable sensors for mental health monitoring. *Sensors*, 23 (3), 1330.

Graham, S., C. Depp, E.E. Lee, C. Nebeker, X. Tu, H.-C. Kim, D.V. Jeste. 2019. Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports*, 21 1-18.

Greenberg, P., A. Chitnis, D. Louie, E. Suthoff, S.-Y. Chen, J. Maitland, P. Gagnon-Sanschagrin, A.-A. Fournier, R.C. Kessler. 2023. The economic burden of adults with major depressive disorder in the United States (2019). *Advances in Therapy*, 40 (10), 4460-4479.

Harrison, J.M., A. Zeevi. 2004. Dynamic scheduling of a multiclass queue in the halfin-whitt heavy traffic regime. *Operations Research*, 52 (2), 243-257.

Hedman, E., E. Andersson, B. Ljótsson, G. Andersson, C. Rück, N. Lindefors. 2011. Cost-effectiveness of internet-based cognitive behavior therapy vs. cognitive behavioral group therapy for social anxiety disorder: Results from a randomized controlled trial. *Behaviour Research and Therapy*, 49 (11), 729-736.

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

Inkster, B., M. Kadaba, V. Subramanian. 2023. Understanding the impact of an ai-enabled conversational agent mobile app on users' mental health and wellbeing with a self-reported maternal event: A mixed method real-world data mHealth study. *Frontiers in Global Women's Health*, 4 1084302.

Kavelaars, RuthAnne, Haley Ward, Kushal M Modi, Anita Mohandas, et al. 2023. The burden of anxiety among a nationally representative us adult population. *Journal of Affective Disorders*, 336 81-91.

Koizumi, N., E. Kuno, T.E. Smith. 2005. Modeling patient flows using a queuing network with blocking. *Health care management science*, 8 49-60.

Krendl, A.C., L. Lorenzo-Luaces. 2022. Identifying peaks in attrition after clients initiate mental health treatment in a university training clinic. *Psychological Services*, 19 (3), 519.

Lancet editorial. 2020. Mental health matters. *The Lancet. Global Health*, 8 (11), e1352.

Lee, E.E., J. Torous, M. De Choudhury, C.A. Depp, S.A. Graham, H.-C. Kim, M.P. Paulus, J.H. Krystal, D.V. Jeste. 2021. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 6 (9), 856-864.

Lee, Eunjung, Jessie Faber, Kathryn Bowles. 2022. A review of trauma specific treatments (tsts) for post-traumatic stress disorder (ptsd). *Clinical Social Work Journal*, 50 (2), 147-159.

Leff, H.S., M. Dada, S.C. Graves. 1986. An LP planning model for a mental health community support system. *Management Science*, 32 (2), 139-155.

Levin, M.E., E.T. Hicks, J. Krafft. 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.

Lewin, S., J. Dick, P. Pond, M. Zwarenstein, G.N. Aja, B.E. van Wyk, X. Bosch-Capblanch, M. Patrick. 2005. Lay health workers in primary and community health care. *Cochrane Database of Systematic Reviews*, (1),.

Lewis, C., N.P. Roberts, S. Gibson, J.I. Bisson. 2020. Dropout from psychological therapies for post-traumatic stress disorder (ptsd) in adults: Systematic review and meta-analysis. *European journal of psychotraumatology*, 11 (1), 1709709.

Long, K.M., G.N. Meadows. 2018. Simulation modelling in mental health: A systematic review. *Journal of Simulation*, 12 (1), 76-85.

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

Managed Healthcare Executive. 2023. Analyzing cost-effectiveness thresholds over time. Available at: https://www.managedhealthcareexecutive.com/view/analyzing-cost-effectiveness-thresholds-o

ver-time.

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

Mavranezouli, Ifigeneia, Odette Megnin-Viggars, Nick Grey, Gita Bhutani, Jonathan Leach, Caitlin Daly, Sofia Dias, Nicky J Welton, Cornelius Katona, Sharif El-Leithy, et al. 2020. Cost-effectiveness of psychological treatments for post-traumatic stress disorder in adults. *PloS one*, 15 (4), e0232245.

McCormick, Garth P. 1976. Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems. *Mathematical Programming*, 10 (1), 147-175.

Milicevic, A. Sasha, K. Mitsantisuk, A. Tjader, D.L. Vargas, T.L. Hubert, B. Scott. 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.

Mohr, D.c., P. Cuijpers, K. Lehman. 2011. Supportive accountability: A model for providing human support to enhance adherence to ehealth interventions. *Journal of medical Internet research*, 13 (1), e1602.

Mohr, D.C., S.M. Schueller, E. Montague, M.N. Burns, P. Rashidi. 2014. The behavioral intervention technology model: an integrated conceptual and technological framework for ehealth and mhealth interventions. *Journal of medical Internet research*, 16 (6), e146.

NICE. 2020. Generalised anxiety disorder and panic disorder in adults: management. Available at: https://www.nice.org.uk/guidance/cg113/chapter/Recommendations.

Noorain, S., K. Kotiadis, M.P. Scaparra. 2019. Application of discrete-event simulation for planning and operations issues in mental healthcare. *2019 Winter Simulation Conference (WSC)*. IEEE, 1184-1195.

Noorain, S., Paola S.M., K. Kotiadis. 2023. Mind the gap: a review of optimisation in mental healthcare service delivery. *Health Systems*, 12 (2), 133-166.

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

Ontario, Health Quality, et al. 2017. Psychotherapy for major depressive disorder and generalized anxiety disorder: a health technology assessment. *Ontario health technology assessment series*, 17 (15), 1.

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

PerfectionGeeks Technologies. 2024. Cost to develop a mental therapy app. URL https://www.perfectiongeeks.com/cost-to-develop-a-mental-therapy-app.

Pfefferbaum, B., C.S. North. 2020. Mental health and the COVID-19 pandemic. *New England journal of medicine*, 383 (6), 510-512.

Smith, C., M. Boden, J. Trafton. 2023. Veterans Health Administration outpatient psychiatry staffing model: Longitudinal analysis on mental health performance. *Journal of General Internal Medicine*, 38 (Suppl 3), 814-820.

Steinert, C., K. Stadter, R. Stark, F. Leichsenring. 2017. The effects of waiting for treatment: A meta-analysis of waitlist control groups in randomized controlled trials for social anxiety disorder. *Clinical Psychology & Psychotherapy*, 24 (3), 649-660.

Stringer, H. 2024. Mental health care is in high demand. psychologists are leveraging tech and peers to meet the need. *Monit Psychol*, 55 (1), 1-60.

Swift, J.K., R.P. Greenberg. 2012. Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of consulting and clinical psychology*, 80 (4), 547.

Swift, J.K., J.L. Whipple, P. Sandberg. 2012. A prediction of initial appointment attendance and initial outcome expectations. *Psychotherapy*, 49 (4), 549.

U.S. Bureau of Labor Statistics. 2023a. Occupational Employment and Wage Statistics, May 2023: Mental Health Counselors. Available at: https://www.bls.gov/oes/current/oes211091.htm.

U.S. Bureau of Labor Statistics. 2023b. Occupational employment and wages, may 2023: Psychologists. https://www.bls.gov/oes/current/oes193031.htm. Accessed 2025.

U.S. Department of Veterans Affairs. 2023a. Mental health conditions among VHA patients. https://www.mentalhealth.va.gov/suicide_prevention/docs/FSTP-Mental-Health-Conditions-Among-VHA-Patients.pdf.

U.S. Department of Veterans Affairs. 2023b. Research on mental health conditions in VHA. https://www.research.va.gov/topics/mental_health.cfm.

U.S. Department of Veterans Affairs. 2024a. Harry s. truman memorial veterans' hospital. https://www.va.gov/columbia-missouri-health-care/about-us/. Reports serving more than 40,000 veterans annually; accessed 2025.

U.S. Department of Veterans Affairs. 2024b. Va Salt Lake City health care. https://www.va.gov/salt-lake-city-health-care/about-us/. Reports serving more than 72,000 enrolled veterans annually; accessed 2025.

U.S. Department of Veterans Affairs. 2024c. Veterans health administration: About VHA. https://www.va.gov/health/aboutvha.asp.

U.S. Government Accountability Office. 2021. Va health care: Improvements needed in monitoring of clinical productivity and staffing levels within veterans health administration. https://www.gao.gov/products/gao-21-545sp.

Van Mieghem, J.A. 1995. Dynamic scheduling with convex delay costs: The generalized cμ rule. *The Annals of Applied Probability*, 809-833.

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

Williams, M.E., J. Latta, P. Conversano. 2008. Eliminating the wait for mental health services. *The Journal of Behavioral Health Services & Research*, 35 (1), 107-114.

World Health Organization. 2024. Opioid overdose. https://www.who.int/news-room/fact-sheets/detail/opioid-overdose Retrieved November 23, 2025.

Xaba, F., M.P Lowane, P.K. Chelule, H.N. Shilubane. 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.

Zao-Sanders, M. 2025. How people are really using gen AI in 2025. *Harvard Business Review*, URL https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025.

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

Zychlinski, N. 2024. Managing queues with reentrant customers in support of hybrid healthcare. *Stochastic Systems*, 14 (2), 167-190.

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

Zychlinski, Noa. 2026. An operational view on managing mass trauma events. *Manufacturing & Service Operations Management*, 28 (1), 76-99.

Appendices

Appendix A: MDP Formulation and Average-Reward Bellman Equations

This section provides the average-reward Bellman equations and summarizes the MDP structure used to compute the optimal policy in Section 5. Our objective is the long-run average expected net benefit in (1).

Uniformization and state-space truncation. To compute the MDP numerically, we apply uniformization to the continuous-time controlled Markov process. The resulting discrete-time MDP has the same average reward (up to time scaling). It suffices to use X as the state, since the queue lengths are determined by the action via $Q_i = X_i - Z_i$, $i \in \mathcal{I}$.

Since the state space is countably infinite, we truncate it by imposing upper bounds $X_i \le X_{\max} = 10N$ for each class $i \in \mathcal{I}$. At the truncation boundary, arrival rates to Class i are set to zero when $X_i = X_{\max}$, so that the truncated process remains well defined and stable. We verified that increasing $X_{\max}$ does not materially affect the computed optimal policy.

For each state X, the feasible server-allocation set is

$$\mathcal{Z}(X) := \left\{ Z \in \mathbb{Z}_+^{|\mathcal{I}|} : 0 \le Z_i \le X_i,\ \sum_{i\in\mathcal{I}} Z_i \le N \right\}.$$

Average-reward Bellman equation (fixed W). Fix a waitlist-treatment configuration $W \in \{0,1\}^{|\mathcal{I}|}$. Let g^W denote the optimal long-run average reward under W, and let $h^W(X)$ denote the associated relative value (bias) function. The average-reward Bellman optimality equations are

$$g^W + h^W(X) = \max_{Z\in\mathcal{Z}(X)} \left\{ r^W(X,Z) + \sum_{X'} P^W(X' \mid X, Z)\, h^W(X') \right\}, \tag{A.1}$$

where

$$r^W(X,Z) = \sum_{i\in\mathcal{I}} \Big[r_i Z_i - \big(c_i - W_i(c_i - c_i^w)\big)(X_i - Z_i) \Big] - \sum_{i\in\mathcal{I}} \bar{f}_i W_i,$$

and $P^W(\cdot \mid X, Z)$ are the transition probabilities of the uniformized chain.

Uniformized relative value iteration (RVI) operator. In computation, we apply uniformization with constant Λ and use the equivalent RVI operator. Let e_i denote the unit vector in $\mathbb{R}^{|\mathcal{I}|}$, and Λ denote the upper bound on the total transition rate over all admissible states and actions. The RVI operator takes the explicit form

$$\begin{aligned}\Xi(X) = \frac{1}{\Lambda} \max_{Z\in\mathcal{Z}(X)} \Bigg\{ & r^W(X,Z) + \sum_{i\in\mathcal{I}} \lambda_i\, \Xi(X + e_i) + \sum_{i\in\mathcal{I}} (\mu_i + \gamma_i) Z_i\, \Xi(X - e_i) + \sum_{i\in\mathcal{I}} \theta_i^W (X_i - Z_i)\, \Xi(X - e_i) \\ & + \left(\Lambda - \sum_{i\in\mathcal{I}} \lambda_i - \sum_{i\in\mathcal{I}} (\mu_i + \gamma_i) Z_i - \sum_{i\in\mathcal{I}} \theta_i^W (X_i - Z_i) \right) \Xi(X) \Bigg\},\end{aligned} \tag{A.2}$$

where $\theta_i^W := (1 - W_i)\theta_i^n + W_i \theta_i^w$.

The optimal average reward g^W is obtained from the convergence of the RVI algorithm.

MDP structure and solution method. Under any stationary policy, the induced Markov chain is unichain: from any state there is a positive probability of reaching the empty state due to service completions and abandonment, while arrivals ensure recurrence of the reachable state space. Consequently, the long-run average reward is independent of the initial state. For each fixed W, we solve the resulting finite-state average-reward MDP using RVI, and select the configuration W that yields the highest average benefit.

Appendix B: Proof of Proposition 1

We begin with Problem (2) and substitute the expression for $\bar{q}_i$ from the constraint (4) into the objective function. This yields

$$\begin{aligned}
&\sum_{i\in\mathcal{I}}\left[r_i\bar{z}_i-\left(c_i-w_i(c_i-c_i^w)\right)\left(\frac{1-w_i}{\theta_i^n}+\frac{w_i}{\theta_i^w}\right)\left(\lambda_i-(\mu_i+\gamma_i)\bar{z}_i\right)-\bar{f}_iw_i\right]\\
=&\sum_{i\in\mathcal{I}}\left[r_i\bar{z}_i+(\mu_i+\gamma_i)\bar{z}_i\left(\frac{c_i}{\theta_i^n}(1-w_i)+\frac{c_i^w}{\theta_i^w}w_i\right)-\lambda_i\left(\frac{c_i}{\theta_i^n}(1-w_i)+\frac{c_i^w}{\theta_i^w}w_i\right)-\bar{f}_iw_i\right] \qquad \text{(B.3)}\\
=&\sum_{i\in\mathcal{I}}\left[\left(r_i+(\mu_i+\gamma_i)\left(\frac{c_i}{\theta_i^n}(1-w_i)+\frac{c_i^w}{\theta_i^w}w_i\right)\right)\bar{z}_i-\lambda_i\left(\frac{c_i}{\theta_i^n}(1-w_i)+\frac{c_i^w}{\theta_i^w}w_i\right)-\bar{f}_iw_i\right],
\end{aligned}$$

where the first equality uses the fact that $w_i\in\{0,1\}$.

To verify this step, consider the two possible cases:

- If $w_i=0$ (no waitlist treatment), then

$$c_i-w_i(c_i-c_i^w)=c_i,\qquad \frac{1-w_i}{\theta_i^n}+\frac{w_i}{\theta_i^w}=\frac{1}{\theta_i^n},$$

and hence

$$\left(c_i-w_i(c_i-c_i^w)\right)\left(\frac{1-w_i}{\theta_i^n}+\frac{w_i}{\theta_i^w}\right)=\frac{c_i}{\theta_i^n}.$$

- If $w_i=1$ (waitlist treatment is provided), then

$$c_i-w_i(c_i-c_i^w)=c_i^w,\qquad \frac{1-w_i}{\theta_i^n}+\frac{w_i}{\theta_i^w}=\frac{1}{\theta_i^w},$$

and therefore

$$\left(c_i-w_i(c_i-c_i^w)\right)\left(\frac{1-w_i}{\theta_i^n}+\frac{w_i}{\theta_i^w}\right)=\frac{c_i^w}{\theta_i^w}.$$

Combining both cases yields

$$\frac{c_i}{\theta_i^n}(1-w_i)+\frac{c_i^w}{\theta_i^w}w_i,$$

which justifies the first equality in (B.3). The second equality follows by grouping the terms that multiply $\bar{z}_i$.

Using the $\mathcal{P}_i$ indices defined in (4), the objective in (B.3) can be rewritten as

$$\sum_{i\in\mathcal{I}}\left[\left(\mathcal{P}_i(1-w_i)+\mathcal{P}_i^ww_i\right)\bar{z}_i-\lambda_i\left(\frac{c_i}{\theta_i^n}(1-w_i)+\frac{c_i^w}{\theta_i^w}w_i\right)-\bar{f}_iw_i\right]. \qquad \text{(B.4)}$$

We now apply McCormick linearization (McCormick 1976). Introduce the auxiliary variable $y_i:=w_i\bar{z}_i$ for each $i\in\mathcal{I}$. Since w_i is binary, we have $y_i=0$ when $w_i=0$ and $y_i=\bar{z}_i$ when $w_i=1$.

The McCormick envelope yields the linear constraints appearing in Problem (3). The first two constraints ensure $\bar{q}_i\geq 0$ and enforce the capacity constraint, while the remaining constraints implement the exact linearization of $y_i=w_i\bar{z}_i$ using the bound $\bar{z}_i\leq\lambda_i/(\mu_i+\gamma_i)$. Because $w_i\in\{0,1\}$, these constraints guarantee $y_i=w_i\bar{z}_i$ exactly.

Therefore, Problem (3) is an equivalent MILP reformulation of Problem (2), and can be solved using standard MILP solvers. Q.E.D.

Appendix C: Parameter information for numerical experiments

In all numerical experiments presented in Section 5, we consider two representative patient classes corresponding to MDD and AD.

Table C.1 summarizes the parameter values for each class, for the experiments corresponding to Table 1. Here, N denotes the number of servers (therapists) in the system, and arrival rates are scaled proportionally with N to preserve the relative class proportions from the VHA calibration.

Table C.1 Parameter values used in the numerical experiments of Section 5. Class A corresponds to Major Depressive Disorder (MDD) and Class B corresponds to Anxiety Disorders (AD). All parameters are derived from the VHA calibration in Section 6.

Parameter	Description	Class A (MDD)	Class B (AD)
λ_i	Arrival rate (weekly)	$81N/50$	$57N/50$
μ_i^b	Baseline service rate	2.96	2.96
β_i	Show-up probability	0.82	0.82
$\mu_i = \beta_i \mu_i^b$	Effective service rate	2.43	2.43
γ_i	Dropout rate (weekly)	0.0067	0.0067
θ_i^n	Baseline abandonment rate	0.04	0.04
$\theta_i^{w,p}$	Positive abandonment rate	0.06	0.06
$\theta_i^{w,n}$	Negative abandonment rate	0.0005	0.0005
b_i	Treatment benefit	12,000	9,000
h_i	Holding cost (weekly)	324	196
h_i^w	Holding cost under waitlist treatment	32.4	19.6
a_i	Supervision cost (weekly)	1	1
α_i^n	Abandonment penalty	240	180
$\alpha_i^{w,n}$	Abandonment penalty (waitlist)	120	90
h_i^d	Dropout cost	120	90
$\bar{f}_i$	Weekly overhead cost	640	640

In the experiments corresponding to Figures 2 and 3, the arrival rates are set to $\lambda_A = 81N/45$ and $\lambda_B = 57N/20$, with $N \in \{5, 10, 15, 20, 25\}$; the treatment benefit of Class A is set to $b_A = 6{,}000$, and the overhead cost of Class A is set to $\bar{f}_A = 128{,}000$. This specification reflects resource-intensive AI interventions for Class A with limited clinical effectiveness. In the experiment presented in Figure 4, the arrival rates are set to $\lambda_A = 27$ and $\lambda_B = 28.5$, and the treatment benefit of Class A is set to $b_A = 3{,}000$. All other parameters remain consistent with those reported in Table C.1.