When AI Is Not Enough: Reducing Diagnostic Errors with Radiologist Oversight

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

Artificial intelligence (AI) is becoming increasingly prevalent, particularly in healthcare, where it is shaping the future of decision-making processes. In radiology, AI has revolutionized diagnostics by enabling rapid analysis of patient imaging. However, the consequences of AI misdiagnoses can be significant. For example, an incorrect result can unnecessarily flag a healthy patient for treatment, while a missed detection may fail to identify a serious condition that requires immediate intervention. To mitigate such risks, most diagnostic systems combine AI analysis with radiologist review: AI first classifies cases, and then radiologists review and confirm or modify the initial diagnosis of AI. Effective radiology scheduling must account for the likelihood and cost of false negatives and false positives, as well as AI characteristics such as sensitivity and specificity.

To address the limitations of AI predictions, we develop a multi-server queuing model with separate queues for suspected-positive and suspected-negative cases. Using a fluid approximation, we derive an index-based policy, a modified version of the $c\mu/\theta$ rule, to optimally schedule and allocate resources, taking into account AI characteristics and potential misclassifications. Our proposed policy naturally incorporates the anchoring effect, causing radiologists to devote more time to misclassified cases. As the anchoring effect is incorporated into the classes' indexes, it may change the classes' prioritization and significantly influence overall system performance. Furthermore, to prevent excessive waiting times, even for patients diagnosed as negative, we extend our model to incorporate diagnosis-based service level requirements established by hospitals and regulators. Numerical results demonstrate the effectiveness and superiority of our policy compared to a widely used benchmark, underscoring its potential to improve diagnostic accuracy and efficiency.

Key words: AI diagnosis, radiology, multi-server scheduling, anchoring effect, resource allocation, fluid model

1. Introduction

AI is rapidly advancing in healthcare, revolutionizing diagnostic processes and enhancing the way medical decisions are made (Dai and Abràmoff 2023). The integration of AI into radiology has the potential to profoundly transform diagnostic workflows and improve patient care. Radiology departments are often overwhelmed by the massive volume of imaging studies that require timely diagnosis (Hosny et al. 2018). This leads to extended patient wait times and radiologist burnout, which in turn increases the risk of diagnostic

errors (Lee et al. 2013). The adoption of AI technology offers a promising solution, enabling faster analysis of medical images, improving radiologist efficiency, and reducing delays (He et al. 2024).

Indeed, many hospitals around the world have implemented AI triage systems, such as Aidoc, to assist in preliminary diagnoses by processing large volumes of imaging data swiftly and effectively (Ariel and Susan 2024). Radiologists then review, confirm, or correct the AI-generated initial results to finalize imaging reports, ensuring high-quality medical services.

The rapid response of AI triage systems provides a significant advantage in prioritizing care and reducing waiting times. By quickly analyzing patient images, the AI triage system sorts patients into different categories or queues based on the severity of their condition (Batra et al. 2023). The system then creates a prioritized worklist, alerting radiologists to first review positively suspected cases (Batra et al. 2023). While this method ensures that many of the high-risk patients receive timely attention, it may inadvertently delay the review of negatively suspected cases, extending their wait times. Furthermore, since AI may not always accurately diagnose every case, some patients may be falsely classified and consequently assigned to the wrong queue. This misclassification can have significant consequences. For instance, an ill patient may be incorrectly flagged as healthy, or a healthy patient may be mistakenly flagged as requiring immediate intervention. Although certain AI triage algorithms are designed to prioritize patients sensitively (Li et al. 2013), the risk of missed diagnoses cannot be fully eliminated.

Scheduling AI-diagnosed cases and allocating the necessary resources for each suspected type must take false negative and false positive classifications into account to ensure a cost-effective diagnostic process and minimize waiting times for all patients.

The queue-based server allocation method also appears to align well with radiologists' work habits. Research by Ibanez et al. (2018) indicates that radiologists often prefer focusing on similar tasks in a concentrated way. In other words, radiologists may be more accustomed to reviewing negative or positive patient reports in batches, rather than alternating between the two. If the predictive accuracy of AI triage systems is relatively high, their integration could further complement these work preferences, streamlining radiologists' workflows effectively.

To gain insights into such systems and their dynamics, we study a multi-class queuing system in which incoming patients/scans are classified as either positively or negatively suspected. Each category includes not only correctly diagnosed cases but also incorrect ones. Importantly, while the true diagnosis of each case is unknown at the outset, the diagnosis time itself may depend on both the actual correct diagnosis and the initial diagnosis made by the AI triage system.

Our model also enables the analysis of an *anchoring effect* (Gaube et al. 2021, Dai and Tayur 2022), a common psychological phenomenon that can cause radiologists to spend more time revising an incorrect initial diagnosis made by AI. This effect is likely to increase the time required to correct a false diagnosis by the AI triage system, as it necessitates a careful review of the initial diagnosis as well as the documentation of the correct diagnosis.

Furthermore, the AI triage system is characterized by its sensitivity and specificity (Monaghan et al. 2021). Sensitivity measures the ability of the AI system to correctly identify true positives, while specificity measures its ability to correctly identify true negatives.

These two characteristics affect the diagnostic accuracy of the AI triage system. Our model incorporates these factors, as well as the prevalence of each diagnosed disease in the population, into multi-server scheduling and resource allocation decisions.

Using a deterministic fluid approximation, we derived an index-based policy that integrates all these factors to maximize the long-run average benefit. In addition to holding and abandonment cost, our policy also considers the AI-triage characteristics, disease prevalence, and system incentives for correcting AI errors. Given that the initial results generated by AI might influence the subsequent judgment of radiologists, the indexes also take into account differences in service rates when reviewing correctly classified versus misclassified cases.

Finally, we offer two policy generalizations. The first incorporates service level requirements for actual positive and actual negative cases, while the second extends the model to a multi-disease setting.

Extensive numerical experiments demonstrate the effectiveness and robustness of our proposed policy, highlighting the complex effects of sensitivity, specificity, and prevalence on performance. Comparing our policy to a widely used benchmark shows its superiority, especially when AI prediction accuracy is moderate.

We find that the anchoring effect of AI's preliminary diagnosis can substantially influence scheduling and resource allocation decisions. Specifically, we model the anchoring effect by assigning extended service times to cases where the AI's initial diagnosis is incorrect, compared to cases with accurate diagnoses. Misclassified cases by the AI triage system require longer processing times. As the anchoring effect intensifies, these service time disparities become more pronounced, impacting the indexes in our proposed policy and ultimately altering prioritization, resource allocation and benefit.

The rest of this paper is organized as follows. Section 2 provides a brief review of the related literature. In Section 3, we present the basic diagnosis model and optimization problem. Section 4 develops the corresponding fluid model and derives the index-based policy. Section 5 presents the analysis of the AI prediction characteristics along with numerical experiments. In Section 6, we generalize the model to incorporate service level requirements for positive and negative cases. Finally, Section 7 concludes the paper and suggests promising directions for future research. All mathematical proofs are provided in the online supplementary material.

2. Literature Review

Our research is relevant to two streams of OR/OM literature. The first stream includes triage models with misclassifications, specifically prioritization problems where customer classes may be imperfectly observed. The second stream relates to scheduling multi-server, multi-class queues. In what follows, we review the relevant literature in each stream.

2.1. Triage Models with Misclassifications

In healthcare, triage is the process of prioritizing patients based on the severity of their condition to ensure that those who need urgent care receive it promptly (Lidal et al. 2013). Triage plays a critical role in hospital systems and emergency response settings. Yet because of the uncertainty of patients' actual condition, they may be misclassified.

Most of the literature on triage models focus on a single-server setting. Van der Zee and Theil (1961) is one of the earlier works focusing on the multi-class queuing system under the uncertainty of customer type information. They found that the wrong classifications caused by uncertain customer type may prolong the expected waiting time and proposed to place customers who are likely to be misclassified by the classifier into a separate class.

Sun et al. (2018) addressed a patient triage problem under severe conditions. By developing a Markov Decision Process (MDP) for the triage system, they assessed its impact on

performance. Their findings suggest that triage is beneficial when triage time is short and patient volume is large; otherwise, introducing triage may increase expected costs. These insights strongly support the adoption of a rapid-response AI-driven triage system, particularly in radiology departments where large volumes of images need to be interpreted. Building on Sun et al. (2018)'s work, Sun et al. (2022) investigated a triage system with hidden customer class identities. By constructing a stylized queuing model and an MDP, they concluded that triage is most suitable when traffic intensity is moderate rather than high, especially considering that the triage process itself can introduce additional waiting costs.

Saghafian et al. (2014) developed a MDP to investigate a complexity-augmented triage system which dose not solely rely on the urgency levels to classify patients. They found the complexity-augmented triage is more reliable and robust, even when the misclassification rate reaches up to 25%. Saghafian et al. (2018) proposed a partially observable MDP for telemedical physician triage, with a single upper-level physician (similar to radiologists in our paper) and multiple lower-level triage nurses (analogous to AI-based triage system in our paper). They analyzed classification errors and found that referring patients to the more experienced upper level ensures a lower long-run average total cost if the cost of misclassification is high.

Kamali et al. (2019) established an analytical framework to examine the applicable scenarios of provider triage (i.e., physician triage) rather than nurse triage. They leverage fluid approximation to obtain the optimal triage policy and found that misclassification may harm the performance of provider triage as well. Focusing on a single-server queueing model, Argon and Ziya (2009) considered the case where customer type identities are not directly observable. They found that increasing the number of priority classes can reduce the long-run average waiting cost. Additionally, they pointed out that when waiting costs are linear, the optimal policy is to assign higher priority to customers with a higher probability of being classified into a high-priority class.

Recently, Singh et al. (2024) examined AI-driven triage in radiology departments. Focusing on a single server setting, they developed an analytical model that optimizes AI classifiers by using patient features to predict priority levels. They suggest a "direct approach" to feature-driven priority queuing, where the classifier predicts the priority queue probabilities from features and is optimized to minimize the average waiting cost.

We identify four key differences relative to previous work. First, we extend the analysis beyond single-server models to a multi-server system, reflecting the typical setup in hospital radiology departments where multiple radiologists work in parallel. We also incorporate abandonments when waiting times are excessive. Second, we focus on AI-driven triage, explicitly incorporating AI-specific characteristics, sensitivity and specificity, which influence prediction accuracy. Third, our model accounts for service level requirements for each classification category, as well as service times that depend on both AI classification and actual diagnosis. Fourth, we examine the system under the anchoring effect, analyzing its impact on scheduling and decision making. Dai and Singh (2024) examined how to integrate AI into healthcare – either as a gatekeeper, directing patients to specialists, or as a second opinion to support diagnoses. Their model accounts for the anchoring effect, where the order of diagnosis influences decision making. If AI first assesses the patient, the specialist may be biased by the AI result. In contrast, if the specialist assesses first, their diagnosis may influence the AI algorithm output.

Similarly to our findings, Dai and Singh (2024) suggest that physician-AI collaboration can, in some instances, yield worse outcomes than physician decision-making alone. In any case, it is crucial to incorporate the anchoring effect into decision-making processes that involve AI-human diagnosis.

2.2. Scheduling of Multi-class Queues

The scheduling of multi-class queues has been extensively studied. The classical $c\mu$ rule was shown by Cox and Smith (1961) to be optimal in a single-server queue with linear holding costs. Various extensions to this rule have been proposed. Van Mieghem (1995) demonstrated that a generalization of the rule for convex holding costs is asymptotically optimal. Later, Mandelbaum and Stolyar (2004) extended this work to incorporate multiple classes of servers.

In multi-server systems, Harrison and Zeevi (2004) and Atar et al. (2004) studied the scheduling of multiple customer classes with abandonment under the critically loaded regime. Atar et al. (2010) subsequently derived the asymptotic optimality of the $c\mu/\theta$ rule for many-server queues with abandonment under the many-server heavy traffic regime. Later, Long et al. (2020) proposed an extension of the rule to handle general queue length cost functions and customer patience time distributions.

Motivated by proactive healthcare, Hu et al. (2022) studied the scheduling of a multi-server queue with two customer classes – urgent and moderate– who can improve or deteriorate while waiting. Zychlinski (2024) examined the scheduling of a hybrid healthcare system with in-person, virtual, and supplementary in-person channels for patients who may require follow-up after a virtual visit. Zhong et al. (2024) considered the scheduling and capacity allocation of a two-stage service system – the first with a single station and the second with multiple parallel stations.

Our research contributes to this field by incorporating the accuracy of AI predictions (e.g., sensitivity and specificity), which affects both types of misclassifications and their associated costs. Additionally, our model accounts for the anchoring effect of the initial AI diagnosis and incorporates service level requirements, which are essential for ensuring high performance in AI-driven triage systems.

2.3. Contributions

The paper's contributions, as we see them, are fourfold:

- **Modeling a radiology department with AI-assisted triage.** We develop a multi-server queuing model for a diagnostic system that integrates AI-based case classification with radiologist review, allowing for confirmation or modification of the AI's initial diagnosis. Our model explicitly incorporates AI characteristics such as sensitivity and specificity, along with the benefits of identifying false negatives and false positives.

- **A modified $c\mu/\theta$ rule for maximizing overall net benefit.** The stochastic scheduling problem we formulate is an MDP with a large state space and policy complexity. To gain structural insights into the optimal scheduling policy, we use a fluid approximation and derive an index-based scheduling policy for suspected positive and suspected negative cases, incorporating the benefits associated with correcting misclassifications. We further prove that the proposed rule is asymptotically optimal in the many-server heavy-traffic regime. Simulation experiments confirm its effectiveness in small and moderate-sized systems.

- **Impact of the anchoring effect on scheduling decisions.** Radiologists tend to spend more time revising incorrect AI diagnoses than reviewing correct ones of the same type. Since the anchoring effect is embedded in the class-specific indices, it can alter prioritization and significantly impact system performance. In extreme cases, avoiding AI triage altogether may be preferable.

• **Integrating service level requirements into scheduling.** These requirements, common in diagnostic settings—including radiology—impose waiting time limits for positive and non-positive cases. Due to AI misclassification, each suspected queue contains both case types. To address this, we integrate these requirements into our model and scheduling framework, ensuring that the average waiting time for both actual negative and actual positive individuals remains within a predefined threshold.

3. The Diagnosis Model

We model the diagnosis system as illustrated in Figure 1. The imaging results are first assessed by an AI triage system, which classifies each image as either suspected positive, denoted by p, or suspected negative, denoted by n. These classification outcomes are then routed to separate queues. We refer to the suspected positive queue as Queue p and the suspected negative queue as Queue n. At the next stage, N radiologists (servers) review each image and AI diagnosis and provide a final diagnosis. In Appendix A we extend the model to a multi-disease setting.

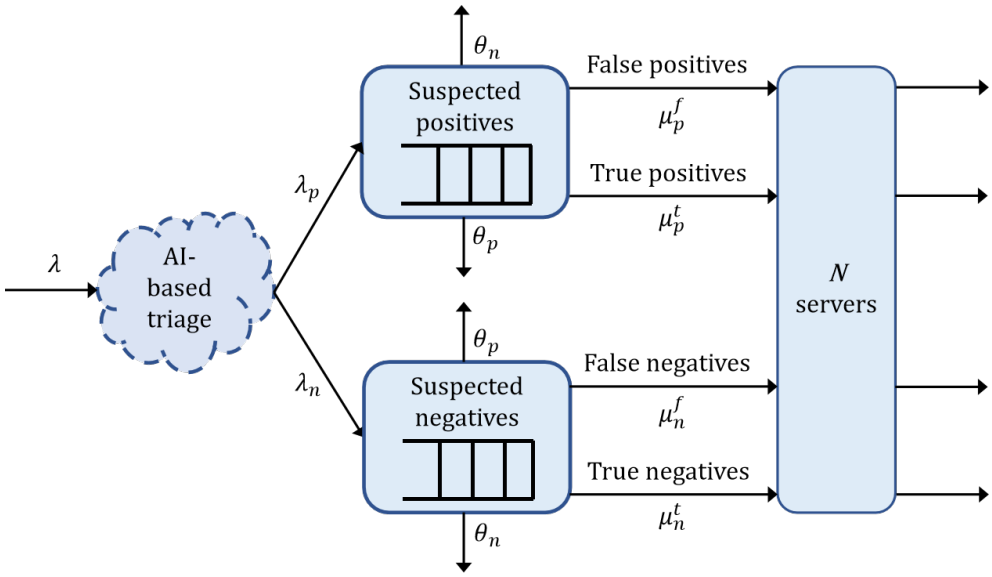

Figure 1 **An illustration of the diagnosis model.**

Due to misclassifications caused by the AI-based triage, each of the queues includes both true and false diagnoses. That is, Queue p includes true positives as well as false positives, while Queue n includes true negatives as well as false negatives. In general, radiologists tend to spend more time diagnosing cases that are actually positive compared to those that

are negative, as the former often involve higher health risks and require greater caution. However, to capture the anchoring effect caused by the initial AI triage, we define a separate service rate for each sub-group. Specifically, let μ_p^t and μ_p^f denote the service rates for true and false positives, respectively, and μ_n^t and μ_n^f denote the service rates for true and false negatives.

Note that a stronger anchoring effect is reflected through a longer diagnosis time for positive cases that were diagnosed as negative compared to true positives ($\mu_n^f < \mu_p^t$). Similarly, a stronger anchoring effect can also be reflected through a longer diagnosis time for negative cases that were diagnosed as positive compared to true negatives ($\mu_p^f < \mu_n^t$).

When waiting times are long, patients might abandon the system; such patients are sometimes referred to as discharged against medical advice (AMA). Let θ_p and θ_n denote the abandonment rates of actual positives and actual negatives, respectively. We assume that service and abandonment times are exponentially distributed.

3.1. Characteristics of the AI Triage System

We denote by $\mathcal{P} \in [0,1]$, the prevalence/proportion of patients in a scanned population who actually have the disease for which they were diagnosed for. The prevalence affects how often the AI encounters positive cases in a given population.

The AI triage system is characterized by its sensitivity and specificity (Monaghan et al. 2021). Sensitivity measures the ability of the AI-triage system to correctly identify those with the disease (true positives). High sensitivity means that the system well detects cases of the disease, minimizing false negative cases. We denote by $\epsilon \in [0,1]$ the sensitivity of our AI triage system. Let T_p and F_n denote the proportion of true positives and false negatives produced by the AI triage system. We, therefore, have

$$T_p = \epsilon\mathcal{P} \text{ and } F_n = (1-\epsilon)\mathcal{P}.$$

Specificity measures the ability of the AI system to correctly identify those without the disease (true negatives). High specificity means the AI triage system well excludes patients who do not have the disease, minimizing false positives. We denote by $\phi \in [0,1]$ the specificity of our AI triage system. Let F_p and T_n denote the proportion of false positives and true negatives produced by the AI triage system. We, therefore, have

$$T_n = \phi(1-\mathcal{P}) \text{ and } F_p = (1-\phi)(1-\mathcal{P}).$$

The imaging arrival process to the radiologists upon completion of their AI diagnosis is a homogeneous Poisson process with rate λ. The arrival rate to Queue p and Queue n are, therefore,

$$\lambda_p = (T_p + F_p)\lambda \quad \text{and} \quad \lambda_n = (T_n + F_n)\lambda.$$

The ratio of false positive cases in Queue p and false negative cases in Queue n can be respectively expressed as

$$f_p = \frac{F_p}{T_p + F_p} \quad \text{and} \quad f_n = \frac{F_n}{T_n + F_n}.$$

Lemma 1 *The following relations hold:*

$$\frac{\mathrm{d}f_p}{\mathrm{d}\mathcal{P}} \leq 0, \quad \frac{\mathrm{d}f_n}{\mathrm{d}\mathcal{P}} \geq 0; \quad \frac{\mathrm{d}f_p}{\mathrm{d}\epsilon} \leq 0, \quad \frac{\mathrm{d}f_n}{\mathrm{d}\epsilon} \leq 0; \quad \frac{\mathrm{d}f_p}{\mathrm{d}\phi} \leq 0, \quad \frac{\mathrm{d}f_n}{\mathrm{d}\phi} \leq 0.$$

Lemma 1 indicates that as the AI triage sensitivity or specificity increases, the proportion of misclassified cases in both queues decreases. The intuition behind this result is clear. For instance, when the sensitivity ϵ increases, the increase in true positive cases in Queue p dilutes the proportion of false positives, while the reduction in false negative cases lowers their proportion in Queue n. Furthermore, we observe that a rise in prevalence increases the number of true positive cases misclassified as negative, leading to a rise in f_n.

3.2. Components and Formulation of the Optimization Problem

We denote by $X_i(t)$, and $Q_i(t)$, $i = p, n$, the number of suspected positive or suspected negative cases in the system and queue, respectively, at time t, $t \geq 0$. Let $Z_i(t)$ denote the number of allocated resources to Queue i at time t. We use the notation $X(t) = (X_p(t), X_n(t))$ and $Q(t) = (Q_p(t), Q_n(t))$, and thereby the system state at time t is $(X(t), Q(t))$. The non-anticipating scheduling policy we seek determines the number of allocated servers to each class. That is, the decision variables are $Z(t) = (Z_p(t), Z_n(t))$, where $Z_p(t) + Z_n(t) \leq N$, and $Q_i(t) = X_i(t) - Z_i(t) \geq 0$, $i = p, n$. Note that here N is the total number of available radiologists/servers. Under these scheduling policies, the process $\{X(t), Q(t) : t \geq 0\}$ is a Markov process.

The operator's goal is to maximize the total expected long-term average net benefit, which consists of three components: holding costs, abandonment costs, and the benefits of correcting misclassifications.

Specifically, let h_p and h_n denote the holding cost per unit time for actual positive and negative patients, respectively. It is reasonable to assume that $h_p > h_n$ because actual positive cases may be more dangerous.

The total expected holding cost up to time T is given by the following equation

$$\mathbb{E}\left[\int_0^T \left[\left(h_p(1-f_p)+h_n f_p\right) Q_p(t)+\left(h_p f_n+h_n(1-f_n)\right) Q_n(t)\right] \mathrm{d}t\right], \tag{1}$$

where the first term represents the holding cost of the suspected positive queue, which includes both true positive and false positive cases. The second term represents the holding cost of the suspected negative queue, which includes both true negative and false negative cases.

Let $R_i(t)$, $i=p,n$, denote the cumulative number of actual positive and actual negative cases that have abandoned the queue by time t. Denote by γ_i, $i=p,n$, the abandonment cost for each such patient. Thus, the expected total abandonment cost can be expressed as

$$\gamma_p \mathbb{E}\left[\int_0^T R_p(t)\mathrm{d}t\right]+\gamma_n \mathbb{E}\left[\int_0^T R_n(t)\mathrm{d}t\right]. \tag{2}$$

Based on the Markovian modeling assumption, (2) can be rewritten as

$$\left(\gamma_p\theta_p\left(1-f_p\right)+\gamma_n\theta_n f_p\right)\mathbb{E}\left[\int_0^T Q_p(t)\mathrm{d}t\right]+\left(\gamma_p\theta_p f_n+\gamma_n\theta_n\left(1-f_n\right)\right)\mathbb{E}\left[\int_0^T Q_n(t)\mathrm{d}t\right]. \tag{3}$$

Then, by combining it with (1), we get that

$$\bar{c}_p \mathbb{E}\left[\int_0^T Q_p(t)\mathrm{d}t\right]+\bar{c}_n \mathbb{E}\left[\int_0^T Q_n(t)\mathrm{d}t\right], \tag{4}$$

where

$$\begin{aligned} \bar{c}_p &:= \left(h_p+\gamma_p\theta_p\right)\left(1-f_p\right)+\left(h_n+\gamma_n\theta_n\right) f_p; \\ \bar{c}_n &:= \left(h_p+\gamma_p\theta_p\right) f_n+\left(h_n+\gamma_n\theta_n\right)\left(1-f_n\right). \end{aligned} \tag{5}$$

Next, we consider the benefits obtained by correcting the AI triage misclassifications. Indeed, if radiologists overlook the problematic diagnoses generated by AI, both patients and hospitals could incur significant losses. In the case of undetected false negatives, patients may experience condition deterioration due to delayed treatment; for undetected false positives, they might undergo unnecessary over-treatments. Clearly, identifying false diagnoses made by AI can help avoid unnecessary medical complications and prevent hospitals from facing claims brought by patients. Let $D_p^f(t)$ denote the cumulative number

of identified false positive cases in Queue p by time t, and $D_n^f(t)$ denote the cumulative number of identified false negative cases in Queue n by time t. The associated benefit of identifying a falsely classified positive is b_p^f, and the associated benefit of identifying a falsely classified negative is b_n^f. We assume that radiologists can correct all error diagnostics provided by AI. Then, the expected total benefit from correct classifications is

$$b_p^f \mathbb{E}\left[\int_0^T D_p^f(t)\mathrm{d}t\right] + b_n^f \mathbb{E}\left[\int_0^T D_n^f(t)\mathrm{d}t\right]. \tag{6}$$

Based on the Markovian modeling assumption (6) can be rewritten as

$$b_p^f f_p \mu_p^f \mathbb{E}\left[\int_0^T Z_p(t)\mathrm{d}t\right] + b_n^f f_n \mu_n^f \mathbb{E}\left[\int_0^T Z_n(t)\mathrm{d}t\right]. \tag{7}$$

Finally, by subtracting (4) from (7), we obtain the total net benefit:

$$\mathbb{E}\left[\int_0^T \left[b_p^f f_p \mu_p^f Z_p(t) + b_n^f f_n \mu_n^f Z_n(t) - \bar{c}_p Q_p(t) - \bar{c}_n Q_n(t)\right] \mathrm{d}t\right].$$

Our objective is, therefore, to find a scheduling policy π that maximizes the total long-run average net benefit, specifically,

$$\max_{\pi\in\Omega} \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T \left[b_p^f f_p \mu_p^f Z_p(t) + b_n^f f_n \mu_n^f Z_n(t) - \bar{c}_p Q_p(t) - \bar{c}_n Q_n(t)\right] \mathrm{d}t\right], \tag{8}$$

where Ω is the set of admissible controls. Problem (8) is formulated as an MDP. To address the curse of dimensionality (Papadimitriou and Tsitsiklis 1999) – characterized by a large (often infinite) state and policy space – and to gain structural insights, we employ a fluid approximation, transforming the stochastic processes into continuous functions. Fluid approximation methods are widely used in service operations management (Whitt 2002, Zychlinski 2023).

4. The Fluid Model

In the fluid model, the stochastic arrival, service, and abandonment processes are replaced by their corresponding deterministic flow rates. We use the lowercase $\bar{q}_p$ and $\bar{q}_n$ to denote the long-run average fluid queue content of suspected positive and negative scans. A fluid scheduling policy π specifies the long-run average service capacity allocation, where $(\bar{z}_p, \bar{z}_n)$ denotes the long-run average *fraction* of servers allocated to each queue. These quantities satisfy the following relations:

$$\begin{aligned}
\lambda_p &= \left(\mu_p^t(1-f_p) + \mu_p^f f_p\right)\bar{z}_p + (\theta_p(1-f_p) + \theta_n f_p)\bar{q}_p, \\
\lambda_n &= \left(\mu_n^t(1-f_n) + \mu_n^f f_n\right)\bar{z}_n + (\theta_p f_n + \theta_n(1-f_n))\bar{q}_n,
\end{aligned} \tag{9}$$

where the first equation – for Queue p – captures the different service rates between true and false positive patients. Since the proportion of false positives, f_p, is hidden within Queue p, and we have previously assumed that radiologists can correct all incorrect conclusions made by the AI triage system, radiologists handle false positive cases with a probability of f_p at a service rate of μ_p^f. They, meanwhile, review the image results of true positive patients with a probability of $(1-f_p)$ at a service rate of μ_p^t. The second equation captures the same only for Queue n.

The fluid analogous problem to the MDP in (8) is the following linear program (LP):

$$\begin{aligned} \max_{\bar{z}_p,\bar{z}_n,\bar{q}_p,\bar{q}_n\geq 0} \quad & b_p^f f_p \mu_p^f \bar{z}_p + b_n^f f_n \mu_n^f \bar{z}_n - \bar{c}_p \bar{q}_p - \bar{c}_n \bar{q}_n \\ \text{s.t.} \quad & \bar{q}_p = \frac{\lambda_p - \left(\mu_p^t(1-f_p) + \mu_p^f f_p\right)\bar{z}_p}{\theta_p(1-f_p)+\theta_n f_p}, \\ & \bar{q}_n = \frac{\lambda_n - \left(\mu_n^t(1-f_n) + \mu_n^f f_n\right)\bar{z}_n}{\theta_p f_n+\theta_n(1-f_n)}, \\ & \bar{z}_p + \bar{z}_n \leq 1. \end{aligned} \tag{10}$$

For convenience of notation, we define the following weighted averages:

$$\begin{aligned} \bar{\mu}_p &= \mu_p^t(1-f_p) + \mu_p^f f_p, \quad \bar{\mu}_n = \mu_n^t(1-f_n) + \mu_n^f f_n, \\ \bar{\theta}_p &= \theta_p(1-f_p) + \theta_n f_p, \quad \bar{\theta}_n = \theta_n(1-f_n) + \theta_p f_n, \end{aligned}$$

Then, by rearranging (10) and omitting the constants that do not affect the optimization, we obtain the following LP:

$$\begin{aligned} \max_{\bar{z}_p,\bar{z}_n\geq 0} \quad & \mathcal{R}_p \bar{z}_p + \mathcal{R}_n \bar{z}_n \\ \text{s.t.} \quad & 0 \leq \bar{z}_p \leq \lambda_p/\bar{\mu}_p, \\ & 0 \leq \bar{z}_n \leq \lambda_n/\bar{\mu}_n, \\ & \bar{z}_p + \bar{z}_n \leq 1, \end{aligned} \tag{11}$$

where the adjusted $c\mu/\theta$ indexes are:

$$\begin{aligned} \mathcal{R}_p &:= \frac{\bar{\mu}_p}{\bar{\theta}_p}\bar{c}_p + b_p^f f_p \mu_p^f = \frac{\mu_p^t(1-f_p)+\mu_p^f f_p}{\theta_p(1-f_p)+\theta_n f_p}\left(\left(h_p+\gamma_p\theta_p\right)(1-f_p) + \left(h_n+\gamma_n\theta_n\right) f_p\right) + b_p^f f_p \mu_p^f, \\ \mathcal{R}_n &:= \frac{\bar{\mu}_n}{\bar{\theta}_n}\bar{c}_n + b_n^f f_n \mu_n^f = \frac{\mu_n^t(1-f_n)+\mu_n^f f_n}{\theta_p f_n+\theta_n(1-f_n)}\left(\left(h_p+\gamma_p\theta_p\right) f_n + \left(h_n+\gamma_n\theta_n\right)(1-f_n)\right) + b_n^f f_n \mu_n^f. \end{aligned}$$

The first two constraints in (11) impose upper bounds on the number of servers allocated to each queue, ensuring that the allocated servers for each queue do not exceed the number of cases.

Note that the adjusted indexes differ from the standard $c\mu/\theta$ indexes in (i) their weighted averages of service and abandonment rates and costs, and (ii) in the benefit they account for in correcting false negative and false positive misclassifications.

In the case where the AI triage system achieves complete accuracy (i.e., both sensitivity ϵ and specificity ϕ equal 1), the $\mathcal{R}$ indexes reduce to the standard $c\mu/\theta$ indexes (Atar et al. 2010).

4.1. The Asymptotic Optimality of the the adjusted $c\mu/\theta$ Rule

In this section, we establish the asymptotic optimality of our adjusted $c\mu/\theta$ rule. We start by introducing the scaling process and other necessary notation.

Consider a sequence of systems indexed by the number of servers, N, and scale up both the number of servers and the arrival rate. We refer to the system with N servers as the N-th system. Denote by $X_i^N(t)$, $Q_i^N(t)$, and $Z_i^N(t)$, $i=p,n$, the corresponding processes in the N-th system. Next, we consider a Poisson arrival process $A^N(t)$ with rate λ^N. Therefore, the arrival processes to Queue p and to Queue n are Poisson processes with rates $\lambda_p^N=(T_p+F_p)\lambda^N$ and $\lambda_n^N=(T_n+F_n)\lambda^N$, respectively.

Let $R_n^{f,N}(t)$ and $R_n^{t,N}(t)$ denote the cumulative number of actual positives and actual negatives that have been abandoned from Queue n by time t in the N-th system, respectively. Similarly, let $R_p^{f,N}(t)$ and $R_p^{t,N}(t)$ be the cumulative number of actual negatives and actual positives that have been abandoned from Queue p by time t, respectively. We denote by $D_p^{f,N}(t)$, $D_p^{t,N}(t)$, $D_n^{f,N}(t)$, $D_n^{t,N}(t)$ the number of service completions for false positives, true positives, false negatives and true negatives by time t in the N-th system, respectively. Then, for every $t\geq 0$, we have

$$
\begin{aligned}
&Z_p^N(t)+Z_n^N(t)\leq N,\\
&X_i^N(t)-Z_i^N(t)=Q_i^N(t),\ i=p,n;\\
&X_i^N(t)=X_i^N(0)+A_i^N(t)-R_i^{f,N}(t)-R_i^{t,N}(t)-D_i^{f,N}(t)-D_i^{t,N}(t),\ i=p,n.
\end{aligned}
$$

Finally, let $\mu_i^{f,N}$, $\mu_i^{t,N}$, and θ_i^N, $i=p,n$, be the corresponding service and abandonment rates in the N-th system.

The normalized average benefit is, therefore,

$$
B_{N,T}(\pi^N)=\frac{1}{NT}\mathbb{E}\left[\int_0^T\left[b_p^N Z_p^N(t)+b_n^N Z_n^N(t)-\bar{c}_p^N Q_p^N(t)-\bar{c}_n^N Q_n^N(t)\right]\mathrm{d}t\right],
$$

where

$$b_p^N = b_p^f f_p \mu_p^{f,N},\ b_n^N = b_n^f f_n \mu_n^{f,N},$$
$$\bar{c}_p^N = \left(h_p + \gamma_p \theta_p^N\right)(1 - f_p) + \left(h_n + \gamma_n \theta_n^N\right) f_p,$$
$$\bar{c}_n^N = \left(h_p + \gamma_p \theta_p^N\right) f_n + \left(h_n + \gamma_n \theta_n^N\right)(1 - f_n).$$

and

$$U_{N,T} = \sup_{\pi^N \in \Omega^N} B_{N,T}(\pi^N).$$

Lastly, denote by $q^* = (\bar{q}_p^*, \bar{q}_n^*)$ and $z^* = (\bar{z}_p^*, \bar{z}_n^*)$ the corresponding values in the fluid model which satisfy $x^* = q^* + z^*$. Then, the optimal value function is $U^* = bz^* - \bar{c}q^*$, where $b = (b_p,\ b_n) = (b_p^f f_p \mu_p^f,\ b_n^f f_n \mu_n^f)$ and $\bar{c} := (\bar{c}_p,\ \bar{c}_n)$.

Assumption 1 *(i) There are positive constants λ, μ_i^f, μ_i^t, and θ_i, $i = p, n$, such that, as $N \to \infty$,*

$$\lambda^N / N \to \lambda,\ \mu_i^{f,N} \to \mu_i^f,\ \mu_i^{t,N} \to \mu_i^t,\ \theta_i^N \to \theta_i.$$

(ii) The random variables $X_i^N(0)/N$, $i = p, n$, are uniformly bounded by a constant M.

Note that per Assumption 1(i), $\lambda_i^N / N \to (T_i + F_i)\lambda$, $i = p, n$, as $N \to \infty$.

Proposition 1 establishes that the adjusted $c\mu/\theta$ rule is asymptotically optimal.

Proposition 1 (asymptotic optimality) *Denote by π_p^N the policy of our adjusted $c\mu/\theta$ rule in the N-th system. Then, under Assumption 1, we have*

(i) $\limsup_{T\to\infty} \limsup_{N\to\infty} U_{N,T} \le U^*$,

(ii) $\limsup_{T\to\infty} \limsup_{N\to\infty} U_{N,T} = \liminf_{T\to\infty} \liminf_{N\to\infty} B_{N,T}(\pi_p^N) = U^*$.

Section 5 examines the impact of the AI Triage system's characteristics on the optimal policy and overall net benefit.

5. Impact of AI Triage Characteristics

The characteristics of the AI triage system affect the optimal allocation and performance. In this section, we use numerical experiments to examine this effect. Section 5.1 analyzes the impact of sensitivity and specificity. Section 5.2 focuses on the effect of prevalence on optimal scheduling, resource allocation, and net benefit. In Section 5.3, we investigate the anchoring effect. Finally, in Section 5.4, we compare the performance of our proposed policy with the standard $c\mu/\theta$ rule.

5.1. The Sensitivity and Specificity Effects

Recall that sensitivity measures the ability of the AI system to correctly identify true positives, while specificity measures its ability to correctly identify true negatives. We consider two scenarios that differ in their consequences of the two misclassification types: false negatives and false positives. These differences are reflected in our model as the benefit of correcting AI misclassifications, specifically the values of b_n^f and b_p^f.

The first scenario includes the case where a false-negative diagnosis could have severe consequences. For example, in the context of a highly infectious disease like COVID-19, misidentifying positive lung imaging as negative could lead to significant costs for the healthcare system. Therefore, we assign a substantial reward for correcting false-negative cases compared to false positives, specifically $b_n^f = 150$ and $b_p^f = 10$.

The second scenario includes the case where uncorrected false-positive results can cause severe psychological or financial distress for patients, as well as unnecessary treatment, such as in the case of multiple sclerosis (Li et al. 2013). In this context, identifying false-positive cases becomes critical. Thus, we set $b_p^f = 150$ and $b_n^f = 10$.

Figure 2 shows the optimal resource allocation (right plots) and the optimal long-term average benefit (left plots) as functions of different sensitivity levels for these two scenarios. We set the service rate μ_n^t for reviewing true negatives at 2 and the service rate μ_n^f for reviewing false negatives at 0.5, considering that radiologists usually spend more time thoroughly evaluating images of actual positive cases. Because of the anchoring effect, identifying false positives and false negatives may take more time than identifying true negatives and true positives, respectively. Thus, we have $\mu_n^f < \mu_p^t$ and $\mu_p^f < \mu_n^t$. Additionally, we set $h_p = 2$ and $h_n = 1$, given that time is more costly for actual positive cases.

In the first scenario (top plots, where $b_n^f > b_p^f$), the allocation shifts from the negative suspected Queue n to the positive suspected Queue p as sensitivity ϵ increases. Specifically, when sensitivity is low, priority is given to Queue n to identify as many false negatives as possible. As sensitivity increases and the AI triage system produces fewer false negatives, more resources can be allocated to Queue p. The jump in optimal allocation occurs exactly at the switching point in priority. Moreover, the optimal long-run average benefit decreases linearly with sensitivity, which can be attributed to the reduced number of corrected false negatives in Queue n.

In the second scenario (bottom plots, where $b_n^f < b_p^f$), the same gradual shift in allocation from Queue n to Queue p occurs. However, since there is no switch in priority ($\mathcal{R}_p > \mathcal{R}_n$ for all sensitivity levels), there are no jumps in allocation. As higher sensitivity increases the proportion of true positives in Queue p and decreases the proportion of false negatives in Queue n (Lemma 1), these effects, combined with a relatively high holding cost h_p for positive cases, increase $\mathcal{R}_p$ relative to $\mathcal{R}_n$.

Unlike the first scenario, in the second scenario, the optimal long-run average benefit slightly increases with sensitivity, as more resources are allocated to Queue p, allowing for the correction of more false positives. Since the benefit from such corrections is high, the total net benefit increases.

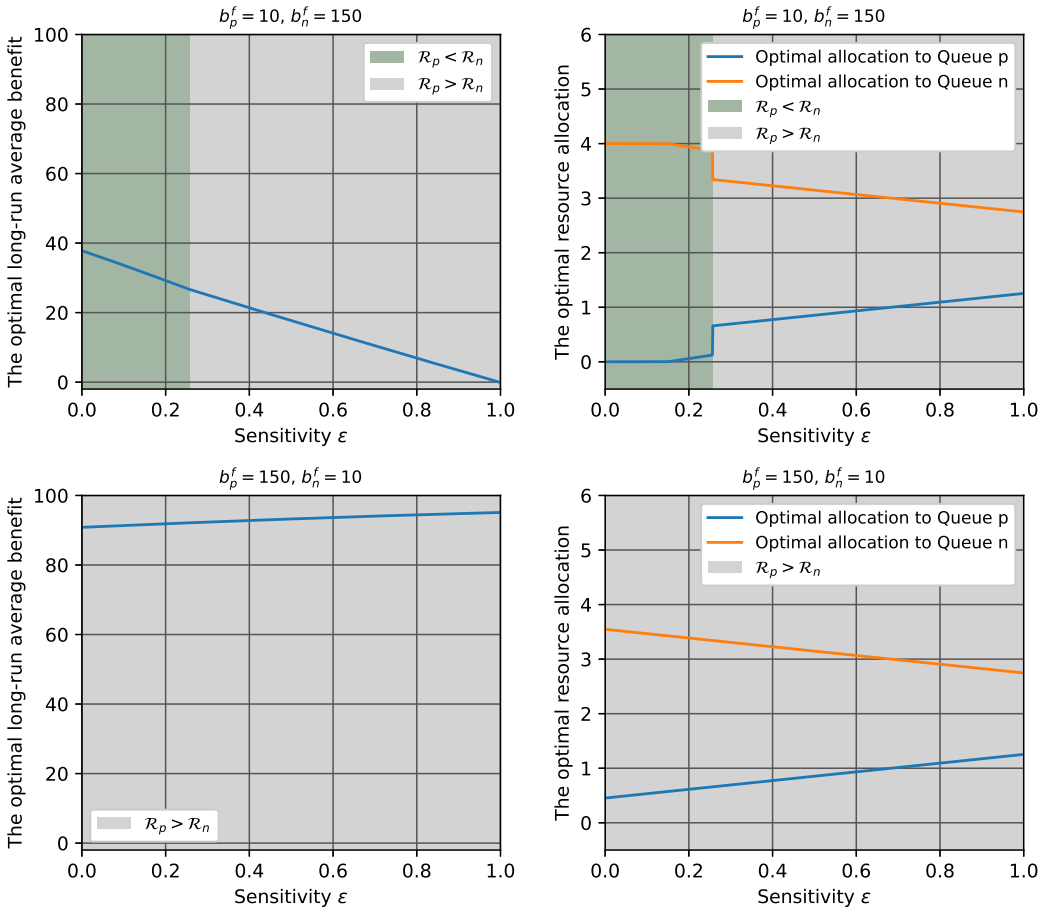

Figure 2 **The sensitivity effect on the optimal long-term average benefit and optimal resource allocation. The parameters are:** $\phi = 0.90$, $\mathcal{P} = 0.15$, $h_p = 2$, $h_n = 1$, $\theta_p = 0.1$, $\theta_n = 0.1$, $\gamma_p = 2$, $\gamma_n = 1$, $\mu_p^f = 1.5$, $\mu_n^f = 0.5$, $\mu_n^t = 2$, $\mu_p^t = 1.5$, $N = 4$, $\lambda = 8$.

Figure 3 shows the optimal resource allocation (right plots) and the optimal long-term average benefit (left plots) versus different specificity levels for these two scenarios.

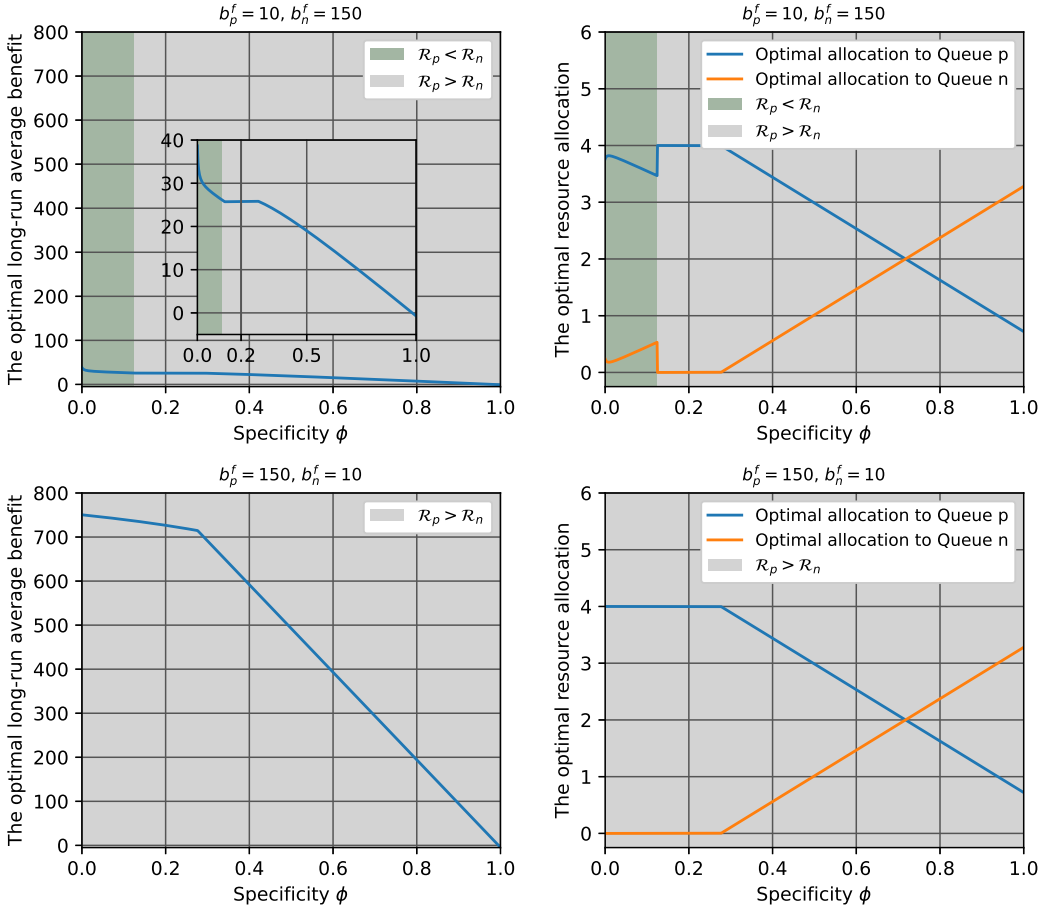

Figure 3 **The specificity effect on the optimal long-term average benefit and optimal resource allocation. The left top plot includes a zoomed-in version of the figure. The parameters are:** $\epsilon = 0.90$, $\mathcal{P} = 0.15$, $h_p = 2$, $h_n = 1$, $\theta_p = 0.1$, $\theta_n = 0.1$, $\gamma_p = 1.5$, $\gamma_n = 1$, $\mu_p^f = 2$, $\mu_n^f = 0.5$, $\mu_n^t = 2$, $\mu_p^t = 1.5$, $N = 4$, $\lambda = 8$.

In the first scenario (top plots, where $b_p^f = 10 < b_n^f = 150$), the priority switches from Queue n to Queue p when $\phi \approx 0.125$. Within each of these two ranges (excluding the switching point itself), the optimal allocation to Queue n increases, while the allocation to Queue p decreases.

The optimal long-term average net benefit generally declines with specificity: higher specificity means fewer false positives, reducing the potential benefits from correcting them. However, there exists a narrow interval around $\phi = 0.2$ where the optimal long-term average benefit slightly increases. This occurs because Queue p is prioritized when specificity is

low, resulting in a relatively large Queue p. Within this interval, all servers are allocated to Queue p, reducing the holding and abandonment costs for this queue. Additionally, since no servers are allocated to Queue n during this interval, the possibility of substantial benefits from correcting false negatives is avoided.

In the second scenario (bottom plots, where $b_p^f = 150 > b_n^f = 10$), there is no switch in priority. When specificity is low, a substantial portion of actual negative patients accumulates in Queue p, increasing the required number of servers. Under these low-specificity conditions, Queue p requires the total number of servers. As specificity ϕ increases, the opportunity to gain substantial benefits from correcting false positives decreases, resulting in a consistent decline in the optimal long-term average benefit.

5.2. The Prevalence Effect

In this section, we examine the effect of disease prevalence on the optimal long-term average benefit and the corresponding optimal resource allocation. We consider an effective AI triage system with high sensitivity ($\epsilon = 0.90$) and specificity ($\phi = 0.90$). Figure 4 shows the optimal net benefit and resource allocation across different prevalence levels. In the first scenario (top plots), $b_p^f < b_n^f$, whereas in the second scenario (bottom plots), $b_p^f > b_n^f$.

In the first scenario, priority shifts from Queue p to Queue n as prevalence increases. As the number of false-negative patients misclassified into Queue p grows, prioritizing Queue n significantly enhances the potential benefits of correcting these errors. Additionally, as prevalence increases, the number of actual negative cases decreases, reducing the demand for servers in Queue n. Consequently, the number of servers assigned to Queue n may decrease even when Queue n is prioritized. A U-shaped pattern emerges in the optimal long-term average benefit under this scenario. When prevalence is moderate, the benefits from correcting false negatives are insufficient to offset the holding and abandonment costs, leading to a net decline in benefits. However, when prevalence is high, the substantial benefits from correcting false negatives outweigh these costs, resulting in an increase in the net benefit.

In the second scenario, Queue p remains prioritized throughout. As prevalence increases, more servers are allocated to Queue p to address the higher demand.

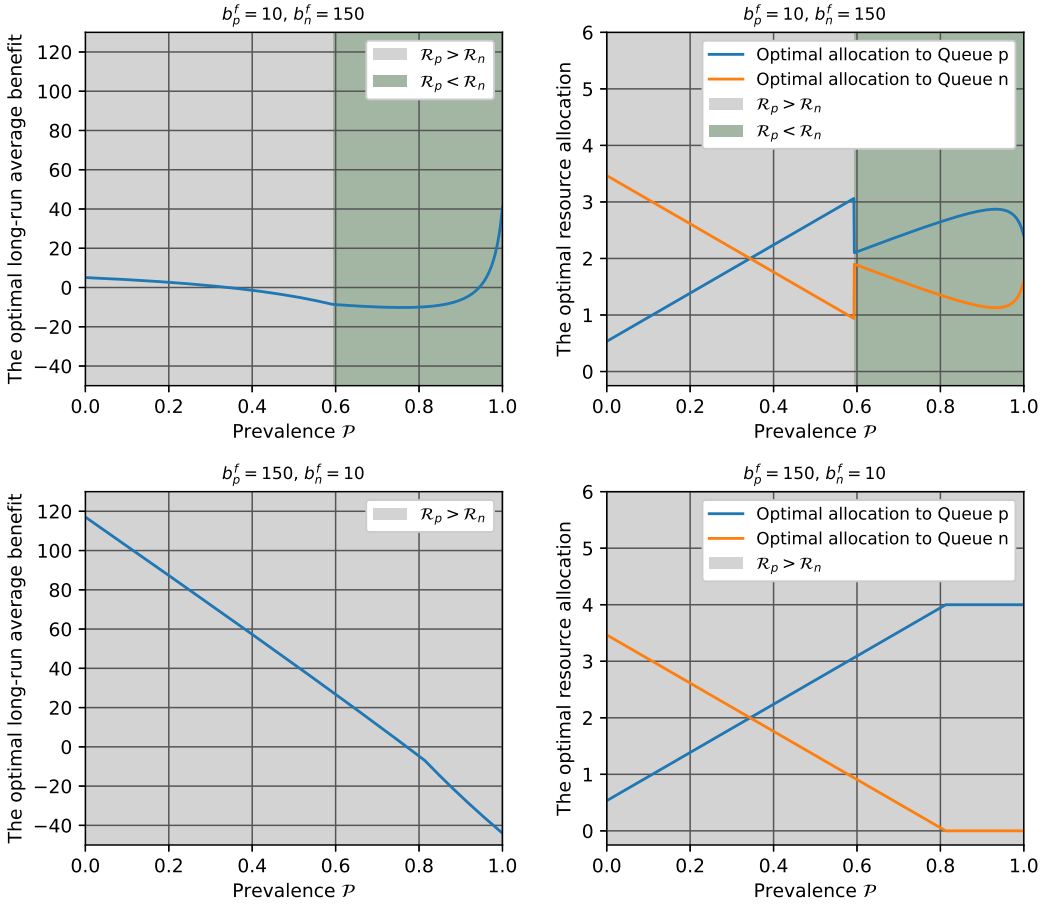

Figure 4 **The prevalence effect on the optimal long-term average benefit and optimal resource allocation. The parameters are:** $\epsilon = 0.90$, $\phi = 0.90$, $h_p = 2$, $h_n = 1$, $\theta_p = 0.1$, $\theta_n = 0.1$, $\gamma_p = 2$, $\gamma_n = 1$, $\mu_p^f = 1.5$, $\mu_n^f = 0.5$, $\mu_n^t = 2$, $\mu_p^t = 1.5$, $N = 4$, $\lambda = 8$.

5.3. The Anchoring Effect

The anchoring effect is a common psychological phenomenon that can cause radiologists to spend more time revising an incorrect initial diagnosis made by the AI triage system than reviewing a correct diagnosis of the same type.

In general, a strong anchoring effect may significantly reduce the system's efficiency by slowing down radiologists when their diagnosis is inconsistent with the AI's initial diagnosis. We denote the level of the anchoring effect by η: the larger η, the stronger the anchoring effect. The anchoring effect modifies the service rates of false positives and false negatives according to the following relation:

$$\eta = \frac{\mu_p^t}{\mu_n^f} = \frac{\mu_n^t}{\mu_p^f}, \quad \eta \geq 1.$$

It is possible to define two separate anchoring effects for each of the above ratios. However, for simplicity, we assume that the anchoring effect has the same impact on both false negatives and false positives.

Figure 5 demonstrates the impact of the anchoring effect on the long-term average net benefit across different prevalence levels. We see that the long-run average net benefit decreases as the anchoring effect becomes larger across all levels of prevalence. Along each level of anchoring effect, the long-run average net benefit increases with prevalence. The anchoring effect may impact the relations between the $\mathcal{R}$ indexes and by that, change the prioritization between the two queues.

Furthermore, when the anchoring effect is significant ($\eta = 2.2$), the average net benefit may become negative as the costs exceed the benefit from correcting misclassifications. In such cases, it may be more beneficial not to use AI triage system at all.

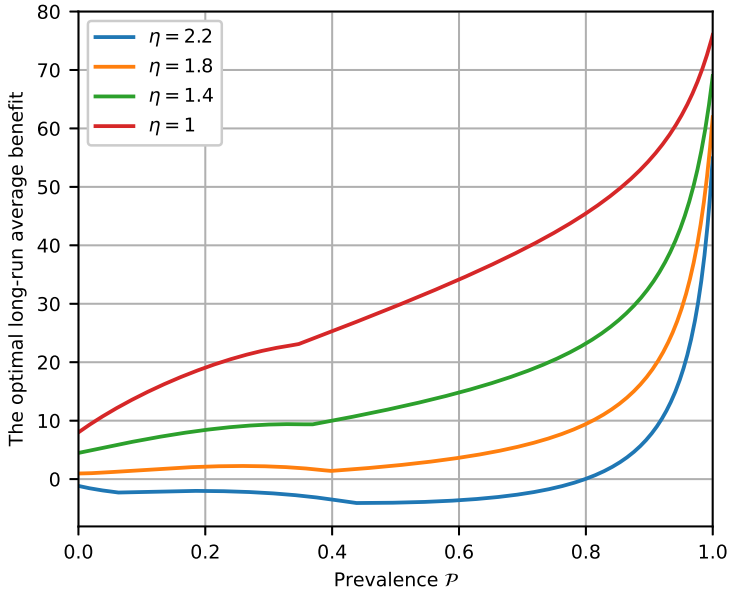

Figure 5 **The anchoring effect for different prevalence levels. The parameters are:** $\phi = 0.90$, $\epsilon = 0.90$, $h_p = 2$, $h_n = 1$, $\theta_p = 0.1$, $\theta_n = 0.1$, $\gamma_p = 2$, $\gamma_n = 1$, $\mu_n^t = 2$, $\mu_p^t = 1.5$, $b_n^f = 150$, $b_p^f = 10$, $N = 4$, $\lambda = 8$.

5.4. Performance Comparison with the $c\mu/\theta$ Rule

To evaluate the performance of our proposed policy, we compare it to a benchmark policy – the standard $c\mu/\theta$ rule – which does not account for the two types of misclassification. Specifically, scheduling according to the standard $c\mu/\theta$ rule treats the parameters for each

queue as if all patients are correctly classified. Consequently, the indices for Queue p and Queue n under the standard $c\mu/\theta$ rule are defined as:

$$(c\mu/\theta)_p := \frac{(h_p + \gamma_p\theta_p)\mu_p^t}{\theta_p} \quad \text{and} \quad (c\mu/\theta)_n := \frac{(h_n + \gamma_n\theta_n)\mu_n^t}{\theta_n}.$$

Figure 6 compares the long-term average net benefit of our policy (referred to as the adjusted $c\mu/\theta$ rule) with the standard $c\mu/\theta$ rule across different scenarios. In the left plots, the system has $N = 3$ servers and an arrival rate of $\lambda = 8$, while in the right plots, the system size is doubled to $N = 6$ servers and $\lambda = 16$. The comparisons are in regard to sensitivity (top plots), specificity (middle plots), and prevalence (bottom plots).

We first note that as sensitivity and specificity improve, the performance of the adjusted $c\mu/\theta$ rule approaches that of the standard $c\mu/\theta$ rule, eventually converging. In other words, when the AI triage system is highly effective, accounting for misclassifications in the scheduling has less impact. For small or moderate values of sensitivity and specificity, however, the adjusted $c\mu/\theta$ rule significantly outperforms the standard $c\mu/\theta$ rule.

With respect to prevalence, while both policies perform similarly for low prevalence, the adjusted $c\mu/\theta$ rule performs significantly better for moderate and high prevalence. This improvement occurs because a higher prevalence indirectly results in more unhealthy patients being misclassified as negative. Finally, we note that the differences between the policies do not diminish with increasing system size.

6. Incorporating Service Level Requirements

Thus far, we have focused on maximizing the overall system net benefit, allowing for prolonged waiting times when the system is overloaded. For instance, when Queue p is prioritized, individuals in Queue n may experience long waiting times. This is particularly concerning for false-negative patients, who need their results in a timely manner. A clinical workflow simulation by Baltruschat et al. (2021) introduced an upper limit on maximum waiting times to mitigate the effects of false-negative outcomes. Conversely, strict prioritization of Queue n can negatively impact true positives in Queue p when the number of servers is constrained.

To address such issues, many diagnostic procedures, especially in radiology departments, include regulations on allowable waiting times for each type of diagnosis (Huang et al. 2015, Barron and Baron 2022). For example, the NHS England National Imaging Board

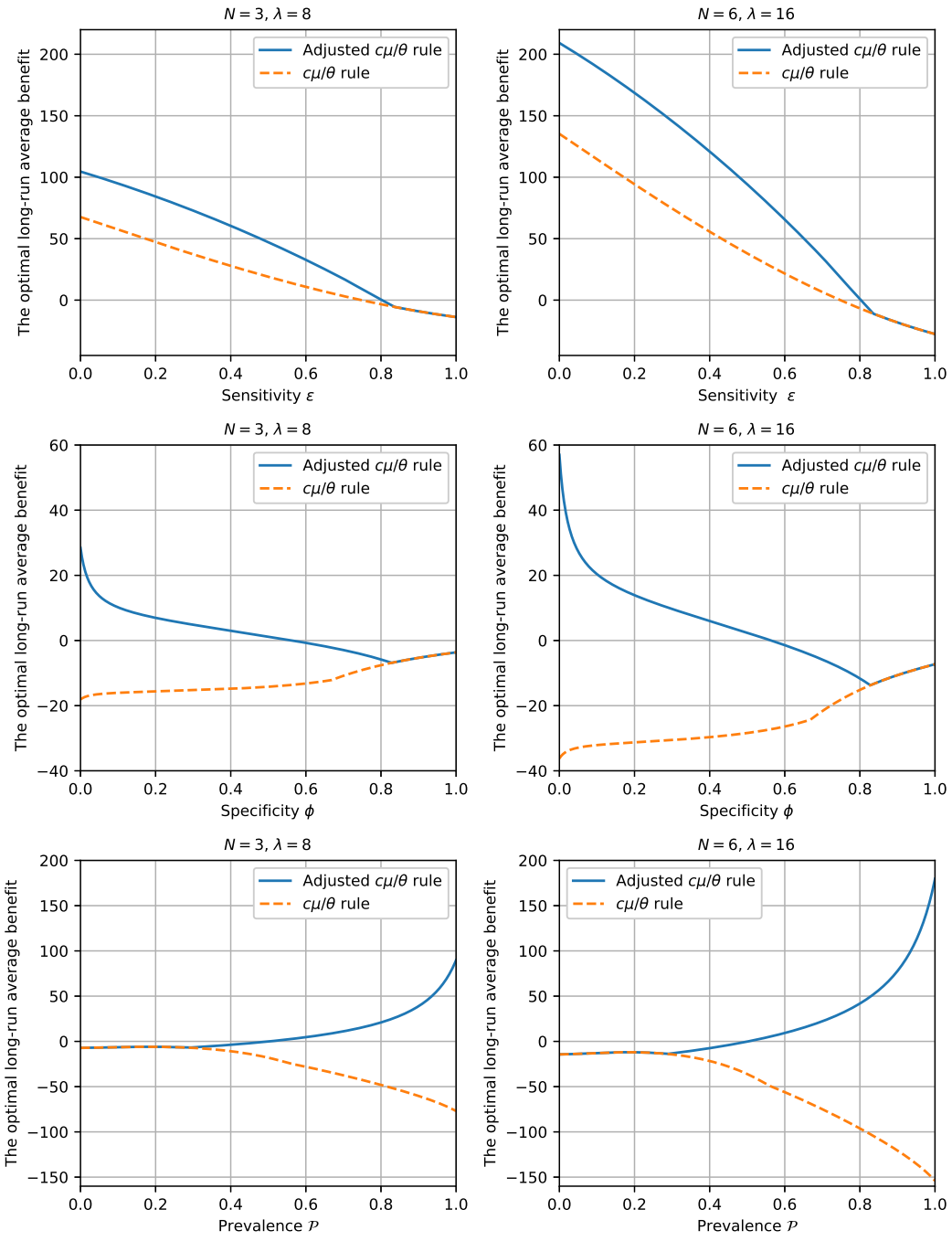

Figure 6 **A comparison to the standard $c\mu/\theta$ rule. The parameters are: $h_p = 2$, $h_n = 1$, $\theta_p = 0.1$, $\theta_n = 0.2$, $\gamma_p = 2$, $\gamma_n = 1$, $\mu_p^f = 1$, $\mu_n^f = 1$, $\mu_n^t = 2$, $\mu_p^t = 1.5$, $b_n^f = 150$, $b_p^f = 10$; in the top plots: $\phi = 0.85$ and $\mathcal{P} = 0.25$; in the middle plots: $\epsilon = 0.85$ and $\mathcal{P} = 0.25$; in the bottom plots: $\phi = 0.85$ and $\epsilon = 0.85$.**

recommends report turnaround times of less than 12 hours for urgent patients and less than 24 hours for non-urgent inpatients.

We incorporate these service level requirements by adding constraints to our optimization problem. Specifically, we ensure that the average waiting time for each actual negative and actual positive individual does not exceed a predetermined threshold.

Let w_p and w_n denote the average waiting time in Queue p, and Queue n, respectively. Let $\bar{W}_p$ and $\bar{W}_n$ denote the average waiting time for *an actual* positive or negative patient, respectively. Therefore,

$$\bar{W}_p = \mathbb{P}\left(\begin{matrix}\text{suspected}\\ \text{positive}\end{matrix}\middle|\ \text{positive}\right) w_p + \mathbb{P}\left(\begin{matrix}\text{suspected}\\ \text{negative}\end{matrix}\middle|\ \text{positive}\right) w_n = \epsilon w_p + (1-\epsilon) w_n,$$
$$\bar{W}_n = \mathbb{P}\left(\begin{matrix}\text{suspected}\\ \text{positive}\end{matrix}\middle|\ \text{negative}\right) w_p + \mathbb{P}\left(\begin{matrix}\text{suspected}\\ \text{negative}\end{matrix}\middle|\ \text{negative}\right) w_n = (1-\phi) w_p + \phi w_n.$$

According to Little's Law (Little 1961), we have $w_p = \bar{q}_p/\lambda_p$ and $w_n = \bar{q}_n/\lambda_n$.

Denote by τ_p and τ_n the waiting times thresholds for positive and negative patients. Then, the optimization problem we consider is

$$\begin{aligned}
\max_{\bar{z}_p, \bar{z}_n, \bar{q}_p, \bar{q}_n} \quad & b_p^f f_p \mu_p^f \bar{z}_p + b_n^f f_n \mu_n^f \bar{z}_n - \bar{c}_p \bar{q}_p - \bar{c}_n \bar{q}_n \\
\text{s.t.} \quad & \bar{q}_p = \frac{\lambda_p - \left(\mu_p^t (1-f_p) + \mu_p^f f_p\right) \bar{z}_p}{\theta_p (1-f_p) + \theta_n f_p}, \\
& \bar{q}_n = \frac{\lambda_n - \left(\mu_n^t (1-f_n) + \mu_n^f f_n\right) \bar{z}_n}{\theta_p f_n + \theta_n (1-f_n)}, \\
& \bar{z}_p \geq 0, \quad \bar{z}_n \geq 0, \\
& \bar{W}_p \leq \tau_p, \quad \bar{W}_n \leq \tau_n, \\
& \bar{z}_p + \bar{z}_n \leq 1.
\end{aligned} \tag{12}$$

We can then explicitly write down the average waiting times w_p, that is,

$$\begin{aligned}
\bar{W}_p &= \epsilon \frac{\left(\lambda_p - \mu_p^t (1-f_p) \bar{z}_p - \mu_p^f f_p \bar{z}_p\right)}{\lambda_p (\theta_p (1-f_p) + \theta_n f_p)} + (1-\epsilon) \frac{\left(\lambda_n - \mu_n^t (1-f_n) \bar{z}_n - \mu_n^f f_n \bar{z}_n\right)}{\lambda_n (\theta_n (1-f_n) + \theta_p f_n)} \\
&= -\bar{z}_p \frac{\epsilon (\mu_p^t (1-f_p) + \mu_p^f f_p)}{\lambda_p (\theta_p (1-f_p) + \theta_n f_p)} - \bar{z}_n \frac{(1-\epsilon)(\mu_n^t (1-f_n) + \mu_n^f f_n)}{\lambda_n (\theta_n (1-f_n) + \theta_p f_n)} \\
&\quad + \frac{\epsilon \lambda_p}{\lambda_p (\theta_p (1-f_p) + \theta_n f_p)} + \frac{(1-\epsilon)\lambda_n}{\lambda_n (\theta_n (1-f_n) + \theta_p f_n)}.
\end{aligned}$$

Therefore, $\bar{W}_p \leq \tau_p$ can be rewritten as $\bar{z}_p \geq \mathcal{B}_1 - \mathcal{A}_1 \bar{z}_n$, where

$$\mathcal{A}_1 = \frac{(\theta_p (1-f_p) + \theta_n f_p) \lambda_p (1-\epsilon)(\mu_n^t (1-f_n) + \mu_n^f f_n)}{(\theta_n (1-f_n) + \theta_p f_n) \lambda_n \epsilon (\mu_p^t (1-f_p) + \mu_p^f f_p)},$$

$$\mathcal{B}_1 = \frac{(\theta_n (1-f_n) + \theta_p f_n) \lambda_n \epsilon \lambda_p + (\theta_p (1-f_p) + \theta_n f_p) \lambda_p (1-\epsilon) \lambda_n}{(\theta_n (1-f_n) + \theta_p f_n) \lambda_n \epsilon (\mu_p^t (1-f_p) + \mu_p^f f_p)} - \frac{(\theta_p (1-f_p) + \theta_n f_p) \lambda_p \tau_p}{\epsilon (\mu_p^t (1-f_p) + \mu_p^f f_p)}.$$

Similarly, for constraint $\bar{W}_n \le \tau_n$ can be rewritten as $\bar{z}_p \ge \mathcal{B}_2 - \mathcal{A}_2 \bar{z}_n$, where

$$\mathcal{A}_2 = \frac{(\theta_p(1-f_p)+\theta_n f_p)\lambda_p \phi(\mu_n^t(1-f_n)+\mu_n^f f_n)}{(\theta_n(1-f_n)+\theta_p f_n)\lambda_n(1-\phi)(\mu_p^t(1-f_p)+\mu_p^f f_p)},$$

$$\mathcal{B}_2 = \frac{(\theta_n(1-f_n)+\theta_p f_n)\lambda_n(1-\phi)\lambda_p + (\theta_p(1-f_p)+\theta_n f_p)\lambda_p \phi \lambda_n}{(\theta_n(1-f_n)+\theta_p f_n)\lambda_n(1-\phi)(\mu_p^t(1-f_p)+\mu_p^f f_p)} - \frac{(\theta_p(1-f_p)+\theta_n f_p)\lambda_p \tau_n}{(1-\phi)(\mu_p^t(1-f_p)+\mu_p^f f_p)}.$$

Problem (12) can then be rewritten as the following LP:

$$\begin{aligned} \max_{\bar{z}_p, \bar{z}_n} \quad & \mathcal{R}_p \bar{z}_p + \mathcal{R}_n \bar{z}_n \\ \text{s.t.} \quad & 0 \le \bar{z}_p \le \frac{\lambda_p}{\mu_p^t(1-f_p)+\mu_p^f f_p}, \\ & 0 \le \bar{z}_n \le \frac{\lambda_n}{\mu_p^t(1-f_n)+\mu_n^f f_n}, \\ & \bar{z}_p + \bar{z}_n \le 1, \\ & \bar{z}_p \ge \mathcal{B}_1 - \mathcal{A}_1 \bar{z}_n, \\ & \bar{z}_p \ge \mathcal{B}_2 - \mathcal{A}_2 \bar{z}_n. \end{aligned} \tag{13}$$

The introduction of two new constraints renders certain resource levels infeasible, meaning no scheduling policy can satisfy the service level constraints. Consequently, the adjusted $c\mu/\theta$ rule may require modifications.

Next, we analyze the impact of the service level constraints on the optimal solution and average net benefit. We consider two cases: one where $\mathcal{R}_p < \mathcal{R}_n$ and another where $\mathcal{R}_p > \mathcal{R}_n$. Figure 7 illustrates the relationships between the optimal resource allocation for different service level requirements.

First, we note that very small thresholds are infeasible, while for very large thresholds, the solutions converge to the unconstrained solution in (11).

When $\mathcal{R}_p < \mathcal{R}_n$, we observe that the number of servers assigned to Queue p decreases while the number of servers assigned to Queue n increases as the threshold for actual positives becomes larger. Due to the higher benefits of correcting false negatives, allocating more servers to Queue n becomes advantageous once the constraints are no longer binding (i.e., when $\tau_p \ge 8.7$), allowing Queue n to receive almost four servers. Conversely, when $\mathcal{R}_p > \mathcal{R}_n$, the trend is reversed.

Figure 8 illustrates the optimal long-run average net benefit for different threshold values of τ_p and τ_n. The blank areas represent infeasibility regions. When $\mathcal{R}_p < \mathcal{R}_n$, the optimal

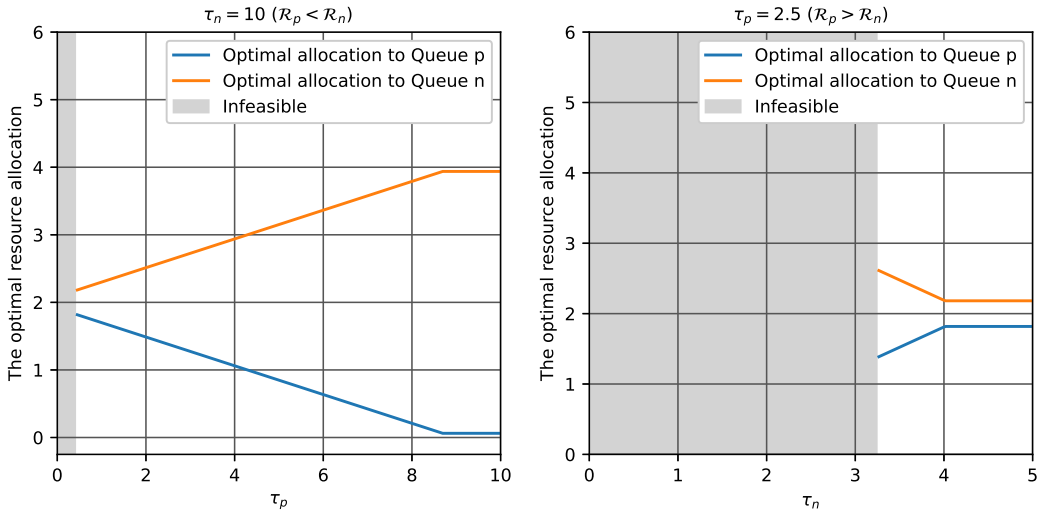

Figure 7 **The optimal resource allocation under varying service level requirements. The parameters are:** $\epsilon = 0.9$, $\phi = 0.9$, $\mathcal{P} = 0.15$, $h_p = 2$, $h_n = 1$, $\theta_p = 0.1$, $\theta_n = 0.1$, $\gamma_p = 2$, $\gamma_n = 1$, $\mu_p^f = 0.75$, $\mu_n^f = 1$, $\mu_n^t = 2$, $\mu_p^t = 1.5$, $N = 4$, $\lambda = 10$, $b_p^f = 10$; **specially, in the left plot,** $b_n^f = 400$ **and in the right plot,** $b_n^f = 100$.

long-run average benefit decreases as τ_p decreases but remains unchanged for varying levels of τ_n. Conversely, when $\mathcal{R}_p > \mathcal{R}_n$, the optimal long-run average benefit decreases as τ_n decreases but remains unaffected by changes in τ_p.

In summary, setting reasonable waiting time thresholds to ensure adequate service levels for correctly classified patients in the lower-priority queue may shift resources and result in a reduced net benefit. This highlights the trade-off between maximizing net benefit and providing better service levels.

7. Concluding Remarks and Future Directions

Motivated by the healthcare industry's move toward integrating AI triage systems, we study this integration in a radiology department. In this setting, upon completing an initial diagnosis, the AI classifies each case into either a positive or negative queue. Radiologists then review the results and correct any misclassifications. We develop a multi-server queuing model aimed at designing efficient scheduling and resource allocation policies for these types of queues. Our model incorporates the characteristics of AI triage systems, capturing the likelihood and costs associated with false positives and false negatives. Using a fluid approximation, we derive an adjusted $c\mu/\theta$ rule that accounts for both types of misclassification and the benefits of correcting them. In addition to being effective and easy to implement, the proposed policy can be adapted to include the anchoring effect and service level requirements for each diagnosis type.

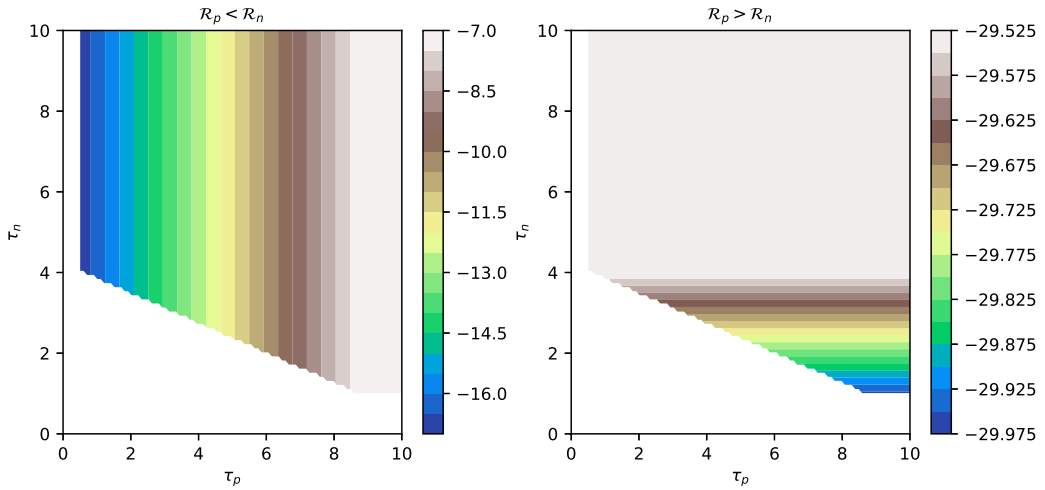

Figure 8 **The effect of service level requirements on long-run average benefit. The parameters are:** $\epsilon = 0.9$, $\phi = 0.9$, $\mathcal{P} = 0.15$, $h_p = 2$, $h_n = 1$, $\theta_p = 0.1$, $\theta_n = 0.1$, $\gamma_p = 2$, $\gamma_n = 1$, $\mu_p^f = 0.75$, $\mu_n^f = 1$, $\mu_n^t = 2$, $\mu_p^t = 1.5$, $N = 4$, $\lambda = 10$, $b_p^f = 10$; **On the left** $b_n^f = 400$ **and on the right** $b_n^f = 100$.

Several directions for future research are worth exploring. One involves considering cases where the anchoring effect not only causes radiologists to spend more time on misclassified cases but also leads them to overlook errors in AI diagnoses. Another potential avenue is extending the model to handle multiple levels of diagnostic outcomes beyond the binary positive and negative classifications considered in this paper. For instance, an AI algorithm provided by Aidoc for detecting acute pulmonary embolism categorizes patients into stat (the most urgent), urgent, and routine categories (Batra et al. 2023). The challenge in incorporating multiple classification outcomes lies in expanding the definitions of false/true positives/negatives to account for varying levels of misclassifications.

Funding. Partial financial support was received from ISF Grant 277/21 and The Bernard M. Gordon Center for Systems Engineering at the Technion.

References

Argon N, Ziya S (2009) Priority assignment under imperfect information on customer type identities. *Manufacturing & Service Operations Management* 11(4):674–693.

Ariel D, Susan P (2024) Aidoc: Building a hospital-centric AI platform. *Harvard Business School Case 624-046* 1–24.

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.

Baltruschat I, Steinmeister L, Nickisch H, Saalbach A, Grass M, Adam G, Knopp T, Ittrich H (2021) Smart chest X-ray worklist prioritization using artificial intelligence: A clinical workflow simulation. *European Radiology* 31:3837–3845.

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

Batra K, Xi Y, Bhagwat S, Espino A, Peshock R (2023) Radiologist worklist reprioritization using artificial intelligence: Impact on report turnaround times for CTPA examinations positive for acute pulmonary embolism. *American Journal of Roentgenology* 221(3):324–333.

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

Dai T, Abràmoff M (2023) Incorporating artificial intelligence into healthcare workflows: Models and insights. *Tutorials in Operations Research: Advancing the Frontiers of OR/MS: From Methodologies to Applications*, 133–155 (INFORMS).

Dai T, Singh S (2024) Using AI as gatekeeper or second opinion: Designing patient pathways for AI-augmented healthcare. *SSRN* (working paper).

Dai T, Tayur S (2022) Designing AI-augmented healthcare delivery systems for physician buy-in and patient acceptance. *Production and Operations Management* 31(12):4443–4451.

Gaube S, Suresh H, Raue M, Merritt A, Berkowitz S, Lermer E, Coughlin J, Guttag JV, Colak E, Ghassemi M (2021) Do as AI say: Susceptibility in deployment of clinical decision-aids. *NPJ Digital Medicine* 4(1):31.

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

He C, Liu W, Xu J, Huang Y, Dong Z, Wu Y, Kharrazi H (2024) Efficiency, accuracy, and health professional's perspectives regarding artificial intelligence in radiology practice: A scoping review. *iRadiology* 2(2):156–172.

Hosny A, Parmar C, Quackenbush J, Schwartz L, Aerts H (2018) Artificial intelligence in radiology. *Nature Reviews Cancer* 18(8):500–510.

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.

Ibanez M, Clark J, Huckman R, Staats B (2018) Discretionary task ordering: Queue management in radiological services. *Management Science* 64(9):4389–4407.

Kamali M, Tezcan T, Yildiz O (2019) When to use provider triage in emergency departments. *Management Science* 65(3):1003–1019.

Lee C, Nagy P, Weaver S, Newman-Toker D (2013) Cognitive and system factors contributing to diagnostic errors in radiology. *American Journal of Roentgenology* 201(3):611–617.

Li D, Shen F, Yin Y, Peng J, Chen P (2013) Weighted youden index and its two-independent-sample comparison based on weighted sensitivity and specificity. *Chinese Medical Journal* 126(6):1150–1154.

Lidal I, Holte H, Vist G (2013) Triage systems for pre-hospital emergency medical services - a systematic review. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 21:1–6.

Little J (1961) A proof for the queuing formula: $L = \lambda W$. *Operations research* 9(3):383–387.

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

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

Monaghan T, Rahman S, Agudelo C, Wein A, Lazar J, Everaert K, Dmochowski R (2021) Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. *Medicina* 57(5):503.

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

Saghafian S, Hopp W, Iravani S, Cheng Y, Diermeier D (2018) Workload management in telemedical physician triage and other knowledge-based service systems. *Management Science* 64(11):5180–5197.

Saghafian S, Hopp W, P VO, Desmond J, Kronick S (2014) Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management* 16(3):329–345.

Singh S, Gurvich I, Van Mieghem J (2024) Feature-driven priority queuing. *SSRN* (working paper).

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

Sun Z, Argon N, Ziya S (2022) When to triage in service systems with hidden customer class identities? *Production and Operations Management* 31(1):172–193.

Van der Zee S, Theil H (1961) Priority assignment in waiting-line problems under conditions of misclassification. *Operations Research* 9(6):875–885.

Van Mieghem J (1995) Dynamic scheduling with convex delay costs: The generalized $c\mu$ 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.

Zhong Z, Cao P, Huang J, Zhou S (2024) Capacity allocation and scheduling in two-stage service systems with multiclass customers. *Manufacturing & Service Operations Management* 26(5):1842–1859.

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.

Online Appendix

Appendix A: An Extension to the Multi-Class Case

In general, radiology departments need to handle multiple disease checks, such as fractures, intracranial bleeding, and pulmonary embolism. In this section, we generalize our diagnostic model to incorporate multiple classes/diseases. We use the subscript $j \in \{1, 2, \ldots, J\}$ to represent Disease j, whose prevalence is $\mathcal{P}_j \in [0, 1]$. The AI algorithm has sensitivity $\epsilon_j \in [0, 1]$ and specificity $\phi_j \in [0, 1]$ for Disease j. The proportions of true positives, false negatives, true negatives, and false positives for Disease j can then be expressed as follows:

$$T_{j,p} = \epsilon_j \mathcal{P}_j, \quad F_{j,n} = (1-\epsilon_j)\mathcal{P}_j, \quad T_{j,n} = \phi_j(1-\mathcal{P}_j), \quad \text{and} \quad F_{j,p} = (1-\phi_j)(1-\mathcal{P}_j).$$

Furthermore, we use the notation $Q_{j,p}$ and $Q_{j,n}$ to represent the suspected positive and suspected negative queues for Disease j, respectively. Figure A.1 illustrates the diagnostic process in the multi-class case.

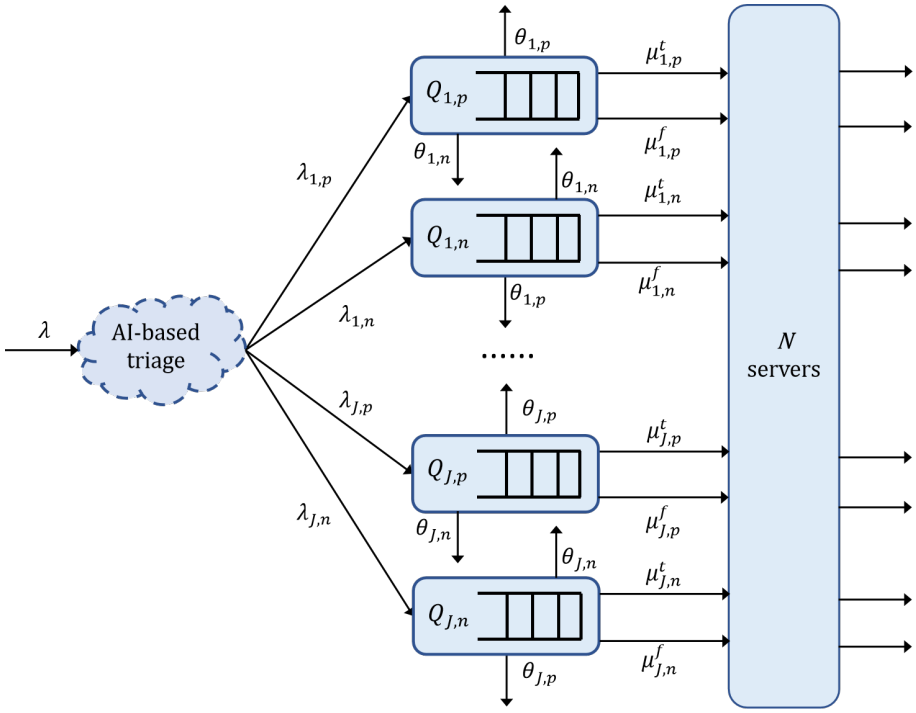

Figure A.1 An Illustration to the multi-class case

Then the arrival rate for queue $Q_{j,p}$ and queue $Q_{j,n}$ respectively are

$$\lambda_{j,p} = \lambda(T_{j,p} + F_{j,p}) \text{ and } \lambda_{j,n} = \lambda(T_{j,n} + F_{j,n}).$$

And in queue $Q_{j,p}$ and $Q_{j,n}$, the proportions of correctly classified positive and negative cases respectively are

$$f_{j,p} = \frac{F_{j,p}}{T_{j,p} + F_{j,p}} \text{ and } f_{j,n} = \frac{F_{j,n}}{T_{j,n} + F_{j,n}}.$$

Let $\mu^t_{j,p}$, $\mu^f_{j,p}$, $\mu^t_{j,n}$, and $\mu^f_{j,n}$ denote the radiologists' service rates for reviewing true positive, false positive, true negative, and false negative initial reports related to Disease j generated by the AI, respectively. Furthermore,

for patients with Disease j, let $h_{j,p}$ and $h_{j,n}$ represent the holding costs per unit time for actual positives and negatives, respectively. Denote $\theta_{j,i}$ as the abandonment rate in queue $\{j,i\}$, $i=p,n$, and $\gamma_{j,i}$ as the abandonment cost per case in $q_{j,i}$.

First, we define the extended version of the weighted average for service and abandonment rates and cost:

$$\bar{\theta}_{j,p}=\theta_{j,p}(1-f_{j,p})+\theta_{j,n}f_{j,p},\quad \bar{\theta}_{j,n}=\theta_{j,n}(1-f_{j,n})+\theta_{j,p}f_{j,n};$$
$$\bar{\mu}_{j,p}=\mu^t_{j,p}(1-f_{j,p})+\mu^f_{j,p}f_{j,p},\quad \bar{\mu}_{j,n}=\mu^t_{j,n}(1-f_{j,n})+\mu^f_{j,n}f_{j,n};$$
$$\bar{c}_{j,p}=(h_{j,p}+\gamma_{j,p}\theta_{j,p})(1-f_{j,p})+(h_{j,n}+\gamma_{j,n}\theta_{j,n})f_{j,p},\quad \bar{c}_{j,n}=(h_{j,p}+\gamma_{j,p}\theta_{j,p})f_{j,n}+(h_{j,n}+\gamma_{j,n}\theta_{j,n})(1-f_{j,n}).$$

Then, the extended fluid optimization problem is the following LP:

$$\begin{aligned}\max_{\bar{z}_p,\bar{z}_n,\bar{q}_p,\bar{q}_n}\quad & \sum_{j=1}^{J}\left(\tilde{R}_p\bar{z}_p+\tilde{R}_n\bar{z}_n\right)\\ \text{s.t.}\quad & 0\le \bar{z}_{j,p}\le \lambda_{j,p}/\bar{\mu}_{j,p},\ j=1,2,...,J,\\ & 0\le \bar{z}_{jn,n}\le \lambda_{j,n}/\bar{\mu}_{j,p},\ j=1,2,...,J,\\ & \bar{z}_{j,p}\ge 0,\bar{z}_{j,n}\ge 0,\ j=1,2,...,J,\\ & \sum_{j=1}^{J}(\bar{z}_{j,p}+\bar{z}_{j,n})\le 1,\end{aligned}$$

where

$$\mathcal{R}_{j,p}:=\frac{\bar{\mu}_{j,p}}{\bar{\theta}_{j,p}}\bar{c}_{j,p}+b^f_{j,p}f_{j,p}\mu^f_{j,p},\quad\text{and}\qquad \mathcal{R}_{j,n}:=\frac{\bar{\mu}_{j,n}}{\bar{\theta}_{n}}\bar{c}_{j,n}+b^f_{j,n}f_{j,n}\mu^f_{j,n}.$$

Using a similar line of reasoning as in the proof of Proposition 1, we can show that following the adjusted $c\mu/\theta$ rule with the extended $\mathcal{R}$ indexes is also asymptotically optimal in the many-server heavy-traffic regime.

Appendix B: Proofs of Theoretical Results

Proof of Lemma 1 Since by definition $\mathcal{P}\in[0,1]$, $\epsilon\in[0,1]$ and $\phi\in[0,1]$, the derivatives of f_p and f_n satisfy the following:

$$\frac{\mathrm{d}f_p}{\mathrm{d}\mathcal{P}}=\frac{\epsilon(\phi-1)}{(\mathcal{P}(\epsilon+\phi-1)-\phi+1)^2}\le 0,\quad \frac{\mathrm{d}f_n}{\mathrm{d}\mathcal{P}}=\frac{\phi(1-\epsilon)}{(\phi-\mathcal{P}(\epsilon+\phi-1))^2}\ge 0;$$

$$\frac{\mathrm{d}f_p}{\mathrm{d}\epsilon}=\frac{(1-\mathcal{P})\mathcal{P}(\phi-1)}{(\mathcal{P}(\epsilon+\phi-1)-\phi+1)^2}\le 0,\quad \frac{\mathrm{d}f_n}{\mathrm{d}\epsilon}=\frac{(\mathcal{P}-1)\mathcal{P}\phi}{(\phi-\mathcal{P}(\epsilon+\phi-1))^2}\le 0;$$

$$\frac{\mathrm{d}f_p}{\mathrm{d}\phi}=\frac{(\mathcal{P}-1)\mathcal{P}\epsilon}{(\mathcal{P}(\epsilon+\phi-1)-\phi+1)^2}\le 0,\quad \frac{\mathrm{d}f_n}{\mathrm{d}\phi}=\frac{(\mathcal{P}-1)\mathcal{P}(1-\epsilon)}{(\phi-\mathcal{P}(\epsilon+\phi-1))^2}\le 0.$$

Q.E.D.

Proof of Proposition 1 We start by defining the following processes for the N-th system: $R^N_i(t)=R^{f,N}_i(t)+R^{t,N}_i(t)$ and $D^N_i(t)=D^{f,N}_i(t)+D^{t,N}_i(t)$. Additionally, the Poisson processes regarding service completion and abandonment satisfy the following:

$$D^{f,N}_i(t)=f_i\tilde{D}^{f,N}_i\left(\int_0^t Z^N_i(s)\mathrm{d}s\right),\ D^{t,N}_i(t)=(1-f_i)\tilde{D}^{t,N}_i\left(\int_0^t Z^N_i(s)\mathrm{d}s\right),\ i=p,n;$$
$$R^{f,N}_i(t)=f_i\tilde{R}^{f,N}_i\left(\int_0^t Q^N_i(s)\mathrm{d}s\right),\ R^{t,N}_i(t)=(1-f_i)\tilde{R}^{t,N}_i\left(\int_0^t Q^N_i(s)\mathrm{d}s\right),\ i=p,n.$$

Note that the Poisson processes $\tilde{R}_i^{f,N}$, $\tilde{R}_i^{t,N}$, $\tilde{D}_i^{f,N}$ and $\tilde{D}_i^{t,N}$ are with rates $\mu_i^{f,N}$, $\mu_i^{t,N}$, $\theta_i^{f,N}$ and $\theta_i^{t,N}$, $i=p,n$, respectively. Then, the linear combination of $\tilde{R}_i^{f,N}$ and $\tilde{R}_i^{t,N}$, $i=p,n$, is also a Poisson processes; namely,

$$\bar{D}_i^N(t)=\tilde{D}_i^N(t)=f_i\tilde{D}_i^{f,N}\left(\int_0^t Z_i^N(s)\mathrm{d}s\right)+(1-f_i)\tilde{D}_i^{t,N}\left(\int_0^t Z_i^N(s)\mathrm{d}s\right)$$

and $\tilde{D}_i^N$ is with rate $\bar{\mu}_i^N:=f_i\mu_i^{f,N}+(1-f_i)\mu_i^{t,N}$. Similarly, we have

$$\bar{R}_i^N(t)=\tilde{R}_i^N(t)=f_i\tilde{R}_i^{f,N}\left(\int_0^t Q_i^N(s)\mathrm{d}s\right)+(1-f_i)\tilde{R}_i^{t,N}\left(\int_0^t Q_i^N(s)\mathrm{d}s\right)$$

and Poisson processes $\tilde{R}_i^N$ is with rate $\bar{\theta}_i^N:=f_i\theta_j^N+(1-f_i)\theta_i^N$, $i,j=p,n$ and $i\neq j$.

Next, we consider the corresponding scaled processes:

$$\begin{aligned}&\bar{X}^N=\frac{1}{N}X^N,\ \bar{A}^N=\frac{1}{N}A^N,\ \bar{Q}^N=\frac{1}{N}Q^N,\ \bar{Z}^N=\frac{1}{N}Z^N,\\&\bar{R}^{f,N}=\frac{1}{N}R^{f,N},\ \bar{R}^{t,N}=\frac{1}{N}R^{t,N},\ \bar{D}^{f,N}=\frac{1}{N}D^{f,N},\ \bar{D}^{t,N}=\frac{1}{N}D^{t,N}.\end{aligned}\tag{B.1}$$

Note that $X^N=(X_p^N(t),X_n^N(t))$, and similarly address the other symbols in (B.1). Then, we have

$$\begin{aligned}&\bar{Z}_p^N(t)+\bar{Z}_n^N(t)\leq 1,\\&\bar{Q}_i^N(t)=\bar{X}_i^N(t)-\bar{Z}_i^N(t),\ i=p,n,\\&\bar{X}_i^N(t)=\bar{X}_i^N(0)+\bar{A}_i^N(t)-\bar{R}_i^{f,N}(t)-\bar{R}_i^{t,N}(t)-\bar{D}_i^{f,N}(t)-\bar{D}_i^{t,N}(t),\ i=p,n.\end{aligned}$$

Similar to Atar et al. (2010), we define the following events, given T and $\delta\in(0,1)$:

$$\begin{aligned}E_A^n&=\left\{\max_{i=p,n}\sup_{t\in[0,T]}\left|\bar{A}_i^N(t)-\frac{\lambda_i^N}{N}t\right|<\delta\right\},\\E_D^n&=\left\{\max_{i=p,n}\sup_{t\in[0,T]}\left|\frac{\tilde{D}_i^N(Nt)}{N}-\bar{\mu}_i^N t\right|<\delta\right\},\\E_R^n&=\left\{\max_{i=p,n}\sup_{t\in[0,KT]}\left|\frac{\tilde{R}_i^N(Nt)}{N}-\bar{\theta}_i^N t\right|<\delta\right\}.\end{aligned}$$

and $E^N=E_{\delta,T}^N:=E_A^n\cap E_D^n\cap E_R^n$, where $K=K(T)=(c_\lambda T)/2+M+1$ and $c_\lambda=\sup_N\max_{i\in\{p,n\}}(\lambda_i^N/N)<\infty$.

The following Lemma B.1 is similar to Lemma 2.1 in Atar et al. (2010), where the main difference is that we consider linear combinations of Poisson processes (that is, $\bar{D}_i^N(t)$ and $\bar{R}_i^N(t)$) rather than a single Poisson process.

Lemma B.1 *Let $T>0$ and $\delta\in(0,1)$ be given. Fix a sequence of policies $\pi^N\in\Omega^N, N\in\mathbb{N}$. Then on the event E^N, one has, for every N,*

$$\left|\bar{D}_i^N(t)-\bar{\mu}_i^N\int_0^t\bar{Z}_i^N(s)\,\mathrm{d}s\right|\vee\left|\bar{R}_i^N(t)-\bar{\theta}_i^N\int_0^t\bar{Q}_i^N(s)\,\mathrm{d}s\right|<\delta,\quad i=p,n;\ t\in[0,T],\tag{B.2}$$

and, with $K=K(T)$ as above,

$$\int_0^T\bar{Q}_i^N(s)\,\mathrm{d}s\leq KT,\quad i=p,n.$$

Furthermore, $\mathbb{P}(E^N)\to 1$ as $N\to\infty$.

Our proof follows a similar line of argument as Proposition 4.1 and Theorem 5.1 in Atar et al. (2010). Before formally presenting the proof of the proposition, we summarize the main adjustments made with respect to Atar et al. (2010):

1. In our study, patients of mixed types coexist within the same queue. We, therefore, take this into account when incorporating abandonments and service completion. For example, for Queue p (Queue n), servers provide services for true positives (true negatives) and false positives (false negatives) at different rates. We thus use $\bar{D}_i^N(t)$ to accommodate all the service completions involving heterogeneous patients by time t for Queue i, $i=p,n$ (rather than $D_i^N(t)$ in Atar et al. (2010), which related to a homogeneous queue). We show that this process, with a weighted average rate $\bar{\mu}_i^N$, $i=p,n$ satisfied $\mathbb{P}(E_D^N)=1$ as $N\to\infty$ (see Lemma B.1 and its proof for details). Similarly, we apply the same approach to heterogeneous patients who abandon the queue while waiting.

2. Our objective function includes benefit terms directly tied to allocation decisions. The normalized average benefit $B_{N,T}(\pi^N)$, is
$$\frac{1}{NT}\mathbb{E}\left[\int_0^T \left(b_p^N Z_p^N(t)+b_n^N Z_n^N(t)\right)\mathrm{d}t\right].$$
In the fluid objective function, this corresponds to $b_p^f f_p \mu_p^f \bar{z}_p + b_n^f f_n \mu_n^f \bar{z}_n$.

3. Our problem is formulated as a benefit maximization problem, whereas the problem studied in Atar et al. (2010) is a minimization problem. Therefore, we first transform our benefit maximization problem into an equivalent cost minimization problem. Following this transformation, the proof proceeds similarly to that in Atar et al. (2010).

Proof of Proposition 1(i). We start by providing the following notation
$$C_{N,T}(\pi^N) := -B_{N,T}(\pi^N) = -\frac{1}{NT}\mathbb{E}\left[\int_0^T \left[b_p^N Z_p^N(t)+b_n^N Z_n^N(t)-\bar{c}_p^N Q_p^N(t)-\bar{c}_n^N Q_n^N(t)\right]\mathrm{d}t\right],$$
$V_{N,T}=-U_{N,T}=\inf_{\pi^N\in\Omega^N} C_{N,T}(\pi^N)$, and $V^*=-U^*=\bar{c}q^*-bz^*$.

To prove $\limsup_{T\to\infty}\limsup_{N\to\infty} U_{N,T}\le U^*$, it suffices to prove that
$$v:=\liminf_{T\to\infty}\liminf_{N\to\infty} V_{N,T}\ge V^*.$$

Let ε be an arbitrary and positive number. Fix $T\ge c_0/\varepsilon$ and $c_0\ge M$, and we can find a sequence of policies $\pi^N\in\Omega^N$ to ensure $\liminf_{N\to\infty} C_{N,T}(\pi^N)\le v+2\varepsilon$. Let $\rho(\varepsilon)$ be a function which vanishes at 0 when $\varepsilon=0$. Next, we just need to show that
$$\liminf_{N\to\infty} C_{N,T}(\pi^N)\ge V^*-\rho(\varepsilon) \tag{B.3}$$
is true, and then we have
$$V^*-\rho(\varepsilon)\le\liminf_{N\to\infty} C_{N,T}(\pi^N)\le v+2\varepsilon.$$
That is, $V^*\le v$ because ε is arbitrary.

Next, we show that (B.3) holds. Denote by $q_i^N=\frac{1}{T}\int_0^T \bar{Q}_i^N(s)\mathrm{d}s$ and $z_i^N=\frac{1}{T}\int_0^T \bar{Z}_i^N(s)\mathrm{d}s$, $i=p,n$. Following the proof of Proposition 4.1 by Atar et al. (2010), there is a $N_0=N_0(\varepsilon)$ satisfying that
$$\bar{c}q^N-bz^N\ge V^*-\rho(\varepsilon)$$
for all $N>N_0$ based on the event E^N. Therefore, we have
$$C_{N,T}(\pi^N)=\mathbb{E}\left[\bar{c}q^N-bz^N\right]\ge\mathbb{E}\left[\mathbf{I}_{E^N}(\bar{c}q^N-bz^N)\right]$$

and

$$\liminf_{N\to\infty} C_{N,T}(\pi^N) \geq (V^* - \rho(\varepsilon)) \liminf_{N\to\infty} \mathbb{P}(E^N) = V^* - \rho(\varepsilon)$$

by $\mathbb{P}(E^N) \to 1$ as $N \to \infty$ in Lemma B.1, where $\mathbf{1}_{(\cdot)}$ is an indicator function (that is, $\mathbf{1}_{(\cdot)} = 1$ when event $(\cdot)$ holds, otherwise 0). This proves that (B.3) holds, and concludes the proof of Proposition 1(i).

Proof of Proposition 1(ii). This proof, which includes two parts, is similar to the proof of Theorem 5.1 in Atar et al. (2010). In the first step, we establish that

$$u_p := \liminf_{T\to\infty} \liminf_{N\to\infty} B_{N,T}(\pi_p^N) \geq U^*. \tag{B.4}$$

– ***Step 1.*** First, we have

$$\begin{aligned} B_{N,T}(\pi_p^N) = &\frac{1}{T}\mathbb{E}\left[\mathbf{1}_{E^N}\left[\int_0^T \left[b^N \bar{Z}^N(t) - \bar{c}^N \bar{Q}^N(t)\right] \mathrm{d}t\right]\right] \\ &+ \frac{1}{T}\mathbb{E}\left[\mathbf{1}_{(E^N_{\delta,T})^c}\left[\int_0^T \left[b^N \bar{Z}^N(t) - \bar{c}^N \bar{Q}^N(t)\right] \mathrm{d}t\right]\right]. \end{aligned} \tag{B.5}$$

Based on Lemma 5.1 in Atar et al. (2010), there exists T_ϵ satisfying

$$\sup_{t\in[T_\epsilon,T]} \|\bar{Z}^N(t) - z^*\| \vee \|\bar{Q}^N(t) - q^*\| < \varepsilon,$$

for every positive ε and $N > N_0$ on the event $E^N_{\delta,T}$. Note that δ and N_0 rely on ε and T. Therefore, for the first term on the right-hand side of (B.5), we have

$$\begin{aligned} &\frac{1}{T}\mathbb{E}\left[\mathbf{1}_{E^N}\left[\int_0^T \left[b^N \bar{Z}^N(t) - \bar{c}^N \bar{Q}^N(t)\right] \mathrm{d}t\right]\right] \\ =&\frac{1}{T}\mathbb{E}\mathbf{1}_{E^N}\left[\left[\int_0^{T_\varepsilon} \left[b^N \bar{Z}^N(t) - \bar{c}^N \bar{Q}^N(t)\right] \mathrm{d}t\right] + \left[\int_{T_\varepsilon}^T \left[b^N \bar{Z}^N(t) - \bar{c}^N \bar{Q}^N(t)\right] \mathrm{d}t\right]\right] \\ \geq &-\frac{||\bar{c}^N||K(T_\varepsilon)T_\varepsilon}{T} + \frac{||b^N||T_\varepsilon}{T} - \bar{c}^N q^* + b^N z^* - ||\bar{c}^N||\varepsilon - ||b^N||\varepsilon, \end{aligned} \tag{B.6}$$

where $||\mathbf{x}||$ represents the 1-norm of $\mathbf{x}$. The inequality in (B.6) is from Lemma B.1. For the second term on the right-hand side of (B.5), we have

$$\frac{1}{T}\mathbb{E}\left[\mathbf{1}_{(E^N_{\delta,T})^c}\left[\int_0^T \left[b^N \bar{Z}^N(t) - \bar{c}^N \bar{Q}^N(t)\right] \mathrm{d}t\right]\right] \geq \mathbb{E}\left[\mathbf{1}_{(E^N_{\delta,T})^c}(||b^N \bar{Z}^N - \bar{c}^N \bar{Q}^N||^*_T)\right], \tag{B.7}$$

where $||f||^*_T$ represents $\inf_{0\leq t\leq T} ||f(t)||$. Notice that due to the uniform integrability of $||\bar{X}^N||^*_T$ and $\bar{Z}^N_p(t) + \bar{Z}^N_n(t) \leq 1$, $||b^N \bar{Z}^N - \bar{c}^N \bar{Q}^N||^*_T$ is also uniformly integrable. Then, from (B.5), (B.6), (B.7) and Lemma B.1, considering $N \to \infty$ and $T \to \infty$, we have

$$u_p \geq -\bar{c}q^* + bz^* - ||\bar{c}||\varepsilon - ||b||\varepsilon.$$

Since ε can be arbitrarily small, we get that $u_p \geq U^*$.

– ***Step 2.*** In Proposition 1(i), we established that $\limsup_{T\to\infty} \limsup_{N\to\infty} U_{N,T} \leq U^*$ as well as

$$\limsup_{T\to\infty} \limsup_{N\to\infty} U_{N,T} \geq \limsup_{T\to\infty} \limsup_{N\to\infty} B_{N,T}(\pi_p^N), \tag{B.8}$$

and thus we have

$$\limsup_{T\to\infty} \limsup_{N\to\infty} B_{N,T}(\pi_p^N) \leq U^*. \tag{B.9}$$

Combining (B.9) and (B.4), we obtain

$$\limsup_{T\to\infty}\limsup_{N\to\infty} B_{N,T}(\pi_p^N) = \liminf_{T\to\infty}\liminf_{N\to\infty} B_{N,T}(\pi_p^N) = U^*. \tag{B.10}$$

Consequently, from (B.8), (B.10) and Proposition 1(i), we get

$$\limsup_{T\to\infty}\limsup_{N\to\infty} U_{N,T} = U^*.$$

Finally,

$$\limsup_{T\to\infty}\limsup_{N\to\infty} U_{N,T} = \liminf_{T\to\infty}\liminf_{N\to\infty} B_{N,T}(\pi_p^N) = U^*,$$

which concludes the proof of Proposition 1(ii). Q.E.D.

Proof of Lemma B.1 For a fixed i, the inequality $\left|\bar{D}_i^N(t) - \bar{\mu}_i^N \int_0^t \bar{Z}_i^N(s)\,\mathrm{d}s\right| < \delta$ holds for every t satisfying $\int_0^t \bar{Z}_i^N(s)\mathrm{d}s < T$ by the definition of event E_D^n. Considering $\bar{Z}_p^N(t) + \bar{Z}_n^N(t) \le 1$, thus $\int_0^t \bar{Z}_i^N(s)\mathrm{d}s \le T$ naturally holds.

Similarly, the inequality $\left|\bar{R}_i^N(t) - \bar{\theta}_i^N \int_0^t \bar{Q}_i^N(s)\,\mathrm{d}s\right| < \delta$ holds for every t satisfying $\int_0^t \bar{Q}_i^N(s)\mathrm{d}s \le KT$ by the definition of event E_R^n. Next, we prove $\int_0^t \bar{Q}_i^N(s)\mathrm{d}s \le KT$. According to

$$\bar{Q}_i^N(t) = \bar{X}_i^N(t) - \bar{Z}_i^N(t) = \bar{X}_i^N(0) + \bar{A}_i^N(t) - \bar{R}_i^N(t) - \bar{D}_i^N(t) - \bar{Z}_i^N(t),$$

we have $\bar{Q}_i^N(t) \le \bar{X}_i^N(0) + \bar{A}_i^N(t)$ because $\bar{Z}_i^N(t)$, $\bar{R}_i^N(t)$, and $\bar{D}_i^N(t)$ are non-negative. Due to $\bar{X}_i^N(0) \le M$ (the Assumption 1(ii)) and $\bar{A}_i^N(t) \le \delta + \bar{\lambda}_i^N t$ (the definition of event E_A^N), we then obtain

$$\int_0^t \bar{Q}_i^N(s)\mathrm{d}s \le \int_0^t (M + \delta + \lambda_i^N t)\mathrm{d}s = Mt + \delta t + \frac{1}{2}\bar{\lambda}_i^N t^2 \le KT,$$

Consequently, $\left|\bar{R}_i^N(t) - \bar{\theta}_i^N \int_0^t \bar{Q}_i^N(s)\mathrm{d}s\right| < \delta$ holds and inequality (B.2) is true.

Finally, since $\bar{A}_i^N(t) = \frac{1}{N}A_i^N(t)$, where $A_i^N(t)$ is a Poisson process with rate λ_i^N, and $\lim_{N\to\infty}\frac{1}{N}\lambda_i^N = \lambda_i$, it follows that $\mathbb{P}(E_A^N) \to 1$ as $N \to \infty$. Furthermore, because both $\tilde{D}_i^{f,N}(t)$ and $\tilde{D}_i^{t,N}(t)$ are Poisson processes, the linear combination of them, $\tilde{D}_i^N(t)$, is also a Poisson process with rate $\bar{\mu}_i^N$. Meanwhile, we have,

$$\lim_{N\to\infty} \bar{\mu}_i^N = \lim_{N\to\infty} (f_i\mu_i^{f,N} + (1-f_i)\mu_i^{t,N}) = f_i\mu_i^f + (1-f_i)\mu_i^t = \bar{\mu}_i.$$

Thus, $\mathbb{P}(E_D^N) \to 1$ as $N \to \infty$. For a similar reason, we have $\mathbb{P}(E_R^N) \to 1$ as $N \to \infty$. Therefore, $\mathbb{P}(E^N) \to 1$ as $N \to \infty$ considering that events E_A^N, E_D^N, E_R^N are mutually independent.

Q.E.D.