Article
http://justc.ustc.edu.cn
Received: March 30, 2023; Accepted: May 20, 2023

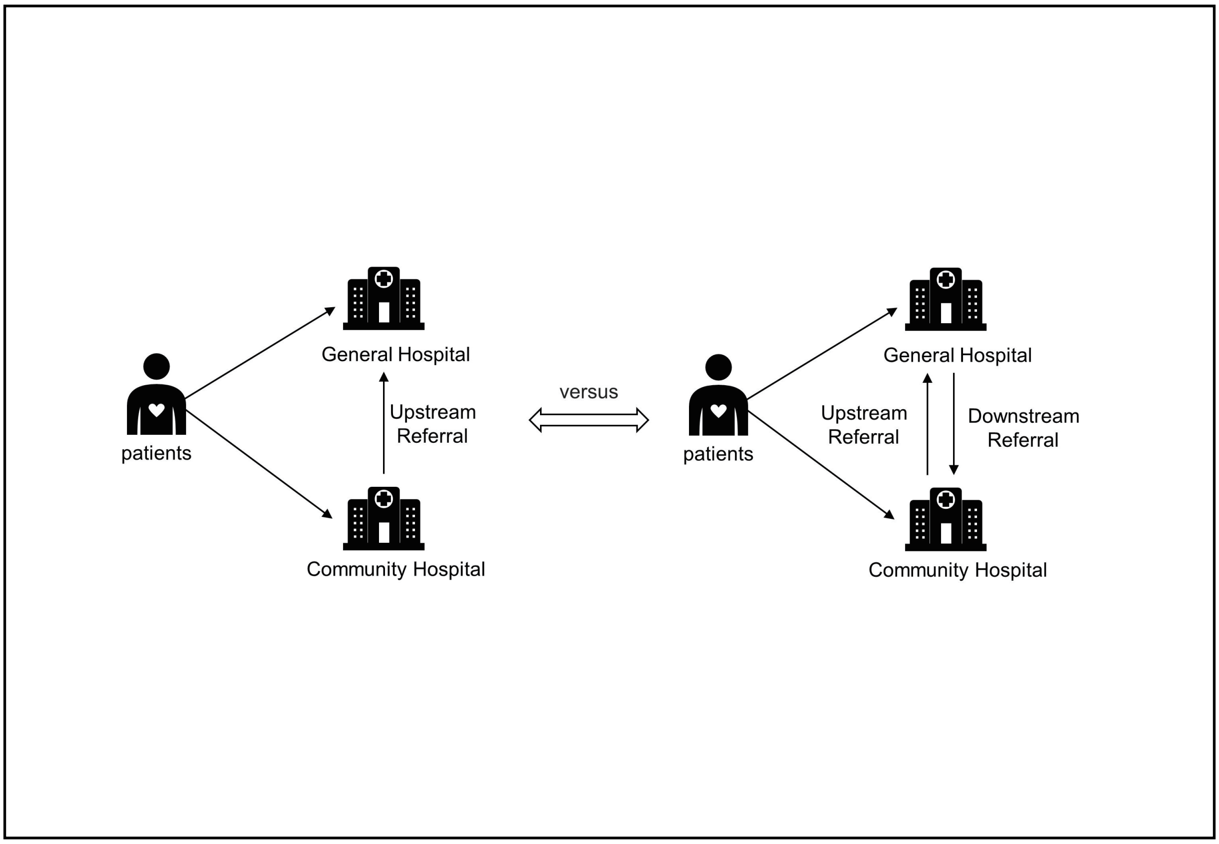
Referral coordination mechanisms in healthcare management

Xiaobei Shen[1], Chunfang Yuan[1] ✉, Xiangyin Kong[1], and Yue Wang[2]

[1]School of Management, University of Science and Technology of China, Hefei 230026, China;
[2]Business School, The Hong Kong University of Science and Technology, Hong Kong 999077, China

✉Correspondence: Chunfang Yuan, E-mail: fangjs@mail.ustc.edu.cn

Graphical abstract



A comparison of the efficacy between one-way referral and two-way referral mechanisms in a two-tier hospital system.

Public summary

- When the capacity cost rate of the treatment procedure in the general hospital is not significantly higher than that in the community healthcare center, the one-way referral mechanism can improve system profit and is superior to the two-way referral mechanism.
- A grand coalition involving all cooperating hospitals leads to the highest system profit, creating a Pareto-improving outcome that benefits both types of hospitals and patients.
- The two-way referral mechanism is unequivocally superior to both the one-way referral mechanism and the non-coordinated baseline when considering heterogeneous patients and the referral-scheme-dependent arrival rate in a congested system.

Citation: Shen X B, Yuan C F, Kong X Y, et al. Referral coordination mechanisms in healthcare management. *JUSTC*, **2025**, 55(8): 0803. CSTR: 32290.14.JUSTC-2023-0052. DOI: 10.52396/JUSTC-2023-0052

JUSTC
Article
http://justc.ustc.edu.cn
Received: March 30, 2023; Accepted: May 20, 2023

Referral coordination mechanisms in healthcare management

Xiaobei Shen[1], Chunfang Yuan[1] ✉, Xiangyin Kong[1], and Yue Wang[2]

[1]*School of Management, University of Science and Technology of China, Hefei 230026, China;*
[2]*Business School, The Hong Kong University of Science and Technology, Hong Kong 999077, China*

✉Correspondence: Chunfang Yuan, E-mail: fangjs@mail.ustc.edu.cn

Cite This: *JUSTC*, **2025**, 55(8): 0803 (13pp)   Read Online   SI Supporting Information

Abstract: Referral systems are widely used to coordinate heterogeneous healthcare providers (e.g., general hospitals (GHs) and community healthcare centers (CHCs)) for improved efficiency. This paper investigates referral coordination within a typical two-tiered system centered around a general hospital (GH) and a community healthcare center (CHC). Specifically, we compare the coordination value of two prevalent mechanisms: one-way referral and two-way referral. We develop a queueing-theoretic model to derive optimal capacity and pricing decisions for the GH and the CHC under each mechanism and then evaluate their relative effectiveness, with key metrics including total system profit, healthcare service prices, and patient waiting times. Our base model yields two key findings. First, counterintuitively, under certain conditions, the one-way referral mechanism can outperform both the two-way mechanism and a non-coordinated baseline. Second, within the one-way framework, full cooperation between the GH and CHCs can lead to a Pareto improvement, benefiting all stakeholders (i.e., the GH, the CHCs, and the patients). This finding is based on an analysis extended to a system of one GH and multiple CHCs, where we show how a profit allocation scheme can be designed to foster such cooperation. Further analysis of a congested system with referral-dependent arrival rates reveals that the two-way mechanism becomes unequivocally superior. Finally, numerical studies confirm that optimal profits across all scenarios increase with the arrival rates of both severe patients in the GH and common patients in the CHCs.

Keywords: referral mechanism; capacity investment; pricing; coordination

CLC number: R197; C935   **Document code:** A

1 Introduction

Effective management of patient flow is critical to improving the efficiency of a healthcare system. A typical system comprises two tiers of hospitals: general hospitals (GHs) and community healthcare centers (CHCs). GHs, equipped with superior medical resources, provide more comprehensive and thorough treatments. In contrast, CHCs, despite their limited resources, focus on offering more convenient and faster routine care. Due to this inherent difference in service capability and quality, GHs generally enjoy greater public trust than CHCs, leading patients to prefer GHs even for minor care needs[1,2]. Consequently, GHs often experience overcrowding, while CHCs remain underutilized. The mismatch between resources and demand not only reduces systemic efficiency but also imposes burdens on patients, such as longer waiting times.

To alleviate the aforementioned resource-demand mismatch, referral systems are widely used to coordinate the two tiers of hospitals. Based on the direction of patient transfer, referrals can be categorized into one-way and two-way mechanisms. In Western systems, such as those in the United Kingdom and the United States, healthcare is typically structured in two levels: primary care physicians (e.g., family doctors) who act as gatekeepers for initial contact and referral, and specialists who provide advanced care. A patient first visits a primary care physician, who then assesses the condition and decides whether to make a referral. Patients with non-severe conditions are treated by the primary physician, while those requiring specialized care are referred to a specialist. The one-way referral mechanism examined in this paper is analogous to this gatekeeper system but is not identical. The two-way referral mechanism, in contrast, implements downward referral (from higher- to lower-level providers) in addition to upward referral[3]. In China and many other Asian countries, however, public preference traditionally favors general hospitals for nearly all conditions, even minor ones such as the common cold. This leads to overcrowding at general hospitals, causing delays for patients with critical or incurable illnesses who need timely, advanced care. To free up capacity for these severe cases, general hospitals may transfer recovering or less ill patients to primary healthcare providers after emergency treatment. Consequently, the public often perceives that the two-way referral system can save patient waiting times and medical costs while creating greater value for providers compared to the one-way system.

Although these referral practices are widely observed, it remains unclear which mechanism is preferable under given conditions. A deeper understanding of the coordination value provided by each mechanism is therefore both necessary and important.

In this study, we first develop a queueing model to capture a two-tier healthcare system where hospitals differ in their medical service levels and locations. Each hospital decides on

CSTR: 32290.14.JUSTC-2023-0052
DOI: 10.52396/JUSTC-2023-0052
JUSTC, **2025**, 55(8): 0803

its service capacity and price. We characterize the optimal capacity investment and pricing decisions for each hospital under both one-way and two-way referral mechanisms. Subsequently, we explore the value of coordination by comparing the optimal decisions and system profit under a coordinated system with those under a non-coordinated baseline. Our main findings are summarized as follows.

First, we find that the one-way referral mechanism can improve system profit and outperform the two-way mechanism when the capacity cost rate of the treatment procedure in the GH is not significantly higher than that in the CHC. This result is counterintuitive, given that the one-way mechanism is a special case of the two-way mechanism. The explanation lies in the trade-off between capacity pooling and operational flexibility. In our model, the profit-enhancing effect of capacity pooling dominates that of flexibility.

Second, focusing on the one-way mechanism, we find that a grand coalition involving all cooperating hospitals leads to the highest system profit, creating a Pareto-improving outcome that benefits both types of hospitals and patients. For hospitals, the grand coalition reduces the investment in medical resources required by treatment procedures. This reduction becomes more pronounced with higher patient arrival rates. For patients, coordination results in better service and shorter waiting times.

We then extend the analysis to a congested system with heterogeneous patients and referral-scheme-dependent arrival rates. With congestion, capacity planning is no longer a decision variable (i.e., each hospital's capacity is exogenous). Patient heterogeneity stems from differing health conditions, which influence medical service pricing. Given that a patient's health condition is the dominant factor, we assume price is a function thereof. Under this setting, we find that the two-way referral mechanism is always superior to both the one-way mechanism and the non-coordinated baseline. Through numerical simulation, we investigate the impact of arrival rates on the total system profit under three scenarios: no coordination, one-way coordination, and two-way coordination. The optimal profit in all three scenarios increases with the arrival rate of severe patients in the GH and the arrival rate of common patients in the CHC. However, it does not increase monotonically with the arrival rate of common patients in the GH.

The remainder of the paper is organized as follows. The relevant literature is reviewed in Section 2. Section 3 introduces the settings of the stable queueing system and characterizes the optimal strategies for capacity investment and pricing in the three systems. The relative merits of the referral mechanisms and the value of coordination are explored in Section 4. A congested system with heterogeneous patients and scheme-dependent arrival rates is examined in Section 5. Concluding remarks are provided in Section 6.

2 Literature review

This study is related to three streams of literature on healthcare service systems: healthcare referral, capacity/pricing level, and performance comparison between systems in health economics.

The first stream of literature mainly explores the referral process in two-tier healthcare systems. Ref. [4] describes a two-level service system in which the first-level CHC serves as a gatekeeper for the second-level GH. Similar literature has mushroomed but can be grouped into two categories: referral rate decisions and incentive mechanism problems. Ref. [4] investigates the optimal referral threshold in a principal-agent model and analyzes the impact of referral-based incentives on the gatekeeper's referral decisions. Taking the impact of waiting times into account, Ref. [5] extends the model proposed by Ref. [4] and investigates the optimal staffing levels and referral rates. These studies focus on the upstream referral process, i.e., patients are diverted from the CHC to the GH. However, in countries such as China, referral flow from the GH to the CHC is more common to alleviate the unbalanced utilization of medical resources. Ref. [2] captures the patient flow related to reverse referrals, where hospitals determine the referral rate for their patients to maximize their own profit, and examines the implementability and effectiveness of the partnership by comparing the two identified rates. Ref. [6] considers the two-way referral problem and designs some contracts to achieve the optimal mutual referral strategy. These papers mainly study the referral situation using queueing models, where service providers seek to maximize/minimize their own profits/costs. Ref. [7] focuses on the dynamic interactions among the GH, the CHC, and the patients by establishing a three-stage Stackelberg game. Instead of focusing on the equilibrium decisions in a decentralized system, we explore the optimal strategies of capacity investment and pricing in a centralized system resulting from a grand coalition.

The second stream is on decisions regarding capacity level and pricing. Most relevant studies have focused on the capacity and pricing strategy of one monopolist service provider among heterogeneous or homogeneous delay-sensitive customers[8–15]. Another stream considers the capacity level and pricing problem involving two or more asymmetric providers[15–19]. Refs. [14, 17] analyze the cooperation of service providers in capacity and price with the objective of minimizing the sum of operating costs or maximizing the joint payoff. Ref. [15] investigates the optimal capacity decision of the GH for referral patients in interaction with the CHC through a two-stage game theoretical approach. All these papers mainly focus on the capacity level and pricing, with the objective of maximizing the provider's utility. However, we explore not only the optimal strategies of capacity investment and pricing for systems with/without coordination but also the performance of different referral mechanisms.

At the same time, there is an extensive body of research that compares benefits between different healthcare service systems. Ref. [20] empirically compares patients' satisfaction with GP services in countries with different direct accessibility scales. Ref. [21] builds a model to compare the performance of a gatekeeping system with that of a nongatekeeping system, where performance is measured by an index of quality net of provision costs. The results show that gatekeeping is superior wherever GP incentives matter. Considering customer delay, Ref. [22] explores the optimal capacity level for primary hospitals in both settings and evaluates the relative

CSTR: 32290.14.JUSTC-2023-0052
DOI: 10.52396/JUSTC-2023-0052
JUSTC, **2025**, 55(8): 0803

merit of each setting in terms of social welfare. Our work differs from these studies in that we do not focus on the direct accessibility scale, but pay attention to the relative merits of different referral mechanisms.

3 Model setting

Consider a system involving a GH with more adequate healthcare resources and higher service quality and a CHC with limited resources and lower quality. A more general system with multiple CHCs can be found in S2 of Supporting Information. Table 1 is a summary of notation.

We use hospital g to denote the GH and hospital c to denote the CHC. Following Ref. [4], we consider that each hospital has two queues, one for the diagnosis procedure and the other for the treatment procedure. For simplicity, we use M/M/1 queues to model each procedure and $k=1$ (resp. $k=2$) to denote the diagnosis procedure (resp. the treatment procedure), which has been widely applied to study stochastic service processes in healthcare operations research[23–28]. We assume that the diagnosis procedures in the GH and CHC are identical. This aligns with practical observation: Diagnosis procedures in hospitals are typically performed using comparable medical equipment, resulting in a homogeneous experience for patients. In contrast, the treatment procedure delivers significantly different value to patients, owing to factors such as varying levels of service. For model tractability, we assume that patients are not able to learn the state (including the price, capacity, waiting times) of each hospital in the system, and they prefer the specific hospital near their homes because they are more familiar with the professionals (doctors, nurses) and the layout of that hospital or due to the high travel costs. We thus make the following assumption throughout the base model.

Hypothesis 1. Patients in location κ, $\kappa \in \mathcal{N} = \{g, c\}$, have a medical diagnosis only in their local hospital κ.

In fact, this assumption holds when the distance between the two-level hospitals is sufficiently large, i.e., $d \geq \tilde{d}$, where the threshold $\tilde{d}$ is characterized in S1 of Supporting Information.

For hospital κ, $\kappa \in \mathcal{N}$, patients arrive according to a Poisson process with the rate λ_k^κ in the procedure k. As we assume that patients have a medical diagnosis only in their local hospital, λ_k^κ is exogenous. In each procedure, a hospital serves patients on a first-come, first-served (FCFS) basis. The service time of the procedure k is exponentially distributed with the mean $1/\mu_k^\kappa$, where μ_k^κ ($\mu_k^\kappa > 0$, $k = 1, 2$) is a scaling parameter corresponding to the service capacity. Then, the expected delay for patients, denoted by $\mathbb{E}(w_k)$, is $\mathbb{E}(w_k) = \dfrac{1}{\mu_k^\kappa - \lambda_k^\kappa}$.

To facilitate the derivation, we consider the case where the marginal cost is constant in this study. Thus, the cost of

Table 1. List of notations.

κ	$\kappa = \{g, c\}$, g denotes the GH and c denotes the CHC
k	$k = 1$ (resp. $k = 2$) denotes diagnosis (resp. treatment) procedure
$G_1(G_2)$	Constant utility of discharged (diagnosed) patients
$H(L)$	Utility for a patient receiving treatment from the GH (CHC)
d	Disutility of distance
$\mu_k^\kappa/\mu_k^{\kappa S}/\mu_k^{\kappa T}$	Service capacity of procedure k in hospital κ in non-coordinated system/one-way referral system/two-way referral system
h	Waiting cost rate of patients
$\mathbb{E}(w_k^\kappa)$	Expected waiting time for hospital κ in procedure k
p_k^κ	Price of procedure k in hospital κ
λ_k^κ	Arrival rate of procedure k in hospital κ in the base model
c_k^κ	Capacity cost rate of procedure k in hospital κ
q_κ	Proportion of patients continuing into treatment procedure in hospital κ
r	Referral threshold
$\lambda^d(r)$	Downstream referral rate
$\lambda^u(r)$	Upstream referral rate
π_κ	Expected profit earned by hospital κ in non-coordinated system
$\pi^S(\pi^T)$	Total profit of the one-way (two-way) referral system
$\mathcal{N}$	The grand coalition
$\mathcal{J}$	A coalition
$\bar{\pi}_\kappa$	Allocated profit of hospital κ under the grand coalition
λ_s	Arrival rate of severe patients in the expanded model
$\lambda_g(\lambda_c)$	Arrival rate of patients visiting the GH (CHC) for first diagnosis
$\mu_g(\mu_c)$	Total capacity of the GH (CHC) in the expanded model
$t_1(t_2)$	Capacity of upstream (downstream) green channel
$\pi_n/\pi_s/\pi_t$	Total profit of the non-coordinated system/one-way referral system/two-way referral system

CSTR: 32290.14.JUSTC-2023-0052
DOI: 10.52396/JUSTC-2023-0052
JUSTC, 2025, 55(8): 0803

hospital κ incurred by the service capacity in procedure k is represented by $c_k\mu_k^\kappa$. As mentioned above, the diagnosis procedures in GH and CHC are identical. Therefore, we make the following assumption.

Assumption 2. The capacity cost rate of the diagnosis procedure in the GH is equivalent to that of the CHC, i.e., $c_1^g = c_1^c = c_1$. The capacity cost rates of the treatment procedure in the GH and CHC satisfy the property $c_2^g \geq c_2^c$. In addition, there exists an upper bound $\bar{c}_2^g$ satisfying the following inequalities:

(i)

$$\bar{c}_2^g q_c\lambda_c + 2\sqrt{h\bar{c}_2^g(q_g\lambda_g + q_c\lambda_c)} - 2\sqrt{h\bar{c}_2^g q_g\lambda_g} \leq q_c\lambda_c(H - L - d + c_2^c) + 2\sqrt{hc_2^c q_c\lambda_c};$$

(ii)

$$\sqrt{h\bar{c}_2^g(q_g\lambda_g + q_c\lambda_c)} - \sqrt{h\bar{c}_2^g(1-r)(q_g\lambda_g + q_c\lambda_c)} + \bar{c}_2^g r(q_g\lambda_g + q_c\lambda_c) \leq \sqrt{hc_2^c r(q_g\lambda_g + q_c\lambda_c)} + c_2^c r(q_g\lambda_g + q_c\lambda_c);$$

(iii) $\bar{c}_2^g \leq c_2^c + \dfrac{q_g\lambda_g}{q_c\lambda_c}c_2^c$.

This assumption indicates that primary care is typically less expensive than specialist care[29]. We also assume that the capacity cost rate of the treatment procedure (e.g., the unit cost of medical equipment) in the GH is moderately higher than that in the CHC. In practice, the capacity cost rate in the GH is unlikely to be at either extreme (very high or very low), making this a reasonable simplification. These assumptions serve to simplify the analysis and yield tractable, insightful theoretical results.

For this system, we analyze two prevalent referral mechanisms. In the one-way referral mechanism, the CHC transfers diagnosed patients to the GH. Under the two-way mechanism, we additionally allow the GH to transfer diagnosed patients with less severe conditions back to the CHC.

Our objective is to determine the service price and capacity of each hospital so as to maximize the total system profit, subject to patient utility constraints. Clearly, a patient from region κ will visit hospital κ for diagnosis only if their utility is non-negative. In our model, the utility of a diagnosed patient depends on the following factors: 1) the medical service level of the hospital; 2) the distance between the GH and the CHC; 3) the waiting times before being treated; and 4) the service price in each procedure. For patients discharged after diagnosis, their utility depends on: 1) the waiting times in the diagnosis procedure, and 2) the service price in the diagnosis procedure. Let q_κ $(0 \leq q_\kappa \leq 1)$ denote the diagnosis rate of hospital κ. Since patients are assumed to be homogeneous across hospitals, the probability of a patient being diagnosed is $1 - q_\kappa$. We assume that discharged patients receive a constant utility, G_1, while diagnosed patients receive a constant utility, G_2, upon confirming their health status. Moreover, we assume that the utility for a patient receiving treatment from hospital $\kappa = g$ (i.e., the GH) is H, and that from hospital $\kappa = c$ (i.e., the CHC) is L. If a patient seeks treatment at a non-local hospital, they incur an additional disutility d, representing the distance or travel cost between the two tiers of hospitals. Regarding the difference in service levels between the GH and CHC, $H - L$, and the inter-hospital distance d, we make the following assumption.

Assumption 3. $H - L \geq d$.

This assumption is reasonable according to Ref. [30], as distance is generally less critical to patients compared to the disparity in service levels. We further assume that h is the waiting cost rate. Given that the expected waiting time in procedure k is $\mathbb{E}(w_k)$, the resulting disutility from waiting in that procedure is $h\mathbb{E}(w_k) = \dfrac{h}{\mu_k^\kappa - \lambda_k^\kappa}$.

4 Optimal strategies

In this section, we characterize the optimal strategies of capacity investment and pricing for each hospital in different systems. Regarding the motivation of healthcare providers, Ref. [31] suggests that the ownership and profit status of providers do not differ. Ref. [32] finds that for-profit providers and public providers are just as likely as one another to overprescribe drugs, and for-profit providers are just as likely to deliver preventive activities as public providers. Therefore, we set profit maximization as the optimization goal of the whole system.

4.1 Non-coordinated system

We first consider the baseline case in which hospitals do not collaborate, and hence, each hospital maximizes its profit individually. We use p_1^κ and p_2^κ to denote the prices of the diagnosis and treatment procedures in the non-coordinated system, respectively. We illustrate this system in Fig. 1.

Let $\lambda_1^\kappa = \lambda_\kappa$. We can then simplify λ_2^κ to $q_\kappa\lambda_\kappa$. Let $\pi_\kappa(\mu_k^\kappa, p_k^\kappa)$

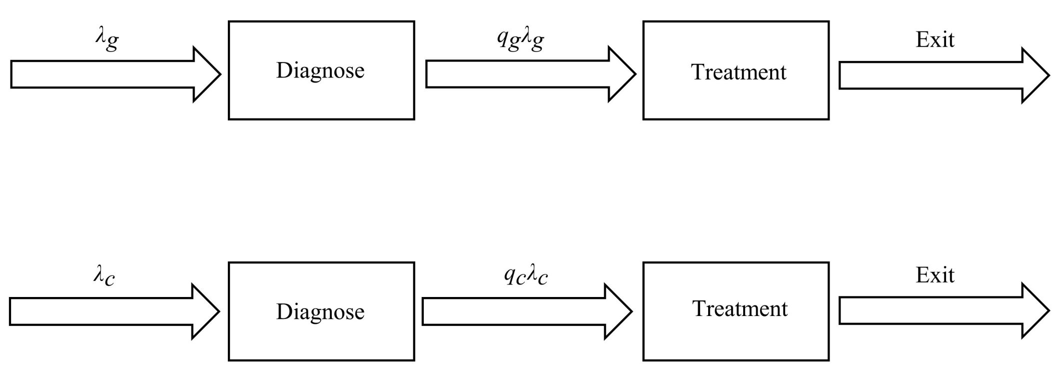


Fig. 1. M/M/1 queues in the non-coordinated system.

CSTR: 32290.14.JUSTC-2023-0052
DOI: 10.52396/JUSTC-2023-0052
JUSTC, **2025**, 55(8): 0803

denote the expected profit earned by hospital κ. The profit maximization problem of hospital κ can then be formulated as follows.

$$\max \pi_\kappa(\mu_k^\kappa, p_k^\kappa) = p_1^\kappa\lambda_\kappa + p_2^\kappa q_\kappa\lambda_\kappa - c_1\mu_1^\kappa - c_2^\kappa\mu_2^\kappa \tag{1}$$

s.t.

$$(1-q_\kappa)\left(G_1 - p_1^\kappa - \frac{h}{\mu_1^\kappa - \lambda_\kappa}\right) + q_\kappa\left(G_2 + V - p_1^\kappa - \frac{h}{\mu_1^\kappa - \lambda_\kappa} - \frac{h}{\mu_2^\kappa - q_\kappa\lambda_\kappa} - p_2^\kappa\right) \geq 0, \tag{2}$$

$$\lambda_\kappa \leq \mu_1^\kappa, \tag{3}$$

$$q_\kappa\lambda_\kappa \leq \mu_2^\kappa. \tag{4}$$

Below, we present the optimal solution to the above maximization problem.

Lemma 1. Given λ_κ, the optimal capacities for the diagnosis procedure and treatment procedure of hospital κ are, respectively,

$$\mu_1^{\kappa*} = \lambda_\kappa + \sqrt{\frac{h\lambda_\kappa}{c_1}}, \tag{5}$$

$$\mu_2^{\kappa*} = q_\kappa\lambda_\kappa + \sqrt{\frac{hq_\kappa\lambda_\kappa}{c_2^\kappa}}. \tag{6}$$

In addition, the optimal prices for the diagnosis procedure and treatment procedure of hospital κ satisfy the following formula:

$$p_1^{\kappa*} + q_\kappa p_2^{\kappa*} = (1-q_\kappa)G_1 + q_\kappa(G_2+V) - \sqrt{\frac{hc_1}{\lambda_\kappa}} - \sqrt{\frac{hq_\kappa c_2^\kappa}{\lambda_\kappa}}, \tag{7}$$

where $V = H$ for $\kappa = g$ and $V = L$ for $\kappa = c$. Hospital κ's optimal profit is

$$\pi_\kappa^* = (G_1 - c_1)\lambda_\kappa + (G_2 + V - G_1 - c_2^\kappa)q_\kappa\lambda_\kappa - 2\sqrt{hc_1\lambda_\kappa} - 2\sqrt{hc_2^\kappa q_\kappa\lambda_\kappa}. \tag{8}$$

The optimal service capacity of each procedure is increasing with the patient arrival rates λ_k^κ (i.e., λ_κ and q_κ) while decreasing with the capacity cost rate c_k. Therefore, the expected delay for patients $\frac{1}{\mu_\kappa - \lambda_\kappa}$ decreases with λ_κ. The expected revenue obtained from each patient, $p_1^{\kappa*} + q_\kappa p_2^{\kappa*}$, is equivalent to patient's expected utility and increases with the arrival rate. These results are consistent with our intuition. However, the effect of the arrival rate λ_κ on the optimal profit of a hospital is not straightforward, as stated in the following proposition.

Proposition 1. For hospital κ, $\kappa \in \mathcal{N}$, its optimal profit, π_κ^*, is a convex function of λ_κ and achieves its minimum at

$$\lambda_{\kappa,\min} = \left[\frac{\sqrt{hc_1} + \sqrt{hc_2^\kappa q_\kappa}}{G_1 - c_1 + (G_2 + V - G_1 - c_2^\kappa)q_\kappa}\right]^2. \tag{9}$$

Moreover, $\pi_\kappa^* \geq 0$ when $G_1 - c_1 + (G_2 + V - G_1 - c_2^\kappa)q_\kappa > 0$ and $\lambda_\kappa \geq 4\lambda_{\kappa,\min}$.

Proposition 1 indicates that a hospital has an incentive to serve patients only when the arrival rate of patients is sufficiently large. Note that, in practice, CHCs usually have low arrival rates. Low arrival rates usually correspond to low incomes. In this case, medical institutions will have insufficient funds to motivate their professionals (through, e.g., performance compensation), purchase more advanced medical equipment, and/or hire more qualified physicians. This result partially explains why the professionality and services in CHCs are usually not well recognized[33].

4.2 Coordinated system

This section analyzes optimal strategies under coordinated systems, assuming the GH and CHC are willing to cooperate. With only two players, collaboration is assured provided it increases both the total system profit and each player's individual profit. The case of incentivizing a grand coalition among more than two players is discussed in Section 4.4.

4.2.1 One-way referral system

First, we consider a one-way referral system, as shown in Fig. 2, under which all patients diagnosed in hospital c are transferred to hospital g for treatment. We use $p_1^{\kappa S}$ and $\mu_1^{\kappa S}$ to denote the price and capacity of the diagnosis procedure in hospital κ, respectively. The price and service capacity of the treatment procedure in hospital g are denoted by p_2^S and μ_2^S, respectively.

Let $\lambda_2^S = q_g\lambda_g + q_c\lambda_c$ and $\pi^S(\mu_1^{\kappa S}, p_1^{\kappa S}, \mu_2^S, p_2^S)$ denote the total profit of the system. The profit maximization problem can be formulated as follows.

$$\max \pi_\kappa(\mu_k^\kappa,\ p_k^\kappa) = p_1^\kappa\lambda_\kappa + p_2^\kappa q_\kappa\lambda_\kappa - c_1\mu_1^\kappa - c_2^\kappa\mu_2^\kappa \tag{10}$$

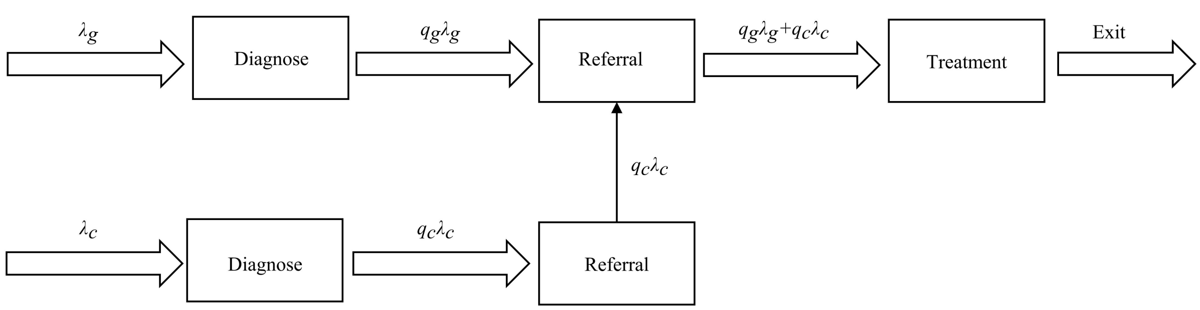


Fig. 2. M/M/1 queues in the one-way referral system.

CSTR: 32290.14.JUSTC-2023-0052
DOI: 10.52396/JUSTC-2023-0052
JUSTC, **2025**, 55(8): 0803

s.t.

$$(1-q_g)(G_1-p_1^{gS}-\frac{h}{\mu_1^{gS}-\lambda_g})+ q_g(G_2+H-p_1^{gS}-\frac{h}{\mu_1^{gS}-\lambda_g}-\frac{h}{\mu_2^S-\lambda_2^S}-p_2^S)\geq 0, \tag{11}$$

$$(1-q_c)(G_1-p_1^{cS}-\frac{h}{\mu_1^{cS}-\lambda_c})+ q_c(G_2+H-d-p_1^{cS}-\frac{h}{\mu_1^{cS}-\lambda_c}-\frac{h}{\mu_2^S-\lambda_2^S}-p_2^S)\geq 0, \tag{12}$$

$$\lambda_1^\kappa \leq \mu_1^{\kappa S},\ \kappa\in\mathcal{N}, \tag{13}$$

$$\lambda_2^S\leq \mu_2^S. \tag{14}$$

The optimal solution is as follows.

Lemma 2. Given λ_κ $(\kappa=g,c)$, the optimal capacities for the diagnosis procedure and treatment procedure are, respectively,

$$\mu_1^{\kappa S*}=\lambda_\kappa+\sqrt{\frac{h\lambda_\kappa}{c_1}}, \tag{15}$$

$$\mu_2^{S*}=\lambda_2^S+\sqrt{\frac{h\lambda_2^S}{c_2^g}}. \tag{16}$$

In addition, the optimal prices of hospital κ satisfy:

$$p_1^{gS*}+q_gp_2^{S*}=(1-q_g)G_1+q_g(G_2+H)-\sqrt{\frac{hc_1}{\lambda_g}}-q_g\sqrt{\frac{hc_2^g}{\lambda_2^S}}, \tag{17}$$

$$p_1^{cS*}+q_cp_2^{S*}=(1-q_c)G_1+q_c(G_2+H-d)-\sqrt{\frac{hc_1}{\lambda_c}}-q_c\sqrt{\frac{hc_2^g}{\lambda_2^S}}. \tag{18}$$

The optimal profit of the system is

$$\pi^{S*}=(G_1-c_1)(\lambda_g+\lambda_c)+dq_g\lambda_g+(G_2+H-G_1-c_2^g-d)\lambda_2^S- 2\sqrt{hc_1\lambda_g}-2\sqrt{hc_1\lambda_c}-2\sqrt{hc_2^g\lambda_2^S}. \tag{19}$$

4.2.2 Two-way referral system

In this subsection, we consider a two-way referral mechanism. We use $p_k^{\kappa T}$ and $\mu_k^{\kappa T}$ to denote the price and capacity, respectively, of the diagnosis and treatment procedure in hospital κ. All other parameters remain the same as those in the one-way referral system, with the introduction of one new parameter: the referral threshold r $(0\leq r\leq 1)$. The two-way referral system is illustrated in Fig. 3. Let λ_g and λ_c denote the arrival rates for the diagnosis procedure in the GH and CHC, respectively. Following diagnosis, a portion of patients is discharged. For each patient proceeding to treatment, the hospital assesses the health condition and makes a referral decision based on a threshold policy. We model a patient's health condition as a random variable uniformly distributed on [0, 1], where a higher value indicates greater severity. A patient with a health condition exceeding the threshold r is referred to the GH for treatment; conversely, the patient is treated in the CHC. The upstream referral rate and downstream referral rate are denoted by $\lambda^u(r)$ and $\lambda^d(r)$, respectively. Then, $\lambda^d(r)=rq_g\lambda_g$ and $\lambda^u(r)=(1-r)q_c\lambda_c$.

Let $\lambda_2^g(r)$ (resp. $\lambda_2^c(r)$) be the arrival rate of the treatment procedure in GH (resp. CHC). Then, $\lambda_2^g(r)=(1-r)\cdot(q_g\lambda_g+q_c\lambda_c)$, and $\lambda_2^c(r)=r(q_g\lambda_g+q_c\lambda_c)$. We use $\pi^T(\mu_k^{\kappa T},p_k^{\kappa T})$ to denote the total profit of the system. The optimization problem can then be expressed as follows.

$$\max \pi^T(\mu_k^{\kappa T},p_k^{\kappa T})=p_1^{gT}\lambda_g+p_1^{cT}\lambda_c+p_2^{gT}(1-r)(q_g\lambda_g+q_c\lambda_c)+ p_2^{cT}r(q_g\lambda_g+q_c\lambda_c)-c_1\mu_1^{gT}-c_1\mu_1^{cT}-c_2^g\mu_2^{gT}-c_2^c\mu_2^{cT} \tag{20}$$

s.t.

$$(1-q_g)\left(G_1-p_1^{gT}-\frac{h}{\mu_1^{gT}-\lambda_g}\right)+ rq_g\left[G_2+L-d-p_1^{gT}-\frac{h}{\mu_1^{gT}-\lambda_g}-p_2^{cT}-\frac{h}{\mu_2^{cT}-r(q_g\lambda_g+q_c\lambda_c)}\right]+ q_g(1-r)\left[G_2+H-\frac{h}{\mu_1^{gT}-\lambda_g}-p_1^{gT}-p_2^{gT}- \frac{h}{\mu_2^{gT}-(1-r)(q_g\lambda_g+q_c\lambda_c)}\right]\geq 0, \tag{21}$$

$$(1-q_c)\left(G_1-p_1^{cT}-\frac{h}{\mu_1^{cT}-\lambda_c}\right)+ rq_c\left[G_2+L-p_1^{cT}-\frac{h}{\mu_1^{cT}-\lambda_c}-p_2^{cT}-\frac{h}{\mu_2^{cT}-r(q_g\lambda_g+q_c\lambda_c)}\right]+ q_c(1-r)\left[G_2+H-\frac{h}{\mu_1^{cT}-\lambda_c}-p_1^{cT}-p_2^{gT}-d- \frac{h}{\mu_2^{gT}-(1-r)(q_g\lambda_g+q_c\lambda_c)}\right]\geq 0, \tag{22}$$

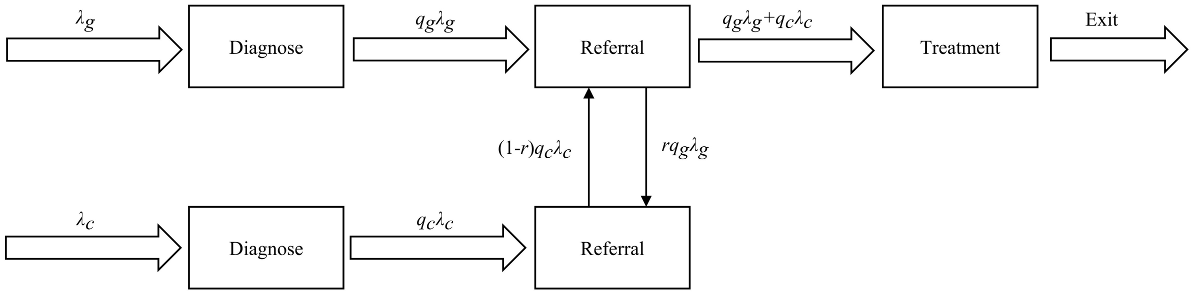


Fig. 3. M/M/1 queues in the two-way referral system.

CSTR: 32290.14.JUSTC-2023-0052
DOI: 10.52396/JUSTC-2023-0052
JUSTC, **2025**, 55(8): 0803

$$\lambda_c \le \mu_1^{cT}, \tag{23}$$

$$\lambda_g \le \mu_1^{gT}, \tag{24}$$

$$(1-r)(q_g\lambda_g + q_c\lambda_c) \le \mu_2^{gT}, \tag{25}$$

$$r(q_g\lambda_g + q_c\lambda_c) \le \mu_2^{cT}. \tag{26}$$

Lemma 3. Given λ_κ ($\kappa = g, c$) and r, the optimal capacities for the diagnosis procedure and treatment procedure are, respectively,

$$\mu_1^{\kappa T*} = \lambda_\kappa + \sqrt{\frac{h\lambda_\kappa}{c_1}}, \tag{27}$$

$$\mu_2^{gT*} = \sqrt{\frac{(1-r)h(q_g\lambda_g + q_c\lambda_c)}{c_2^g}} + (1-r)(q_g\lambda_g + q_c\lambda_c), \tag{28}$$

$$\mu_2^{cT*} = \sqrt{\frac{rh(q_g\lambda_g + q_c\lambda_c)}{c_2^c}} + r(q_g\lambda_g + q_c\lambda_c). \tag{29}$$

In addition, the optimal prices of hospital κ satisfy the following formula:

$$\begin{aligned} & p_1^{gT*} + (1-r)q_g p_2^{gT*} + rq_g p_2^{cT*} = \\ & (1-q_g)G_1 + q_g G_2 + (1-r)q_g H + rq_g(L-d) - \\ & \sqrt{\frac{hc_1}{\lambda_g}} - q_g\sqrt{\frac{hc_2^g(1-r)}{(q_g\lambda_g + q_c\lambda_c)}} - q_g\sqrt{\frac{hc_2^c r}{(q_g\lambda_g + q_c\lambda_c)}}, \end{aligned} \tag{30}$$

$$\begin{aligned} & p_1^{cT*} + (1-r)q_c p_2^{gT*} + rq_c p_2^{cT*} = \\ & (1-q_c)G_1 + q_c G_2 + (1-r)q_c(H-d) + rq_c L - \\ & \sqrt{\frac{hc_1}{\lambda_c}} - q_c\sqrt{\frac{hc_2^g(1-r)}{(q_g\lambda_g + q_c\lambda_c)}} - q_c\sqrt{\frac{hc_2^c r}{(q_g\lambda_g + q_c\lambda_c)}}. \end{aligned} \tag{31}$$

The optimal profit of the system is

$$\begin{aligned} \pi^{T*} = & (\lambda_g + \lambda_c)(G_1 - c_1) + \\ & (q_g\lambda_g + q_c\lambda_c)[G_2 - G_1 - c_2^g(1-r) - c_2^c r + (1-r)H + rL] - \\ & (1-r)\lambda_c d q_c - r\lambda_g q_g d - 2\sqrt{hc_1\lambda_g} - 2\sqrt{hc_1\lambda_c} - \\ & 2\sqrt{hc_2^c r(q_g\lambda_g + q_c\lambda_c)} - 2\sqrt{hc_2^g(1-r)(q_g\lambda_g + q_c\lambda_c)}. \end{aligned} \tag{32}$$

Proposition 2. The optimal profit is first decreasing and then increasing with the referral threshold r.

Proposition 2 indicates that the optimal profit is maximized at the extreme values of the referral threshold, specifically when $r = 0$ or $r = 1$. This implies that it is optimal to either transfer all diagnosed patients from the CHC to the GH for treatment, or to transfer all diagnosed patients from the GH to the CHC for treatment. In other words, consolidating service capacity at one facility yields the highest profit.

4.3 Comparisons

In this section, we first evaluate the relative efficacy of the two referral mechanisms by comparing their total system profit. We then quantify the value of coordination by comparing the superior mechanism with the non-coordinated baseline.

4.3.1 One-way referral vs. two-way referral

This section compares the one-way and two-way referral mechanisms. Observe that when the referral threshold $r = 1$, the two-way mechanism reduces to the one-way case. While the greater generality of the two-way mechanism might suggest superior profitability, our analysis reveals that this is not always true.

Proposition 3. Under Assumption 2, the one-way referral mechanism yields a higher total system profit and shorter patient waiting times, at the expense of a higher medical service price, compared to the two-way mechanism.

Proposition 3 indicates that, under Assumption 2, the one-way mechanism is more advantageous for a service provider as it generates higher profit. This finding is counterintuitive given that the one-way mechanism is a special case of its two-way counterpart. The result stems from a trade-off between capacity pooling and operational flexibility. In the one-way system, all treatment resources are pooled in the GH. In contrast, the two-way system offers greater flexibility by allowing bidirectional patient flow. According to Lemmas 2 and 3, the expected revenue per patient equals the patient's expected utility and increases with the arrival rate λ_κ. Capacity pooling reduces waiting times, lowers required capacity, and improves service quality. However, it also increases the unit capacity cost rate as $c_2^g \ge c_2^c$, leading to higher service prices. Crucially, under Assumption 2 c_2^g is not significantly higher than c_2^c, the cost savings from reduced capacity investment in the one-way system outweigh the associated cost increase. Consequently, the one-way mechanism delivers a higher total system profit.

4.3.2 The value of coordination

Up to this point, we have demonstrated the superior effectiveness of the one-way referral mechanism. In the following, we further explore the value of coordination by focusing on this mechanism.

Proposition 4. When c_2^g is higher but not significantly higher than c_2^c, the inequality $\pi^{S*} \ge \sum_{\kappa\in\mathcal{N}} \pi_\kappa^*$ holds.

Proposition 4 proves that the optimal profit under the one-way mechanism can exceed that of the non-coordinated baseline. Recall that in the baseline, hospital κ may earn a negative profit, eliminating its incentive to serve patients. This property implies that the one-way mechanism can elevate the profit of hospital κ from negative to positive, thereby motivating it to improve service.

By comparing patient waiting times between the one-way system and the baseline, we derive the following result.

Proposition 5. Under Assumption 2, the one-way referral mechanism reduces the waiting times for patients in the treatment procedure.

As $\mathbb{E}^* = \dfrac{1}{\mu_k^{\kappa*} - \lambda_k^{\kappa*}} = \dfrac{hc_k^\kappa}{\lambda_k^\kappa}$, the expected delay for patients increases with the capacity cost rate c_k^κ and decreases with the arrival rate λ_k^κ. For hospital g, the one-way mechanism

CSTR: 32290.14.JUSTC-2023-0052
DOI: 10.52396/JUSTC-2023-0052
JUSTC, **2025**, 55(8): 0803

increases the arrival rate of its treatment procedure, thereby reducing patient waiting times. For hospital c, both the arrival rate and the capacity cost rate of treatment procedure increase under the one-way mechanism. Consequently, patient waiting times of treatment procedure reduced when c_2^g is higher but not significantly higher than c_2^c.

Furthermore, we derive the following comparative results regarding the optimal capacity and pricing decisions between the one-way system and the non-coordinated baseline.

Proposition 6. Under Assumption 2, we have the following results:

(i) The inequality $c_2^g\mu_2^{S*} \le c_2^g\mu_2^{g*} + c_2^c\mu_2^{c*}$ holds. Let $\Delta_c = c_2^g\mu_2^{g*} + c_2^c\mu_2^{c*} - c_2^g\mu_2^{S*}$. Then, $\partial\Delta_c/\partial\lambda_\kappa \ge 0$, i.e., Δ_c is increasing with λ_κ for any $\kappa \in \mathcal{N}$.

(ii) Let $\Delta_p = (p_1^{\kappa S*} + q_\kappa p_2^{S*}) - (p_1^{\kappa*} + q_\kappa p_2^{\kappa*})$. Then, the inequalities $p_1^{\kappa S*} + q_\kappa p_2^{S*} \ge p_1^{\kappa*} + q_\kappa p_2^{\kappa*}$ and $\partial\Delta_p^g/\partial\lambda_c \le 0$ hold.

This proposition demonstrates that coordination reduces the investment in medical resources required by treatment procedures, with the reduction being more pronounced for higher patient arrival rates. Compared to the non-coordinated baseline, coordination also yields a higher expected revenue per patient, although this additional profit decreases with the arrival rate λ_κ. As established in Lemma 1 and Proposition 5, since the expected revenue equals the patient's expected utility and patient waiting time are shorter in the one-way system, patients effectively pay more for reduced delay, indicating their willingness to do so.

Thus far, the base model demonstrates that the one-way referral mechanism outperforms both the non-coordinated baseline and the two-way mechanism. To examine the robustness of this finding, Section 5 extends the comparison to a congested system where arrival rates are influenced by the referral scheme itself.

4.4 Profit allocation

Our base model considers a system with one GH and one CHC, where fostering cooperation through a profit allocation scheme is straightforward. In this section, we extend the analysis to a system comprising one GH and multiple CHCs, demonstrating how to design such a scheme. We employ the one-way referral mechanism for our analysis; the corresponding scheme for the two-way mechanism can be derived following a similar approach.

Let $\bar{\pi}_\kappa$ be the profit of hospital κ under the grand coalition $\mathcal{N}$; then, $\bar{\pi}_\kappa(\kappa \in \mathcal{N})$ is a profit allocation scheme for the grand coalition. If $\sum_{\kappa\in\mathcal{N}} \bar{\pi}_\kappa = \pi^{S*}$, where π^{S*} is the optimal profit of the system with the one-way referral system, then the allocation scheme is sufficient. An allocation scheme is individually rational if $\bar{\pi}_\kappa \ge \pi_\kappa^*$ and is stable for a coalition $\mathcal{J}$ if $\sum_{\kappa\in\mathcal{J}} \bar{\pi}_\kappa \ge \bar{\pi}_{\mathcal{J}}$. In addition, an allocation scheme is said to be a member of the core if it satisfies the following inequalities:

$$\sum_{\kappa\in\mathcal{J}} \bar{\pi}_\kappa \ge \bar{\pi}_{\mathcal{J}},\ \forall \mathcal{J} \subseteq \mathcal{N};$$

$$\sum_{\kappa\in\mathcal{N}} \bar{\pi}_\kappa = \pi^{S*}.$$

When an allocation scheme is in the core, no subset of players can secede from the grand coalition and form smaller coalitions. Hence, the existence of an allocation scheme that is in the core (the core is nonempty) is sufficient in our context to show that it is optimal for all hospitals to form the grand coalition. We consider a profit allocation scheme built on the Shapley value due to the following property.

Proposition 7. For our one-way referral system, the Shapley value is at the core.

In our system, the Shapley value of hospital κ, $\kappa \in \mathcal{N}$, is defined as follows:

$$\bar{\pi}_\kappa = \sum_{\mathcal{J}\subseteq\mathcal{N}\setminus\{\kappa\}} \frac{|\mathcal{J}|!(n-|\mathcal{J}|)!}{(n+1)!}(\bar{\pi}_{\mathcal{J}\cup\{\kappa\}} - \bar{\pi}_{\mathcal{J}}), \tag{33}$$

where $\mathcal{J}$ denotes the subset of hospitals in the coalition without hospital κ, and $\mathcal{N}$ denotes the complete set.

For a subset $\mathcal{J} \subseteq \mathcal{N}$, if it involves the GH and m CHCs for $0 \le m \le n$, then

$$\begin{aligned}\bar{\pi}_{\mathcal{J}} = &\left(\pi_g^* + 2\sqrt{hc_2 q_g \lambda_g}\right) + \\ &\sum_{i\in\mathcal{J}\setminus g}\left[\pi_i^* + (H-L-d)q_i\lambda_i + 2\sqrt{hc_2 q_i\lambda_i}\right] - \\ &2\sqrt{hc_2\Big(q_g\lambda_g + \sum_{i\in\mathcal{J}\setminus g} q_i\lambda_i\Big)},\end{aligned}$$

where $\pi_g^*(\pi_i^*)$ is the optimal profit of the GH(CHC) in the non-coordinated system. To simplify the expression, we define $C_g = 2\sqrt{hc_2 q_g\lambda_g}$, $C_i = 2\sqrt{hc_2 q_i\lambda_i}$, $E_i = (H-L-d)q_i\lambda_i$, and $B_m\{\mathcal{J}\setminus g\} = 2\sqrt{hc_2\big(q_g\lambda_g + \sum_{i\in\mathcal{J}\setminus g} q_i\lambda_i\big)}$, all of which are nonnegative. Then, the expression of $\bar{\pi}_{\mathcal{J}}$ above can be simplified to

$$\bar{\pi}_{\mathcal{J}} = \pi_g^* + C_g + \sum_{i\in\mathcal{J}\setminus g}(\pi_i^* + E_i + C_i) - B_m\{\mathcal{J}\setminus g\}.$$

However, if the subset $\mathcal{J}$ only involves m CHCs, then

$$\bar{\pi}_{\mathcal{J}} = \sum_{\kappa\in\mathcal{J}} \pi_\kappa^*.$$

Now let $A_m = \left\{\sum_{k_1=1}^{n-m+1}\sum_{k_2=k_1+1}^{n-m+2}\cdots\sum_{k_m=k_{m-1}+1}^{n} B_m\{k_1,\cdots,k_m\}\right\}/C_n^m$ be the average value of the $B_m\{\cdot\}$'s with $m \le n$, where $C_n^m = n!/[m!(n-m)!]$, and define $A_e = \left\{\sum_{i=1}^{n}(E_i + C_i)\right\}/n$ as the average value of the terms $(E_i + C_i)$'s over all i.

Then, we have the following result for the GH.

Lemma 4. Based on the Shapley value, the GH is allocated the following profit in the grand coalition:

$$\bar{\pi}_g = \pi_g^* + \frac{n}{n+1}C_g + \frac{n}{2}A_e - \frac{1}{n+1}\sum_{m=1}^{n} A_m. \tag{34}$$

For the n CHCs, we denote by $\overline{\mathcal{M}_i}$ the set of CHCs excluding hospital i, i.e., $\overline{\mathcal{M}_i} = \mathcal{N}\setminus\{g,i\}$. There are $n-1$ elements in the set $\overline{\mathcal{M}_i}$. Define

CSTR: 32290.14.JUSTC-2023-0052
DOI: 10.52396/JUSTC-2023-0052
JUSTC, **2025**, 55(8): 0803

$$f_i = \frac{1}{(n+1)n} B_1\{i\} + \frac{1}{n+1}(B_n\{1,\cdots,n\} - B_{n-1}\{1,\cdots,i-1,i+1,\cdots,n\}) + \sum_{m=2}^{n-1} \frac{1}{n+1}\frac{1}{C_n^m} \sum_{k_1=1}^{n-m+1} \cdots \sum_{k_{m-1}=k_{m-2}+1}^{n-1} (B_m\{k_1,\cdots,k_{m-1},i\} - B_{m-1}\{k_1,\cdots,k_{m-1}\}),$$

where $k_1,\cdots,k_{m-1}$ are the elements of the set $\overline{\mathcal{M}_i}$ such that $k_j \neq i$ for any $j \in \{1,\cdots,m-1\}$. Then, we have the following result for the profit of hospital i.

Lemma 5. Based on the Shapley value, hospital i for $i \in \mathcal{N} \setminus g$ is allocated the following profit:

$$\bar{\pi}_i = \pi_i^* + \frac{1}{2}(C_i + E_i) + \frac{1}{n(n+1)} C_g - f_i. \tag{35}$$

Then, we have the following property.

Proposition 8. If $\pi_\kappa^* \geq 0$ holds, then π_κ^* is increasing with λ_κ for any $\kappa \in \mathcal{N}$.

This proposition implies that within the grand coalition, hospital κ is allocated a higher profit for a larger patient arrival rate λ_κ. This aligns with intuition that a hospital contributing more patients to the system should receive greater compensation. However, this intuition does not hold universally. When $\pi_\kappa^* < 0$, a larger λ_κ may indicate that hospital κ incurs disproportionately high costs for diagnostic capacity, thereby depressing the overall system profit. In such scenarios, in fact, π_κ^* can be decreasing with λ_κ.

5 A congested system

In this section, we consider heterogeneous patients and the influence of the referral scheme on arrival rates in a congested system. Our model setting is similar to Ref. [15], in which the heterogeneity of patients stems from differing health conditions.

In this congested system, patients with common conditions may choose either the GH or the CHC for first diagnosis, whereas severe cases must be treated by the GH due to its superior medical capability. However, congestion in the GH may prevent some patients from securing registration. Conversely, the CHC face underutilization, yet they carry a risk of potential misdiagnosis or inadequate care for complex cases. Crucially, the implemented referral scheme directly influences patient arrival rates. The presence of an upward referral channel (from CHC to GH) makes patients more willing to visit a CHC initially, knowing that subsequent referral is an option. A similar effect applies when a downward referral channel (from GH to CHC) exists. Based on this setting, we analyze and compare three systems sequentially: the non-coordinated baseline, the one-way referral mechanism, and the two-way referral mechanism.

5.1 System without coordination

As illustrated in Fig. 4, parameters λ_g and λ_c denote the mean arrival rate of patients (per unit time) in the GH and CHC, respectively. Severe patients, with an arrival rate of λ_s, must visit the GH for first diagnosis.

Given patient heterogeneity, the probability of proceeding to treatment varies. Therefore, patient flow is modeled as a single procedure with two parallel channels: a common channel and a green channel. In this congested system, service capacity of hospital κ is fixed. Reflecting the GH's overcrowding, we focus on the case where $\lambda_s + \lambda_g \geq \mu_g$. The common channel follows a first-come-first-served (FCFS) rule, while patients entering via the green channel experience no waiting. Since patients in the common channel are identical, each has an equal probability of successful registration in the GH, denoted by $\frac{\mu_g}{\lambda_s + \lambda_g}$. Patients who fail to register cannot receive treatment and must re-enter the system in the next period.

A patient's health condition x is modeled as a random variable uniformly distributed on [0, 1], where a higher x indicates greater severity. For severe patients, $x = 1$. In this setting, the treatment price depends on x. Consistent with prior research[15], we assume the influence of health condition on price is dominant, as treating more severe symptoms is typically more expensive. Following Ref. [15], we set the price equal to the health condition, i.e., $p(x) = x$. Consequently, the expected profit from treating a common patient at first diagnosis is 1/2. As noted, the CHC cannot treat all patients and risks mismanaging complex cases. Thus, the CHC adopts a treatment threshold k: Only patients with condition $x \leq k$ are treated; those with condition $x > k$ must visit the GH in the next period. Again following Ref. [15], the CHC's revenue in the absence of mistreatment is $\lambda_c \times \int_0^r x\mathrm{d}x = \lambda_c r^2/2$, while the total cost of mistreatment is $\lambda_c r^3/2$, as the additional cost increases with disease severity. Since capacities are exogenous

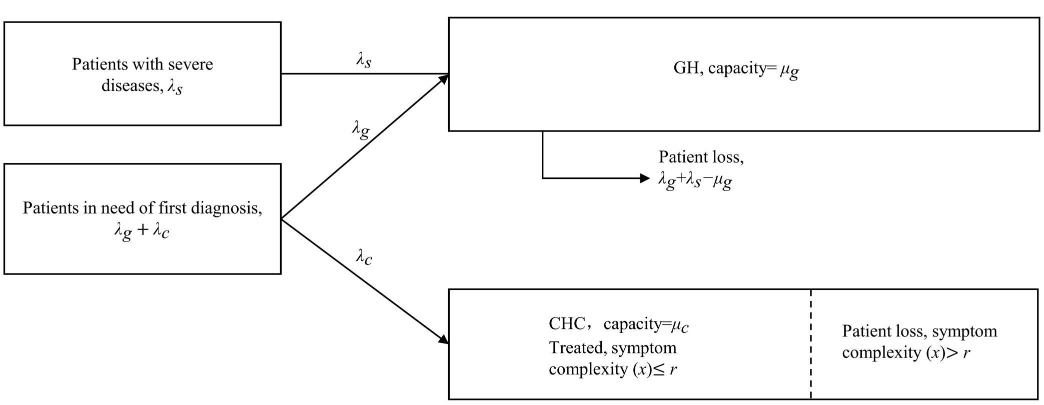


Fig. 4. System without coordination.

CSTR: 32290.14.JUSTC-2023-0052
DOI: 10.52396/JUSTC-2023-0052
JUSTC, **2025**, 55(8): 0803

in this section, the fixed routine operational costs are omitted from the objective function.

$$\pi_n=\left[\lambda_s\times 1+\lambda_g\times\frac{1}{2}\right]\frac{\mu_g}{\lambda_s+\lambda_g}+\lambda_c\frac{r^2-r^3}{2}=\frac{\mu_g(2\lambda_s+\lambda_g)}{2(\lambda_s+\lambda_g)}+\lambda_c\frac{r^2-r^3}{2}. \tag{36}$$

5.2 Coordinated system

As in the base model, the collaboration between the GH and CHC is premised on the condition that it enhances the aggregate system profit while also raising the individual payoff for each player.

5.2.1 One-way referral system

We now consider a one-way referral system, as shown in Fig. 5, in which patients with more serious symptoms in the CHC are transferred to the GH for treatment.

To facilitate the green channel, the GH allocates part of its capacity to treat referred patients, represented by t_1. The remaining capacity μ_g-t_1 is open to patients directly visiting the GH. In the one-way referral system, the probability that one patient could be diagnosed and treated by the GH decreases from $\frac{\mu_g}{\lambda_s+\lambda_g}$ to $\frac{\mu_g-t_1}{\lambda_s+\lambda_g}$. Consequently, some patients visiting the GH for the first diagnosis will switch to CHC. As t_1 increases, more patients flow to the CHC. In extreme cases, no one will choose the GH for the first diagnosis when the green channel occupies all the capacity of the GH, i.e., $t_1=\mu_g$, and no one would alter his/her choice when there is no green channel, i.e., $t_1=0$.

Let the number of patients who seek the first diagnosis and their preferences shift from GH to CHC be $m_1=m_1(\lambda_g,t_1)$. According to the previous analyses, we have $m_1(\lambda_g,0)=0$ and $m_1(\lambda_g,\mu_g)=\lambda_g$. For simplicity, we assume that $m_1=\frac{\lambda_g}{\mu_g}t_1$. The profit of the system can then be expressed as follows.

$$\begin{aligned}\pi_s=&\\ &[\lambda_s\times 1+(\lambda_g-\frac{\lambda_g}{\mu_g}t_1)\times\frac{1}{2}]\frac{\min\{\lambda_s+\lambda_g-\frac{\lambda_g}{\mu_g}t_1,\mu_g-t_1\}}{\lambda_s+\lambda_g-\frac{\lambda_g}{\mu_g}t_1}+\frac{r^2-r^3}{2}\cdot\\ &\min\{\mu_c,\lambda_c+\frac{\lambda_g}{\mu_g}t_1\}+\frac{r+1}{2}\min\{(1-r)\min\{\mu_c,\lambda_c+\frac{\lambda_g}{\mu_g}t_1\},t_1\}=\\ &\frac{\min\{\lambda_s+\lambda_g-\frac{\lambda_g}{\mu_g}t_1,\mu_g-t_1\}}{2}+\frac{\lambda_s}{2(\lambda_s+\lambda_g-\frac{\lambda_g}{\mu_g}t_1)}+\frac{r^2-r^3}{2}\cdot\\ &\min\{\mu_c,\lambda_c+\frac{\lambda_g}{\mu_g}t_1\}+\frac{r+1}{2}\min\{(1-r)\min\{\mu_c,\lambda_c+\frac{\lambda_g}{\mu_g}t_1\},t_1\}.\end{aligned} \tag{37}$$

5.2.2 Two-way referral system

In the two-way referral system, the CHC (resp. GH) allocates t_1 (resp. t_2) of its capacity to facilitate the green channel, as shown in Fig. 6.

Let $m_2(\lambda_c,t_2)$ be the number of patients who seek the first diagnosis, but their preferences shift from CHC to GH. Then, $m_1(\lambda_g,t_1)=\frac{\lambda_g}{\mu_g}t_1$ and $m_2(\lambda_c,t_2)=\frac{\lambda_c}{\mu_c}t_2$. The profit of the system, denoted by π_t, can then be expressed as follows.

$$\begin{aligned}\pi_t=&\frac{[\lambda_s\times 1+(1-r)(\lambda_g-\frac{\lambda_g}{\mu_g}t_1+\frac{\lambda_c}{\mu_c}t_2)\times\frac{r+1}{2}]}{\lambda_s+(1-r)(\lambda_g-\frac{\lambda_g}{\mu_g}t_1+\frac{\lambda_c}{\mu_c}t_2)}\min\{\mu_g-t_1,\lambda_s+(1-r)(\lambda_g-\frac{\lambda_g}{\mu_g}t_1+\frac{\lambda_c}{\mu_c}t_2)\}+\\ &\frac{r+1}{2}\min\{(1-r)\min\{\mu_c-t_2,\lambda_c+\frac{\lambda_g}{\mu_g}t_1-\frac{\lambda_c}{\mu_c}t_2\},t_1\}+\frac{r^2-r^3}{2}\min\{\mu_c-t_2,\lambda_c+\frac{\lambda_g}{\mu_g}t_1-\frac{\lambda_c}{\mu_c}t_2\}+\\ &\frac{r-r^2}{2}\min\{r\min\{\lambda_g-\frac{\lambda_g}{\mu_g}t_1+\frac{\lambda_c}{\mu_c}t_2,\frac{(\mu_g-t_1)(\lambda_g-\frac{\lambda_g}{\mu_g}t_1+\frac{\lambda_c}{\mu_c}t_2)}{\lambda_s+(\lambda_g-\frac{\lambda_g}{\mu_g}t_1+\frac{\lambda_c}{\mu_c}t_2)}\},t_2\}=\\ &\frac{2\lambda_s+(1-r)(\lambda_g-\frac{\lambda_g}{\mu_g}t_1+\frac{\lambda_c}{\mu_c}t_2)(r+1)}{2[\lambda_s+(1-r)(\lambda_g-\frac{\lambda_g}{\mu_g}t_1+\frac{\lambda_c}{\mu_c}t_2)]}\min\{\mu_g-t_1,\lambda_s+(1-r)(\lambda_g-\frac{\lambda_g}{\mu_g}t_1+\frac{\lambda_c}{\mu_c}t_2)\}+\\ &\frac{r+1}{2}\min\{(1-r)\min\{\mu_c-t_2,\lambda_c+\frac{\lambda_g}{\mu_g}t_1-\frac{\lambda_c}{\mu_c}t_2\},t_1\}+\frac{r^2-r^3}{2}\min\{\mu_c-t_2,\lambda_c+\frac{\lambda_g}{\mu_g}t_1-\frac{\lambda_c}{\mu_c}t_2\}+\\ &\frac{r-r^2}{2}\min\{r\min\{\lambda_g-\frac{\lambda_g}{\mu_g}t_1+\frac{\lambda_c}{\mu_c}t_2,\frac{(\mu_g-t_1)(\lambda_g-\frac{\lambda_g}{\mu_g}t_1+\frac{\lambda_c}{\mu_c}t_2)}{\lambda_s+(\lambda_g-\frac{\lambda_g}{\mu_g}t_1+\frac{\lambda_c}{\mu_c}t_2)}\},t_2\}.\end{aligned} \tag{38}$$

CSTR: 32290.14.JUSTC-2023-0052
DOI: 10.52396/JUSTC-2023-0052
JUSTC, **2025**, 55(8): 0803

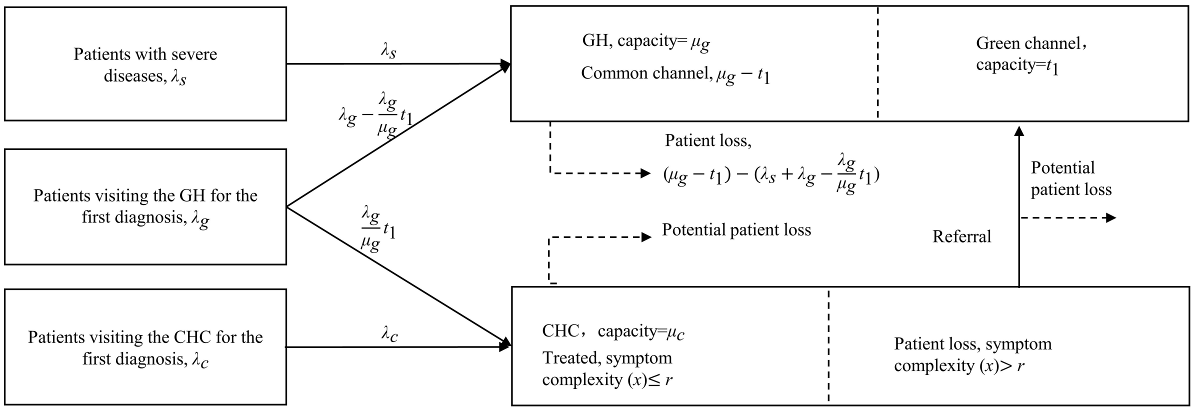


Fig. 5. One-way referral system.

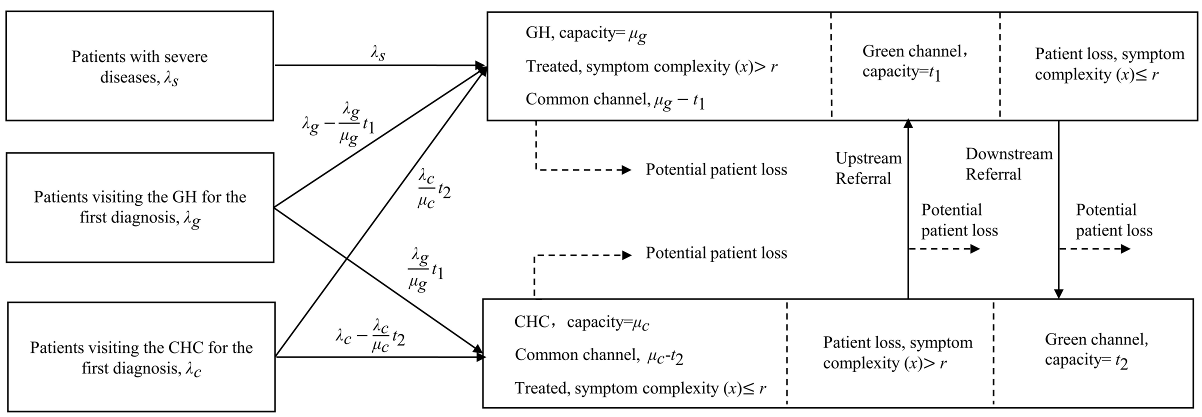


Fig. 6. System with a two-way referral mechanism.

5.3 Simulation study

This section presents a numerical study to investigate how patient arrival rates affect total system profit under three regimes: the non-coordinated baseline, the one-way coordinated system, and the two-way coordinated system.

Fig. 7 validates the value of coordination. To calculate the optimal profit of each system, we optimize over the referral threshold and the capacity of the green channels. A comparison between the optimal profits reveals that forming a coordinated alliance consistently yields a profit gain for the system. Furthermore, the two-way referral mechanism invariably generates a higher alliance profit than the one-way mechanism.

Fig. 7a demonstrates that the optimal profit is increasing with the arrival rate of patients requiring further treatment. This trend arises because treating severe patients, who constitute this flow, is more profitable than treating common patients.

Fig. 7b depicts that the optimal profit of the non-coordinated system decreases with the arrival rate of patients visiting the GH for first diagnosis, i.e., λ_g. This is because the limited

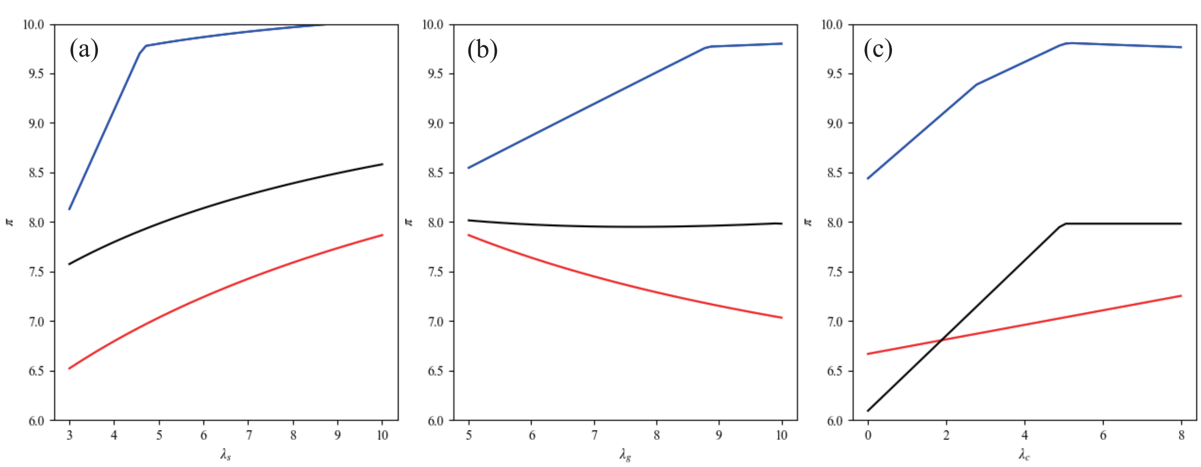


Fig. 7. Comparison of the total profits among the three different scenarios: (a) $\lambda_s \in [3,10]$, $\lambda_g = 10$, $\lambda_c = 5$, $\mu_g = 10$, $\mu_c = 8$; (b) $\lambda_s = 5$, $\lambda_g \in [5,10]$, $\lambda_c = 5$, $\mu_g = 10$, $\mu_c = 8$; (c) $\lambda_s = 5$, $\lambda_g = 10$, $\lambda_c \in [0,8]$, $\mu_g = 10$, $\mu_c = 8$.

CSTR: 32290.14.JUSTC-2023-0052
DOI: 10.52396/JUSTC-2023-0052
JUSTC, **2025**, 55(8): 0803

capacity of the GH causes registration failures. A higher λ_g means these limited resources are occupied by less profitable (common) patients, thereby reducing total revenue. For the one-way referral system, when λ_g is relatively low, its increase raises the arrival rate of the CHC and the upstream green channel, while reducing the registration success rate for severe patients. As λ_g increases beyond a point, the revenue from both the CHC and the upstream green channel stabilizes due to their capacity constraints.

Fig. 7c indicates that both the non-coordinated system profit (π_n) and the one-way referral system profit (π_s) increase with the arrival rate of patients visiting the CHC for first diagnosis (λ_c), but π_s grows faster. Notably, when λ_c is small, π_n can be higher than π_s. This occurs because the upstream green channel consumes GH resources that would otherwise serve severe patients. In our setting, capacity of the CHC (μ_c) is greater than λ_c, so π_n increases linearly with λ_c within this constraint. In the one-way referral system, an increase in λ_c initially boosts arrival rates of both the CHC and the upstream green channel. However, once λ_c exceeds the combined capacity of the CHC (μ_c) and the upstream referral channel (t_1), the flow of successfully transferred patients reaches a ceiling, stabilizing the system's optimal profit.

In the two-way referral system, when λ_c is small, the increase in λ_c raises not only the arrival rate of the CHC but also the patient flow from the CHC to the GH (upstream green channel) and from the GH to the CHC (downstream green channel). This shift in patient choice reduces the registration success probability of severe patients in the GH. When λ_c exceeds the combined capacity of the CHC (μ_c) and the green channels (t_1), the negative effect of reduced severe patient registration gradually dominates the positive effect of increased volume, curbing further profit growth. Synthesizing the insights from Fig. 7a and 7c, we observe that the optimal profit increases with the arrival rates of severe patients (λ_s) and common patients in the CHC (λ_c), but does not increase monotonically with the arrival rate of common patients in the GH (λ_g). Therefore, management emphasis should be placed on guiding common patients to the CHC for first diagnosis to optimize system effectiveness.

In summary, our findings reveal a contingent superiority of referral mechanisms. The one-way referral mechanism can be more profitable than its two-way counterpart in a stable queueing system with homogeneous patients and exogenous arrival rates. However, the conclusion reverses in a congested system with heterogeneous patients and endogenous (referral-scheme-dependent) arrival rates. Here, the resource constraint of the GH becomes binding. The downward referral pathway in the two-way mechanism creates capacity for high-severity patients by redirecting common cases, leading to higher-value utilization of medical resources. Consequently, the two-way system achieves a strictly higher optimal profit than both the one-way system and the non-coordinated baseline. As homogeneous patients and stable queueing systems are both common model settings[7,22], our base model provides valuable and non-obvious insights within the classical theoretical setting, challenging the intuitive presumption that greater flexibility (two-way) always dominates.

6 Conclusions

In this paper, we adopt the non-coordinated system as the baseline and investigate two referral mechanisms: the one-way and two-way referrals. To the best of our knowledge, this study is the first to systematically examine and compare the efficacy of these two mechanisms from the perspectives of both healthcare providers and patients.

We characterize the optimal capacity investment and pricing strategies under the three systems and derive the following key findings. First, the two-way referral mechanism may underperform the one-way mechanism when the capacity cost rate of the treatment procedure in the GH is not significantly higher than that of the CHC. Second, we demonstrate the value of coordination, showing that the one-way mechanism benefits both types of hospitals as well as patients. For this superior mechanism, we prove that a grand coalition of all hospitals maximizes total system profit and propose a profit allocation scheme based on the Shapley value to incentivize coalition formation. Finally, we extend the analysis to a more general setting: a congested system with heterogeneous patients and endogenous (referral-scheme-dependent) arrival rates. In this context, the two-way referral system emerges as optimal. Although the results differ from those of the base model, both offer valuable insights for distinct practical scenarios. For congested systems with endogenous arrivals, management should prioritize guiding common patients to the CHC for first diagnosis.

While our work provides important managerial insights into hospital coordination, it is not without limitations. To facilitate analysis, we made several assumptions, such as patients being unable to observe hospital states (e.g., price, capacity, waiting time) and constant marginal costs. Promising directions for future research include investigating coordination when patients choose hospitals based on expected utility and vary in acuity, designing referral mechanisms within medical alliances comprising multiple upper- and lower-level hospitals or multi-tier networks, and exploring systems where one-way referrals flow from GHs to CHCs.

Supporting information

The supporting information for this article can be found online at https://doi.org/10.52396/JUSTC-2023-0052. It includes two parts: S1 and S2. S1 is proofs of assumptions, lemmas, and propositions; S2 is the case with multiple CHCs.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (72222015, 72171215, 72101243) and Key Funds for Youth Innovation of University of Science and Technology of China (YD2040002018).

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Xiaobei Shen is a Special Researcher at the School of Management,

CSTR: 32290.14.JUSTC-2023-0052
DOI: 10.52396/JUSTC-2023-0052
JUSTC, **2025**, 55(8): 0803

University of Science and Technology of China (USTC). She received her Ph.D. degree from The Hong Kong University of Science and Technology in 2013. From 2013 to 2016, she did postdoctoral research at City University of Hong Kong. In 2016, she joined USTC. Her research mainly focuses on data analysis and decision optimization.

Chunfang Yuan is currently an Operations Specialist at Full Truck Alliance Group (Manbang Group). She received her master's degree from the University of Science and Technology of China. Her research mainly focuses on operations management.

References

[1] Song J, Qiu Y Z, Liu Z K. Integrating optimal simulation budget allocation and genetic algorithm to find the approximate Pareto patient flow distribution. *IEEE Transactions on Automation Science and Engineering,* **2015**, *13* (1): 149–159.

[2] Li N, Kong N, Li Q L, et al. Evaluation of reverse referral partnership in a tiered hospital system: A queuing-based approach. *International Journal of Production Research,* **2017**, *55* (19): 5647–5663.

[3] Li M N, Zhang Y, Lu Y, et al. Factors influencing two-way referral between hospitals and the community in China: A system dynamics simulation model. *Simulation,* **2018**, *94* (9): 765–782.

[4] Shumsky R A, Pinker E J. Gatekeepers and referrals in services. *Management Science,* **2003**, *49* (7): 839–856.

[5] Hasija S, Pinker E J, Shumsky R A. Staffing and routing in a two-tier call centre. *International Journal of Operational Research,* **2005**, *1*: 8–29.

[6] Liu X Y, Cai X Q, Zhao R Q, et al. Mutual referral policy for coordinating health care systems of different scales. *International Journal of Production Research,* **2015**, *53* (24): 7411–7433.

[7] Wang J J, Li Z P, Shi J J. Hospital referral and capacity strategies in the two-tier healthcare systems. *Omega,* **2021**, *100*: 102229.

[8] Deo S, Iravani S, Jiang T T, et al. Improving health outcomes through better capacity allocation in a community-based chronic care model. *Operations Research,* **2013**, *61* (6): 1277–1294.

[9] Oh J, Su X M. Reservation policies in queues: Advance deposits, spot prices, and capacity allocation. *Production and Operations Management,* **2018**, *27* (4): 680–695.

[10] Xiang Y S, Zhuang J. A medical resource allocation model for serving emergency victims with deteriorating health conditions. *Annals of Operations Research,* **2016**, *236* (1): 177–196.

[11] Barrera J, Garcia A. Auction design for the efficient allocation of service capacity under congestion. *Operations Research,* **2015**, *63* (1): 151–165.

[12] Çil E B, Lariviere M A. Saving seats for strategic customers. *Operations Research,* **2013**, *61* (6): 1321–1332.

[13] Maglaras C, Yao J, Zeevi A. Optimal price and delay differentiation in large-scale queueing systems. *Management Science,* **2018**, *64* (5): 2427–2444.

[14] Zeng Y L, Zhang L M, Cai X Q, et al. Cost sharing for capacity transfer in cooperating queueing systems. *Production and Operations Management,* **2018**, *27* (4): 644–662.

[15] Gao X Q, Wen J P, Song J. Capacity allocation and revenue sharing in healthcare alliances. *Flexible Services and Manufacturing Journal,* **2020**, *32* (4): 829–851.

[16] Wang H Y, Olsen T L, Liu G Q. Service capacity competition with peak arrivals and delay sensitive customers. *Omega,* **2018**, *77*: 80–95.

[17] Ravner L, Shamir N. Pricing strategy and collusion in a market with delay sensitivity. *Naval Research Logistics,* **2021**, *68* (5): 577–602.

[18] Chen W H, Zhang Z G, Chen X H. On two-tier healthcare system under capacity constraint. *International Journal of Production Research,* **2020**, *58* (12): 3744–3764.

[19] Chen W H, Zhang Z G, Hua Z S. Analysis of price competition in two-tier service systems. *Journal of the Operational Research Society,* **2019**, *70* (11): 1938–1950.

[20] Kroneman M W, Maarse H, van der Zee J. Direct access in primary care and patient satisfaction: A European study. *Health Policy,* **2006**, *76* (1): 72–79.

[21] Mariñoso B G, Jelovac I. GPs' payment contracts and their referral practice. *Journal of Health Economics,* **2003**, *22* (4): 617–635.

[22] Zhou W H, Li X Z, Qian Q. Comparison of gatekeeping and non-gatekeeping designs in a service system with delay-sensitive customers. *Journal of Systems Science and Systems Engineering,* **2021**, *30* (2): 125–150.

[23] Chao X L, Liu L M, Zheng S H. Resource allocation in multisite service systems with intersite customer flows. *Management Science,* **2003**, *49* (12): 1739–1752.

[24] Bekker R, de Bruin A M. Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research,* **2010**, *178* (1): 45–65.

[25] Dobson G, Lee H H, Sainathan A, et al. A queueing model to evaluate the impact of patient "batching" on throughput and flow time in a medical teaching facility. *Manufacturing & Service Operations Management,* **2012**, *14* (4): 584–599.

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

[27] Dobson G, Tezcan T, Tilson V. Optimal workflow decisions for investigators in systems with interruptions. *Management Science,* **2013**, *59* (5): 1125–1141.

[28] Bhattacharjee P, Ray P K. Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: A review and reflections. *Computers & Industrial Engineering,* **2014**, *78*: 299–312.

[29] Wen J P, Jiang H Y, Song J. A stochastic queueing model for capacity allocation in the hierarchical healthcare delivery system. *Asia Pacific Journal of Operational Research,* **2019**, *36* (1): 1950005.

[30] Serra D, Marianov V, ReVelle C. The maximum-capture hierarchical location problem. *European Journal of Operational Research,* **1992**, *62* (3): 363–371.

[31] Sloan F A. Not-for-profit ownership and hospital behavior. In: Handbook of Health Economics. Berlin: Elsevier, **2000**, *1*: 1141–1174.

[32] Meng Q Y, Liu X Z, Shi J S. Comparing the services and quality of private and public clinics in rural China. *Health Policy and Planning,* **2000**, *15* (4): 349–356.

[33] Pan X L, Dib H H, Wang X H, et al. Service utilization in community health centers in China: A comparison analysis with local hospitals. *BMC Health Services Research,* **2006**, *6*: 93.

CSTR: 32290.14.JUSTC-2023-0052
DOI: 10.52396/JUSTC-2023-0052
JUSTC, **2025**, 55(8): 0803