Strategic Queueing Behavior of Two Groups of Patients in a Healthcare System

Youxin Liu 1,2,*, Liwei Liu 1,*, Tao Jiang 3 and Xudong Chai 4

1 School of Mathematics and Statistics, Nanjing University of Science and Technology, Nanjing 210094, China
2 Department of Elementary Teaching, Wuhu Institute of Technology, Wuhu 241003, China
3 College of Economics and Management, Shandong University of Science and Technology, Qingdao 266590, China; jiangt@sdust.edu.cn
1 School of Mathematics-Physics and Finance, Anhui Polytechnic University, Wuhu 241000, China; ahpuxcd@mail.ahpu.edu.cn
* Correspondence: yxliu@whit.edu.cn (Y.L.); lwliu@njust.edu.cn (L.L.)

Abstract: Long waiting times and crowded services are the current medical situation in China. Especially in hierarchic healthcare systems, as high-quality medical resources are mainly concentrated in comprehensive hospitals, patients are too concentrated in these hospitals, which leads to overcrowding. This paper constructs a game-theoretical queueing model to analyze the strategic queueing behavior of patients. In such hospitals, patients are divided into first-visit and referred patients, and the hospitals provide patients with two service phases of “diagnosis” and “treatment”. We first obtain the expected sojourn time. By defining the patience level of patients, the queueing behavior of patients in equilibrium is studied. The results suggest that as long as the patients with low patience levels join the queue, the patients with high patience levels also join the queue. As more patients arrive at the hospitals, the queueing behavior of patients with high patience levels may have a negative effect on that of patients with low patience levels. The numerical results also show that the equilibrium behavior deviates from a socially optimal solution; therefore, to reach maximal social welfare, the social planner should adopt some regulatory policies to control the arrival rates of patients.

Keywords: healthcare system; comprehensive hospital; patients; rational queueing behavior; equilibrium analysis

MSC: 90B22; 60K25

1. Introduction

For a long time, in China, “being difficult and expensive to see a doctor” has been an important problem troubling the government, medical service institutions and patients. However, due to the inherent advantages of comprehensive hospitals in medical quality and service capacity, they usually act as high-level healthcare providers, which leads to an excessive concentration of patients in these hospitals, while the number of patients in community hospitals is relatively small, and part of the resources are scarce and part of the resources are idle. As a result, patients need to wait for a long time to see a doctor, leading to overcrowding in such hospitals.

This paper studies a queueing setting of comprehensive hospitals in which two groups of delay-sensitive patients (first-visit and referred patients) with common or chronic diseases arrive at comprehensive hospitals to get their conditions diagnosed and also receive the required treatment. Both groups line up in a common queue for service based on a first-come, first-served (FCFS) discipline. Our motivation to analyze this setting...
is overcrowding in comprehensive hospitals, as they serve both their own patients and also the patients referred to them, where the referred patients come from the community hospitals for treatments. For example, in comprehensive hospitals for common or chronic diseases, their own patients (e.g., the first-visit patients) need to go through two service phases of “diagnosis” and “treatment”, but the referred patients (e.g., the patients who firstly seek service at the community hospitals and cannot be cured) need to line up in a common queue with the first-visit patients, bypass the “diagnosis” and go directly to “treatment”. For these patients with common or chronic diseases, they are more concerned about their waiting time in hospitals. Therefore, service delay is an important indicator of patients perceived value, and the length of waiting time has an important effect on patients' strategic decisions. Since queueing theory is the main tool for solving the queueing problem in healthcare systems, we could incorporate queueing economics in comprehensive hospitals. On the basis of the queueing setting in comprehensive hospitals, in this study, we model the hospitals as queueing systems for a mixture of two types of patients with different delay sensitivities and service rewards, where the first-visit patients go through two service phases of “diagnosis” and “treatment”, and the referred patients only need “treatment”. To the best of our knowledge, the strategic behavior of patients in such queueing systems has not been analyzed and will be a good research direction.

We present a queueing-theoretic formulation of healthcare systems. In healthcare systems, healthcare providers with a limited capacity are often faced with delay-sensitive patients. For the sake of analytical tractability, in our work, we assume that patients arrive at the system according to a Poisson process. Patients are divided into two groups according to their medical condition and delay sensitivities as well as their valuation of service offerings. More generally, we do not set a limit on the relative magnitude of the delay sensitivity and the valuation of service offerings between the two groups of patients. Since the cost of diagnosis is relatively lower than that of treatment, it is assumed that the cost of service is the same for the two types of patients, and the influence of service cost on the decision-making of delay-sensitive patients is not considered in this paper. Therefore, based on the valuation of service offerings and the expected cost of waiting in hospitals, each patient who arrives at the comprehensive hospital decides whether to join the queue or balk. Different from the assumption in Wen et al. (2019) that the hospital provides two separate queues [1], namely a common line for first-visit patients and a green line for referred patients, in this paper, we assume that the two groups of patients line up in a common queue. Therefore, in this paper, we adopt the analysis method of Tang et al. (2018) [2] to deal with the queueing problem in comprehensive hospitals, study how the delay of the pooled queue affects patients' queueing decisions and derive the queueing behavior of patients in equilibrium.

There are three streams in the literature related to our study: queueing behavior of customers, various customer classifications in behavioral queueing models and queueing economics in healthcare systems.

The first related stream of research is the analysis of queueing behavior of customers. In general, the literature on the queueing behavior of customers can be roughly divided into two categories: observable and unobservable queues. In the observable queues, the problem that needs to be analyzed is the dynamic control of queues, whereas in the unobservable queues, customers rely on their net service utilities to make queueing decisions. Naor (1969) [3] is the first scholar to study customers' strategic behavior in queueing systems. In Naor's model, the author considers an observable M/M/1 queue, where customers decide to either join the queue or balk based on the service reward and the waiting cost. Edelson and Hilderbrand (1975) [4] extended Naor's model and considered the strategic queueing behavior of customers in an unobservable M/M/1 queueing system. Since then, by allowing the customers to make their queueing decisions, a large amount of the literature has studied both the observable and unobservable queues, and studied the relationship between socially and individually optimal queueing decisions; interested readers
may refer to the good survey of studies; see, e.g., Hassin and Haviv (2003) [5], Hassin (2016) [6], Ibrahim (2018) [7] and Economou (2022) [8]. In healthcare systems, service delay and waiting time are major concerns for delay-sensitive patients; hence, we need to introduce a queueing model in the healthcare systems and study the patients queueing decisions.

The second related stream of research studies various customer classifications in behavioral queueing models. Roughly speaking, depending on the number of customer types in the queue, the literature on behavioral queueing models can also be divided into two categories: homogeneous and heterogeneous customers. In the first category, there is no difference between customers, i.e., all customers have the same service reward, the same sensitivity to delay, share the same information level, etc. Therefore, all customers are indistinguishable; in the observable queue, the queueing decision of a customer is independent of the strategies of all other customers, and in the unobservable queue, the game among customers is a symmetric game. The classic results can be seen in the review literature: Hassin and Haviv (2003) [5] and Hassin (2016) [6]. However, in the second category, customers are heterogeneous as they may have different sensitivities to delay, different service rewards, different psychological characteristics, heterogeneous information, etc. So far, there is also a large number of studies in the literature on various customer classifications in behavioral queueing models. For example, Ni et al. (2013) [9] classified customers according to “customer intensity” and studied the revenue-maximization service provider’s price–speed decision in customer-intensive service systems. In Zhou et al. (2014) [10], the customers are categorized into two types based on the valuation for service and sensitivity to delay, and the authors solved the optimal uniform pricing problem by using a queueing model with two types of customers. Considering whether customers have sufficient information about the service, Zhou et al. (2014) [11] divided customers into informed and uninformed customers and solved the problem of when service enterprises should provide free experience service to uninformed customers. Tang et al. (2018) [2] classified customers into two types based on workload and delay sensitivity and obtained the equilibrium queueing strategies of the two types of customers. Based on the availability of queue length information and system state, Hu et al. (2018) [12] and Wang and Wang (2019) [13] classified customers into informed and uninformed customers. In the M/M/1 queue and retrial queue, the authors, respectively, studied the equilibrium strategies of customers and the effect of information heterogeneity on throughput and social welfare. Wang and Fang (2022) [14] took the heterogeneity of priority awareness as a classification criterion; Wang and Sun (2022) [15] classified customers according to customer service experience, according to whether customers choose to stay on-site when the system provides services; and Hanukov et al. (2023) [16] also divided customers into questioning customers and trusting customers; then, the authors studied the customer equilibrium strategy and system optimization decision-making in the priority queueing system, online service queueing system, queue-inventory system and other related systems.

In comprehensive hospitals for common or chronic diseases, as they serve both their own patients and also the patients referred to them, we need to categorize patients into first-visit and referred patients, so that we can investigate the mutual influence mechanism of the decision-making process of two types of customers and deduce the equilibrium queueing strategies of patients.

Our study is also very relevant to queueing economics in healthcare systems. A lot of researchers are interested in using queueing models to study patient queueing problems in hierarchic healthcare systems. So far, according to the composition of the institution in hierarchic healthcare systems, these systems can be broadly divided into two categories: horizontal and vertical hierarchic healthcare systems. In the first category, the hierarchic healthcare systems consist of public hospitals offering free services and private hospitals offering toll services. In the second category, the hierarchic healthcare systems consist of two types of hospitals, namely comprehensive hospitals with high-quality medical resources and community or primary hospitals with low-quality medical resources, where
uncured patients could be referred to comprehensive hospitals from community hospitals and cured patients could be transferred from comprehensive hospitals to community hospitals for further rehabilitation therapy.

For the first category, the related literature includes Guo et al. (2014) [17], Chen et al. (2015) [18], Hua et al. (2016) [19], Qian and Zhuang (2017) [20], Wan and Wang (2017) [21], Qian et al. (2017) [22], Zhang and Yin (2021) [23,24], Zhou et al. (2022) [25], Chen (2023) [26], and Hu et al. (2024) [27]. In Guo et al. (2014) [17], the authors analyzed the pricing and capacity decisions of a two-tier medical service system under the condition of ensuring self-financing. In Chen et al. (2015) [18] and Hua et al. (2016) [19], the authors considered two subsidy schemes and analyzed the competition between public and private hospitals in a two-tier medical service system. The main results showed that a relatively small rate could perfectly harmonize the two-tier medical service system, and at the same time, the subsidy coordination method can effectively reduce the waiting time in public hospitals and improve social welfare. In Qian and Zhuang (2017) [20], the authors studied the coordination mechanism of tax/subsidy and service capacity planning of a two-tier medical service system in terms of welfare redistribution. The results showed that tax subsidies or capacity planning for hospitals can induce patients with different time-delay sensitivities to choose different hospitals. In Wan and Wang (2017) [21], from the perspectives of patient-waiting-time minimization and social welfare maximization, the authors analyzed the optimal decision of a two-tier medical service system. In Qian et al. (2017) [22], the authors studied the strategy of reducing patients waiting time through a subsidy mechanism in the public medical service system. By analyzing the public medical service system, the authors obtained the optimal subsidy mechanism for subsidy and waiting time. In Zhang and Yin (2021) [23], the authors defined the mixed information and non-real-time information cases, and then based on the matrix analytic method, they proposed a computational approach to analyze the system performance and examine the joint effect of delay information and pricing on the system performance. In Zhang and Yin (2021) [24], the authors investigated a two-tier service system with customers asymmetric preference for charged-service and free-service providers. By constructing an M/M/1 queueing model, they derived the customers choice in equilibrium and found that in some cases, the two-tier service system could solve the over-congestion problem and reduce the total social cost. Moreover, in Zhou et al. (2022) [25], the authors studied a mixed duopoly service system with private and public service providers. In Chen (2023) [26], the author investigated a two-tier co-payment healthcare system under a uniform pricing and subsidy coordination mechanism. In Hu et al. (2024) [27], the authors modeled privatized public service systems as queueing systems and displayed whether the government adopting myopic adjustment plays a critical role in choosing the regulation instrument.

For the second category, the related literature includes Li et al. (2017) [28], Li et al. (2019) [29], Wen et al. (2019) [1], Zhou et al. (2021) [30], Li et al. (2021) [31], Li et al. (2021) [32], Wang et al. (2021) [33, Rajan et al. (2019) [34], and Li et al. (2023) [35,36]. For example, Li et al. (2017) [28] considered the reverse referral (upstream referral) in the tiered healthcare system with delay-sensitive patients and used the queueing approach to examine the effect of reverse referral partnerships. Li et al. (2019) [29] considered a gatekeeping system with heterogeneous patients and investigated the effects of online inquiry service on performance by using the queueing-game theory. In the hierarchical healthcare system, Wen et al. (2019) [1] proposed a stochastic tandem queueing model to obtain the optimal capacity allocation. In a referral healthcare system, Zhou et al. (2021) [30] considered gatekeeping and non-gatekeeping settings. By using the queueing approach, the authors compared the effect of the two settings on a social planner’s capacity decision. Li et al. (2021) [31] considered an operational-level control agreement framework and developed a multifidelity model-based optimization approach to solve the Pareto optimization for control agreement in patient referral coordination. Li et al. (2021) [32] developed two payment schemes to facilitate capacity sinking. The authors constructed a four-stage game model.
under the queueing-game theory framework to study the capacity reallocation in a down-
stream referral healthcare system. Wang et al. (2021) [33] investigated the hospital referral
and capacity strategies in a downstream referral healthcare system, in which the authors
obtain the equilibrium strategy of the comprehensive hospital provider’s referral rate and
primary hospital provider’s capacity level. Furthermore, the authors also explored the im-
 pact of some parameters on the referral healthcare system. In addition to the literature on
referral healthcare systems, Rajan et al. (2019) [34] also considered a healthcare system
with heterogeneous patients. In the healthcare system, the authors investigated the effects
of telemedicine technology on patient utility and healthcare provider’s operating deci-
sions. Moreover, the authors presented some policy implications for facilitating the fur-
ther development of telemedicine in chronic care. Li et al. (2023) [35,36] proposed some
contract mechanisms to solve healthcare imbalance in hierarchical healthcare systems.

In contrast to these studies, we model the comprehensive hospital as a queueing sys-
tem with a mixture of two types of patients and two service phases, rather than the classic
M/M/1 queue with homogeneous patients, which makes the analysis more challenging.
Moreover, different from the homogeneous patients who have either the same service re-
ward or the same sensitivity for delay, we assume that the two types of patients are heter-
ogeneous in both delay sensitivity and service reward, and the patients line up in a
pooled queue rather than two dedicated queues. Under this assumption, either group of
patients may have a higher incentive to join the hospital, rather than one group of patients
always having a higher incentive than the other. As a result, we need to distinguish dif-
ferent scenarios to solve the problem under consideration. Finally, we assume that the two
groups of patients arriving at the comprehensive hospital are two independent Poisson
flows, so that, the proportion of one type of patient is no longer regarded as exogenous
but endogenous through the two independent Poisson flows. Based on these differences,
we aim to find how the delay of the pooled queue affects patients’ queueing decisions and
derive the queueing behavior of patients in equilibrium.

This paper is structured as follows: Section 2 incorporates the heterogeneity of pa-
tients into a queueing system and describes the queueing model. Section 3 gives the con-
crete expression of the expected sojourn time of an arbitrary patient after deriving the
steady-state probabilities of the system. Section 4 derives the equilibrium queueing strat-
gegies of the two groups of patients. Section 5 presents some numerical examples to verify
the correctness of the theoretical results. Section 6 is a separate discussion section to brief
the research findings. Section 7 concludes this study.

2. Model Formulation

Consider a monopolistic comprehensive hospital in a healthcare system that provides
patients with two service phases of “diagnosis” and “treatment”. The service settings in
the comprehensive hospital are shown in Figure 1. Patients are divided into two groups:
the first-visit patients (labeled d) and the referred patients (labeled t). The arrival process
of first-visit and referred patients is independent of each other; they arrive at the system
according to a Poisson process with rates \( \lambda_d \) and \( \lambda_t \). As a result, the total arrival rate of
patients is \( \lambda = \lambda_d + \lambda_t \), the fraction of first-visit patients is \( \gamma = \frac{\lambda_d}{\lambda_d + \lambda_t} \) and the fraction of
referred patients is \( 1 - \gamma \), which is endogenous via the two independent Poisson
flows.

The two groups of patients differ in three ways. First, the two groups of patients come
from different sources and have different medical experiences. The first-visit patients do
not receive any medical diagnosis and treatment before reaching the comprehensive hos-
pital. After arriving at the hospital, they need to line up in the queue and go through two
service phases of “diagnosis” and “treatment”. However, the referred patients who have
received medical diagnosis and treatment at community hospitals cannot be cured and
need to be referred to the comprehensive hospital, where they queue up with the first-visit patients, bypass the “diagnosis” and go directly to “treatment”. Second, the two groups of patients are heterogeneous in their delay sensitivities, denoted by $\theta_d$, $\theta_t$. Third, the valuation of service offerings (the service reward obtained after the service is completed) for both groups of patients is different, denoted by $R_d$, $R_t$. For the last two aspects, we do not set a limit on the relative magnitude of the delay sensitivity and the valuation of service offerings between the two groups of patients.

Moreover, the “diagnosis” and “treatment” are independent. It is assumed that the service time in the “diagnosis” and “treatment” service phases follows the exponential distribution with the parameters $\mu_d$ and $\mu_t$, respectively. The hypotheses of the Poisson arrival process and exponential service time in hospitals have been empirically tested by Kim et al. (1999) [37] and have been widely used in the relevant medical operations management literature. An arriving patient, if required, joins the queue. The discipline is FCFS and is not related to the type of service.

Figure 1. Queueing model settings for the comprehensive hospital.

Actually, by this assumption, the service time in our model is type-dependent, i.e., the service time of first-visit patients is the sum of two exponential distributions of mean $1/\mu_d$ and $1/\mu_t$, and that of referred patients follows an exponential distribution with mean $1/\mu_t$. Therefore, the problem studied can be modeled as a special M/H2/1 queue, where the unconditional service time does not follow an exponential distribution, but rather the sum of two exponential distributions of mean $1/\mu_d$ and $1/\mu_t$ with probability $\gamma$, and an exponential distribution of mean $1/\mu_t$ with $1-\gamma$, which is a weighting of two distributions and represents a hyper-exponential distribution. By Pollaczeck–Khintchine lemma, we may obtain the expected sojourn time of an arbitrary patient. Next, we first obtain the stability condition, then we use the queueing theory and construct a continuous-time Markov chain to obtain the expected sojourn time of an arbitrary patient. Tables 1 and 2 list the established model parameters and the system variables in the subsequent content.

Table 1. The established model parameters.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_d$ ($\lambda_t$)</td>
<td>The arrival rate of first-visit patients (referred patients)</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>The total arrival rate of patients</td>
</tr>
<tr>
<td>$\gamma$ ($\overline{\gamma}$)</td>
<td>The fraction of first-visit patients (referred patients)</td>
</tr>
<tr>
<td>$\theta_d$ ($\theta_t$)</td>
<td>The unit-time waiting cost of first-visit patients (referred patients)</td>
</tr>
</tbody>
</table>
The service reward received by a first-visit patient (referred patient) once the service is complete, or the perceived value for first-visit patients (referred patients) 

\[ R_d(R_i) \]

The expected service time of a first-visit patient (referred patient) 

\[ \frac{1}{\mu_d} \quad \text{or} \quad \frac{1}{\mu_i} \]

The expected sojourn time of an arbitrary patient 

\[ W(\lambda_d, \lambda_i) \]

The utility of an arriving first-visit patient (referred patient) who decides to join the queue 

\[ U_d(U_i) \]

The patience level of first-visit patients (referred patients) 

\[ \beta_d(\beta_i) \]

---

**Table 2. Notation and system variables.**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q_{n,j}^d )</td>
<td>The unique individual optimal pure strategy of first-visit patients (referred patients) in the fully observable case</td>
</tr>
<tr>
<td>( \lambda^e_d(\lambda^e_i) )</td>
<td>The equilibrium queueing behavior of first-visit patients (referred patients) in the unobservable case</td>
</tr>
<tr>
<td>( q^e_d ) ( (q^e_i) )</td>
<td>The equilibrium joining probability of first-visit patients (referred patients) in the unobservable case</td>
</tr>
<tr>
<td>( q^{soc}_{d} ) ( (q^{soc}_i) )</td>
<td>The social optimal strategies of first-visit patients (referred patients)</td>
</tr>
</tbody>
</table>

---

3. Performance Analysis

3.1. Stability Condition

Let \( L(t) \) denote the number of patients in the system at time \( t \) and \( J(t) \) denote the service phase provided by the hospital at time \( t \), where 

\[
J(t) = \begin{cases} 
0, & \text{the hospital is in the service phase of "diagnosis" at time } t, \\
1, & \text{the hospital is in the service phase of "treatment" at time } t.
\end{cases}
\]

Obviously, \( X(t) = \{(L(t), J(t)), t \geq 0]\) is a continuous-time Markov chain with state space \( \Omega=\{(0)\} \cup \{(k, j) : k \geq 1, j = 0,1\} \). The transition rate diagram is given in Figure 2.

![Figure 2. Transition rate diagram of \( \{(L(t), J(t)), t \geq 0]\) .](image)

Considering the state change of the Markov chain, we have the following state-transition-rate matrix:
We first use Theorem 1.7.1 of Neuts (1981) [38] to derive the stability condition of the underlying queueing system.

Lemma 1. The queueing system is stable if and only if 
$$
\lambda_d \mu_d + \lambda_e (\mu_e + \mu_d) < \mu_e \mu_d.
$$

Proof of Lemma 1. The proof process is provided in the Appendix A.

3.2. Expected Sojourn Time

To obtain the expected sojourn time of an arbitrary patient, we first define the steady-state probabilities by

$$
\pi_{k,j} = \lim_{t \to +\infty} P(L(t) = k, J(t) = j), k \geq 1, j = 0,1, \pi_0 = \lim_{t \to +\infty} P(L(t) = 0)
$$

Then, by constructing the balance equations and obtaining the steady-state probabilities, we could derive the expected sojourn time of an arbitrary patient. The following proposition gives the explicit solution of the expected sojourn time:

Proposition 1. The expected sojourn time of an arbitrary patient is given by

$$
W(\lambda_d, \lambda_e) = \left(\frac{\lambda_d}{\mu_d (\lambda_d + \lambda_e)} + \frac{1}{\mu_e} \right) \left[ 1 + \frac{\lambda_d + \lambda_e}{\mu e} \right] - \frac{\lambda_d + \lambda_e}{\mu_e \mu_d}.
$$

Proposition 1 identifies the expected sojourn time of an arbitrary patient given the arrival rates $\lambda_d$ and $\lambda_e$. From the expression, we find that the expected sojourn time should be a weighted average of the sojourn time for the two types of patients. Next, we discuss the properties of the result by the following proposition.

Proposition 2. (i) If $\lambda_d$ and $\lambda_e$ satisfy $\lambda_d > 0$, $\lambda_e > 0$ and $\lambda_d \mu_d + \lambda_e (\mu_e + \mu_d) < \mu_e \mu_d$, then

$$
\frac{\partial W(\lambda_d, \lambda_e)}{\partial \lambda_d} > 0;
$$
(ii) If \( \lambda_d \) and \( \lambda_i \) satisfy \( \lambda_d > 0 \), \( \lambda_i > 0 \), \( \lambda_d \mu_i + \lambda_i (\mu_i + \mu_d) < \mu_i \mu_d \), and 
\[ \lambda_d \in [\bar{\lambda}_d, \frac{\mu_i \mu_d}{\mu_i + \mu_d}], \] 
where \( \bar{\lambda}_d \) satisfies the equation 
\[ \frac{\mu_i^2 + \bar{\lambda}_d \mu_i}{(\mu_i \mu_d - \bar{\lambda}_d \mu_i - \bar{\lambda}_d \mu_d)^2} = \frac{1}{\bar{\lambda}_d \mu_d}, \] 
then 
\[ \frac{\partial W(\lambda_d, \lambda_i)}{\partial \lambda_i} > 0. \]

**Proof of Proposition 2.** The proof process is provided in the Appendix A.

Proposition 2 shows how the arrival rates \( \lambda_d \) and \( \lambda_i \) affect the expected sojourn time of patients in a hospital. See Figure 3 for the illustration. As shown in Figure 3, with different arrival rates \( \lambda_i \), \( W(\lambda_d, \lambda_i) \) is monotonically increasing in \( \lambda_d \). However, with different arrival rates \( \lambda_d \), as \( \lambda_i \) increases, the expected sojourn time no longer increases monotonously. This interesting counter-intuitive phenomenon may be attributed to two factors: (1) the differences in service time between two groups of patients; and (2) whether the hospital is congested. When the arrival rates of two types of patients are very small, the hospital is not congested, and relatively speaking, the hospital has enough service capacity at this time; therefore, the waiting time of patients in the queue (queueing time) is almost negligible, and the sojourn time (the sum of waiting time in the queue and the service time) is only related to the service time. Since the service time of first-visit patients is longer than that of referred patients, and the expected sojourn time of patients is a weighted average of the two types of patients sojourn time, when the arrival rate of referred patients increases gradually from zero, the proportion of referred patients increases, which leads to a gradual decrease in the weighted average sojourn time. However, when the arrival rate of referred patients exceeds a threshold, the expected sojourn time increases as the arrival rate of referred patients increases.

Proposition 2 indicates that the waiting time and sojourn time have different properties in our queueing model. When the hospital has enough service capacity, patients can receive service immediately without waiting in the queue; in this case, the waiting time in the queue can be ignored, and the sojourn time is equal to the service time. Only when the hospital is congested, that is, patients need to wait for service, the sojourn time of patients in the hospital is mainly the waiting time in the queue, and the more congested queue always leads to the longer expected waiting time and expected sojourn time. In order to reduce the sojourn time of patients in hospitals, the result could provide the following management implications for medical service managers: (1) when the arrival rates of the two types of patients remain unchanged, the most direct way is to increase investment and expand service capacity; and (2) without changing the service capacity, comprehensive hospitals should control the arrival rate of first-visit patients below a certain level and adjust the arrival rate of acceptable referred patients based on the service capacity and specific threshold.
Remark 1. \( \frac{\partial W(\lambda_d, \lambda_i)}{\partial \lambda_i} < \frac{\partial W(\lambda_d, \lambda_i)}{\partial \lambda_d} \), which indicates that the marginal impact of first-visit patients on the expected sojourn time is larger than that of referred patients. This is because first-visit patients need to go through two service phases of “diagnosis” and “treatment”, so they have a stronger impact on the sojourn time than referred patients.

4. Equilibrium Queueing Behavior of Two Groups of Patients

4.1. Analysis of Fully Observable Case

We begin the analysis by studying the fully observable case. In this case, each patient knows their own type and also receives exact information about the state of the queueing system upon arrival, i.e., they get informed about the number of patients in the system and the service phase being provided by the hospital.

For a first-visit patient, when they find the system at a state \((0)\) upon their arrival, then, if they join the queue, their expected sojourn time is equal to \(\frac{1}{\mu_i} + \frac{1}{\mu_d}\). When they find the system at a state \((n, 0), n > 0\), their expected sojourn time \(W_{(n,0)}^d\) is equal to

\[
[(n-1)\gamma + 2] \left( \frac{1}{\mu_i} + \frac{1}{\mu_d} \right) + (n-1)(1-\gamma) \frac{1}{\mu_i}.
\]

When they find the system at a state \((n, 1), n > 0\), their expected sojourn time \(W_{(n,1)}^d\) is equal to

\[
[(n-1)\gamma + 1] \left( \frac{1}{\mu_i} + \frac{1}{\mu_d} \right) + [(n-1)(1-\gamma) + 1] \frac{1}{\mu_i}.
\]

For a referred patient, when they find the system at a state \((0)\) upon their arrival, then, if they join the queue, their expected sojourn time is equal to \(\frac{1}{\mu_d}\). When they find the system at a state \((n, 0), n > 0\), their expected sojourn time \(W_{(n,0)}^r\) is equal to

\[
[(n-1)\gamma + 1] \left( \frac{1}{\mu_i} + \frac{1}{\mu_d} \right) + [(n-1)(1-\gamma) + 1] \frac{1}{\mu_i}.
\]
Their expected sojourn time $t_{nW}$ is equal to
\[(n-1)\gamma\left(\frac{1}{\mu_t} + \frac{1}{\mu_d}\right) + [(n-1)(1-\gamma) + 2]\frac{1}{\mu_t}.

**Proposition 3.** (1) In the fully observable case, the unique individual optimal pure strategy of first-visit patients is as follows:

Case 1: If $R_d < \Theta_d\left(\frac{1}{\mu_t} + \frac{1}{\mu_d}\right)$, balking is the unique individual optimal pure strategy.

Case 2: If $R_d \geq \Theta_d\left(\frac{1}{\mu_t} + \frac{1}{\mu_d}\right)$, there exists a unique individual optimal pure strategy which has the following forms:

\[q_{n,j}^d = \begin{cases} 1, & n \leq n_d^j, j = 0,1, \\ 0, & n > n_d^j, \end{cases}
\]

where $n_d^0 = \left[n_0^*\right]$, $n_0^* = \frac{R_d - 2\left(\frac{1}{\mu_t} + \frac{1}{\mu_d}\right)}{\Theta_d - 2\left(\frac{1}{\mu_t} + \frac{1}{\mu_d}\right)} + 1$, $n_d^1 = \left[n_1^*\right]$, $n_1^* = \frac{R_d - \Theta_d\left(\frac{2}{\mu_t} + \frac{1}{\mu_d}\right)}{\Theta_d - \Theta_d\left(\frac{2}{\mu_t} + \frac{1}{\mu_d}\right)} + 1$.

(2) In the fully observable case, the unique individual optimal pure strategy of referred patients is as follows:

Case 1: If $R_i < \frac{\Theta_i}{\mu_t}$, balking is the unique individual optimal pure strategy.

Case 2: If $R_i \geq \frac{\Theta_i}{\mu_t}$, there exists a unique individual optimal pure strategy which has the following forms:

\[q_{n,j}^i = \begin{cases} 1, & n \leq n_i^j, j = 0,1, \\ 0, & n > n_i^j, \end{cases}
\]

where $n_i^0 = \left[n_0^{**}\right]$, $n_0^{**} = \frac{R_i - 2\left(\frac{1}{\mu_t} + \frac{1}{\mu_d}\right)}{\Theta_i - 2\left(\frac{1}{\mu_t} + \frac{1}{\mu_d}\right)} + 1$, $n_i^1 = \left[n_1^{**}\right]$, $n_1^{**} = \frac{R_i - \Theta_i\left(\frac{2}{\mu_t} + \frac{1}{\mu_d}\right)}{\Theta_i - \Theta_i\left(\frac{2}{\mu_t} + \frac{1}{\mu_d}\right)} + 1$.

**Proof of Proposition 3.** The proof process is provided in the Appendix A.

Proposition 3 presents the explicit equilibrium strategies of patients in the fully observable case, which is a four-threshold strategy $[n_0^0, n_1^0, n_0^1, n_1^1]$. From the results, we observe that the service time, the delay sensitivity, the service reward and the fraction of
first-visit patients have an impact on the four-threshold strategy. The difference between \( n_d^j \) and \( n_i^j \), \( j = 0, 1 \) depends on the magnitude of the relationship between \( R_d / \theta_d \) and \( R_i / \theta_i \).

4.2. Analysis of Fully Unobservable Case

Next, we focus on the fully unobservable case. In the unobservable case, patients cannot observe the number of patients in the system and the service phase being provided by the hospital. Thus, all patients decide whether to join the queue or balk based on the evaluation of service offerings and the expected waiting cost. Given the arrival rates \( \lambda_d \) and \( \lambda_i \), the utility of a tagged type-\( i \) patient who decides to join the queue can be defined as

\[
U_i = R_i - \theta W(\lambda_d, \lambda_i), i = d, t .
\]

Define \( W(\lambda_d) = W(0, \lambda_d) = \frac{1}{\mu_i - \lambda_i} \) and \( W(\lambda_i, 0) = \frac{\mu_i + \mu_d - \lambda_d}{\mu_i (\mu_d - \lambda_d \mu_i - \lambda_d \mu_d)} \), we first characterize the rational equilibrium behavior of patients under two extreme cases.

**Lemma 2.** If \( \lambda_d = 0 \), i.e., only the referred patients arrive at the comprehensive hospital, then the equilibrium queueing behavior of referred patients is given by

\[
\lambda_i^e = \begin{cases} 
\lambda_i, & \beta_i \geq \frac{1}{\mu_i - \lambda_i} \\
\mu_i - \frac{1}{\beta_i} \frac{1}{\mu_i} \leq \beta_i < \frac{1}{\mu_i - \lambda_i}. & \\
0, & \beta_i < \frac{1}{\mu_i} 
\end{cases}
\]

**Lemma 3.** If \( \lambda_i = 0 \), i.e., only the first-visit patients arrive at the comprehensive hospital, then the equilibrium queueing behavior of first-visit patients is given by

\[
\lambda_d^e = \begin{cases} 
\lambda_d, & \beta_d \geq \frac{\mu_i + \mu_d - \lambda_d}{\mu_i (\mu_d - \lambda_d \mu_i - \lambda_d \mu_d)} \\
\frac{\beta_d \mu_i \mu_d - (\mu_i + \mu_d)}{\beta_d (\mu_i + \mu_d) - 1} + \frac{1}{\mu_i} \frac{1}{\mu_d} \leq \beta_d < \frac{\mu_i + \mu_d - \lambda_d}{\mu_i \mu_d - \lambda_d \mu_i - \lambda_d \mu_d}. & \\
0, & \beta_d < \frac{1}{\mu_i} + \frac{1}{\mu_d} 
\end{cases}
\]

**Proof of Lemma 3.** The proof process is provided in the Appendix A.

Lemma 2 and Lemma 3 identify the equilibrium form of patients for the extreme cases \( (\lambda_d = 0 \text{ and } \lambda_i = 0) \). Under the two extreme cases, the underlying queueing system respectively degrades to a classical M/M/1 queue and an M/M/1 queue with two service phases; therefore, the equilibrium arrival rate can be easily obtained.

Next, we consider the general case. Using the method of Tang et al. (2018) [2], we define \( \beta_i = \frac{R_i}{\theta_i} \), \( i = d, t \), which represents the maximum time that a type-\( i \) patient is willing to stay in the comprehensive hospital. According to the model description, although
the service time in our model is type-dependent, the problem studied (the M/M/1 queue with two groups of patients) is also equivalent to an M/H₂/1 queue with two groups of patients, where H₂ represents that the service time follows a hyper-exponential distribution. Since the two groups of patients queue up in the same queue, the expected sojourn time is the same for both groups. At this time, as long as the group of patients with low patience levels choose to join the queue, the group of patients with high patience levels will join the queue (they think they will also receive service soon). For instance, \( \beta_d > \beta_i \) (respectively, \( \beta_d < \beta_i \)) means that the first-visit patients (referred patients) have a relatively higher level of patience than the referred patients (first-visit patients), as long as the referred patients (first-visit patients) decide to join the queue, then all the first-visit patients (referred patients) also join the queue.

It is worth noting that, if the two groups of patients differ in only one aspect, either in the service reward or in the sensitivity to delay, then one group of patients always has a higher patience level (a higher incentive to join the queue) than the other, i.e., either \( \beta_d > \beta_i \) or \( \beta_d < \beta_i \); however, if the two groups of patients differ in two aspects (service reward and the sensitivity to delay), either group of patients may have a higher patience level, i.e., the relationship between \( \beta_d \) and \( \beta_i \) is uncertain. Therefore, we need to analyze the rational equilibrium queueing strategies of the two groups of patients by distinguishing the relationship between \( \beta_d \) and \( \beta_i \).

It should be noted that, if \( \beta_d = \beta_i \), since the two groups of patients queue up in the same queue, there is no difference in decision-making behavior between the two groups of patients. Therefore, when \( R_i - \theta W(\lambda_d, \lambda_i) > 0 \), i.e., \( \beta_i > W(\lambda_d, \lambda_i) \), all patients join the queue, the corresponding equilibrium arrival rates \((\lambda_d^e, \lambda_i^e) = (\lambda_d, \lambda_i)\); when \( R_i - \theta \lim_{\lambda_i \to 0} W(\lambda_i) < 0 \), i.e., \( \beta_i < 1/\mu \), all patients balk the queue, the corresponding equilibrium arrival rates \((\lambda_d^e, \lambda_i^e) = (0, 0) \). When \( R_i - \theta W(\lambda_d, \lambda_i) \leq 0 \) and \( R_i - \theta \lim_{\lambda_i \to 0} W(\lambda_i) \geq 0 \), there exists a unique equilibrium joining probability of patients \( q^e \) that satisfies \( R_i - \theta W(\lambda_d^e, \lambda_i^e) = 0 \). The corresponding equilibrium arrival rates \((\lambda_d^e, \lambda_i^e) = (\lambda_d^e, \lambda_i^e) \). Below, we study the equilibrium behavior of patients for the case \( \beta_d \neq \beta_i \).

**Proposition 4.** When \( \lambda_d \), \( \lambda_i \) satisfy \( \lambda_d > 0, \lambda_i > 0 \) and \( \lambda_d \mu_d + \lambda_i (\mu_i + \mu_d) < \mu_d \mu_i \), if \( \beta_d < \beta_i \), the equilibrium queueing behavior of patients can be obtained as follows.

(i) If \( \beta_d < W(\lambda_i) \leq \beta_i \), the equilibrium arrival rates of the two groups of patients \( \lambda_d^e = 0, \lambda_i^e = \lambda_i \);

(ii) If \( \beta_d < \beta_i < W(\lambda_i) \), the equilibrium arrival rates of the two groups of patients \( \lambda_d^e = 0, \lambda_i^e = \max \left( \mu_i - \frac{1}{\beta_i}, 0 \right) \);

(iii) If \( \beta_i > \beta_d \geq W(\lambda_d, \lambda_i) > W(\lambda_i) \), the equilibrium arrival rates of the two groups of patients \( \lambda_d^e = \lambda_d, \lambda_i^e = \lambda_i \);

(iv) If \( \beta_i > \beta_d \geq W(\lambda_i) \) and \( \beta_d < W(\lambda_d, \lambda_i) \), the equilibrium arrival rates of the two groups of patients \( \lambda_d^e = \lambda_i, \lambda_i^e \) satisfies the following equation:
Proof of Proposition 4. The proof process is provided in the Appendix A.

Proposition 5. When \( \lambda_d, \lambda_i \) satisfy \( \lambda_d > 0, \lambda_i > 0 \), \( \lambda_d\mu_d + \lambda_i\mu_i < \mu_d \mu_i \) and
\[
\lambda_d \leq \frac{\mu_i \mu_d}{\mu_i + \mu_d},
\]
where \( \lambda_d \) satisfies the equation
\[
\frac{\mu_i + \lambda_d\mu_i}{(\mu_i\mu_d - \lambda_d\mu_i - \lambda_d\mu_i)^2} = \frac{1}{\lambda_d\mu_d},
\]
if \( \beta_d > \beta_i \), the equilibrium queueing behavior of patients can be obtained as follows:

(i) If \( \beta_i < W(\lambda_d) \leq \beta_d \), the equilibrium arrival rates of the two groups of patients
\[
\lambda_d^* = \lambda_d, \quad \lambda_i^* = 0;
\]

(ii) If \( \beta_i < \beta_d < W(\lambda_d) \), the equilibrium arrival rates of the two groups of patients
\[
\lambda_d^* = \max\left(\frac{\beta_d\mu_i\mu_d - (\mu_i + \mu_d)}{\beta_d(\mu_i + \mu_d) - 1}, 0\right), \lambda_i^* = 0;
\]

(iii) If \( \beta_d > \beta_i \geq W(\lambda_d), \lambda_i \), the equilibrium arrival rates of the two groups of patients
\[
\lambda_d^* = \lambda_d, \quad \lambda_i^* = \lambda_i;
\]

(iv) If \( \beta_i > \beta_d \geq W(\lambda_d) \) and \( \beta_d < W(\lambda_d), \lambda_i \), the equilibrium arrival rates of the two groups of patients
\[
\lambda_d^* = \lambda_d - \lambda_d, \quad \lambda_i^* = \lambda_i;
\]

satisfies the following equation:
\[
\left(\frac{\lambda_d}{\mu_d(\lambda_d + \lambda_i^*)} + \frac{1}{\mu_i}\right)\left[1 + \frac{(\lambda_d + \lambda_i^*)(\mu_d + \mu_i - (\lambda_d + \lambda_i^*))}{\mu_d\mu_i - \lambda_d\mu_i - (\lambda_d + \lambda_i^*)\mu_d}\right] - \frac{\lambda_d + \lambda_i^*}{\mu_d\mu_i} = \beta_d,
\]

where
\[
W(\lambda_d) = \frac{\mu_i + \mu_d - \lambda_d}{\mu_i\mu_d - \lambda_d\mu_i - \lambda_d\mu_d}.
\]

Proof of Proposition 5. The proof process is provided in the Appendix A.

In general, Propositions 4 and 5 give the queueing behavior of patients in equilibrium. According to the above results, we observe that all groups of patients decide to join the queue when their patience levels are very high. When the patience level of one group of patients is higher than the other, if the group of patients with low patience levels does not join the queue, i.e., they adopt the “all balk” strategy, then the group of patients with high patience levels might be more likely to decide to join the queue, that is, they might adopt either the “all join” strategy or a “mixed strategy”. However, if a fraction of patients with low patience levels joins the queue, i.e., the group of patients with low patience levels adopt a “mixed strategy”, then the group of patients with high patience levels must adopt the “all join” strategy. These results imply that the queueing behavior of one group of patients hinges not only on their patience level but also on their sojourn time in the system, which in turn is influenced by the joining-or-balking decision of the other group of patients.

Next, we turn to a social planner’s point of view. The problem is the maximization of social welfare, with respect to the strategy space \( (q_d, q_i) \in [0, 1] \times [0, 1] \) that should be imposed on the patients. Define \( SW_d(q_d) \) and \( SW_i(q_i) \) as the net utilities accruing to first-visit and referred patients in unit time, respectively, we then have
Social welfare is the sum of all patients net utilities, that is,

\[ SW(q_d, q_r) = \lambda_d q_d R_d + \lambda_r q_r R_r - (\lambda_d q_d \theta_d + \lambda_r q_r \theta_r) W(\lambda_d, \lambda_r). \]

The social welfare maximization problem for social planners can be established as

\[ \max_{q_d, q_r} SW(q_d, q_r) = \lambda_d q_d R_d + \lambda_r q_r R_r - (\lambda_d q_d \theta_d + \lambda_r q_r \theta_r) W(\lambda_d, \lambda_r). \]

Because of the very involved nature of social welfare, it would be difficult to derive the socially optimal solutions \((q_d^{soc}, q_r^{soc})\) in closed form; thus, the result can only be achieved numerically. For example, the two-variable optimization problem can be solved by using a sequential optimization approach. First, for any given joining probability \(q_d \in [0,1]\), we derive the optimal conditional probability \(q_t^{soc}(q_d)\), then by plugging \(q_t^{soc}(q_d)\) into \(SW(q_d, q_r)\), the two-variable optimization problem becomes a univariate optimization problem, and the optimal \(q_d\) is established on the basis of \(q_t^{soc}(q_d)\). Of course, the two-variable optimization problem can also be solved by other optimization algorithms (e.g., particle swarm optimization and genetic algorithm).

### 5. Numerical Examples

In this section, we will give some numerical examples to verify the accuracy of the theoretical results and provide some management implications.

Define \(q_i^e = \frac{\tilde{\lambda}_i}{\lambda_i}, i = d, r\) as the equilibrium joining probabilities of two groups of patients. Assume that \(\beta_d = 1, \beta_r = 2, \lambda_d = 0.1 (\lambda_r = 1), \mu_d = 4, \mu_r = 4\), i.e., \(\beta_d < \beta_r\). Proposition 4 is pictorially shown in Figure 4. From Figure 4, we find that the equilibrium joining probabilities of both groups of patients decrease as the arrival rates \(\lambda_i, \lambda_d\) increase, and the equilibrium joining probability \(q_i^e\) is no less than \(q_d^e\) as \(\beta_i > \beta_d\), which is consistent with the conclusion of Proposition 4. When the first-visit patients adopt the “all balk” strategy, the referred patients might adopt either the “all join” strategy or a “mix strategy” (choose randomly between balking and joining), when the first-visit patients adopt a “mixed strategy”, the referred patients adopt the “all join” strategy. The reason is that the maximum patience time could affect the equilibrium joining probabilities of patients. As long as the expected sojourn time is less than the maximum patience time, patients are more likely to join the queue. Finally, from Figure 4a,b, as \(\lambda_i\) and \(\lambda_d\) are greater than certain thresholds, a slight increase in \(\lambda_i\) or \(\lambda_d\) may lead to a relatively large change in the equilibrium joining probabilities of first-visit patients, and compared with \(\lambda_i\), the increase in \(\lambda_d\) has less impact on \(q_d^e\).
Figure 4. The equilibrium joining probabilities of patients versus $\lambda_i$ and $\lambda_j$ ($\beta_i > \beta_j$).

Assume that $\beta_d = 2$, $\beta_i = 1.5$ ($\beta_i = 1$), $\lambda_d = 2$ ($\lambda_i = 1$), $\mu_d = 4$, $\mu_i = 8$, i.e., $\beta_i < \beta_d$. Proposition 5 is pictorially shown in Figure 5. From Figure 5, we still find that the equilibrium joining probabilities of both groups of patients decrease as the arrival rates $\lambda_i$ and $\lambda_j$ increase, and the equilibrium joining probability $q_{d\beta}^e$ is no less than $q_{i\beta}^e$ as $\beta_d > \beta_i$, which is also consistent with the conclusion of Proposition 5. When the referred patients adopt the “all balk” strategy, the first-visit patients might adopt either the “all join” strategy or a “mix strategy” (randomize between balking and joining), when the referred patients adopt a “mixed strategy”, the first-visit patients adopt the “all join” strategy. The results are also attributed to the relationship between the maximum patience time and the expected sojourn time. Moreover, from Figure 5a,b, as $\lambda_i$ and $\lambda_j$ are greater than certain thresholds, a slight increase in $\lambda_i$ or $\lambda_j$ may also lead to a relatively large change in the equilibrium joining probabilities of referred patients, and compared with $\lambda_i$, the increase in $\lambda_j$ has a greater impact on $q_{d\beta}^e$, which is contrary to the conclusion of Figure 4.
Figure 5. The equilibrium joining probabilities of patients versus $\lambda_i$ and $\lambda_d$ ($\beta_i < \beta_d$).

From Figures 4 and 5, we conclude that, with the increase in the arrival rates $\lambda_i$ and $\lambda_d$, the group of patients with high patience levels may have a negative impact on the joining decision of patients with low patience levels, i.e., the rational queueing strategies of patients with high patience levels could discourage patients with low patience levels from making the joining decision. In brief, as more patients arrive at the comprehensive hospital, patients with a high patience level may crowd out the other group of patients and push them to balk the queue.

In Figures 6 and 7, we, respectively, assume that $\lambda_d = 0.8$, $\mu_d = 5$, $\mu_i = 3$ and $\lambda_i = 1$, $\beta_i = 0.4$, $\mu_i = 4$, $\mu_i = 4$, $R_i = 1$, $\theta_i = 2$, $\theta_i = 1$, then Figures 6 and 7 depict the effect of $\lambda_i$ and $R_d$ on the individual equilibrium strategies and social optimal strategies. From Figure 6, it can be deduced that under the cases $\beta_i < \beta_d$ and $\beta_i > \beta_d$, the individual equilibrium strategies and social optimal strategies decrease as $\lambda_i$ increases. It makes sense that decreasing the arrival rate could benefit patients by reducing their sojourn time; meanwhile, the new arriving patient who knows a high arrival rate can predict the higher load of the hospital, which increases their waiting cost and makes them reluctant to join the queue. However, as shown in Figure 7, the individual equilibrium strategy and social optimal strategy of first-visit patients increase with $R_d$, whereas the individual equilibrium strategy and social optimal strategy of referred patients decrease with $R_d$. This may be caused by the fact that the patients with a high patience level may crowd out the other group of patients and push them to balk the queue. Finally, from Figures 6 and 7, we note that the inequality $q_{i, soc}^d \leq q_i^d, i = d, t$ holds. The main reason for the results is that multiple self-interested patients in a queueing game ignore the negative externalities of their actions on each other (e.g., the joining decision of new arriving patients would prolong the sojourn time and increase the delay cost of future patients), and one consequence of this is an equilibrium that deviates from the socially optimal solution. Therefore, when maximizing social welfare, we should consider the negative externalities, that directly lead to inequality. The results could provide managerial insights to medical managers, especially the government. Given the service reward, the delay sensitivity and the service time, in order to maximize social welfare, medical managers need
to allocate patients, which will break the individual optimal equilibrium between patients, and take measures to eliminate the deviation between the individual equilibrium strategy and social optimal strategy, i.e., to guide some patients to choose primary hospitals for diagnosis and treatment so as to reduce the number of patients in comprehensive hospitals. Specific measures are as follows: First, the medical service managers could adopt a gatekeeping design to reduce the number of first-visit patients in comprehensive hospitals. For example, in China, through the medical insurance system, the government uses appropriate cost adjustment mechanisms to guide first-visit patients to preferentially choose community or primary hospitals for diagnosis and treatment. In addition, the government has improved the family doctor signing services and adopted the priority mechanism to regulate the flow of first-visit patients. Moreover, the government could also set reasonable prices for first-visit patients who first choose comprehensive hospitals so as to reduce the proportion of first-visit patients to comprehensive hospitals. Second, the medical service managers could adopt measures to reduce the number of referred patients in comprehensive hospitals. For example, the government should allocate medical resources rationally, improve the medical level of community hospitals and the cure rate of patients by the sinking of high-quality resources, and then reduce the number of patients referred to comprehensive hospitals.

Figure 6. The individual equilibrium strategies and social optimal strategies versus $\lambda$.

(a) $\beta_d = 1 < \beta_i = 1.5$

(b) $\beta_i = 1 < \beta_d = 2.5$
6. Research Findings

The theoretical results and numerical results are summarized as follows: (1) Both groups of patients decide to join the queue when their patience levels are very high; (2) when the patience level of one group of patients are higher than the other, if the group of patients with low patience levels adopt the “all balk” strategy, then the group of patients with high patience levels might adopt either the “all join” strategy or a “mixed strategy”, whereas, if the group of patients with low patience levels adopt a “mixed strategy”, then the group of patients with high patience levels must adopt the “all join” strategy; (3) as more patients arrive at the comprehensive hospital, the group of patients with high patience levels may crowd out the other group of patients and push them to balk the queue; (4) the equilibrium behavior deviates from a socially optimal solution; therefore, to reach the maximal social welfare, the social planner should adopt some regulation measures, such as appropriate cost adjustment mechanisms, the family doctor signing services, the priority mechanism and the sinking of high-quality resources, etc., to control the arrival rates of patients in comprehensive hospitals. By doing so, they can mitigate overcrowding and reduce waiting times, ultimately improving the efficiency of the healthcare system.

7. Concluding Remarks

Motivated by the queueing problems arising in comprehensive hospitals, we constructed a game-theoretical queueing model to analyze the equilibrium queueing behavior of patients in comprehensive hospitals. We first derived the stability condition, so that we could obtain the expected sojourn time of an arbitrary patient under the stability condition. Based on the expected sojourn time, we defined the utility of patients who join the queue and derived the strategic queueing behavior of patients in equilibrium under the fully observable and unobservable cases. Finally, we derived the socially optimal solutions numerically and provided managerial insights according to the theoretical results and numerical results.

In this current work, we assume that patients could share the same information level. However, due to the lack of access to information, in today’s service industries, information homogeneity is difficult to determine, and the queueing problem with information heterogeneity is often encountered; in future work, analyzing the strategic behavior of
patients with information heterogeneity would be an interesting direction. Another direction is that strategic patients exhibit bounded rationality instead of full rationality, so studying the effect of bounded rationality on the queueing behavior of patients would also be an interesting and important direction.

Author Contributions: Conceptualization, Y.L. and L.L.; methodology, T.J. and X.C.; software, T.J. and X.C.; validation, Y.L. and L.L.; formal analysis, T.J.; investigation, Y.L.; resources, Y.L.; data curation, Y.L.; writing—original draft preparation, Y.L. and T.J.; writing—review and editing, T.J. and L.L.; visualization, Y.L.; supervision, L.L.; project administration, Y.L. and L.L.; funding acquisition, Y.L., L.L., T.J. and X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 61773014, Grant No. 12001329), National Natural Science Foundation of China (Grant No. 11901307) and Natural Science Foundation of Jiangsu Province, China (Grants BK20180783 and 18KJB110021), Anhui University Natural Science Research Project of China (Grant No. 2022AH052209), and Scientific Research Foundation of Anhui Polytechnic University (Grant No. 2022YQQ026).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Proof of Lemma 1. Based on Theorem 1.7.1 in Neuts (1981), the system is stable if and only if \( xCe < xBe \), where \( e = (1, 1)^T \), \( x = (x_0, x_1) \) is the invariant probability vector of

\[
Y = A + B + C = \begin{pmatrix} -\mu_d & \mu_d \\ \mu_i \gamma & -\mu_i \gamma \end{pmatrix},
\]

which satisfies \( xY = 0 \) and \( xe = 1 \). From the equations, we have \( x_0 = \frac{\mu_i \gamma}{\mu_d + \mu_i \gamma} \),

\[
x_1 = \frac{\mu_d}{\mu_d + \mu_i \gamma},
\]

then, the stability condition \( xCe < xBe \) converts into \( \lambda < \frac{\mu_i \mu_d}{\mu_d + \mu_i \gamma} \),

i.e., \( \hat{\lambda} \mu_d + \lambda_d (\mu_i + \mu_d) < \mu_i \mu_d \).

Proof of Proposition 1. According to the special structure of the state-transition-rate matrix, the balance equation can be obtained as follows:

\[
\lambda \pi_0 = \mu_i \pi_{1,1},
\]

\[
(\lambda + \mu_d) \pi_{1,0} = \lambda \gamma \pi_0 + \mu_i \gamma \pi_{2,1},
\]

\[
(\lambda + \mu_d) \pi_{k,0} = \lambda \pi_{k-1,0} + \mu_i \gamma \pi_{k+1,1}, \quad k \geq 2,
\]

\[
(\lambda + \mu_i) \pi_{1,1} = \lambda \gamma \pi_{0,0} + \mu_d \pi_{1,0} + \mu_i \gamma \pi_{2,1},
\]

\[
(\lambda + \mu_i) \pi_{k,1} = \lambda \pi_{k-1,1} + \mu_d \pi_{k,0} + \mu_i \gamma \pi_{k+1,1}, \quad k \geq 2.
\]

Define \( P_0(z) = \sum_{k=1}^{\infty} \pi_{k,0} z^k \), \( P_1(z) = \sum_{k=1}^{\infty} \pi_{k,1} z^k \), \( (|z| < 1) \) from the above balance equations, we could easily have the following equation set:
\[(\lambda + \mu_d - \lambda z)P_0(z) = \frac{\mu_d}{z} P_1(z) + \pi_0 \lambda \gamma (z-1) \]
\[(\lambda + \mu - \lambda z - \frac{\mu_d}{z})P_1(z) = \mu_d P_0(z) + \pi_0 \lambda \gamma (z-1) \]

Solving the equation set, we can derive the expressions of \(P_0(z), P_1(z)\):

\[ P_0(z) = \frac{\lambda \gamma z^2 (z-1)[(\lambda(1-z) + \mu_d)]}{g_0(z) g_1(z) - \mu_d \mu \gamma z} \]
\[ P_1(z) = \frac{\lambda z^2 (z-1)[(\lambda(1-z) + \mu_d)]}{g_0(z) g_1(z) - \mu_d \mu \gamma z} \]

where \(g_0(z) = \lambda z(1-z) + \mu_d z\), \(g_1(z) = \lambda z(1-z) + \mu(z - \gamma)\).

\[ \pi_0 = 1 - \rho = 1 - \lambda \left( \frac{1}{\mu_d} + \frac{\gamma}{\mu} \right) \]

According to the results, we could easily derive the expected queue length (including the patient being served):

\[ L = \frac{dP_0(z)}{dz} \bigg|_{z=1} + \frac{dP_1(z)}{dz} \bigg|_{z=1} = \lambda \left( \frac{\gamma}{\mu_d} + \frac{1}{\mu} \right) \left[ 1 + \frac{\lambda(\mu_d + \mu - \lambda)}{\mu_d \mu - \lambda \gamma \mu - \lambda \mu_d} \right] - \frac{\lambda^2}{\mu_d \mu} \]

By using Little’s law, we could obtain the expected sojourn time of an arbitrary patient:

\[ W(\lambda, \gamma) = \left( \frac{\gamma}{\mu_d} + \frac{1}{\mu} \right) \left[ 1 + \frac{\lambda(\mu_d + \mu - \lambda)}{\mu_d \mu - \lambda \gamma \mu - \lambda \mu_d} \right] - \frac{\lambda}{\mu_d \mu} \]

Substituting \(\lambda = \lambda_d + \lambda_i, \gamma = \frac{\lambda_d}{\lambda_d + \lambda_i}\) into \(W(\lambda, \gamma)\), we have the final expression

\[ W(\lambda_d, \lambda_i) = \left( \frac{\lambda_d}{\mu_d (\lambda_d + \lambda_i)} + \frac{1}{\mu} \right) \left[ 1 + \frac{(\lambda_d + \lambda_i)[\mu_d + \mu - (\lambda_d + \lambda_i)]}{\mu_d \mu - \lambda_d \mu - (\lambda_d + \lambda_i) \mu_d} \right] - \frac{\lambda_d + \lambda_i}{\mu_d \mu} \]

**Proof of Proposition 2.** Based on the expression of \(W(\lambda_d, \lambda_i)\), we find that it is complicated to determine the relationship between \(W(\lambda_d, \lambda_i)\) and \(\lambda_d, \lambda_i\) directly. In the following content, we will use the chain rule of binary functions to determine the monotonicity of \(W(\lambda_d, \lambda_i)\) with respect to \(\lambda_d, \lambda_i\). First, according to \(W(\lambda, \gamma)\), we have

\[ \frac{\partial W}{\partial \lambda} = \frac{\gamma \mu_i^2 + \mu_d^2 + \gamma \mu \mu_d}{(\mu, \mu - \lambda \gamma \mu - \lambda \mu_d)^2} > 0, \quad \frac{\partial W}{\partial \gamma} = \frac{1}{\mu_d} \left( \frac{\lambda \mu_d \mu_d (\mu_d + \mu - \lambda)}{(\mu, \mu - \lambda \gamma \mu - \lambda \mu_d)^2} \right) > 0. \]

From \(\frac{\partial \lambda}{\partial \lambda_d} = 1, \frac{\partial \gamma}{\partial \lambda_d} = \frac{\lambda_d}{(\lambda_d + \lambda_i)^2}\), we have

\[ \frac{\partial \gamma}{\partial \lambda_i} = -\frac{\lambda_d}{(\lambda_d + \lambda_i)^2}, \quad \frac{\partial \gamma}{\partial \lambda_i} = -\frac{\lambda_i}{(\lambda_d + \lambda_i)^2}, \quad \frac{\partial \gamma}{\partial \lambda_i} = -\frac{\lambda_i}{(\lambda_d + \lambda_i)^2}. \]
According to the chain rule, we could obtain the partial derivative of \(W(\lambda_d, \lambda_i)\) with respect to \(\lambda_d\) and \(\lambda_i\):

\[
\frac{\partial W(\lambda_d, \lambda_i)}{\partial \lambda_d} = \frac{\partial W}{\partial \lambda} \frac{\partial \lambda}{\partial \lambda_d} + \frac{\partial W}{\partial \gamma} \frac{\partial \gamma}{\partial \lambda_d}
\]

\[
= \frac{\gamma^2 + \mu^2 + \gamma \mu_d}{(\mu_d - \lambda \mu_i - \lambda \mu_d)^2} + \frac{1}{\mu_d} \left( 1 + \frac{\lambda \mu_d (\mu_i + \mu_d - \lambda)}{(\mu_d - \lambda \mu_i - \lambda \mu_d)^2} \right) \frac{\lambda_i}{(\lambda_d + \lambda_i)^2} > 0,
\]

\[
\frac{\partial W(\lambda_d, \lambda_i)}{\partial \lambda_i} = \frac{\partial W}{\partial \lambda} \frac{\partial \lambda}{\partial \lambda_i} + \frac{\partial W}{\partial \gamma} \frac{\partial \gamma}{\partial \lambda_i}
\]

\[
= \frac{\gamma^2 + \mu^2 + \gamma \mu_d}{(\mu_d - \lambda \mu_i - \lambda \mu_d)^2} - \frac{1}{\mu_d} \left( 1 + \frac{\lambda \mu_d (\mu_i + \mu_d - \lambda)}{(\mu_d - \lambda \mu_i - \lambda \mu_d)^2} \right) \frac{\lambda_d}{(\lambda_d + \lambda_i)^2} \mu_d.
\]

To determine the value of \(\frac{\partial W(\lambda_d, \lambda_i)}{\partial \lambda_i}\), define

\[
M(\lambda_d, \lambda_i) = \frac{\mu^2 + \lambda \mu_d}{(\mu_d - \lambda \mu_i - \lambda \mu_d)^2} - \frac{\lambda_d}{(\lambda_d + \lambda_i)^2} \mu_d
\]

which is not necessarily continuous at the point \((0,0)\). Since \(\frac{\partial M(\lambda_d, \lambda_i)}{\partial \lambda_i} > 0\), then,

\(M(\lambda_d, \lambda_i)\) is monotonically increasing with respect to \(\lambda_i\), i.e., \(M(\lambda_d, \lambda_i) > M(\lambda_d, 0)\).

From \(M(\lambda_d, 0) = \frac{\mu^2 + \lambda \mu_d}{(\mu_d - \lambda \mu_i - \lambda \mu_d)^2} - \frac{1}{\lambda_d} \mu_d\) and \(\mu_d - \lambda \mu_i - \lambda \mu_d > 0\), we could easily have

\[
\frac{dM(\lambda_d, 0)}{d \lambda_d} = \frac{1}{\lambda_d} (\mu_d - \lambda \mu_i - \lambda \mu_d)^2 + 2(\mu_d - \lambda \mu_i - \lambda \mu_d)(\mu^2 + \lambda \mu_d) + \frac{1}{\lambda_d^2} \mu_d > 0
\]

which means \(M(\lambda_d, 0)\) is monotonically increasing with respect to \(\lambda_d\).

Due to \(\lim_{\lambda_d \to 0} M(\lambda_d, 0) < 0\) and \(\lim_{\lambda_d \to \lambda_d, \lambda_i} M(\lambda_d, 0) > 0\), we find that there exists a unique \(\lambda_d\) that makes \(M(\lambda_d, 0) = 0\), i.e., if \(\lambda_d \in [\bar{\lambda}_d, \frac{\mu_d}{\mu_i + \mu_d}]\), \(M(\lambda_d, 0) \geq 0\).

According to the above analysis, we could obtain the results.

**Proof of Proposition 3.** If an arriving first-visit or referred patient finds the system at a state \((n, i), i = 0, 1\), then, the utilities would be \(U_{obi}(n, i) = R_d - \theta_d W^d(n, i)\) or
$U'_{\text{obs}}(n,i) = R_i - \theta W'_{(n,i)}$. Since $U^d_{\text{obs}}(n,i)$ and $U'_{\text{obs}}(n,i)$ are monotonically decreasing with respect to $n$, we only need to find the unique root that satisfies $U^d_{\text{obs}}(n,i) = 0$ or $U'_{\text{obs}}(n,i) = 0$. The results can be easily obtained.

**Proof of Lemma 2.** If $\lambda_d = 0$, the queueing system degrades into a classical M/M/1 queue, the expected sojourn time of an arbitrary patient is $W(\lambda) = W(\lambda_i,0) = \frac{1}{\mu_i - \lambda_i}$, the utility of a patient who chooses to join the queue is $U_i = R_i - \theta W(\lambda_i)$, then

1. If $U_i = R_i - \theta W(\lambda_i) \geq 0$, i.e., $\beta_i \geq \frac{1}{\mu_i - \lambda_i}$, all the referred patients join the queue, and the corresponding equilibrium arrival rate $\lambda^*_i = \lambda_i$;

2. If $U_i = R_i - \theta W(\lambda_i) < 0$ and $U_i = R_i - \theta \lim_{\lambda \to 0} W(\lambda_i) \geq 0$, i.e., $\frac{1}{\mu_i} \leq \beta_i < \frac{1}{\mu_i - \lambda_i}$, a fraction of the referred patients joins the queue, the corresponding equilibrium arrival rate $\lambda^*_i = \mu_i - \frac{1}{\beta_i}$;

3. If $U_i = R_i - \theta \lim_{\lambda \to 0} W(\lambda_i) < 0$, i.e., $\beta_i < \frac{1}{\mu_i}$, all the referred patients balk the queue, and the corresponding equilibrium arrival rate $\lambda^*_i = 0$.

**Proof of Lemma 3.** If $\lambda = 0$, the queueing system degrades into an M/M/1 queue with two service phases, then the expected sojourn time of an arbitrary patient is $W(\lambda_d) = \frac{1}{\mu, \mu_d - \lambda_d}$, the first derivative

$$
\frac{dW(\lambda_d)}{d\lambda_d} = \frac{\mu^2 + \mu_d^2 + \mu \mu_d}{\mu, \mu_d - \lambda_d \mu_d} > 0,
$$

i.e., $W(\lambda_d)$ is an increasing function with respect to $\lambda_d$, and the utility of a patient who chooses to join the queue is $U_d = R_d - \theta_d W(\lambda_d)$, then

1. If $U_d = R_d - \theta_d W(\lambda_d) \geq 0$, i.e., $\beta_d \geq W(\lambda_d)$, all the first-visit patients join the queue, and the corresponding equilibrium arrival rate $\lambda^*_d = \lambda_d$;

2. If $U_d = R_d - \theta_d W(\lambda_d) < 0$ and $U_d = R_d - \theta_d \lim_{\lambda \to 0} W(\lambda_d) \geq 0$, i.e., $\frac{1}{\mu_i} + \frac{1}{\mu_d} \leq \beta_d < W(\lambda_d)$, a fraction of the first-visit patients joins the queue, and the corresponding equilibrium arrival rate $\lambda^*_d = \frac{\beta_d \mu_d \mu_d - (\mu_i + \mu_d)}{\beta_d (\mu_i + \mu_d) - 1}$;

3. If $U_d = R_d - \theta_d \lim_{\lambda \to 0} W(\lambda_d) < 0$, i.e., $\beta_d < \frac{1}{\mu_i} + \frac{1}{\mu_d}$, all the first-visit patients balk the queue, and the corresponding equilibrium arrival rate $\lambda^*_d = 0$. 
Proof of Proposition 4. If $\lambda > 0, \lambda > 0$, both groups of patients arrive at the comprehensive hospital. Since $\beta < \beta$, as long as the first-visit patients decide to join the queue, then all the referred patients also join the queue. Next, we consider the following two scenarios.

**Scenario 1:** All the first-visit patients balk the queue, i.e., the corresponding equilibrium arrival rate $\lambda^e = 0$. In this scenario, $W(0, \lambda^e) = W(\lambda^e) = \frac{1}{\mu - \lambda^e}$, which is consistent with the expected sojourn time of the classical M/M/1 queue. Next, we consider two subcases.

**Case I:** If $U = R - W(\lambda^e) \geq 0$, i.e., $\beta \geq \frac{1}{\mu - \lambda^e}$, then all the referred patients join the queue. Since all the first-visit patients balk the queue, then, $U = R - W(\lambda^e) < 0$, that is, $\beta < \frac{1}{\mu - \lambda^e}$. Therefore, if $\beta < \frac{1}{\mu - \lambda^e} \leq \beta$, all the first-visit patients balk the queue, and all the referred patients join the queue, i.e., the equilibrium arrival rates $\lambda^e = 0$, $\lambda^e = \lambda^e$.

**Case II:** If $U = R - W(\lambda^e) < 0$, then only a fraction of the referred patients chooses to join the queue while others balk. In this subcase, there exists a unique equilibrium arrival rate of the referred patients makes that $R - W(\lambda^e) = 0$, i.e., $\lambda^e = \mu - \frac{1}{\beta}$. Therefore, if $\beta < \beta < \frac{1}{\mu - \lambda^e}$, all the first-visit patients balk the queue, and a fraction of the referred patients joins the queue, i.e., the corresponding equilibrium arrival rates $\lambda^e = 0, \lambda^e = \mu - \frac{1}{\beta}$.

**Scenario 2.** A fraction of the first-visit patients join the queue, i.e., the equilibrium arrival rate of the first-visit patients $\lambda^e > 0$. In this scenario, all the referred patients join the queue, that is, $\lambda^e = \lambda^e$. Moreover, since there are first-visit patients joining the queue, then $U = R - W(\lambda^e) \geq 0$, i.e., $\beta \geq W(\lambda^e) = \frac{1}{\mu - \lambda^e}$. Next, we also consider two subcases.

**Case I:** If $U = R - W(\lambda^e) \geq 0$, i.e., $W(\lambda^e, \lambda^e) \leq \beta$, since $W(\lambda^e, \lambda^e)$ is monotonically increasing with respect to $\lambda^e$, we then have $W(0, \lambda^e) < W(\lambda^e, \lambda^e)$. Therefore, if $\beta \geq W(\lambda^e, \lambda^e)$, all the first-visit patients join the queue, and all the referred patients join the queue, i.e., the equilibrium arrival rates $\lambda^e = \lambda^e, \lambda^e = \lambda^e$.

**Case II:** If $U = R - W(\lambda^e, \lambda^e) < 0$, i.e., $W(\lambda^e, \lambda^e) > \beta$, since $W(\lambda^e, \lambda^e)$ is monotonically increasing with respect to $\lambda^e$, then, there exists a unique $\lambda^e$ makes that $U = R - W(\lambda^e, \lambda^e) = 0$, i.e., $\lambda^e$ satisfies the equation $W(\lambda^e, \lambda^e) = \beta$. In this subcase, if $\frac{1}{\mu - \lambda^e} \leq \beta < W(\lambda^e, \lambda^e)$, the equilibrium arrival rates are $\lambda^e = \lambda^e, \lambda^e = \lambda^e$, where $\lambda^e$ satisfies the following equation:
By summarizing the above analysis, we could obtain the results in Proposition 4.

**Proof of Proposition 5.** If \( \lambda_d > 0, \lambda_d > 0 \), both two groups of patients arrive at the comprehensive hospital. Since \( \beta_d > \beta_d \), as long as the referred patients decide to join the queue, then all the first-visit patients also join the queue. Similar to the proof of Proposition 4, we consider two scenarios to obtain the results of Proposition 5 next.

**Scenario 1.** All the referred patients balk the queue, i.e., the equilibrium arrival rate \( \lambda_d^e = 0 \). In this scenario, from the stability condition and the expression of the expected sojourn time, we have \( W(\lambda_d, 0) = W(\lambda_d) = \frac{\mu_i + \mu_d - \lambda_d}{\mu_i \mu_d - \lambda_d \mu_i - \lambda_d \mu_d} > 0 \). Next, we consider two subcases as follows:

- **Case I:** If \( U_d = \beta_d W(\lambda_d) \geq 0 \), i.e., \( \beta_d \geq W(\lambda_d) \), all the first-visit patients join the queue. Since all the referred patients balk the queue, then \( U_d = R_d - \theta_d W(\lambda_d) > 0 \), i.e., \( \beta_d < W(\lambda_d) \). Therefore, when \( \beta_d < W(\lambda_d) \leq \beta_d \), in equilibrium, all the referred patients join the queue, and all the first-visit patients balk the queue, the equilibrium arrival rates \( \lambda_d^e = \lambda_d \), \( \lambda_d^e = 0 \).

- **Case II:** If \( U_d = \beta_d W(\lambda_d) < 0 \), a fraction of the first-visit patients decides to join the queue, and others balk. Since \( W(\lambda_d) = \frac{\mu_i + \mu_d - \lambda_d}{\mu_i \mu_d - \lambda_d \mu_i - \lambda_d \mu_d} \) is monotonically increasing with respect to \( \lambda_d \), then, there exists a unique equilibrium arrival rate \( \lambda_d^e = \frac{\beta_d \mu_i \mu_d - (\mu_i + \mu_d)}{\beta_d (\mu_i + \mu_d) - 1} \) makes that \( R_d - \theta_d W(\lambda_d^e) = 0 \). Therefore, when \( \beta_d < \beta_d < W(\lambda_d) \), the equilibrium arrival rates of the two groups of patients \( \lambda_d^e = \lambda_d^e = 0 \).

**Scenario 2.** A fraction of the referred patients joins the queue, i.e., the equilibrium arrival rate \( \lambda_d^e > 0 \), in this scenario, all the first-visit patients choose to join the queue, i.e., \( \lambda_d^e = \lambda_d \). Moreover, since a fraction of the referred patients joins the queue, we have \( U_d = R_d - \theta_d W(\lambda_d) \geq 0 \), that is, \( \beta_d \geq W(\lambda_d) \). Next, we continue to consider two subcases.

- **Case I:** If \( U_d = R_d - \theta_d W(\lambda_d, \lambda_d) \geq 0 \), i.e., \( W(\lambda_d, \lambda_d) \leq \beta_d \), since \( W(\lambda_d, \lambda_d) \) is a monotonically increasing function about \( \lambda_d \) as \( \lambda_d \in [\lambda_d, \frac{\mu_i \mu_d}{\mu_i + \mu_d}] \); therefore, \( W(\lambda_d, 0) < W(\lambda_d, \lambda_d) \). When \( \beta_d > \beta_d \geq W(\lambda_d, \lambda_d) \), then all the first-visit patients join the queue, and all the referred patients join the queue, the equilibrium arrival rates \( \lambda_d^e = \lambda_d \), \( \lambda_d^e = \lambda_d \).

- **Case II:** If \( U_d = R_d - \theta_d W(\lambda_d, \lambda_d) < 0 \), i.e., \( W(\lambda_d, \lambda_d) > \beta_d \). Since \( W(\lambda_d, \lambda_d) \) is a monotonically increasing function about \( \lambda_d \) as \( \lambda_d \in [\lambda_d, \frac{\mu_i \mu_d}{\mu_i + \mu_d}] \), then, there exists a
unique equilibrium arrival rate \( \lambda_i^* \) makes that \( U_i = R_i - \theta W(\lambda_j, \lambda_i^*) = 0 \). In this sub-case, when \( W(\lambda_j) \leq \beta_i < W(\lambda_j, \lambda_i^*) \), the equilibrium arrival rates \( \lambda_j^* = \lambda_j \) and \( \lambda_i^* \) satisfies:

\[
\frac{\lambda_j}{\mu_j(\lambda_j + \lambda_i^*)} + \frac{1}{\mu_i} \left[ 1 + \frac{(\lambda_j + \lambda_i^*)(\mu_j + \mu_i - (\lambda_j + \lambda_i^*)\mu_j)}{\mu_j(\mu_i - \lambda_j)\mu_i - (\lambda_j + \lambda_i^*)\mu_j} \right] \frac{\lambda_j + \lambda_i^*}{\mu_i} = \beta_i.
\]

By summarizing the above analysis, we could obtain the results in Proposition 5.

References

15. Wang, J.; Sun, K. Optimal pricing and capacity sizing for online service systems with free trials. OR Spectr. 2022, 44, 57–86.


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.