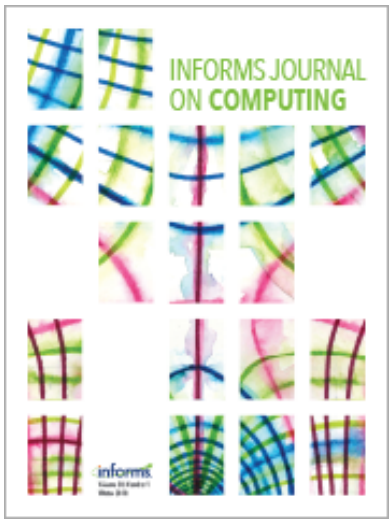


INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Physician Staffing for Emergency Departments with Time-Varying Demand

Ran Liu, Xiaolan Xie

To cite this article:
Ran Liu, Xiaolan Xie (2018) Physician Staffing for Emergency Departments with Time-Varying Demand. INFORMS Journal on Computing 30(3):588-607. https://doi.org/10.1287/ijoc.2017.0799

http://pubsonline.informs.org/journal/ijoc/

INFORMS JOURNAL ON COMPUTING
Vol. 30, No. 3, Summer 2018, pp. 588–607
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

Physician Staffing for Emergency Departments with Time-Varying Demand

Ran Liu,[a] **Xiaolan Xie**[b, c]

[a] Department of Industrial Engineering and Management, Shanghai Jiao Tong University, 200240 Shanghai, China; [b] Antai College of Economics and Management, Shanghai Jiao Tong University, 200052 Shanghai, China; [c] Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Centre CIS, F-42023 Saint-Etienne France

Contact: liuran2009@sjtu.edu.cn, http://orcid.org/0000-0002-7922-8969 (RL); xie@emse.fr, http://orcid.org/0000-0002-6579-1523 (XX)

Received: March 8, 2016
Revised: December 1, 2016; May 31, 2017
Accepted: October 30, 2017
Published Online: October 15, 2018

https://doi.org/10.1287/ijoc.2017.0799

Abstract. Fluctuations in emergency department (ED) patient arrivals during the day are one of the main causes of the long waiting times that are frequently encountered, and ED staffing is one of the key drivers of ED service quality improvement. This paper first proposes discrete-time models for approximating the patient waiting times for any given ED staffing. The waiting time approximation is based on three simple ideas: the separation of patients served in a period and patients overflowed, the combination of $M/M/c$ approximation for patients served and waiting time analysis of overflow patients, and the transformation of the performance evaluation into an optimization problem with the number of overflow patients as decision variables. The resulting waiting time approximations are then integrated into ED staffing optimization models, and variable neighborhood search algorithms are developed to solve the ED staffing models. Numerical experiments with real-life data from Chinese hospitals are performed to validate the proposed models and algorithms. The results show that the proposed methodology is able to significantly reduce the total waiting time of patients without increasing staff capacity.

History: Accepted by Allen Holder, Area Editor for Applications in Biology, Medicine, and Health Care.
Funding: The work was supported by National Natural Science Foundation of China [Grants 71672112, 71671111, and 71432006].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/ijoc.2017.0799.

Keywords: emergency department • time-varying demand • staffing • queueing theory

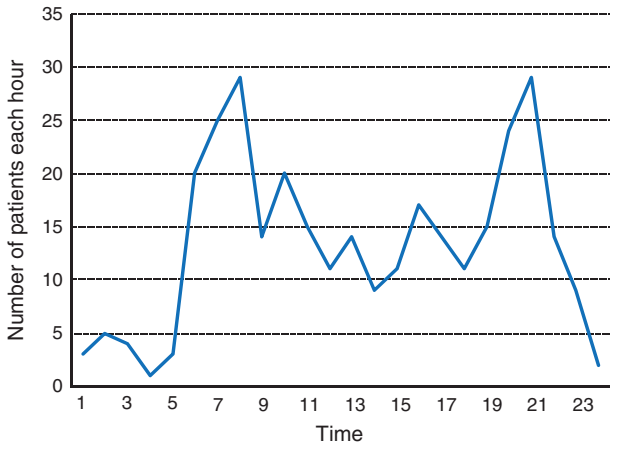
1. Introduction

With the largest population in the world, China has a physician density that is far below that of the leading world economies, such as the United States and France. Chinese hospitals, especially good hospitals in large cities such as Shanghai and Beijing, are always overcrowded and suffer from physician shortage. The emergency department (ED) is the frontline of any hospital, serving as a center of medical treatment for acutely ill and injured patients and providing 24-hour continuous medical care. With the growing demand for ED care, increasing overcrowding, and prolonged waiting times, Chinese hospitals are severely criticized for providing a low quality of service. Such scenarios are not only found in China. For instance, approximately 7.7% of the 36.6 million adults in the United States who have sought care in a hospital ED have reported difficulties in receiving emergency care, and more than half of these adults have cited long waiting times as a cause of the problem (Kennedy et al. 2004).

There are many reasons behind the increasing overcrowding and long waiting times that ED patients experience. One major cause is unpredictable demands for care that vary substantially over the day. ED struggles to provide adequate staff to address time-varying demands. We use Ruijin Hospital, located in Shanghai, China, as an example. Ruijin is one of the best hospitals in China, and its ED consists of 11 departments, such as the emergency medical department (EMD) and resuscitation room. Among these departments, the EMD is the largest (with 10 physicians) and receives the most visits to the Ruijin ED. Compared with other departments in the ED, the EMD has the longest patient waiting time.

Figure 1 shows the typical daily patient arrival pattern of the EMD of Ruijin Hospital. The patient arrival rate is low at night, followed by a sudden increase beginning at approximately 06:00, remaining high during the day and displaying another peak at approximately 21:00. Although the hospital is aware of the variability in patient arrival rate over the day, EMD physician staffing is performed manually and is quite simple: four physicians in the morning, four in the afternoon, and two in the evening. Such fixed physician staffing ignores fluctuations in demand over the day and is an important reason why the EMD is congested. Simple fixed staffing is common in a large number of

Figure 1. (Color online) Hourly ED Patient Arrival over the Course of a Day



healthcare organizations. This is mainly because physician staffing/scheduling has always been a difficult and time-consuming task for the hospital and ED managers (Cheang et al. 2003, De Causmaecker and Vanden Berghe 2011, Green et al. 2007), and the impact of time-varying demand on physician staffing is not well understood.

In this paper, motivated by our collaboration with two hospitals in Shanghai, China, we address the physician staffing problem of the ED to tackle its significantly time-varying and stochastic demands. Because timely access to an emergency provider is a critical dimension of an ED's quality (Green et al. 2007), we use patient waiting time as the efficiency criterion of the ED staffing settings.

Numerous models and algorithms have been developed to investigate various aspects of physician/nurse staffing and scheduling. Mathematical programming is a traditional technique used for physician and nurse staffing and scheduling that aims to identify optimal solutions, and it has been applied to solve relatively simple problems with few constraints (Bard and Purnomo 2005, Beliën and Demeulemeester 2008, Burke and Curtois 2014, Glass and Knight 2010, Jaumard et al. 1998). To solve complicated real-world problems with increasing hard and soft constraints, heuristics such as tabu search (Carter and Lapierre 2001, Ikegami and Niwa 2003), ant colony optimization (Gutjahr and Rauner 2007), genetic algorithm (Aickelin and Dowsland 2004, Moz and Pato 2007, Yeh and Lin 2007), and variable neighborhood search (VNS) (Burke et al. 2008, 2010; Puente et al. 2009) have been proposed. Since the modern hospital and its ED have become an increasingly complex system, it is difficult for analytical models to describe the hospital dynamics and ED reality accurately; simulation offers a framework to address this issue (Jacobson et al. 2006, Jun et al. 1999).

Among the rich body of approaches to physician and nurse staffing and scheduling in the existing literature, few approaches deeply investigate the impact of time-varying demands. Since the time-varying arrival rate is an important characteristic of the demand of an ED, it is inappropriate to staff an ED according to its daily average arrival rate (Green et al. 1991). Green et al. (2006) use a stationary independent period-by-period approach (SIPP) to determine how to vary staffing levels to meet changing demands in an urban hospital ED in the United States. In the SIPP, a separate stationary model is applied to each discrete time interval, with the average arrival rate as the input parameter. Each stationary model is independently solved for the minimum number of servers (physicians) needed to meet the service target in that period. The staffing recommendations of Green et al. (2006) have been implemented in that ED, and the proportion of patients who left without being seen (LWBS) has decreased. Ahmed and Alkhamis (2009) design a decision support system for the operation of an ED unit at a governmental hospital in Kuwait. In their study, the time-varying, nonhomogeneous patient arrival process is considered, and simulation-based heuristics are used to determine the optimal number of staff members required to maximize patient throughput and to reduce patient waiting time. Defraeye and Van Nieuwenhuyse (2013) consider the question of how staffing decisions should be adapted in an ED with time-varying arrivals and customer impatience to control customer waiting time. An extension of the simulation-based algorithm proposed by Feldman et al. (2008) is designed in their work. Zeltyn et al. (2011) apply a simulation model of an ED in a time-varying environment to staff scheduling problems on several different time horizons. Sinreich and Jabali (2007) propose a generic simulation model coupled with a linear programming model for ED staffing and scheduling. Izady and Worthington (2012) use queuing theory and simulation in an iterative staffing scheme to determine schedules for an ED in the United Kingdom.

In addition to hospital EDs, many application areas, such as call centers and teller systems in banks, also face highly time-varying demands (arrival calls, customers). Therefore, many studies in such fields also consider staffing and scheduling problems under time-varying demands. One natural and commonly adopted approach is the use of stationary approximations for performance evaluations in nonstationary systems. For example, given the lag between peak arrival and peak congestion, an effective modification of the above SIPP approach, called lag SIPP (Green et al. 2001), has been proposed to identify the call center staffing levels needed to achieve a specified level of service. The pointwise stationary approximation (PSA) approach uses the instantaneous arrival rate at each time in a separate stationary model (Green and Kolesar 1991). The underlying assumption of PSA is that the steady state

is realized almost immediately, which is the case only if the number of arrivals and service completions is sufficiently high relative to the frequency and magnitude of the arrival rate fluctuations (Whitt 1991). Infinite server (IS) approximations account for congestion lag in a different way by relying on analytically tractable results for infinite server queues (Eick et al. 1993). The time-varying number of customers in an IS approximations system is approximated by its infinite server counterpart; the delay probability is then approximated (Jennings et al. 1996). This approximation is further analyzed by Massey and Whitt (1997), which develops an asymptotic characterization of the time lag between the maximum arrival rate and the maximum number of busy servers, as the arrival changes more slowly. Koole and van der Sluis (2003) propose an approach for joint staffing and shift scheduling for call centers with an overall service-level objective, in which the performance of a shift schedule is evaluated by stationary approximation. Liu and Whitt (2012) derive a delayed-infinite-server approximation model that is applied to a staffing algorithm for a system with a time-varying arrival rate by decomposing the original system into two infinite-server systems, with one representing waiting jobs including abandonment and the other representing the jobs in service.

Another means of accommodating changes in arrival rate is numerical methods. For example, Ingolfsson et al. (2002) investigate exact methods, numerically solving the Chapman–Kolmogorov forward equations (Kleinrock 1974) for time-varying systems to calculate the associated transient system behavior. Ingolfsson et al. (2010) approximate continuously varying parameters with small, discrete intervals and use the randomization method (Grassmann 1977) to explicitly calculate the change in system occupancy from one small interval to the next. Ingolfsson et al. (2007) compare several numerical performance evaluation methods in terms of their accuracy and speed for the time-varying queue system, showing that the randomization method is almost as accurate as the exact method of the Chapman–Kolmogorov forward equations and uses approximately half the computational time. The third method is based on the *fluid model*. Fluid models regard arrival and departure processes as continuous flows rather than discrete processes, and they tend to become more accurate as the number of servers grows large. Whitt (2006) notes that fluid approximations are particularly useful to assess performance in systems that are temporarily overloaded, in which many traditional methods fail. Yom-Tov and Mandelbaum (2014) use the time-dependent number of busy servers in an infinite-server system as the input in their staffing algorithm for a time-varying queue system, in which customers who return to service several times during their sojourn within the system are considered. Meanwhile, simulation or empirical methods are also used in the staffing and shift scheduling literature. For example, Atlason et al. (2004, 2008) propose a cutting plane method to solve the call center staffing and scheduling problems, in which a simulation is used to evaluate the service level in multiple time periods. Castillo et al. (2009) propose a shift scheduling problem with multiple decision criteria. A heuristic is used to generate schedules that are evaluated by a discrete-event simulation. Avramidis et al. (2010) propose an approach that combines simulation with cutting plane methods for solving the shift scheduling problem in a multiskill call center. Nah and Kim (2013) build a mathematical program to obtain a minimum cost shift schedule for a hospital reservation call center, in which empirical data are used to estimate the system performance (e.g., expected waiting time and expected abandonment rate). Defraeye and Van Nieuwenhuyse (2016a) present a simulation-based branch-and-bound algorithm to estimate optimal shift schedules in systems with nonstationary demand and service-level constraints. Given the performance evaluation by simulation, the algorithm can easily be extended to handle a variety of assumptions, such as general service and abandonment processes. For more references on staffing and scheduling problems under time-varying demand for service, we refer the reader to Aksin et al. (2007), Defraeye and Van Nieuwenhuyse (2011, 2016b), Gans et al. (2003), and Green et al. (2007).

In sum, the key advantage of stationary approximation is its simplicity. It can be applied to any system as long as the stationary counterpart is available. However, stationary approximations cannot be applied to overloaded service systems without abandonment because the system is unstable in such cases (Defraeye and Van Nieuwenhuyse 2016b). Numerical methods, such as the randomization method, rely heavily on Markovian assumptions. In this paper, we focus on the ED of a hospital in which the arrival rate of patients varies within and between days. Compared with the extant literature on ED physician staffing, in our problem, during some periods (intervals) of a day, the ED is heavily overloaded and many patients have to wait a long time (i.e., they are overflowed from one period to the next). Rather than using a simulation-based approach, we attempt to analytically describe the patients who are served or overflowed in each period. The ED systems considered in this paper also significantly differ from other service systems, such as call centers. In our ED system, since the number of physicians is limited, we must account for the overload intervals and consider the patients who are overflowed from one period to another (even overflowed two periods). In a large call center system, there may be hundreds of agents and acceptable waits are on the order of minutes. Furthermore, in call centers, the arrival pattern of calls is known to have a similar shape during the

day; however, in our problem, the arrival pattern varies from one day to another, especially between seasons. This makes the problem more difficult to solve. Additionally, an important performance measure of call centers is *telephone service factor*, given by $\Pr\{\text{Wait} < T\}$, for some $T \geq 0$ (Robbins and Harrison 2010). By contrast, in our problem, the objective is to minimize the total waiting time of patients.

In this paper, rather than employing sophisticated nonstationary queuing analysis (Green and Soares 2007; Whitt 2004a, b, 2007), we propose a simple yet efficient approach to evaluate waiting time in nonstationary ED systems with any given staffing. We use a discrete-time model that relies on three simple ideas: (i) divide the patients of each period into patients served and overflow patients, (ii) use simple queuing results to evaluate waiting times, and (iii) transform the performance evaluation into an optimization problem with the numbers of overflow patients as decision variables. A combination of the steady-state waiting time of $M/M/c$ queues and the waiting time analysis of overflow patients leads to two efficient waiting time approximations. These waiting time approximations are then integrated into ED physician staffing models, and an efficient and effective VNS algorithm is designed to solve the problem. Numerical experiments with real-life data collected from Chinese hospitals show that the proposed approximation models can accurately evaluate the performance of ED services, the VNS algorithm is competitive and successful in quick identification of high-quality solutions, and optimized ED physician staffing can effectively reduce patient waiting time without increasing the number of staff or working hours.

The remainder of this paper is organized as follows. An initial ED patient waiting time approximation is addressed in Section 2. Section 3 proposes an improved waiting time approximation that evaluates the exact waiting time of overflow patients. Sections 4 and 5 integrate the two approximations into the ED staffing model, respectively. Section 6 proposes a VNS algorithm for ED staffing optimization, and Section 7 presents the numerical results. Section 8 summarizes our results and discusses future research directions.

2. Waiting Time Approximation

One important feature of an ED such as the ED of Ruijin Hospital is that the arrival rate varies greatly over the day. Such a time-varying arrival rate leads to a nonstationary service system. Meanwhile, the hospital's service capacity is also nonstationary and depends on ED personnel scheduling. In this paper, we focus on ED physicians, the bottleneck of most EDs, and address the daily ED physician staffing problem.

This paper proposes a discrete-time model to capture the nonstationary ED behaviors. We divide the staffing horizon into T periods, each with an equal period length Δ. The time periods are indexed from 1 to T. In each period $t \in T$, we divide the patients into two types: patients served in period t and patients delayed to period $t+1$ or later. Patients who are delayed are referred to as overflow patients. Let u_t be the number of patients served in t, and let q_t be the number of overflow patients.

The ED has a given number of N physicians, and the staffing decision is represented by p_t, denoting the number of physicians available in period t. The goal of our physician staffing problem is to minimize the total patient waiting time.

This section evaluates the waiting time for a given physician staffing decision. The following assumptions made for tractability of the ED model have been validated by practical investigation and data from Ruijin Hospital:

A1. We simplify the ED service process as a single-stage multiserver queuing system: patients arrive, wait in a common queue, consult with a physician, and then leave the ED.

A2. Patients arrive at the ED according to a Poisson process at rate λ_t and are served on a first-come-first-serve (FCFS) basis. As a result, the mean number of arrivals is $\lambda_t\Delta$.

A3. The consultation times are exponentially distributed with service rate μ. As a result, the maximum service capacity of period t is $p_t\mu\Delta$.

A4. The consultation time is short with respect to the period length. Furthermore, if the ED is not overloaded ($\lambda_t < p_t\mu$), it reaches its steady state within the period, and each patient consultation begins and is completed in t.

A5. No patient leaves without being served (LWBS).

A6. The ED has sufficient capacity to clear the workload at the end of the day.

Although the proportion of LWBS patients is an important measure of ED performance and quality of care, few LWBS patients are observed in our field study at Ruijin Hospital. Thus, we ignore LWBS patients. Assumption A6 is often observed because the ED typically has a very low arrival rate at night.

Given the above assumptions, the ED can be considered a standard $M/M/s$ queuing system in each period. The following steady-state waiting time of a stable $M/M/s$ queue with arrival rate λ and c servers is used:

$$W^{M/M/s}(\lambda, c) = \frac{\rho}{\lambda(1-\rho)^2}\pi_c, \tag{1}$$

where

$$\rho = \frac{\lambda}{c\mu}, \quad \pi_c = \frac{(c\rho)^c}{c!}\left[\sum_{k=0}^{c-1}\frac{(c\rho)^k}{k!} + \frac{(c\rho)^c}{c!(1-\rho)}\right]^{-1}.$$

The main problem with this approach is that the stability condition might not hold for periods of peak arrival in which the limited number of physicians is

not sufficient. As a result, the steady-state waiting time of (1) is not directly applicable.

The first idea to overcome the above difficulty is to *use different waiting time estimations for patients served in a period and overflow patients.* More specifically, we use the following estimations:

$$\text{Mean waiting time of patients served in } t \approx W^{M/M/s}(u_t/\Delta, p_t), \tag{2}$$

$$\text{Total waiting time of overflow patients to } t+1 \approx F^{\text{overflow}}(q_t, \lambda_t), \tag{3}$$

where

$$F^{\text{overflow}}(q_t, \lambda_t) \approx \begin{cases} \sum_{i=1}^{q_t} \frac{1}{\lambda_t} \cdot i = \frac{q_t(q_t+1)}{2\lambda_t}, & \text{if } q_t \le \lambda_t \Delta; \\ (q_t - \lambda_t\Delta)\Delta + \sum_{i=1}^{\lambda_t\Delta} \frac{1}{\lambda_t} \cdot i = q_t\Delta + \frac{\Delta - \lambda_t\Delta^2}{2} & \text{otherwise.} \end{cases}$$

In above estimation (3), the waiting time of overflow patients is calculated as follows. For a stationary Poisson arrival process of rate λ_t, the mean elapsed time since the last arrival observed at the end of the period and the mean interarrival time are all equal to $1/\lambda_t$. Thus, if the number q_t of overflow patients is less than the number $\lambda_t\Delta$ of new patients arriving in this period, the waiting time of the ith last overflow patient is i/λ_t, and the total waiting times of all overflow patients is $\sum_{i=1}^{q_t} i/\lambda_t$. Otherwise, there are $q_t - \lambda_t\Delta$ patients overflowed from the previous period with additional waiting time Δ, and $\lambda_t\Delta$ patients arrive in the new period with a total waiting time $\sum_{i=1}^{\lambda_t\Delta} i/\lambda_t$.

The above estimations lead to a total waiting time in period t equal to $u_t W^{M/M/s}(u_t/\Delta, p_t) + F^{\text{overflow}}(q_t, \lambda_t)$. This approach leads to another difficulty because the number of patients served and the number of overflow patients are unknown random variables.

Rather than employing sophisticated nonstationary queuing analysis, *we transform this performance evaluation problem into an optimization problem consisting of selecting u_t and q_t to minimize the total patient waiting time.*

As will be shown in the numerical experiments, this simple idea is surprisingly efficient because the "artificial decision" variables u_t and q_t closely match the related simulated performance measures and, consequently, provide a good estimation of the total waiting time.

The optimization-based waiting time estimation approach, denoted as APP1, is as follows:

APP1:

$$\min_{(u_1,\dots,u_T)} \sum_{t=1}^{T} W_t \tag{4}$$

$$\text{subject to } q_1 = \lambda_1\Delta - u_1, \tag{5}$$

$$q_t = q_{t-1} + \lambda_t\Delta - u_t, \quad \forall t \in T\backslash\{1\}, \tag{6}$$

$$q_T = 0, \tag{7}$$

$$u_t \le p_t\mu\Delta, \quad \forall t \in T, \tag{8}$$

$$W_t = u_t W^{M/M/s}(u_t/\Delta, p_t) + F^{\text{overflow}}(q_t, \lambda_t), \quad \forall t \in T, \tag{9}$$

$$u_t, q_t \in \mathbb{R}^+, \quad \forall t \in T. \tag{10}$$

Objective function (4) minimizes the total waiting time of all patients. Constraints (5) and (6) are flow balance equations. Constraint (7) is from Assumption A6. Constraint (8) is the capacity constraint, and constraint (9) is the waiting time approximation. In Section 4, we will show that APP1 can be transformed into a linear integer program and can, hence, be solved by a standard solver. We will later propose a more efficient dynamic programming algorithm.

3. Improving Waiting Time Approximation

The APP1 approximation relies on the assumption of the stationary arrival of all patients served in each period. While this assumption is reasonable for patients arriving in the period, it is not reasonable for overflow patients. Overflow patients are already waiting at the beginning of the period and are served consecutively by physicians.

The APP2 approximation is based on the exact waiting time estimation of overflow patients, while the waiting time of patients served in the period of their arrival is estimated by the steady-state waiting time.

Let $u_t = u_t^1 + u_t^2$, where u_t^1 is the number of overflow patients served in period t and u_t^2 is the number of patients arriving and served in period t. On the basis of the FCFS service rule, overflow patients are served first. As a result, we obtain

$$u_t^1 = \min(q_{t-1}, u_t). \tag{11}$$

Let W_t^1 be the total waiting time of overflow patients served in period t. First, consider the case of a single physician, and let d_i be the service time of ith overflow patient. Overflow patients are served sequentially, and we obtain

$$W_t^1 = E\left[\sum_{j=2}^{u_t^1}\sum_{i=1}^{j-1}, d_i\right] = \sum_{j=2}^{u_t^1}\sum_{i=1}^{j-1} E[d_i].$$

Because all service times have a mean of $1/\mu$,

$$W_t^1 = \frac{(u_t^1 - 1)u_t^1}{2\mu}.$$

In the general case with any number p_t of physicians, the first p_t patients do not wait, all physicians remain busy, and patients depart according to a Poisson process with rate $p_t\mu$ until the last overflow patients begin

to be served. The mean waiting time for patient $(p_t + i)$ is $i/p_t\mu$, and the total wait time of overflow patients served in t is

$$W_t^1 = \frac{((u_t^1 - p_t)^+ + 1)(u_t^1 - p_t)^+}{2p_t\mu}. \tag{12}$$

More specifically, APP2 uses the following waiting time approximations:

total wait time of overflow patients served in $t = W_t^1$,
wait per patient arriving and served in $t \approx W^{M/M/s}(u_t/\Delta, p_t)$,
total waiting time of patient overflowed to $t+1 \approx F^{\text{overflow}}(q_t, \lambda_t)$.

Thus, we obtain the total waiting time for all patients in period t:

$$W_t = \frac{((u_t^1 - p_t)^+ + 1)(u_t^1 - p_t)^+}{2p_t\mu} + (u_t - u_t^1)W^{M/M/s}(u_t/\Delta, p_t) + F^{\text{overflow}}(q_t, \lambda_t). \tag{13}$$

Similar to APP1, APP2 is solved by the following optimization problem:

$$\text{APP2:} \quad \min_{(u_1,\ldots,u_T)} \sum_{t=1}^{T} W_t \tag{14}$$

subject to constraints (5)–(8), (10), (11), and (13).

4. Staffing Model

On the basis of the waiting time approximation APP1, this section addresses the ED physician staffing problem to determine the number p_t of physicians in each period t. The goal is to provide hospital ED managers with a decision tool to address time-varying demands through appropriate scheduling of physician working time.

In our staffing model, a physician is assumed to be able to start at the beginning of any period and then work continuously until the end of his or her daily duty. We refer to a physician's working periods as his or her working shift. The ED physician staffing problem consists of determining a working shift for each physician to minimize the total patient waiting time. In practice, many hard and soft constraints on working time regulation exist in ED staffing. It is a challenging task for ED managers to identify good physician staffing patterns. According to our discussions with the ED manager and physicians of Ruijin Hospital, the following constraints are important in ED staffing.

A7. Each physician has only one working shift in one day; that is, a physician's daily working time is continuous and cannot be interrupted.

A8. The shift length is more than the lower bound (LBD) and less than the upper bound (UBD).

A9. The total working time of all N physicians should not exceed the ED physician time budget TW.

A10. For each period t with new physicians starting their shifts, there should be at least one physician working both periods $t-1$ and t.

Assumption A10 is called the "handshaking" constraint in hospitals and in this paper. It is required for patient information to be transferred between physicians working different shifts.

In our staffing model, the ED physician staffing decision is represented by two sets of variables:

- s_t: number of physicians starting their shifts in period t, and
- e_t: number of physicians completing their shifts at the beginning of period t.

The staffing model termed MIP1 is summarized as follows:

$$\text{MIP1:} \quad \min \sum_{t=1}^{T} W_t \tag{15}$$

subject to (5)–(10) and

$$p_1 = s_1 - e_1 + p_T, \tag{16}$$

$$p_t = \sum_{i=2}^{t}(s_i - e_i) + p_1, \quad \forall\, t \in T\backslash\{1\}, \tag{17}$$

$$\sum_{t=1}^{T} s_t \le N, \tag{18}$$

$$\sum_{t=1}^{T} p_t \le TW/\Delta, \tag{19}$$

$$p_t \ge s_t + 1, \quad \forall\, t \in T, \tag{20}$$

$$\sum_{t=1}^{T}(s_t - e_t) = 0, \tag{21}$$

$$p_t - s_t \ge \sum_{i=t+1}^{t+\text{LBD}-1} e_{\tau(i)}, \quad \forall\, t \in T, \tag{22}$$

$$p_t \le \sum_{i=t+1}^{t+\text{UBD}} e_{\tau(i)}, \quad \forall\, t \in T, \tag{23}$$

$$e_t, s_t, p_t \in \mathbb{N}, \quad \forall\, t \in T, \tag{24}$$

where $\tau(i) = t$ if $i < T+1$, and $\tau(i) = i - T$ otherwise.

The objective function (15) minimizes the total waiting time of all patients. Relations (16) and (17) determine the number of physicians in each period as a function of staffing variables. Constraint (18) is the constraint of the number of ED physicians, and constraint (19) ensures the total ED physician time budget. Constraint (20) is the handshaking constraint. Constraint (21) imposes that a shift starts and ends within one day. Constraints (22) and (23) are shift length constraints.

Clearly, this MIP1 formulation is nonlinear because of constraint (9), in which the steady-state waiting time $W^{M/M/s}(u_t/\Delta, p_t)$ and the waiting time $F^{\text{overflow}}(q_t, \lambda_t)$ of overflow patients are complex nonlinear functions. Standard solvers such as CPLEX cannot be used to solve it directly. A linearization technique is proposed to transform MIP1 (also APP1) into a linear model.

The linearization is performed by restricting to integer u_t and integer $\lambda_t\Delta$, which are sufficient for our application but can be easily extended to fractional u_t and $\lambda_t\Delta$ as integer multiples of any real base unit.

Let $U = \lfloor N\mu\Delta \rfloor$ be the maximum number of patients who can be served in a period, and let $V = \lfloor \sum_{t\in T} \lambda_t\Delta \rfloor$ be the maximum number of patients who may be overflowed in a period, where $\lfloor x \rfloor$ denotes the greatest integer smaller than or equal to x.

The linearization is performed with the following four sets of binary variables:

- x_{ti}: binary variable equal to 1 iff $u_t = i$,
- y_{tk}: binary variable equal to 1 iff $p_t = k$,
- z_{tki}: binary variable equal to 1 iff $p_t = k$ and $u_t = i$, and
- a_{ti}: binary variable equal to 1 iff $q_t = i$.

As a result, constraint (9) can be replaced by linear constraints (25)–(33), and MIP1 can be transformed into an integer linear program. Constraints (25) and (26) ensure that u_t patients are served in period t. Constraints (27) and (28) guarantee that p_t physicians are available in period t. Constraints (29) and (30) impose for each period t one and only one combination of (u_t, p_t). Constraints (31)–(32) ensure that there are q_t overflowed patients for period t. For each (t, k, i) in constraint (33), we precompute $W^{M/M/s}(i/\Delta, k)$ and $F^{\text{overflow}}(i, \lambda_t)$; thus, (33) is the final linearization of constraint (9):

$$\sum_{i=0}^{U} x_{ti} = 1, \quad \forall t \in T, \tag{25}$$

$$u_t = \sum_{i=0}^{U} i x_{ti}, \quad \forall t \in T, \tag{26}$$

$$\sum_{k=1}^{N} y_{tk} = 1, \quad \forall t \in T, \tag{27}$$

$$p_t = \sum_{k=1}^{N} k y_{tk}, \quad \forall t \in T, \tag{28}$$

$$\sum_{i=0}^{U} \sum_{k=1}^{N} z_{tki} = 1, \quad \forall t \in T, \tag{29}$$

$$z_{tki} \geq x_{ti} + y_{tk} - 1, \quad \forall t \in T, i \in \{0, 1, \ldots, U\}, k \in \{1, \ldots, N\}, \tag{30}$$

$$\sum_{i=0}^{V} a_{ti} = 1, \quad \forall t \in T, \tag{31}$$

$$q_t = \sum_{i=0}^{V} i a_{ti}, \quad \forall t \in T, \tag{32}$$

$$W_t = \sum_{i=0}^{U} \sum_{k=1}^{N} i \cdot z_{tki} \cdot W^{M/M/s}(i/\Delta, k) + \sum_{i=0}^{V} a_{ti} \cdot F^{\text{overflow}}(i, \lambda_t), \quad \forall t \in T. \tag{33}$$

It is not immediately apparent whether shifts can be derived from a feasible staffing decision $\{s_t, e_t\}$. We show that such shifts exist. Consider the following shifts: s_t new shifts start at the beginning of each period; the shift termination is determined from period 1 to period T on a FCFS basis by allowing the p_T physicians who started the previous day to finish first. We need to check whether each shift meets the shortest-time LBD and the longest-time UBD constraints. Assume, by contrast, that a shift starting in period t has a length smaller than *LBD* (i.e., finishes at the beginning of period $t + \text{LBD} - 1$ or earlier). As a result, physicians who work period t but started earlier have all completed their shift at the beginning of $t + \text{LBD} - 1$. This implies $\sum_{i=t+1}^{t+\text{LBD}-1} e_{\tau(i)} > p_t - s_t$, contradicting constraint (22) and proving that shift lengths are at least *LBD*. Similarly, it can be shown that shift lengths are at most *UBD*.

5. Improving the Staffing Model

In this section, we present a new staffing model, denoted as MIP2, based on the improved waiting time approximation APP2. The difference between models MIP1 and MIP2 is the accuracy of the waiting time approximation:

$$\text{MIP2:} \quad \min \sum_{t=1}^{T} W_t \tag{34}$$

subject to constraints (5)–(8), (10), (16)–(24), (11), and (13).

This model is highly nonlinear because of constraints (11) and (13). Constraint (11) $u_t^1 = \min(q_{t-1}, u_t)$ is linearized by introducing a new auxiliary binary variable g_t and constraints (35)–(39):

$$u_t^1 \leq u_t, \quad \forall t \in T, \tag{35}$$

$$u_t^1 \leq q_{t-1}, \quad \forall t \in T, \tag{36}$$

$$u_t^1 \geq u_t - (1 - g_t)M, \quad \forall t \in T, \tag{37}$$

$$u_t^1 \geq q_{t-1} - g_t M, \quad \forall t \in T, \tag{38}$$

$$g_t \in \{0, 1\}, \quad \forall t \in T, \tag{39}$$

where M is a large positive value. The linearization of constraint (13) is similar to the linearized procedure of MIP1. Constraint (13) is replaced by two sets of new binary variables, h_{ti} and l_{tkij}, following constraints (40)–(44), together with constraints (25)–(28). Constraints (40) and (41) ensure that the number of overflow patients served in period t is u_t^1. Constraints (42) and (43) impose that for each period t there is only one combination of (u_t, u_t^1, p_t). Similar to the linearization of constraint (33), for each combination (t, k, i, j) in (44), we precompute the corresponding waiting time:

- h_{ti}: binary variable equal to 1 iff $u_t^1 = i$,
- l_{tkij}: binary variable equal to 1 iff $p_t = k$, $u_t = i$, and $u_t^1 = j$;

$$\sum_{i=0}^{U} h_{ti} = 1, \quad \forall t \in T, \tag{40}$$

$$u_t^1 = \sum_{i=0}^{U} i h_{ti}, \quad \forall t \in T, \tag{41}$$

$$\sum_{i=0}^{U} \sum_{k=1}^{N} \sum_{j=0}^{i} l_{tkij} = 1, \quad \forall t \in T, \tag{42}$$

$$l_{tkij} \geq x_{ti} + y_{tk} + h_{tj} - 2, \quad \forall t \in T,\ i, j \in \{0, 1, \ldots, U\},\ k \in \{1, \ldots, N\}, \tag{43}$$

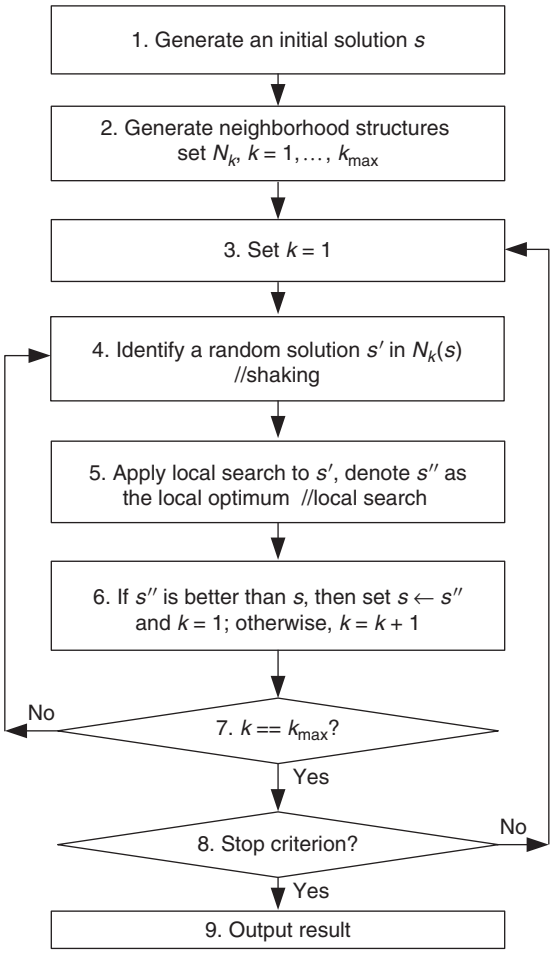
$$W_t = \sum_{i=0}^{U} \sum_{k=1}^{N} \sum_{j=0}^{i} l_{tkij} \cdot \left(\frac{((j-k)^+ + 1)(j-k)^+}{2k\mu} + (i-j) W^{M/M/s}\left(\frac{i}{\Delta}, k\right) \right) + \sum_{i=0}^{V} a_{ti} \cdot F^{\text{overflow}}(i, \lambda_t), \quad \forall t \in T. \tag{44}$$

Although we can obtain linearized models for MIP1 and MIP2, the above mixed-integer programming (MIP) formulations can be used only for problems of a very small size. This is evidenced by the numerical results presented in Section 7, in which we attempt to solve different test instances using the CPLEX 12 solver. CPLEX is not even able to find a feasible solution in a reasonable computational time (i.e., 48 hours) for MIP2 of some real-life instances. Therefore, we propose a heuristic algorithm to address staffing problems.

6. Solution Procedure for the Staffing Problem

In this section, we propose a VNS-based algorithm to solve the staffing problem. The VNS systematically changes neighborhoods during the algorithmic search, and its general structure is outlined in Figure 2. The algorithm starts from an initial solution and a set of neighborhoods N_k, $k = 1, \ldots, k_{\max}$. Then, a random neighborhood solution s' is obtained by a shaking function with respect to the kth neighborhood. Next, a set of local search improvement procedures is applied to the solution s' to obtain a local optimum solution s''. If s'' is better than solution s, then the incumbent solution s is updated by s'', and the algorithm continues with the first neighborhood k_1. Otherwise, the algorithm switches to the next $k+1$th neighborhood. The entire algorithm stops when its termination condition is satisfied. We call the procedure from step 3 to step 8 an algorithm iteration. The entire VNS stops once it reaches the maximum number of iterations.

Figure 2. General Structure of the VNS Algorithm



6.1. Initial Solution

We first design a simple greedy algorithm to generate an initial solution, which consists of three steps. In the initial solution, all the physicians work the same duration $\lfloor TW/(N\Delta) \rfloor$ periods. Step 1 selects a physician to start in period 1. Step 2 adds another physician to start in the last working period of the previous physician to meet the handshaking constraint. Step 2 is repeated until the handshaking constraint is met in all periods. Step 3 adds remaining physicians one by one starting in a period that minimizes the overall objective function.

6.2. Shaking

The shaking procedure is used to properly balance perturbing the incumbent solution and retaining the useful aspects of the incumbent solution. We design a simple and effective shaking procedure in our VNS by randomly removing and then reinserting some physicians. The shaking procedure first randomly selects some physicians one by one and then removes them. These physicians are then reinserted with a random shift start but unchanged shift length. The above procedure is repeated until the handshaking constraint is

met. For each VNS neighborhood N_k with $k = 1, \ldots, k_{\max}$, $k+1$ physicians are removed and reinserted using the shaking procedure; $k_{\max}$ is set to 3 in our numerical experiments.

6.3. Local Search

Each staffing solution obtained by the shaking procedure is improved by local search procedures to obtain a locally optimal solution. Three local search procedures are designed, and the best of the three local optimums is selected. Each local search procedure is a local improvement performed through a complete search of the feasible neighborhoods until no local improvement is possible. The three local procedures are illustrated in Figure 3. We stress that only local moves leading to feasible neighbor staffing solutions are considered.

Local search *Relocation* relocates a physician shift to another shift start but does not change the shift length; $T-1$ relocations are possible for each physician. Local search *Extension* attempts to extend a physician's shift length by adding one extra working period either at the beginning or at the end of the shift.

Local search *Exchange* modifies two physician shifts simultaneously within a move. In each Exchange move, one shift becomes longer by adding one period at the shift start or shift end, and another shift becomes shorter by removing one period at either the start or end.

6.4. Evaluation of Staffing Solutions

The goal of this subsection is to evaluate a specific staffing solution. As the number p_t of physicians in each period is known, APP1 and APP2 can be solved by standard integer linear program solvers. Unfortunately, this method is too time-consuming to be used in VNS, which requires the evaluation of a large number of staffing solutions. This subsection proposes an exact dynamic programming (DP) method and a simple policy-based method.

The DP method decomposes the staffing evaluation into a set of stage-dependent and state-dependent problems. Each stage is related to a period t, and the state is defined as the number q_{t-1} of overflow patients from previous periods. The stage-state subproblem is defined as follows:

$f_t(q_{t-1})$: minimal total waiting time from t to T by starting with q_{t-1} overflow patients.

The following optimality equations can be used to recursively solve all these subproblems:

$$f_t(q_{t-1}) = \min_{0 \le u_t \le \min(q_{t-1}+\lambda_t\Delta, p_t\mu\Delta)} \{W_t(q_{t-1}, u_t) + f_{t+1}(q_{t-1}+\lambda_t\Delta - u_t)|\}, \quad \forall 1 \le t \le T, \quad (45)$$

where $W_t(q_{t-1}, u_t)$ is the total period t waiting time determined by (9) in APP1 and (13) in APP2, and by convention, $f_{T+1}(0) = 0$, $f_{T+1}(q_T) = M$, $\forall\, q_T > 0$, with M being a large number. The total waiting time $\sum_{t=1}^{T} W_t$ is given by $f_1(0)$.

The DP method can be used to evaluate the solution cost but is not sufficiently fast to be applied to the VNS local search procedure, especially for staffing model MIP2. Therefore, we design a simple and quick policy-based method to estimate the objective cost. In each period t, we set the number of patients served:

$$u_t = \min(q_{t-1} + \lambda_t\Delta, p_t\mu\Delta - \gamma), \quad (46)$$

where γ is a predefined parameter that can be considered as a slack capacity to avoid a large waiting time $W^{M/M/s}(u_t/\Delta, p_t)$ of customers served in period t. This expression indicates that, compared with the modified maximum service capacity $p_t\mu\Delta - \gamma$, if the total available patients $q_{t-1} + \lambda_t\Delta$ in period t is small, all $q_{t-1} + \lambda_t\Delta$ patients will be served; otherwise, the system only serves its modified capacity.

Figure 3. Local Search Procedures

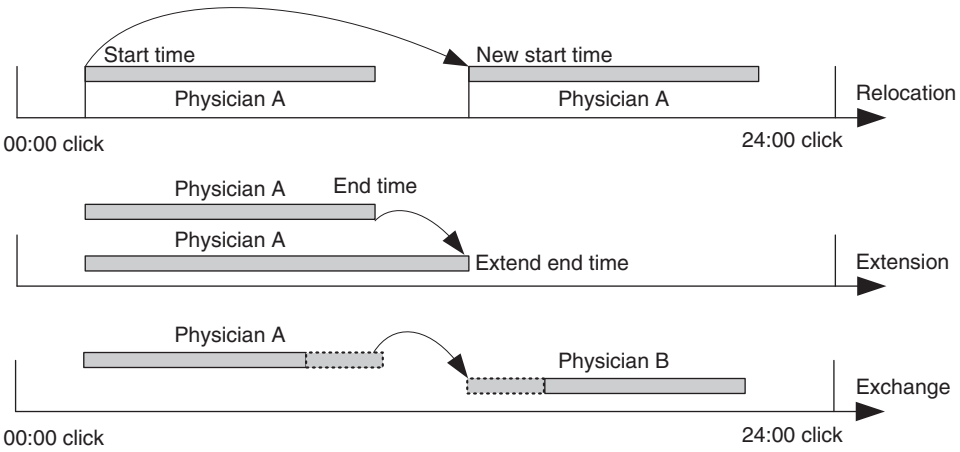
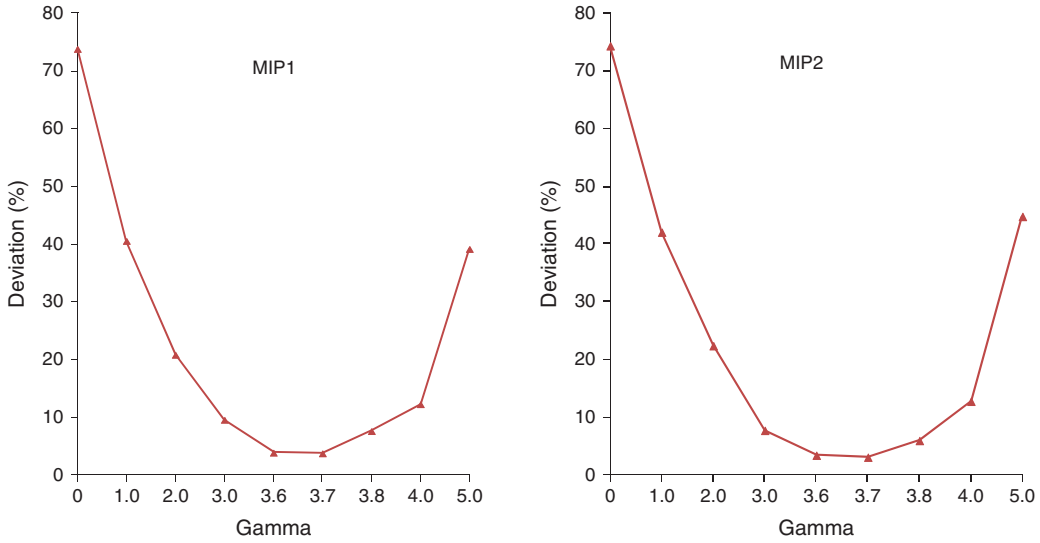


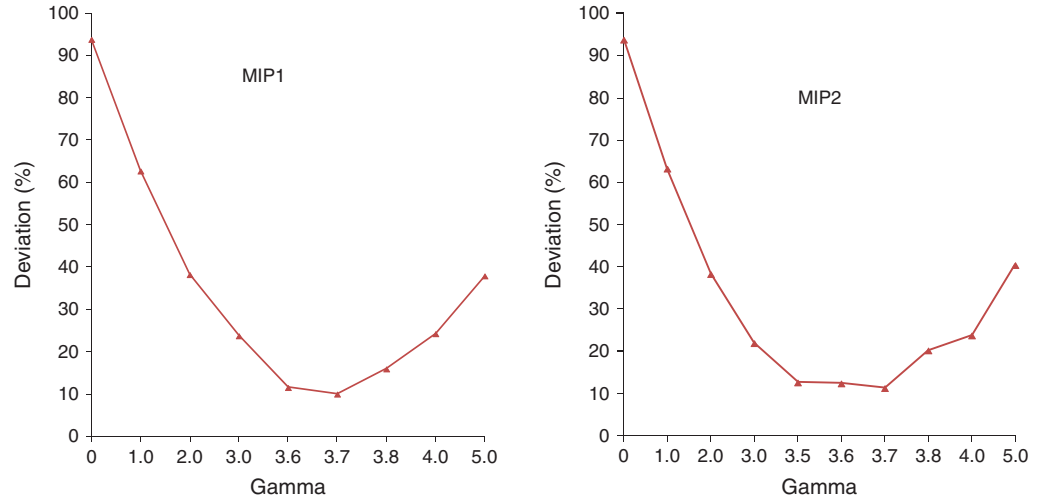
Figure 4. (Color online) Gap of the Policy-Based Method vs. DP for Ruijin Day 3 Scenario



Clearly, this policy-based method cannot ensure an optimal solution cost, but it is rather quick and simple. Not surprisingly, γ is a key factor in the accuracy of this policy-based method. We choose the optimal staffing of the one-week scenario of Ruijin and one-week scenario of Shanghai No Six Hospital in Section 7.2 as parameter tuning test instances. On the basis of the two approximation models APP1 and APP2, we apply two evaluation methods (DP and policy-based method) to the optimal staffing solution and compare their results. Figure 4 illustrates the percentages of deviation of the policy-based estimation from the exact DP method for the day 3 scenario of Ruijin, in which the left part is based on APP1 and the right part is based on APP2. Figure 5 illustrates the corresponding deviations of day 7 of No Six Hospital. Note that $\gamma = 0$ is not a good choice because the resulting cost is more than twice the optimal cost. From $\gamma = 0$, the policy-based solutions improve as γ increases. With $\gamma = 3.7$, for the Ruijin day 3 data, the deviations from the optimum are only 3.86% and 3.02% for MIP1 and MIP2, respectively. More preliminary experiment suggests that $\gamma = 3.7$ can serve as a good default value.

Based on the above evaluation methods, two VNS approaches are considered: VNS-DP using DP to evaluate the solution and VNS-TC using both DP and policy-based methods to evaluate the solution. In VNS-TC, a neighbor solution s is abandoned if its biased policy-based cost $V^p(s)$ degrades the current best solution V^{best} by a percentage of deviation (i.e., $V^p(s) > V^{\text{best}} \times (1+\omega)$); otherwise, its exact cost $V^{\text{DP}}(s)$ is determined by DP and used in VNS. The threshold value ω allows us to maintain a balance between the algorithmic accuracy and speed. This local search mechanism

Figure 5. (Color online) Gap of the Policy-Based Method vs. DP for No Six Hospital Day 7 Scenario



reduces the computation time; however, a poor choice of threshold ω may reduce the quality of the final solutions of the VNS. A numerical comparison of different threshold values ω is shown in Section 7.2.

7. Numerical Experiments

In this section, numerical experiments are conducted to assess the performance of the proposed models and methods. First, the waiting time approximations are examined and compared with simulation results. Next, we test different optimization objectives in the staffing model and check the system performances. Then, computational experiments are performed to assess the performance of the proposed VNS. Algorithms are run in C++ on a 3.2 GHz Dual Core computer with 2 GB memory and Linux operating system. The MIP models are solved using commercial solver CPLEX 12.6 with a maximum time limit of 48 hours and a memory limit of 48 GB for solving each test instance. Each VNS algorithm stops after 300 iterations.

All staffing solutions presented in this paper are precisely evaluated using our event-driven simulation program with 1.6×10^5 replications. Statistics are collected to evaluate key performance indicators, such as the average number of patients served during each period and the total waiting time of all patients on each day.

We select real-life data from two hospitals located in Shanghai (Ruijin Hospital and No Six Hospital) to serve as the basis for the numerical experiment presented in this section. Each hospital provides three weeks of ED operation data; a total of 6×7 instances are considered, each corresponding to one day. The arrival pattern of the ED of Ruijin Hospital is summarized as follows. The arrival rate of patients is low at midnight (01:00–05:00). There is a sudden increase at approximately 06:00; then the arrival rate drops slightly and remains at approximately 10–15 per hour until the second peak at approximately 21:00. The arrival rate decreases thereafter, falling back to the low at midnight. The arrival pattern of No Six Hospital is similar to that of Ruijin Hospital. However, its arrival rate peaks at approximately 07:00 and 19:00. Furthermore, on average, its arrival rate is higher than that of Ruijin Hospital. For example, the number of patients per day in Ruijin Hospital is approximately 300–350, whereas that of the No Six Hospital is about 480–520.

The actual staffing of Ruijin Hospital is as follows: three shifts per day, two physicians for the night shift (22:00–06:00), four for the morning shift (06:00–14:00), and four for the afternoon shift (14:00–22:00). The staffing of the No Six Hospital is similar: three physicians for the night shift (22:00–06:00), six for the morning shift (06:00–14:00), and five for the afternoon shift (14:00–22:00). The key staffing parameters of two hospitals are summarized in Table A in the online supplement. Note that the handshaking constraint is not formally imposed in the actual staffing.

7.1. Waiting Time Approximations vs. Simulation

We first compare the waiting time approximations APP1 and APP2 with the simulation results to examine the accuracies of two approximations. We select two weeks of data (one week from Ruijin and one from No Six Hospital) for the experiment. Tables 1 and 2 show the total waiting time (in hours) of each hospital with the actual staffing. The 95% confidence interval is estimated for the simulation results. We also show the deviation from the simulations of APP1 and APP2. Furthermore, to more clearly compare APP1, APP2, and the simulation results, we chose two days (day 4 from Ruijin and day 2 from No Six Hospital) as examples to illustrate the number of patients served per hour, the hourly total patient waiting time, and the number of overflow patients per hour in Figures 6, 7, and 8, respectively. In these figures, the left part pertains Ruijin data and the right part pertains to No Six Hospital. Note that although the handshaking constraint is not imposed in the actual staffing, this violation does not have a negative impact on the comparisons

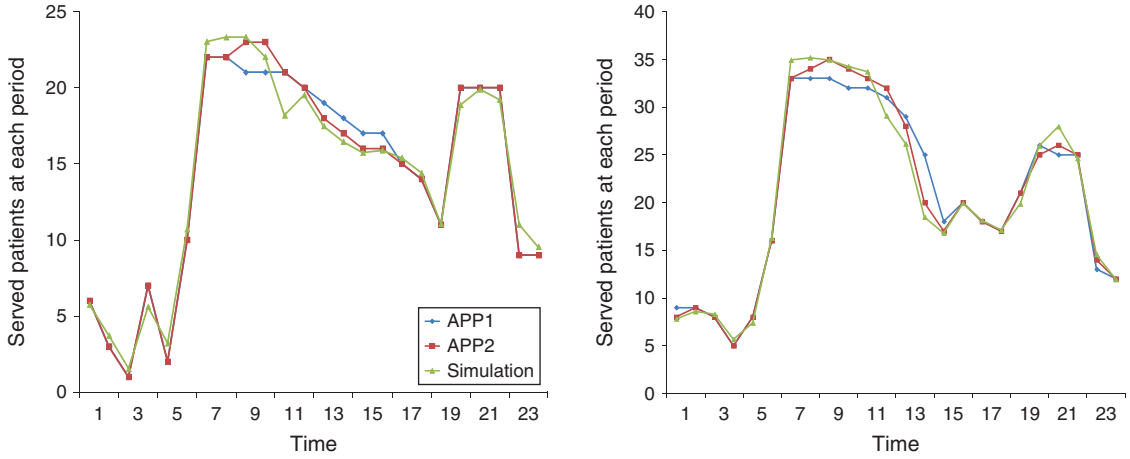
Table 1. Total Daily Waiting Times for the Actual Staffing of Ruijin Hospital

		APP1		APP2	
	Simulation	Waiting time	Dev. (%)	Waiting time	Dev. (%)
Day 1	73.569 ± 0.277	93.050	20.94	80.104	8.16
Day 2	137.842 ± 0.371	187.587	26.52	149.687	7.91
Day 3	84.518 ± 0.157	98.885	14.53	86.658	2.47
Day 4	101.240 ± 0.269	137.112	26.16	111.346	9.08
Day 5	73.634 ± 0.221	92.699	20.57	78.595	6.31
Day 6	96.436 ± 0.326	133.269	27.64	109.196	11.69
Day 7	86.053 ± 0.265	111.835	23.05	95.394	9.79
Avg.	93.327	122.062	23.54	101.569	8.11

Table 2. Total Daily Waiting Times for the Actual Staffing of No Six Hospital

		APP1		APP2	
	Simulation	Waiting time	Dev. (%)	Waiting time	Dev. (%)
Day 1	166.933 ± 0.420	197.974	15.68	174.926	4.57
Day 2	152.995 ± 0.566	175.738	12.94	155.648	1.70
Day 3	181.325 ± 0.281	198.916	8.84	178.834	−1.39
Day 4	109.462 ± 0.354	131.546	16.79	117.607	6.93
Day 5	206.368 ± 0.729	246.464	16.27	209.708	1.59
Day 6	163.323 ± 0.432	197.623	17.36	173.120	5.66
Day 7	154.806 ± 0.539	181.155	14.55	161.460	4.12
Avg.	162.173	189.917	14.61	167.329	3.08

Figure 6. (Color online) Number of Patients Served per Hour from Ruijin (Left) and No Six Hospital (Right) Data



between the waiting time approximations and simulation results because this constraint exists only in the staffing models.

Two conclusions can be drawn. First, in Tables 1 and 2, the daily waiting times of the two optimization-based approximations APP1 and APP2 are reasonably close to the simulation results since the quantities are difficult to estimate. However, both tend to overestimate daily waiting time. Furthermore, Figures 6–8 show that APP1 and APP2 very closely match the evolution of the simulated performances throughout the day for all performance indicators.

Second, as shown in Tables 1 and 2, APP2 is a better approximation than APP1, especially in terms of patient waiting time. For each test instance, the waiting time obtained by APP2 is closer to the simulation result than that of APP1. For seven instances from Ruijin Hospital, the error in the daily total patient waiting time is, on average, 23.54% for APP1 and 8.11% for APP2; for the one week of instances from No Six Hospital, the mean deviation of the daily waiting time is 14.61% and 3.08% for the two models. The better precision of APP2 is also evident in Figures 6–8.

7.2. Tuning VNS Algorithm Parameters

This subsection addresses the parameter tuning of the VNS algorithms to achieve an appropriate balance between precision and speed. Two VNS methods, VNS-DP and VNS-TC, are considered. The VNS-TC depends on a threshold value ω, below which a solution requires the DP evaluation. In our VNS-TC implementation, ω is set as a percentage of the objective function. The values $\omega = 0.1\%, 0.5\%, 1\%, 5\%, 10\%, 20\%,$ and 30% are tested, and the related VNS-TC methods are termed VNS-TC-0.1, VNS-TC-0.5, etc. VNS-DP uses DP to exactly evaluate each staffing solution and is a special case of VNS-TC with $\omega = \infty$.

One week of data from No Six Hospital is used. For each test instance, each method is run 10 times. On the basis of each approximation model (APP1 and

Figure 7. (Color online) Hourly Patient Waiting Time from Ruijin (Left) and No Six Hospital (Right) Data

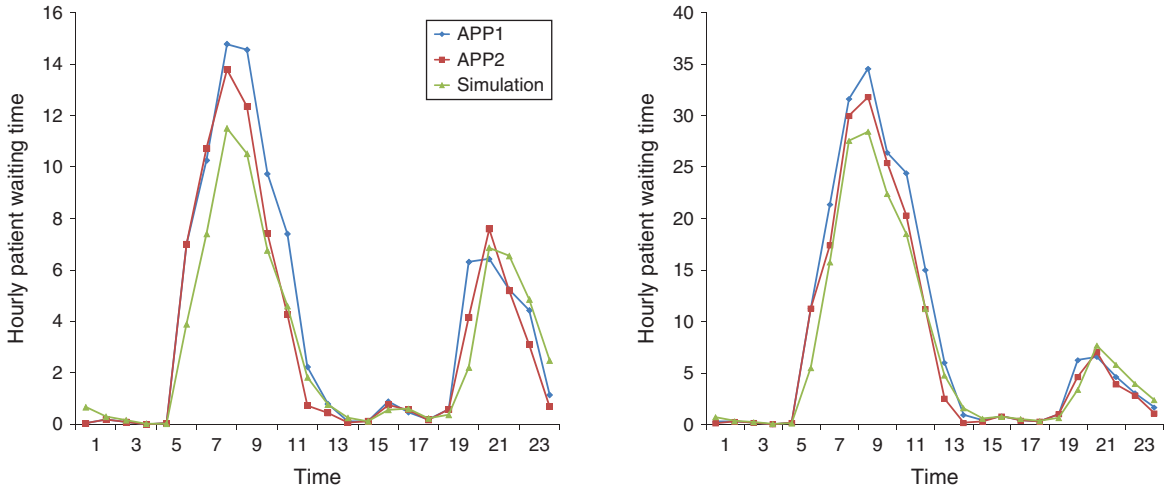


Figure 8. (Color online) Number of Overflow Patients per Hour from Ruijin (Left) and No Six Hospital (Right) Data

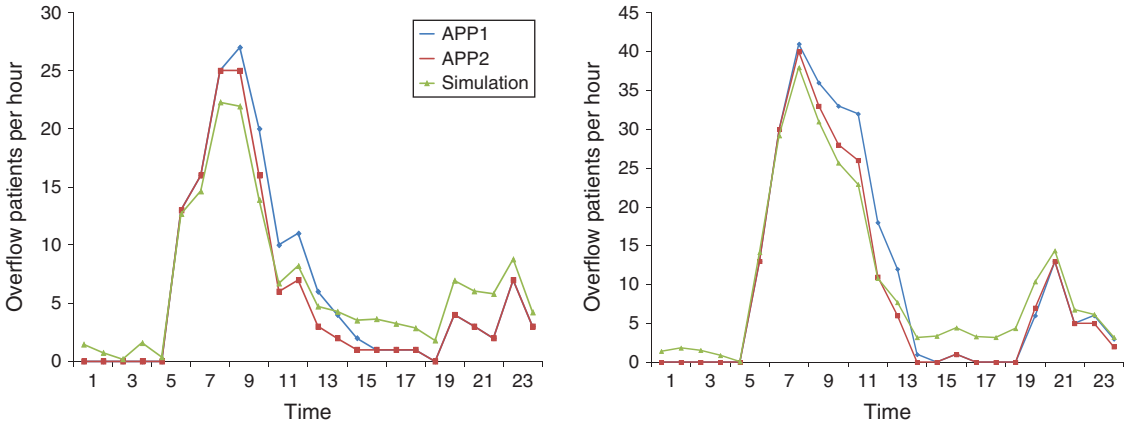
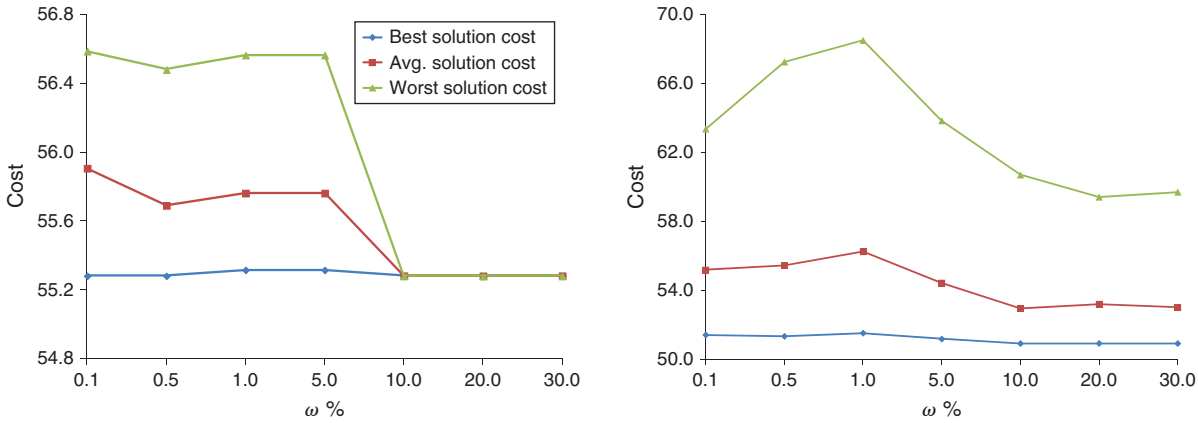


Figure 9. (Color online) The Mean Value of the Best, Average, and Worst Solution Costs of Various ω's (Left: APP1, Right: APP2)



APP2), for each method we obtain the *best*, the *average*, and the *worst* solution costs from 10 runs. The complete results are summarized in Figure 9, in which the left part shows the results of APP1 and the right part shows the result of APP2. Detailed comparisons between five selected methods (VNS-DP, VNS-TC-0.1, VNS-TC-1, VNS-TC-10, and VNS-TC-30) are presented in Tables B and C in the online supplement.

Several conclusions can be drawn. First, VNS-DP does not obviously outperform other methods with respect to solution costs. As shown in Table C in the online supplement, for the hardest test instances (i.e., using data from No Six Hospital and the APP2 model), the mean values of all best, average, and worst solution costs of VNS-DP are 50.931, 52.781, and 60.342, respectively, whereas the corresponding values of VNS-TC-30 are 50.931, 53.144, and 59.703. Meanwhile, VNS-DP requires a much longer computation time than that of other methods and requires 20–30 times more CPU time than VNS-TC-10 for the test instances from No Six Hospital.

Second, the solution quality of VNS-TC increases with the threshold value and becomes stable beyond $\omega = 10\%$. For example, with the APP1 model, the mean values of the best, average, and worst solution costs over seven days decrease from $\omega = 0.1\%$ to 10% but do not significantly improve from $\omega = 10\%$ to 30%.

Third, the running CPU time of VNS-TCs increases with the values of the threshold from 110.1 seconds to 553.8 seconds for APP1 and from 235.5 seconds to 1,337.2 seconds for APP2 because of the more frequent use of the exact DP evaluation. We find the running time is still moderate when $\omega = 10\%$. With APP2, VNS-TC-10 needs 0.14 times more CPU time than VNS-TC-0.1, whereas VNS-TC-30 requires 5 times the CPU time of VNS-TC-10.

In summary, based on the experimental results, $\omega = 10\%$ is a reasonable choice to maintain the balance between speed and accuracy. Therefore, we adopt $\omega = 10\%$ as our default VNS-TC threshold value.

7.3. Optimal Staffing vs. Actual Staffing

This subsection first assesses the optimality of VNS-TC and then compares it with the actual hospital staffing using simulation. To intensively examine the proposed algorithm, we use all six weeks of real-life data from Ruijin and No Six Hospitals. Furthermore, we derive three weeks (3×7 days) of larger-scale test instances

according to the practical arrival pattern of ED: in each test instance the first peak appears at approximately 07:00 to 09:00, and the second crowded period is approximately 19:00 to 21:00. Each large-scale instance consists of 21 available physicians and approximately 1,000 visiting patients. During each hour, patients arrive according to a Poisson process; the service times are exponentially distributed with mean service rate $\mu = 5.872$. VNS runs 10 times for each test instance. The optimality of VNS-TC is assessed by comparing it with CPLEX 12 with a limit of 48 hours and 48 GB memory. Because of space limitations, the detailed computation results are presented in Tables D–F in the online supplement.

First, we compare the performances of VNS and CPLEX. For the Ruijin test instances, on average, CPLEX is able to optimally solve MIP1 in 3.3 hours; it cannot solve MIP2 to optimality in 48 hours for any test instance. MIP1-VNS based on MIP1 always finds the optimal solution in all instances, with an average computation time of 96 seconds. MIP2-VNS based on MIP2 always finds a much better solution than CPLEX within a much shorter computation time. Similar observations can be made for the test instances of No Six Hospital.

We also compare the VNS staffing solution with the actual hospital staffing. As shown in Tables D–F, the MIP staffing solutions are significantly better than the actual hospital staffing: they reduce the total patient waiting time to 30%–50% that of the actual staffing. This result is even more remarkable given that VNS staffing solutions are built under handshaking constraints that are not met by the actual staffing.

Comparing the two staffing models, MIP2 is better than MIP1. Of 21 Ruijin test instances, MIP1-VNS and MIP2-VNS find the same solutions for 14 instances; for 6 test instances, MIP2-VNS obtains better solutions than those of MIP1-VNS. For one test instance, MIP1-VNS very slightly outperforms MIP2-VNS. The same conclusion can be drawn from the results of No Six Hospital (see Table E) and the large-scale data (see Table F).

We now explore the detailed staffing solutions to see how our near-optimal MIP2-VNS staffing matches the arrival pattern. Figures 10 and 11 show staffing solutions of Ruijin week 1, day 1 and No Six Hospital week 1, day 5, with the patient arrival pattern on the left and the number of physicians staffed for each period on the right. More results are given in Figures A and B in the online supplement. We can observe that the staffing

Figure 10. (Color online) Patient Arrival Pattern and Corresponding Staffing (Ruijin, Week 1, Day 1)

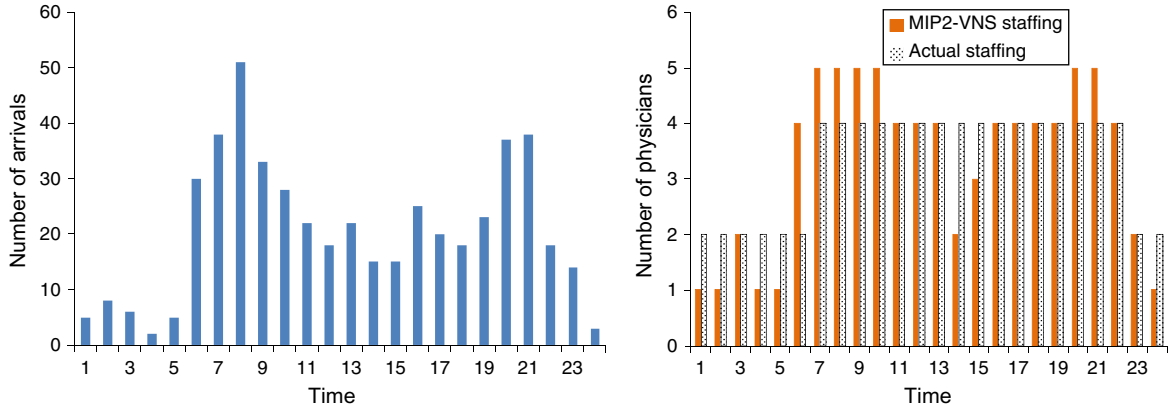
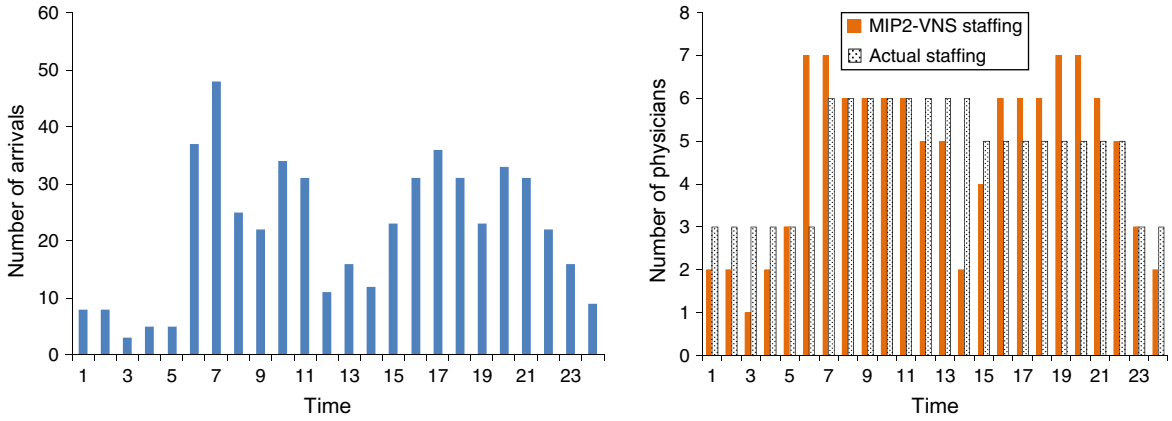


Figure 11. (Color online) Patient Arrival Pattern and Corresponding Staffing (No Six Hospital, Week 1, Day 5)



level of the MIP2-VNS solution better matches the fluctuation in the patient arrival rate. For example, in Figure 10, the MIP2-VNS staffing level reaches the daily maximum of five physicians during the two peak periods of 07:00–09:00 and 20:00–21:00 and appropriately lowers down to two physicians at the daytime trough around 14:00. Another interesting observation is that the staffing level remains high immediately after the peak arrivals to absorb patients overflowed from the peak periods. By contrast, the actual staffing is not as flexible as our staffing and results in a longer patient waiting time. Similar observations can be drawn from the other cases. Therefore, we can assert that in a service system with time-varying and stochastic demands, the better the staffing level matches the demand fluctuation, the better the staffing solution.

Consider now the shift schedules derived from the staffing solution. Figures 12 and 13 give the shift schedules for Ruijin week 1, day 1 and No Six Hospital week 1, day 5, in which each line corresponds to a physician. More results are shown in Figures C and D in the online supplement. The following observations can be made.

First, more shifts are used compared with the actual three standard shifts. There are seven and nine different shifts in Figures 12 and 13, respectively. This is not surprising, as no restriction is made on the type of shifts and the shift starting time, and an additional handshaking constraint is imposed on our solution. Nevertheless, the shifts can be roughly grouped into three types: "morning shifts," "afternoon shifts," and "night shifts." This result coincides with the actual staffing patterns of the hospitals, but shifts of the same type can start at different times in our solution.

Second, compared with the actual shift scheduling, our morning shifts start earlier to better match the morning peak arrivals. For example, the actual Ruijin operation assigns four physicians to the morning shift starting at 06:00, whereas our solution in Figure 12 lets three physicians start at 05:00 and another one start at 06:00.

Third, as our ED staffing models do not consider relevant working time regulations/costs and organization constraints, some shifts may be hard to adopt in practice. For example, there are shifts starting at 02:00 and 03:00 in Figures 12 and 13. Such shift start times might not be reasonable in practice. Future research taking into account such working time constraints is needed to make our results applicable.

7.4. Sensitivity Analysis

This section presents a sensitivity analysis to show how the physician staffing adapts to changes in arrival rate and the number of available physicians. Four days of test data are selected: days 1 and 4 of the first week of practical data from Ruijin Hospital and days 1 and 4 of the first week's data from the large-scale test instances consisting of 21 physicians. Two cases are considered for each test instance: one that varies the arrival rate of the base case and one that varies the number of physicians. The proposed VNS and MIP2 are used to determine physician staffing.

Since Ruijin Hospital only provides the actual hospital staffing for the current 10 physicians, the "actual staffing" for different numbers of physicians is derived from the actual physician staffing. For example, in terms of test instances derived from Ruijin Hospital, staffing with more than 10 physicians is obtained by iteratively adding a physician to the shift with the highest per-physician workload. Staffing with fewer physicians is obtained by iteratively removing a physician from the

Figure 12. (Color online) Physician Working Shift of VNS Solution for Ruijin Hospital (Week 1, Day 1)

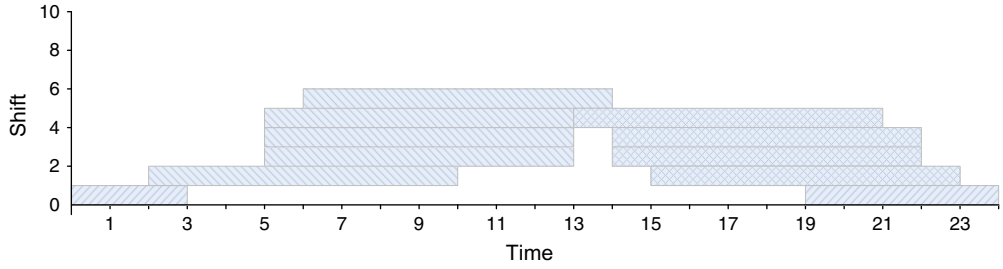


Figure 13. (Color online) Physician Working Shift of VNS Solution for No Six Hospital (Week 1, Day 5)

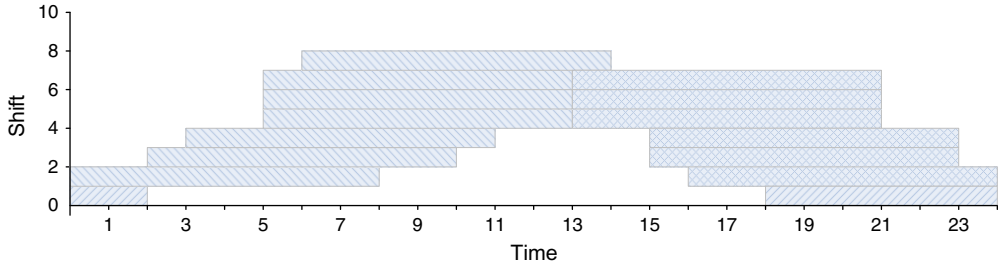


Figure 14. (Color online) Patient Waiting Time vs. Arrival Rate for Ruijin, Week 1, Day 1 (Left) and Large-Scale Instance, Week 1, Day 4 (Right)

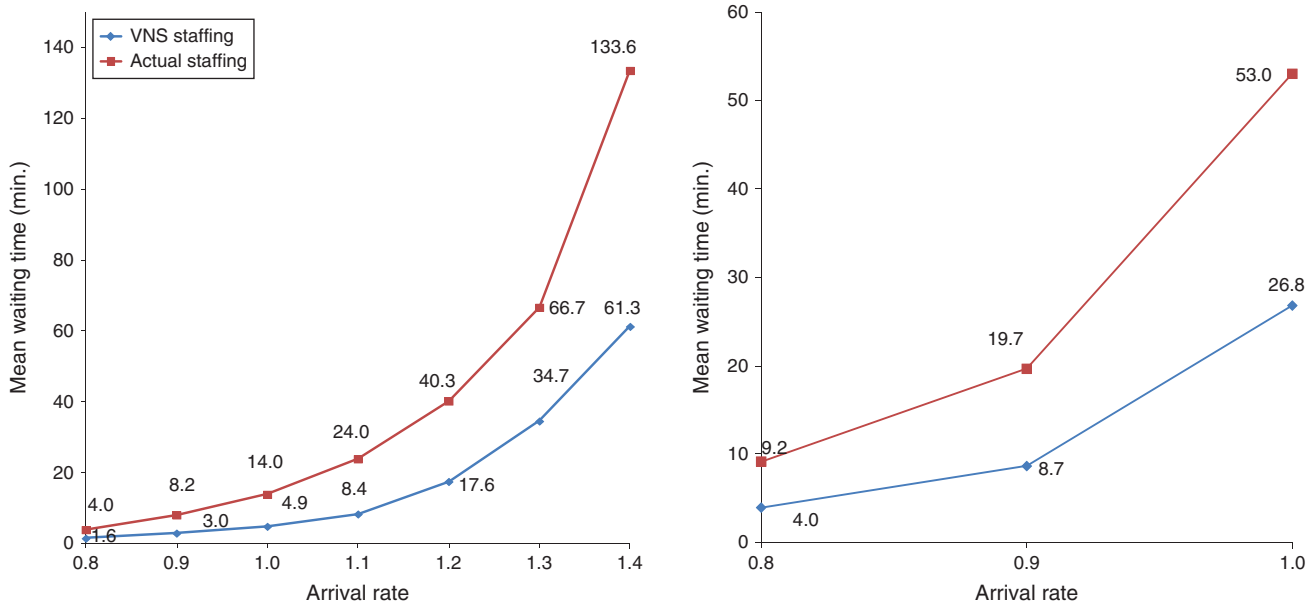
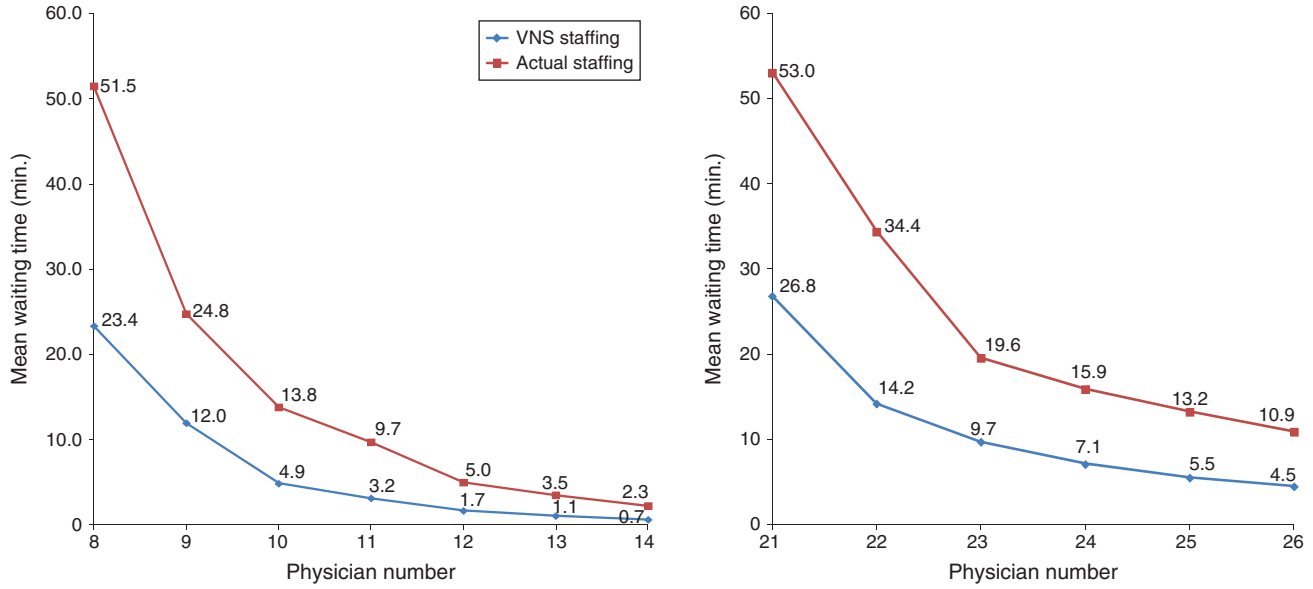


Figure 15. (Color online) Patient Waiting Time vs. Physician Number for Ruijin, Week 1, Day 1 (Left) and Large-Scale Instance, Week 1, Day 4 (Right)



shift with the lowest per-physician workload. A similar method is applied to the large-scale test instances to obtain the actual staffing for each number of physicians.

Figure 14 shows the average waiting time per patient for day 1 at Ruijin Hospital and day 4 in the large-scale test instance, for various arrival rates, for both our VNS and actual hospital staffing. Figure 15 presents the sensitivity results versus the number of physicians. More sensitivity results are given in Figures E and F in the online supplement.

We find that the average patient waiting time increases with the arrival rate. The average patient waiting time is a convex function of the arrival rate, which echoes the convexity properties of the classical Erlang B and C queues (see Whitt 2002). With Ruijin data, there is an important change from 120% to 140% arrival rates. For example, for the Ruijin day 1 test instance, the mean waiting time per patient in actual staffing for the 120%, 130%, and 140% arrival rates is 40, 67, and 134 minutes, respectively. Compared with the actual staffing, our VNS staffing is clearly better: its corresponding mean waiting times are 18, 35, and 61 minutes.

For the large-scale test instance, it cannot serve higher arrival rates since it is already a high-traffic system over the day. For example, the mean traffic intensities for day 1 and day 4 scenarios are both already 94.6%.

Table 3. Simulated Waiting Times of MIP1 and MIP3 Solutions for Ruijin Data

	MIP1			MIP3		
	W_t	Min–max	Max-wt	W_t	Min–max	Max-wt
Day 1	26.003 ± 0.126	0.261 ± 0.011	0.431 ± 0.009	34.317 ± 0.166	0.145 ± 0.006	0.331 ± 0.008
Day 2	52.433 ± 0.155	0.683 ± 0.010	0.703 ± 0.010	61.721 ± 0.172	0.359 ± 0.007	0.540 ± 0.005
Day 3	34.329 ± 0.106	0.201 ± 0.008	0.404 ± 0.007	39.842 ± 0.154	0.200 ± 0.006	0.363 ± 0.006
Day 4	52.171 ± 0.124	0.420 ± 0.006	0.525 ± 0.006	56.957 ± 0.183	0.265 ± 0.006	0.443 ± 0.005
Day 5	25.587 ± 0.087	0.164 ± 0.005	0.336 ± 0.006	32.121 ± 0.145	0.145 ± 0.004	0.306 ± 0.004
Day 6	29.967 ± 0.079	0.264 ± 0.008	0.429 ± 0.007	35.258 ± 0.127	0.181 ± 0.007	0.327 ± 0.005
Day 7	32.091 ± 0.157	0.422 ± 0.009	0.443 ± 0.008	40.763 ± 0.156	0.216 ± 0.005	0.416 ± 0.005

Similarly, the average patient waiting time is also a convex function of the number of physicians. Our VNS staffing is significantly better than the actual hospital staffing. If we set a mean waiting time target of 45 minutes, VNS staffing can handle up to a 20% increase in ED arrival or the absenteeism of two physicians (i.e., with only eight physicians) in Ruijin Hospital data sets. From Figure 15, with Ruijin's week 1, day 1 data and only eight physicians, VNS still provides a staffing solution with a mean waiting time of 23.4 minutes, whereas the actual hospital staffing rule leads to a mean waiting time of 51.5 minutes.

The sensitivity analysis provides a means of managing the ED. With a waiting time target of 45 minutes, at Ruijin Hospital, the VNS staffing is able to handle approximately 20% more ED patients or the absenteeism of two physicians.

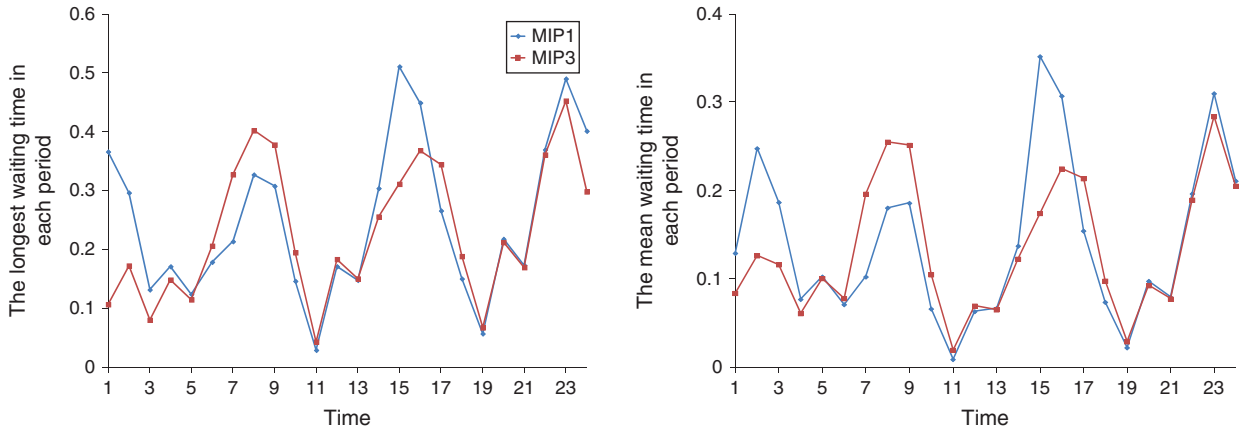
7.5. Smoothing Patient Waiting Times

In two optimization-based waiting time estimation approaches, APP1 and APP2, and two staffing models, MIP1 and MIP2, the objective function attempts to minimize the total waiting time of patients over all periods. In actuality, our methodology is flexible and can be easily extended to different optimization objectives. For example, in the ED time-varying staffing setting, it is sometimes desirable to have smooth waiting times for patients arriving at different times. We use expression (47) to approximately balance the waiting time of patients arriving at different periods. The second term in (47) takes into account the total waiting time as a secondary criterion in which ε is a small positive value. On the basis of constraints (5)–(10) in waiting time approximation APP1, constraints (16)–(24) in model MIP1, and objective (47), we obtain a new staffing model, MIP3:

$$\min\left(\max_{t\in T}(W_t/\lambda_t)\right)+\varepsilon\sum_{t=1}^{T}W_t. \tag{47}$$

In objective (47), the min–max function is nonlinear, but it can be easily linearized. We select one week of data from Ruijin Hospital and solve to optimality MIP1 and MIP3 using CPLEX. The optimal solutions of the two models are then simulated to obtain statistical results, such as the total waiting time of patients and the maximum waiting time among all patients. The final simulation results are presented in Table 3, in which "W_t" represents the total waiting time in hours, "Min–max" is the value of the first term in (47), and "Max-wt" gives the maximum waiting time among all patients on each day. We then investigate the intraday waiting time distribution for the day 4 test instance. Figure 16 illustrates the longest and the mean waiting times of patients who arrive in each period.

Figure 16. (Color online) Longest Waiting Time (Left) and Mean Waiting Time (Right) of Patients Arriving in Each Period (Ruijin, Day 4)



From Figure 16, two staffing solutions have similar intraday waiting time distributions, with longer waiting times for patients arriving at approximately 07:00–09:00, 14:00–16:00, and 23:00–24:00.

As expected, MIP3 provides a better balance of the waiting times among patients. For Ruijin day 4 data, the longest waiting time among all patients is 0.525 hours for the MIP1 solution and is reduced to 0.443 hours for the MIP3 solution. Over the week, the average daily longest waiting time is 0.467 hours for the MIP1 staffing, which is reduced to 0.389 hours by MIP3. Figure 16 also shows that MIP3 better smoothes both the longest and mean waiting times of patients arriving in different periods.

On the other hand, Table 3 shows that the total waiting time of the MIP1 solution is naturally lower than that of the MIP3 solution. Furthermore, the combination of constraints (5)–(10) in approximation APP1 and objective function (47) leads to a new approximation model, APP3. Additional experiments show that, for a given staffing solution, APP1 provides a better approximation than APP3, with results closer to the simulation results.

8. Conclusions and Future Research

In this work, we have addressed the ED staffing problem with nonstationary arrivals to minimize patient waiting time. Because of time-varying patient arrivals, the ED system is critically overloaded during peak hours and cannot ensure the system's stability throughout the day. A simple discrete-time analytical model is proposed to estimate patient waiting time based on two simple but surprisingly efficient ideas: (i) split patients in each period into patients served and overflow patients and (ii) transform the evaluation problem into an optimization problem with the number of overflow patients as a decision variable and with the waiting time of patients served estimated by simple queueing models. These waiting time approximations are then incorporated into two physician staffing models.

We then design VNS algorithms to solve these highly nonlinear optimization models. Numerical experiments with field data produce the following results: (i) the simple waiting time approximations are surprisingly good; (ii) the proposed VNS algorithms are able to find good staffing solutions in a short computation time; (iii) the optimal physician staffing significantly improves the actual hospital staffing; and (iv) for time-varying and stochastic demands, ED staffing should follow the demand fluctuation by increasing the staffing level for peak periods, keeping it higher to absorb overflows, and lowering it at trough periods.

Although the ED staffing techniques proposed in this paper are efficient in improving service quality, a significant gap exists for real-life applications. The basic ED staffing models of this paper should be extended to take into account relevant working time regulations and organization constraints. Shifts cannot be freely chosen as in our paper but from some predefined shift patterns. There might also be restrictions on the starting times of different personnel. ED physicians might have different preferences. The optimization criterion of total patient waiting time should also be extended to take into consideration of ED personnel cost and equal service quality.

Fortunately, our staffing models and VNS algorithms are flexible enough to take these factors into consideration. For example, our method was shown, in Section 7.5, to directly apply to smoothing the waiting times for patients arriving at different times. Our paper provides a basis that is flexible enough to include practical considerations that are required for real-life applications. As a result of this work, now Ruijin Hospital and No Six Hospital both collaborate with us to improve their ED staffing.

This work can be extended in several directions. First, weekly or monthly ED staffing taking into account the patient waiting time is more complicated because of the change of scale and other complex staffing constraints. Second, ED staffing taking into account all human and material resource requirements is an interesting avenue for future research, given that waiting times depend on both nurse staffing and physician staffing. Another interesting avenue for future research is ED staffing for other performance indicators, such as meeting the waiting time target. A more challenging future research topic is the joint optimization of ED staffing and patient flow control to account for heterogeneous ED care requirements and corresponding heterogeneous service-level requirements.

Acknowledgments

The support of the National Natural Science Foundation of China is gratefully acknowledged. The authors thank two anonymous reviewers for their valuable comments on a previous version of the paper.

References

Ahmed MA, Alkhamis TM (2009) Simulation optimization for an emergency department healthcare unit in Kuwait. *Eur. J. Oper. Res.* 198(3):936–942.

Aickelin U, Dowsland KA (2004) An indirect genetic algorithm for a nurse-scheduling problem. *Comput. Oper. Res.* 31(5):761–778.

Aksin Z, Armony M, Mehrotra V (2007) The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* 16(6):665–688.

Atlason J, Epelman MA, Henderson SG (2004) Call center staffing with simulation and cutting plane methods. *Ann. Oper. Res.* 127(1):333–58.

Atlason J, Epelman MA, Henderson SG (2008) Optimizing call center staffing using simulation and analytic center cutting-plane methods. *Management Sci.* 54(2):295–309.

Avramidis AN, Chan W, Gendreau M, Lecuyer P, Pisacane O (2010) Optimizing daily agent scheduling in a multiskill call center. *Eur. J. Oper. Res.* 200(3):822–832.

Bard JF, Purnomo HW (2005) Preference scheduling for nurses using column generation. *Eur. J. Oper. Res.* 164(2):510–534.

Beliën J, Demeulemeester E (2008) A branch-and-price approach for integrating nurse and surgery scheduling. *Eur. J. Oper. Res.* 189(3):652–668.

Burke EK, Curtois T (2014) New approaches to nurse rostering benchmark instances. *Eur. J. Oper. Res.* 237(1):71–81.

Burke EK, Li J, Qu R (2010) A hybrid model of integer programming and variable neighbourhood search for highly-constrained nurse rostering problems. *Eur. J. Oper. Res.* 203(2):484–493.

Burke EK, Curtois T, Post G, Qu R, Veltman B (2008) A hybrid heuristic ordering and variable neighbourhood search for the nurse rostering problem. *Eur. J. Oper. Res.* 188(2):330–341.

Carter M, Lapierre S (2001) Scheduling emergency room physicians. *Health Care Management Sci.* 4(4):347–360.

Castillo I, Joro T, Li YY (2009) Workforce scheduling with multiple objectives. *Eur. J. Oper. Res.* 196(1):162–170.

Cheang B, Li H, Lim A, Rodrigues B (2003) Nurse rostering problems—A bibliographic survey. *Eur. J. Oper. Res.* 151(3): 447–460.

De Causmaecker P, Vanden Berghe G (2011) A categorisation of nurse rostering problems. *J. Scheduling* 14(1):3–16.

Defraeye M, Van Nieuwenhuyse I (2011) Setting staffingng levels in an emergency department: Opportunities and limitations of stationary queuing models. *Rev. Bus. Econom.* 56(1):73–100.

Defraeye M, Van Nieuwenhuyse I (2013) Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm. *Decision Support Systems* 54(4):1558–1567.

Defraeye M, Van Nieuwenhuyse I (2016a) A branch-and-bound algorithm for shift scheduling with stochastic nonstationary demand. *Comput. Oper. Res.* 65(January):149–162.

Defraeye M, Van Nieuwenhuyse I (2016b) Staffing and scheduling under nonstationary demand for service: A literature review. *Omega* 58(1):4–25.

Eick SG, Massey WA, Whitt W (1993) $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci.* 39(2):241–252.

Feldman Z, Mandelbaum A, Massey WA, Whitt W (2008) Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* 54(2):324–338.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Glass CA, Knight RA (2010) The nurse rostering problem: A critical appraisal of the problem structure. *Eur. J. Oper. Res.* 202(2): 379–389.

Grassmann WK (1977) Transient solutions in Markovian queueing systems. *Comput. Oper. Res.* 4(1):47–53.

Green LV, Kolesar PJ (1991) The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.* 37(1):84–97.

Green LV, Soares J (2007) Computing time-dependent waiting time probabilities in $M(t)/M/s(t)$ queuing systems. *Manufacturing Service Oper. Management* 9(1):54–61.

Green LV, Kolesar PJ, Soares J (2001) Improving the SIPP approach for staffing service systems that have cyclic demands. *Oper. Res.* 49(4):549–564.

Green LV, Kolesar PJ, Svoronos A (1991) Some effects of nonstationarity on multiserver Markovian queueing systems. *Oper. Res.* 39(3):502–511.

Green LV, Kolesar PJ, Whitt W (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management* 16(1):13–39.

Green LV, Soares J, Giglio JF, Green RA (2006) Using queueing theory to increase the effectiveness of emergency department provider staffing. *Acad. Emergency Medicine* 13(1):61–68.

Gutjahr WJ, Rauner MS (2007) An ACO algorithm for a dynamic regional nurse-scheduling proble1m in Austria. *Comput. Oper. Res.* 34(3):642–666.

Ikegami A, Niwa A (2003) A subproblem-centric model and approach to the nurse scheduling problem. *Math. Programming* 97(3):517–541.

Ingolfsson A, Haque MA, Umnikov A (2002) Accounting for time varying queueing effects in workforce scheduling. *Eur. J. Oper. Res.* 139(3):585–597.

Ingolfsson A, Akhmetshina E, Budge S, Li Y (2007) A survey and experimental comparison of service-level-approximation methods for nonstationary $M_t/M/s_t$ queueing systems with exhaustive discipline. *INFORMS J. Comput.* 19(2):201–214.

Ingolfsson A, Campello F, Wu XD, Cabral E (2010) Combining integer programming and the randomization method to schedule employees. *Eur. J. Oper. Res.* 202(1):153–163.

Izady N, Worthington DJ (2012) Setting staffing requirements for time-dependent queueing networks: The case of accident and emergency departments. *Eur. J. Oper. Res.* 2193):531–540.

Jacobson SH, Hall SN, Swisher JR (2006) Discrete-event simulation of health care systems. Hall RW, ed. *Patient Flow: Reducing Delay in Healthcare Delivery*, International Series in Operations Research and Management Science, Vol. 91 (Springer, New York), 211–252.

Jaumard B, Semet F, Vovor T (1998) A generalized linear programming model for nurse scheduling. *Eur. J. Oper. Res.* 107(1):1–18.

Jennings OB, Mandelbaum A, Massey WA, Whitt W (1996) Server staffing to meet time-varying demand. *Management Sci.* 42(10): 1383–1394.

Jun JB, Jacobson SH, Swisher JR (1999) Application of discrete-event simulation in health care clinics: A survey. *J. Oper. Res. Soc.* 50(2):109–123.

Kennedy J, Rhodes K, Walls CA, Asplin BR (2004) Access to emergency care: Restricted by long waiting times and cost and coverage concerns. *Ann. Emergency Medicine* 43(5):567–573.

Kleinrock L (1974) *Queueing Systems, Volume 1: Theory* (John Wiley & Sons, New York).

Koole G, van der Sluis E (2003) Optimal shift scheduling with a global service level constraint. *IIE Trans.* 35(11):1049–1055.

Liu Y, Whitt W (2012) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* 60(6): 1551–1564.

Massey WA, Whitt W (1997) Peak congestion in multi-server service systems with slowly varying arrival rates. *Queueing Systems* 25(1–4):157–172.

Moz M, Pato MV (2007) A genetic algorithm approach to a nurse rerostering problem. *Comput. Oper. Res.* 34(3):667–691.

Nah JE, Kim S (2013) Workforce planning and deployment for a hospital reservation call center with abandonment cost and multiple tasks. *Comput. Indust. Engrg.* 65(2):297–309.

Puente J, Gómez A, Fernández I, Priore P (2009) Medical doctor rostering problem in a hospital emergency department by means of genetic algorithms. *Comput. Indust. Engrg.* 56(4):1232–1242.

Robbins TR, Harrison TP (2010) A stochastic programming model for scheduling call centers with global service level agreements. *Eur. J. Oper. Res.* 207(3):1608–1617.

Sinreich D, Jabali O (2007) Staggered work shifts: A way to downsize and restructure an emergency department workforce yet maintain current operational performance. *Health Care Management Sci.* 103):293–308.

Whitt W (1991) The pointwise stationary approximation for $Mt/Mt/s$. *Management Sci.* 37(3):307–14.

Whitt W (2002) Solutions for the Erlang B and C formulas. IEOR 6707 Homework 1e, Columbia University, New York. http://www.columbia.edu/~ww2040/ErlangBandCFormulas.pdf.

Whitt W (2004a) A diffusion approximation for the $G/GI/n/m$ queue. *Oper. Res.* 52(6):922–941.

Whitt W (2004b) Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* 50(10):1449–1461.

Whitt W (2006) Fluid models for multiserver queues with abandonments. *Oper. Res.* 54(1):37–54.

Whitt W (2007) What you should know about queueing models to set staffing requirements in service systems. *Naval Res. Logist.* 54(5):476–484.

Yeh JY, Lin WS (2007) Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department. *Expert Systems Appl.* 32(4):1073–1083.

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

Zeltyn S, Marmor YN, Mandelbaum A, Carmeli B, Greenshpan O, Mesika Y, Wasserkrug S, et al. (2011) Simulation-based models of emergency departments: Operational, tactical, and strategic staffing. *ACM Trans. Modeling Comput. Simulation* 21(4): 1–21.