Setting Staffing Requirements for Time Dependent Queueing Networks: The Case of Accident and Emergency Departments

Navid Izady, Dave Worthington

Management Science Department, Lancaster University, LA1 4YX *

April 7, 2011

Abstract

An incentive scheme aimed at reducing patients' waiting times in accident and emergency departments was introduced by the UK government in 2000. It requires 98 percent of patients to be discharged, transferred, or admitted to inpatient care within 4 hours of arrival. Setting the minimal hour by hour medical staffing levels for achieving the government target, in the presence of complexities like time-varying demand, multiple types of patients, and resource sharing, is the subject of this paper. Building on extensive body of research on time dependent queues, we propose an iterative scheme which uses infinite server networks, the square root staffing law, and simulation to come up with a good solution. The implementation of this algorithm in a typical A&E department suggests that significant improvement on the target can be gained, even without increase in total staff hours.

1 Introduction

Accident and Emergency (A&E) departments are facing increasing pressure to improve the quality of care. 'Timeliness' is one important aspect of emergency care quality and has particularly been reflected in a waiting time target set by the National Health System (NHS) for A&Es across the UK. This target requires 98 percent of A&E patients to have their service completed, i.e. be discharged, transferred, or admitted to inpatient care, within four hours of arrival (see Mayhew and Smith 2008 for an historical account). Since its introduction in 2000, much has been said and written

*n.izady@lancaster.ac.uk

on the requirements and implications of the 4-hour target (see, for example, Munro et al. 2006, Mortimore and Cooper 2007, and references therein). However, one subject appears to have drawn less attention and that is the staffing requirements for achieving the target.

Government collected data shows most A&E departments are making the 4-hour target, and the very few remaining ones are close to making it (Department of Health Statistics 2010). However, increasing numbers of patients visiting A&Es every year (total attendance has risen at an average annual rate of 3.6 percent during the last 12 years according to Department of Health Annual Report 2009), and evidence of some A&Es taking special actions to avoid breaching the target (Gunal and Pidd 2009), suggests that reducing emergency care delays is still a high priority. Staffing algorithms seem appealing in this respect as they enable A&E managers to better match capacity of their resources to patients' needs and thus reduce waits.

All A&E departments exhibit time-dependent behavior; that is, the variation of the patients' arrival rate (the mean number of arrivals per unit time) by time of day and, sometimes, by day of week. The volume of patients is also likely to change in different seasons. In response to these variations in demand, staffing levels are generally varied over the course of a day. Our purpose here is to determine the minimal hour-by-hour levels of medical staff — doctors, emergency nurse practitioners (ENPs), ECG technicians, lab technicians, radiologists, and nurses — needed to meet the 4-hour target.

Setting staffing requirements of a service system with time varying demand is a major challenge; traditional queueing theory formulae cannot be directly applied to this type of system as the parameters (mainly the arrival rates) do not remain constant long enough for the system to settle down to steady state. It has attracted a significant body of research over the last two decades, and a few approaches have been developed as a result; see Green et al. (2007) and Whitt (2007) for a comprehensive review. However, most of the research sofar has concentrated on single service systems, specifically on call centers. Staffing A&E departments is far more complicated due to the 'network nature' of their services; upon arrival to an A&E, patients go through various care processes undertaken by professionals with various skills. There also exist 'multiple' types of patients, each having different resource requirements and pathways through the network. Moreover, some resources are shared among some processes in the network. A staffing algorithm needs to take all these complexities into account.

Vassilacopoulos (1985) used a deterministic model for allocating physicians to weekly shifts in an A&E department. They set physician levels proportional to the hourly mean arrival rates. Coats and Michalis (2001) used simulation to compare two different shift patterns with the existing one in an A&E department. Green et al. (2006) modeled a local emergency department in the US as a single station queueing system and used a *Lagged Stationary Independent Period by Period* (Lag SIPP) approach to determine physicians staffing. Implementation of their suggested physician levels led to a significant improvement in the proportions of patients who *left without being seen* (LWBS).

Sinreich and Jabali (2007) suggested a heuristic algorithm for downsizing emergency departments while maintaining patients' *length of stay* (LOS). Their heuristic method combines a simulation model with a linear programming model in an iterative manner to produce shift schedules for doctors, nurses, and image technicians. It schedules 'one resource' at each stage, where that resource is chosen by a 'delay factor' estimated by the simulation model. Their simulation results indicate that a significant reduction in working capacity can be obtained without causing a statistically significant impact on the patients' LOS.

The 4-hour service quality target concerns the total time a patient spends in the system (the *sojourn time*). It is different from prevalent targets used in other service systems. In call centers, for example, the service quality target is stated in terms of callers' waiting times, i.e. x percent of calls must be responded to within y seconds. Green et al. (2006) used a similar target for their physician staffing model called the time to *first encounter with a doctor* (FED). But, as they have pointed out, FED data are not usually collected by hospital IT systems. On the contrary, the sojourn time data (the patients' arrival and departure times) are registered in most emergency departments and reflects all waits in the system. It is also the performance target set in England.

However, we cannot directly staff the system for the 98 percent sojourn time target nor can we even evaluate the sojourn time distribution under a specific set of staffing levels using analytical or numerical methods. Building on the extensive body of research on time dependent queues, we propose a heuristic iterative approach which combines non-stationary infinite server networks, the square root staffing law, and simulation. In particular, it uses infinite server networks to calculate the time dependent workload imposed on each type of medical staff; the square root staffing law to find the required staffing levels according to a prescribed delay probability, and simulation to

translate the delay probability to sojourn time distribution. The algorithm seeks to stabilize the quality of services delivered by all emergency care providers at all times. It addresses a mid-term planning horizon, during which the total volume of patients is assumed to remain approximately constant.

We tested our staffing algorithm on a 'typical' A&E department in the UK. We then used a modified version of the S-model of Sinreich and Jabali (2007) to produce shift schedules (shifts start times, durations, and number of employees assigned) that match the proposed staffing levels as close as possible. The method developed here is built on a generic A&E conceptual and simulation model and therefore has the potential to be applied in many A&E departments across UK. Given appropriate modifications in the simulation model, it might well be applied to any other emergency department.

It is worth noting that insufficient inpatient beds is frequently cited as the major reason for breaches of the 4-hour target (Cooke et al. 2004). However, based on our experiments here and evidence from non-English emergency departments cited above, we believe that appropriate staffing of A&E workforce can help reduce patients' sojourn time significantly. In comparison to the cost of acquiring and maintaining inpatient beds, this is likely to be a much more cost-effective solution.

The paper is organized as follows. A generic A&E model is discussed in Section 2. The staffing algorithm is explained in Section 3, and is implemented to a typical A&E department in Section 4. This is followed by shift scheduling and conclusions in Sections 5 and 7.

2 A Generic A&E Department Model

Many simulation models of emergency departments have been built. A few of these models, like Sinreich and Marmor (2005), Gunal and Pidd (2006), and Fletcher et al. (2006), are generic. The generic simulation model of Fletcher et al. (2006) was built for the UK Department of Health to inform policy makers of the barriers in implementing the 4-hour target, and was also used to aid local hospitals in improving their emergency services. We have based our study on an updated version of this model.

Three types of patients, minor, major, and admitted, and six types of medical staff are considered in this model. The process charts are depicted in Figure 1 for the minor type of patients and in Figure 2 for the major and admitted types. It identifies the pathway each patient type goes

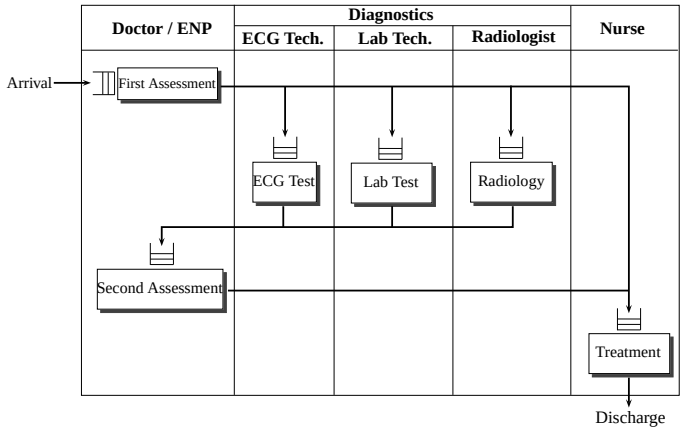

Figure 1: Process chart of minor patients.

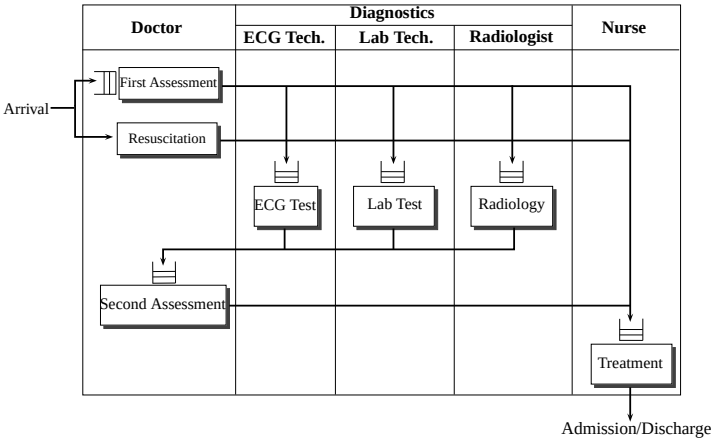

Figure 2: Process chart of major and admitted patients.

through and the corresponding resources. Notice that clerical tasks and related staff are not built into this model as their impact on the total sojourn time is deemed negligible.

Admitted and major patients are first assessed by a doctor. They may require some diagnostic tests — ECG test, lab test, and/or radiology — in which case the results must be interpreted again by a doctor. This is followed by some sort of treatment performed by a nurse. Then a decision as to admit the patient to a hospital ward or to discharge him/her will be made. Admitted type of patients are assumed to be admitted, while major patients are discharged home.

There always exists a very small proportion of admitted and major patients who arrive with severe conditions and require immediate care. They are first put in a resuscitation room, where at

least two doctors (or more if available) are called to perform life-saving actions. Then they proceed along the normal admitted/major patient route. Minor patients have a similar route through the network, but their first and second assessments can be done by either a doctor or by an ENP. All minor patients are discharged home.

As illustrated in the process charts, a queue is formed in front of each task except for the resuscitation. Since patients placed in this unit need to be dealt with immediately, two doctors would have to interrupt their tasks temporarily if all of them are busy when a patient arrives. Diagnostics are assumed to be dedicated to the A&E department. The workload coming from other hospital wards must be counted if it was not the case (see Sinreich and Marmor 2005 on modeling arrivals to diagnostics). Admitted patients would sometimes have to wait until an inpatient bed becomes available (the so-called 'trolley wait'). We have not considered trolley waits in our model as it is outside the control of A&E department. Nevertheless, it prolongs the patients' LOS.

3 A Heuristic Staffing Algorithm

Our objective here is to determine the required number of each type of medical staff — doctors, ENPs, ECG technicians, lab technicians, radiologists, and nurses — during each 'staffing interval' so that the 4-hour sojourn time target is met. 'Staffing interval' refers to the period during which the number of staff remains constant. It might be one or two hours in emergency departments.

One simple idea is to allocate resources per staffing interval approximately proportionate to the corresponding average arrival rate. This approach, as used by Vassilacopoulos (1985) for staffing A&E physicians, matches the capacity with the arrival rate but not with the actual workloads. Experiments with non-stationary single service queues show that the actual congestion levels experienced by customers typically lag behind the arrival rates. The congestion peak time, for example, occurs sometime after the arrival rate peaks. An estimate of this time lag is the mean service time (see Green and Kolesar 1997 and Massey and Whitt 1997 for more details). The time lag is expected to increase in networks of services (like the A&E network) as some services (like the second assessment or treatment in the A&E model) might be delivered a long time after patients' arrival to the system. The heuristic algorithm we propose here employs queueing models to estimate the size and timing of the workloads imposed on all types of resources in the system and so provides a better matching.

We build on Jennings et al. (1996) staffing method and expand it for networks. They proposed a method for staffing a single service queueing system for achieving a relatively stable service quality at all times. Sensitivity of the services provided in emergency care and the fact that patients (some with life-threatening conditions) may arrive at any time, day or night, amplifies the significance of providing a consistently high service level in an A&E department. Hence, we set staffing levels so as to maintain the quality of services given by all types of resources approximately at some constant level at all times. The measure of service quality underlying this statement is the 'probability of delay', i.e. the probability of a customer having to wait before beginning service.

We chose to work with delay probability as research on non-stationary single service queues suggests that achieving a time-stable delay probability is possible. Moreover, with a stable delay probability, other performance metrics, such as utilization, average queue length, and average waiting time, show some time-stability as well (Jennings et al. 1996, Feldman et al. 2008). Having staffed the A&E network for achieving a chosen delay probability, simulation is used to estimate the percentage discharged during 4 hours. If the 4-hour target is not met or is over achieved, the above process is repeated for a lower or higher delay probability until the target is satisfied.

Below we first review the staffing method of Jennings et al. (1996) for achieving a target delay probability and then extend it to networks of services. These approaches are then combined in a heuristic iterative algorithm in the final section for achieving the 4-hour completion time target.

3.1 Staffing Single Service Queues

Consider an $M_t/G/s(t)$ queueing system in which the arrival process is a non-homogeneous Poisson process (the M_t) with deterministic time dependent arrival rate function $\{\lambda(t), t \geq 0\}$, service times follow a general distribution (the G) with cumulative distribution function $G(x)$, and number of servers is a time dependent function $s(t)$. In order to achieve the target delay probability of α, Jennings et al. (1996) proposed using the *square root staffing law* as follows

$$s(t) = \lceil m_\infty(t) + \beta\sqrt{m_\infty(t)} \rceil, \tag{1}$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to x, $m_\infty(t)$ is the *time dependent offered load*, and β is a quality of service (QoS) parameter. The time dependent offered load function $m_\infty(t)$ is a measure of the workload in the system at any time t, estimated by the mean number of busy servers in the corresponding $M_t/G/\infty$ queue, with the same arrival and service processes as

the original system but with infinitely many servers. This estimation is motivated by the fact that the offered load in stationary queues coincides with mean busy servers in the related stationary infinite server queue. It also allows for the time lag existing between congestion levels and the arrival rate. See Feldman et al. (2008) for a detailed discussion. Mean busy servers in an $M_t/G/\infty$ queue is computed as follows (Eick et al. 1993)

$$m_\infty(t) = \int_0^t \lambda(u) G^c(t-u) du, \tag{2}$$

where $G^c(t) = 1 - G(t)$. The QoS parameter β is chosen according to the targeted delay probability α. Based on a heavy traffic limit theorem, Halfin and Whitt (1981) established the following relation between β and α

$$\alpha = \left[1 + \beta \frac{\Phi(\beta)}{\phi(\beta)}\right]^{-1}, \tag{3}$$

where ϕ and Φ are, respectively, the density function and the cdf of the standard normal distribution. Since the number of servers needs to remain constant during each staffing interval, Jennings et al. (1996) proposed using the maximum offered load over each interval in equation (1) to ensure maintaining the targeted service level at all times.

For an $M_t/M/s(t)$ system (with Exponential service times), Halfin-Whitt limiting regime suggests that setting the staffing function according to (1) results in the following distribution for the virtual waiting time W_t of a fictitious customer arriving at time t (Whitt 1992)

$$\Pr(W_t > x) = \Pr(W_t > 0)\, e^{-\beta\mu x\sqrt{s(t)}}, \tag{4}$$

and so,

$$E[W_t] = \frac{\Pr(W_t > 0)}{\beta\mu\sqrt{s(t)}}, \tag{5}$$

where $\Pr(W_t > 0)$ is the delay probability at t (replaced with α in (3)), and $1/\mu$ is the mean service time. The above equations suggest that virtual waiting time distribution depends upon number of servers, mean service time, and delay probability. Hence, even if delay probability is stable, the variations in $s(t)$ coming from using the square root staffing law with a fixed value for the QoS parameter at all times will cause oscillations in the waiting times.

3.2 Extension to Networks

A natural generalization of the square root staffing law to non-stationary networks is as follows. Consider an $(M_t/G/s_k(t))^K/M$ network with K service stations indexed by $k = 1, 2, \cdots, K$. The

final M represents a stationary Markovian routing process (fixed probabilities assigned to different routes) through the network. For each time t, $s_k(t)$ units of resource type k are assumed to serve in station k. Note that each resource type here is assumed to be merely responsible for one service station (no resource pooling). We relax this constraint in the next section.

We want to determine the set of staffing functions $\{s_k(t), k = 1, \cdots, K\}$ so as to achieve a target delay probability α across all service stations at all times. To do so, we use the square root staffing function

$$s_k(t) = \lceil m_\infty^k(t) + \beta\sqrt{m_\infty^k(t)} \rceil, \tag{6}$$

where $m_\infty^k(t)$ is the time dependent offered load function of service station (or resource type) k. In line with Green et al. (2007), we propose estimating $m_\infty^k(t)$ by the mean number of busy servers of resource type k in the associated $(M_t/G/\infty)^K/M$ network, with the same arrival, service, and routing processes as the original network, but with infinitely many servers of all K types of resources. Infinite server networks are analytically tractable, and the required equations for computing mean busy servers are given in Theorem 1.2 of Massey and Whitt (1993).

In order to compute mean busy servers of the uncapacitated network, we decided to decompose the $(M_t/GI/\infty)^K/M$ network with stochastic routes into a number of $(M_t/GI/\infty)^{K_1}/D$ tandem networks, in which a series of K_1 stations are visited successively (the D means deterministic routing). Note that K_1 can be greater than K, depending on the number of feedbacks in the original networks. As pointed out by Massey and Whitt (1993), it is quite natural to think of any network with stochastic routes as a multi-class network, where each class of customers follows a deterministic path in the network, and its external arrival rate function is the original arrival rate multiplied by the probability that customers take that specific path. Adopting such a decomposition approach makes infinite server networks analysis easier for two reasons. First, computing mean busy servers in an uncapacitated tandem network is straight forward knowing that departure times of an $M_t/G/\infty$ queue form a non-homogeneous Poisson process with the following rate function (Eick et al. 1993)

$$d(t) = \int_0^\infty \lambda(t-u)dG(u), \tag{7}$$

where $\lambda(t)$ is the arrival rate. Then, as departures from one station would be arrivals to the next one, mean busy servers of each station can be easily computed with the successive use of Equations (2) and (7). Second, as customers do not interact with each other in infinite server networks, adding

up the mean busy servers of each station across all tandem models yields the mean busy servers of that station in the original $(M_t/GI/\infty)^K/M$ network.

A spreadsheet program for computing the offered load of a network with time-varying demand and multiple types of customers has been developed using the decomposition method described above. It can be obtained from the authors upon request.

Based on single service results, we expect staffing a network using (6) with a common value for the QoS parameter β for all service stations (resource types), results in relatively time-stable delay probabilities at all stations. However, the instability of waiting time distributions over time (suggested by (4) and (5)) raises the important question of whether the sojourn time related measures become time-stable or not. This will be investigated empirically in Section 4.

3.3 Staffing A&E Department

The process charts of a typical A&E department in Figures 1 and 2 characterize a network with seven stations and six types of resources providing service for three types of patients. The basic algorithm for staffing A&E department for achieving the 4-hour target is outlined below.

Step 1. Set $\alpha = 1.0$, and calculate the maximum offered load imposed on each type of resource during each staffing interval using the method described in the previous section.

Step 2. Find the value of the QoS parameter β satisfying Equation (3).

Step 3. For each type of resource, use the square root staffing function (6) to obtain the staffing levels of all staffing intervals.

Step 4. Given the resulting staffing levels, estimate the percentage discharged within 4 hours using simulation.

step 5. If the percentage discharged is less (more) than 98 percent, decrease (increase) α and go back to Step 2; otherwise stop and return the staffing functions.

However, in this particular case, there is some pooling of resources; doctors are pooled among the first assessment, resuscitation and second assessment, and ENPs are shared between the first and second assessments. Moreover, doctors are able to handle all types of patients, while ENPs can deal only with minor patients (the so-called 'skill-based routing'). These two phenomena must

be accounted for in setting staffing requirements. For step 1 of the above algorithm, we assume a different type of resource is allocated to each of the seven stations of the A&E network (no pooling). Minor patients are also assumed to be dealt with only by ENPs. This latter assumption reflects the higher cost associated with doctors and so the preference of assigning them to only major and admitted types of patients. Having calculated the offered load of all seven stations for the three types of patients as described in Section 3.2, we add up the offered load of the first assessment, resuscitation, and second assessment of major and admitted patient types to obtain the workload of doctors, denoted by $m_{\infty}^{\mathrm{D}}(t)$, and add up the offered load of the first and second assessments of minor patients to obtain the workload of ENPs, denoted by $m_{\infty}^{\mathrm{E}}(t)$. The total workload of doctors and nurses together therefore would be $m_{\infty}^{\mathrm{D\&E}}(t) = m_{\infty}^{\mathrm{D}}(t) + m_{\infty}^{\mathrm{E}}(t)$. The workload of other resources is obtained in the normal way by adding up the offered load of the corresponding stations for all types of patients.

In step 3, the total staffing of doctors and ENPs is collectively set by using $m_{\infty}^{\mathrm{D\&E}}(t)$ in the square root rule. Subtracting the required number of doctors, obtained by using $m_{\infty}^{\mathrm{D}}(t)$ in the square root law, yields the ENP staffing levels. This method ensures that doctors, as the main resource of the system, are sufficiently assigned to fulfill the prescribed service quality. In addition, the combined resource of doctors plus ENPs should be sufficient to achieve the prescribed service level for their joint workload. This is validated by our experiments in the next section. For more complicated skill based staffing methods, see Wallace and Whitt (2005) and references therein.

4 Case Study

To test our approach, we apply it for a 'typical' UK A&E department, based on information extracted from a 7-day survey undertaken in 12 hospital trusts by Fletcher et al. (2006). Based on the survey data, minor, major, and admitted patients constitute (on average) 57, 19, and 24 percents of total attendees. Demand profiles (percentage of attendees of each patient type during each hour) were extracted from survey data. Hourly arrival rates and its breakdown to patient types are illustrated in Figure 3, assuming a total annual attendance of 87,000 patients (an average sized A&E in the UK). A day of week effect was not observed in the survey data. Average service times were estimated by A&E experts and for the purpose of this demonstration were assumed to be Exponential. Percentage of patients of each type requiring diagnostic tests were acquired from

local sources.

To demonstrate the advantage of stabilizing performance using our algorithm, we compare results with a 'baseline' staffing profile based on three simple 8-hour shifts, with the staffing levels allocated pro-rata to an 'expected' workload during each 8-hour period. The expected workload of each resource is simply calculated as the arrival rate of patients requiring that resource multiplied by the average service time taken from that resource. The pro-rata factor simply depends on the average utilization level we choose for each staff type. Based on the Fletcher et al. (2006) simulation model, we set average utilization levels of resources as in Table 1.

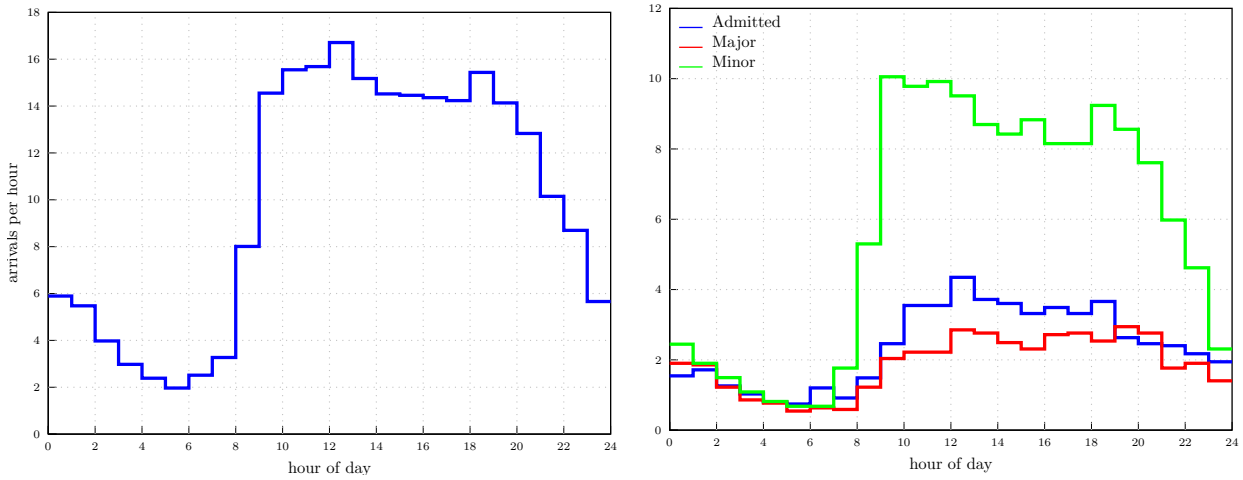

Figure 3: Daily arrival rates to the A&E department (left), and its breakdown to patient types (right).

Table 1: Baseline profiles target utilization levels

Resource Type	Doctors	ENPs	ECG Tech.	Lab Tech.	Radiologists	Nurses
Target Utilization	0.80	0.55	0.40	0.40	0.50	0.65

Now we use the algorithm proposed in Section 3.3 to construct staffing profiles of all resources for achieving the 4-hour target. Having calculated the offered load in Step 1 of the algorithm, we iterated steps 2, 3, 4, and 5 until the 98 percent target was hit after 6 iterations. The percentage discharged over 4 hours, estimated by simulating the system for 100 weeks, versus the probability of delay is plotted in Figure 4. Each point in this figure corresponds to one iteration of the algorithm

(more points are plotted for illustrative purposes). It is clear from the plot that for achieving the 98 percent discharge target, each type of resource needs to be staffed according to a delay probability of 75 percent, which is equivalent to the QoS parameter $\beta = 0.221$. We refer to the resulting staffing profiles as the 'balanced' profiles.

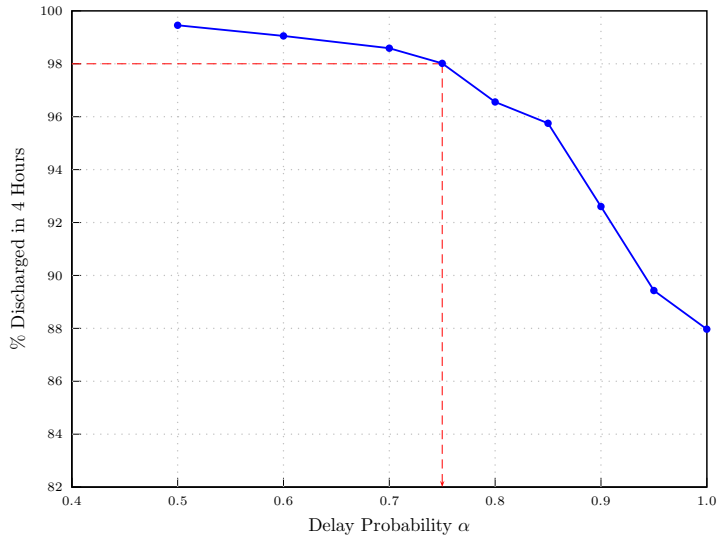

Figure 4: Percentage discharged in 4 hours vs. delay probability.

The balanced and baseline staffing profiles for all resource types are depicted in Figure 5. Running the simulation model with the baseline profiles results in 96 percent of patients discharged within 4 hours. Comparing the total resource hours of the optimal profiles with the baseline (given in Table 2) reveals that the 4-hour target can be achieved with the same total hours of doctors, 1h (3%) more of ECG technicians, and 5h (13%), 4h (13%), 4h (7%), and 8h (8%) less of ENPs, lab technicians, radiologists, and nurses, respectively. Hence, the performance is improved with the balanced profiles despite a total reduction of 20 staff hours (6%) in the size of workforce.

Table 2: Total staff hours (utilizations) of resources with baseline and balanced profiles

Resource Type	Doctors	ENPs	ECG Tech.	Lab Tech.	Radiologists	Nurses
Baseline	72 (80%)	40 (56%)	32 (42%)	40 (39%)	56 (51%)	104 (65%)
Balanced	72 (74%)	35 (65%)	33 (39%)	36 (44%)	52 (54%)	96 (70%)

The average resource utilization levels in Table 2 are generally higher for the balanced profiles,

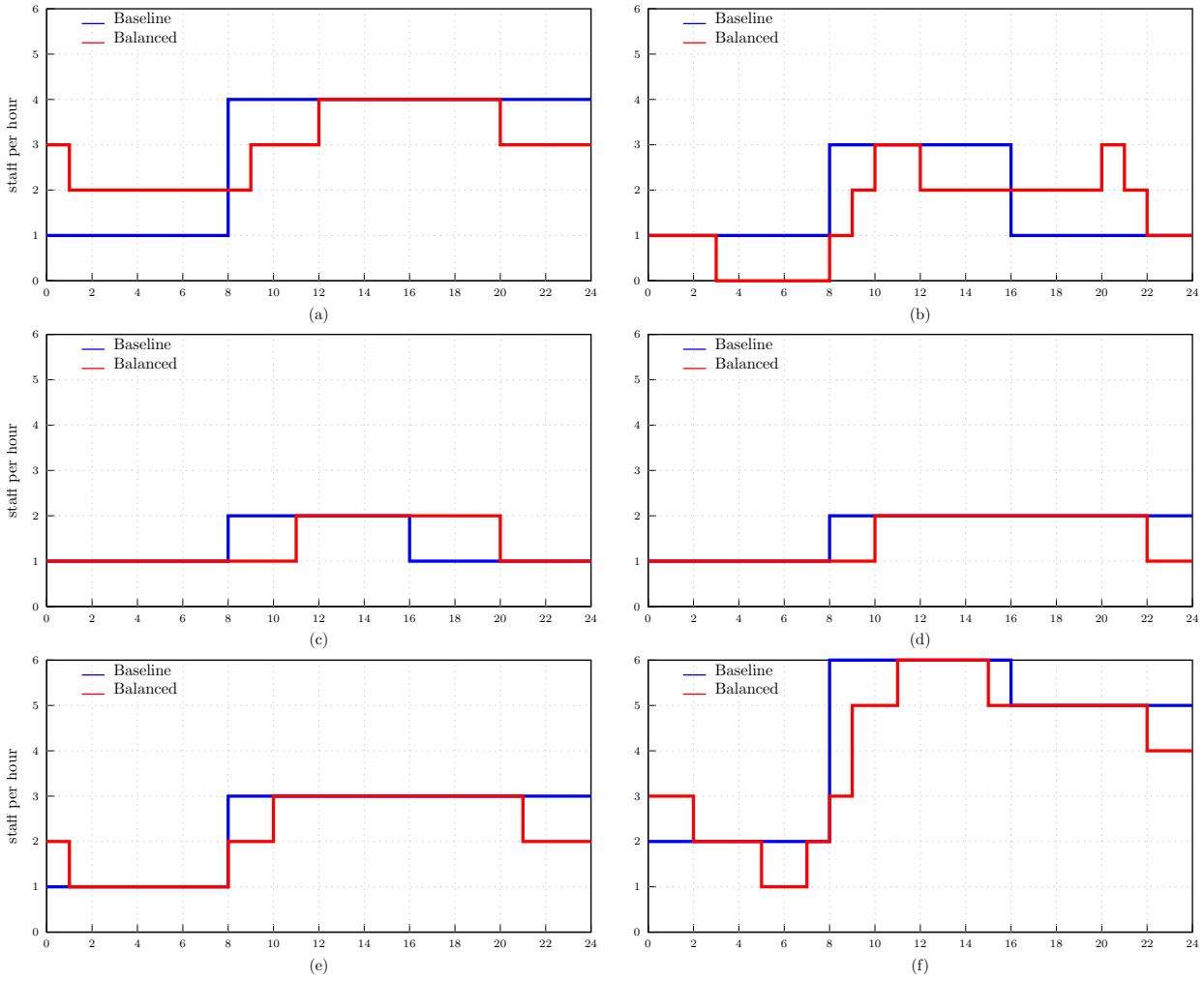

Figure 5: Baseline and balanced staffing levels for : (a) doctors, (b) ENPs, (c) ECG technicians, (d) lab technicians, (e) radiologists, and (f) nurses.

consistent with the overall reduction in staff hours. However, more important in this context is the time dependent analysis of utilization. For example, Figure 6 reveals that doctor utilization is always between 50 and 90 percent with the balanced profile, whereas it exceeds this range for a significant period of time with the baseline profile. With the baseline profile, ENP's utilization levels drop below 30 percent during early hours of the morning. This problem has been overcome in the balanced profile by removing ENPs from those hours. The range of utilization levels with the balanced profiles for ECG-technicians, lab technicians, radiologists, and nurses are 20-60, 25-60, 30-65, and 40-85 percent, respectively, whereas the corresponding levels with the baseline profiles are 10-70, 15-50, 20-65, and 20-80 percent. Overall, a more time-stable utilization is observed for all types of resources with the balanced profiles.

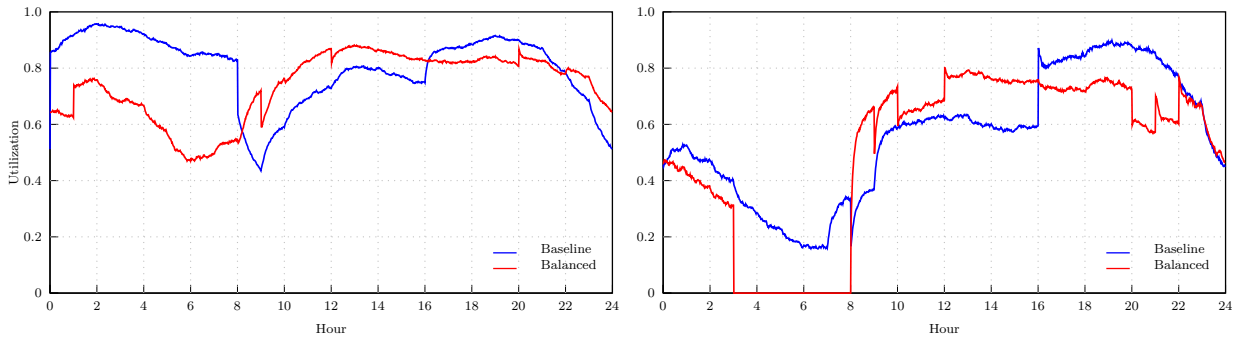

Figure 6: Utilization of doctors (left) and ENPs (right).

The average waiting times and delay probabilities of patients under both baseline and balanced scenarios are given in Table 3, along with the long-term proportions of patients going through each clinical task. Major improvements in waiting times are observed in the first and second assessments. Notice that higher average delay probabilities associated with these two tasks under the balanced scenario indicates larger proportions of patients would have to wait to receive these services. However, because their waiting times are spread more evenly across time, the average waits decline by almost 40 percent in both tasks compared to the baseline scenario. The even performance of the system under balanced profiles is clearly demonstrated in Figure 7, which plots average waiting times and delay probabilities of patients for the first and second assessments against their arrival times to the queue. Similar plots, not included here, show stability of average queueing times and delay probabilities for all the other.

Table 3: Average waiting times and delay probabilities of patients

Measure	Scenario	First Assess.	ECG Test	Lab Test	Radiology	Second Assess.	Treatment
Proportion	—	99%	16%	27%	29%	55%	100%
Waiting Time (min)	Baseline	12.70	17.36	4.23	7.83	13.32	6.95
	Balanced	7.76	9.55	6.87	10.11	8.24	10.1
Delay Probability	Baseline	49%	40%	28%	32%	41%	53%
	Balanced	51%	34%	35%	38%	50%	54%

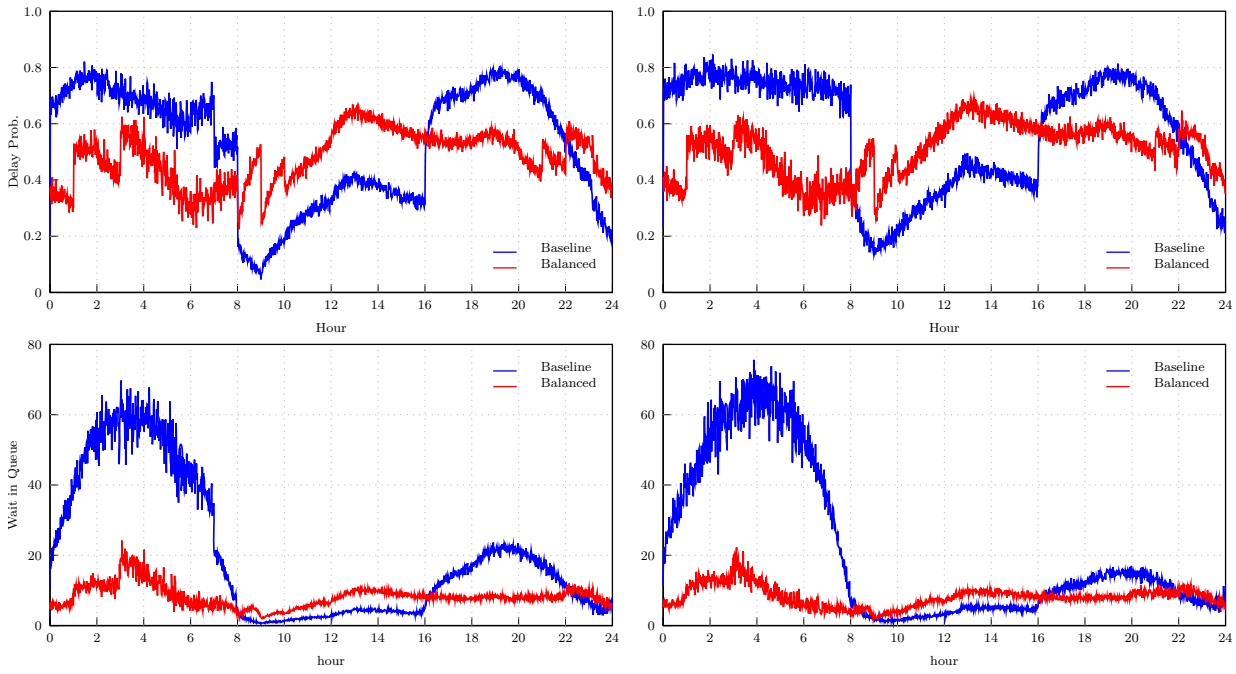

Figure 7: Average delay probabilities (top) and queueing times (below) for the first (left) and second (right) assessments.

Mean sojourn times and percentage of patients staying in the system more than 4 hours are plotted versus arrival times in Figure 8, which again show a high level of time-stability with the balanced profiles. We conclude that staffing a queueing network using the square root staffing law, with offered load values computed from the associated infinite server networks and a common value for the QoS parameter for all resources, evens out not only the individual delay probabilities, but also other performance metrics, especially the sojourn time related measures, over time.

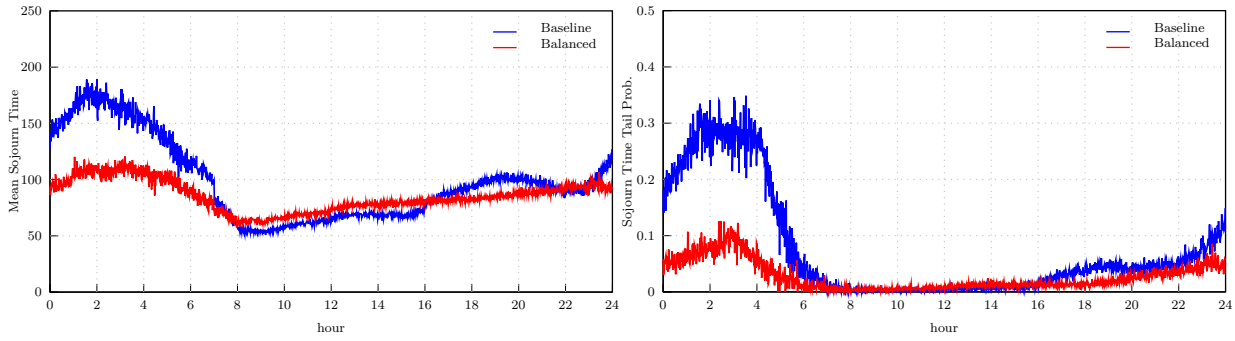

Figure 8: Mean sojourn time (left) and percentage discharged after four hours (right).

Though using a common value for the QoS parameter for staffing all resources simplifies the search process, it is not necessary. In fact, it can remain at the discretion of the management to impose different service levels on different resources depending on their availability, relative cost, and/or clinical priorities. The algorithm we proposed here produces a feasible starting point (in terms of the 4-hour target), based on which other alternative profiles can be developed by dictating higher service levels for some resources and lower for some others. In our case study , for example, suppose we need to reduce the total doctor hours in the system. This reduction must be compensated for with some extra hours of another resource in order to maintain the performance at the 98 percent discharge level. In our example, reducing the QoS parameter of doctors from 0.221 to 0.202 and keeping everything else unchanged result in 69 doctor hours and 38 ENP hours, which compared to the balanced profiles has 3 doctor hours less and 3 ENP hours more. Alternatively, we can keep the ENPs profile unchanged and add 4 nurse hours instead to compensate for the reduction in doctor hours. This can be achieved by increasing the nurses' QoS parameter from 0.221 to 0.305.

5 Shift Scheduling

Having set the ideal staffing requirements, a major concern is to produce shift schedules that comply with the proposed staffing levels and legal constraints. In particular, it may be that feasible shift schedules would push actual staffing levels much closer to our baseline staffing profiles. We therefore investigate this concern by showing how shift scheduling in this context can be addresses. In line with most research in this area (Green et al. 2001, Sinreich and Jabali 2007, Ingolfsson et al. 2010), we use an integer programming approach to schedule shifts with least deviations from the proposed staffing levels. We have modified the S-model of Sinreich and Jabali (2007) to produce shift patterns for each resource type as follows.

Let $(s_0, s_1, \ldots, s_{23})$ denote the set of staffing levels set by the staffing algorithm. Let I denote the set of permissible shift patterns constructed in compliance with legal constraints. Each shift pattern $i \in I$ is represented by a binary vector $p_i \equiv (p_{ij}, j = 0, 1, \ldots, 23)$, where p_{ij} equals 1 if shift i includes hour j as a working period and 0 otherwise. Let $x_i, i \in I$ be the decision variable denoting the number of employees scheduled to work on shift i. Hence, $\sum_{i \in I} p_{ij} x_i$ represents the total number of employees working at hour j for $j = 0, \ldots, 23$. We assume penalty costs of P^o and P^u are associated, respectively, with each resource hour over-staffing and under-staffing. The integer programming model is as below.

$$\min \left[P^o \sum_{j=0}^{23} \Delta_j^+ + P^u \sum_{j=0}^{23} \Delta_j^- \right] \tag{8}$$

st.

$$\sum_{i \in I} p_{ij} x_i = a_j, \qquad j = 0, \ldots, 23, \tag{9}$$

$$a_j - s_j = \Delta_j^+ - \Delta_j^-, \qquad j = 0, \ldots, 23, \tag{10}$$

$$\sum_{i \in I} y_i \leq k, \tag{11}$$

$$x_i \leq M y_i, \qquad i \in I, \tag{12}$$

$$x_i \geq 0, x_i \in \text{integers, and } y_i = 0, 1, \qquad i \in I, \tag{13}$$

$$\Delta_j^+, \Delta_j^- \geq 0, \qquad j = 0, \ldots, 23, \tag{14}$$

where Δ_j^+ and Δ_j^- denote, respectively, over-staffing and under-staffing at hour j, M is a large

number, and k is the maximum number of allocated shifts. The total number of allocated shifts in our baseline profiles of Section 4 was three. In order to keep the total number of shifts to which employees are assigned by the integer program less than a specified threshold, we have included constraint (11). This constraint uses dummy variable y_i, which is set to 1 (by constraint 12) if at least one employee works on shift i for $i \in I$.

We applied the above optimization model to the balanced staffing profiles of our case study (given in figure 5), assuming 7, 8, 9, and 10 hour long shifts. We set the maximum number k of allocated shifts to 4, over-staffing penalty $P^o = 1$, and under-staffing penalty $P^u = 2$. The resulting shift schedules are given in Table 4. The staffing requirements of doctors, ECG technicians, and lab technicians are exactly followed by these allocated shifts. The new profiles of ENPs, radiologists, and nurses are given in Figure 9, which have 2 radiologist hours and 3 nurse hours more than the balanced profiles. The total saving compared to the baseline profiles would be 15 hours. Hence, we see that, as expected, the constraints of shift scheduling have reduced the net staff savings from the original 20 hours. However, it is still possible to save 15 hours per day with only 4 shift patterns for each staff type.

6 Discussion

In this section, we discuss the possibility of applying the algorithm for other emergency departments and address some practical issues that might arise when implementing the reuslts. The method we have proposed here is fairly general and, having made the necessary modifications to the conceptual and simulation model, can be applied to a wide range of emergency departments.

Apart from hour of day effect that we considered in our modeling scheme, a day of week effect might also be observed in the arrival processes of patients to other emergency departments. In those situations, using hourly arrival rates of patients over a week, the iterative staffing algorithm produces weekly staffing profiles. The S-model presented above would then need to be changed to allow for considering different shifts in different days of a week. As the total volume of patients is also likely to change in different seasons, e.g. higher admissions in flu seasons, we recommend using the method once for every season where the forecast of arrival rates for that season has been obtained from the corresponding seasons in previous years. If a significant change in the arrival patterns is expected to happen during a season, the algorithm should be applied once again with

Table 4: Work shift schedules of A&E staff

Resource	Shift 1			Shift 2			Shift 3			Shift 4		
	Start	End	N.	Start	End	N.	Start	End	N.	Start	end	N.
Doctors	1:00	9:00	2	9:00	17:00	3	12:00	20:00	1	17:00	1:00	3
ENPs	8:00	15:00	2	15:00	22:00	2	20:00	3:00	1	-	-	-
ECG Tech.	3:00	12:00	1	11:00	18:00	1	12:00	20:00	1	18:00	3:00	1
Lab Tech.	6:00	15:00	1	10:00	20:00	1	15:00	22:00	1	20:00	6:00	1
Radiologists	1:00	11:00	1	8:00	18:00	2	11:00	21:00	1	18:00	1:00	2
Nurses	2:00	9:00	2	9:00	16:00	5	12:00	22:00	1	16:00	2:00	4

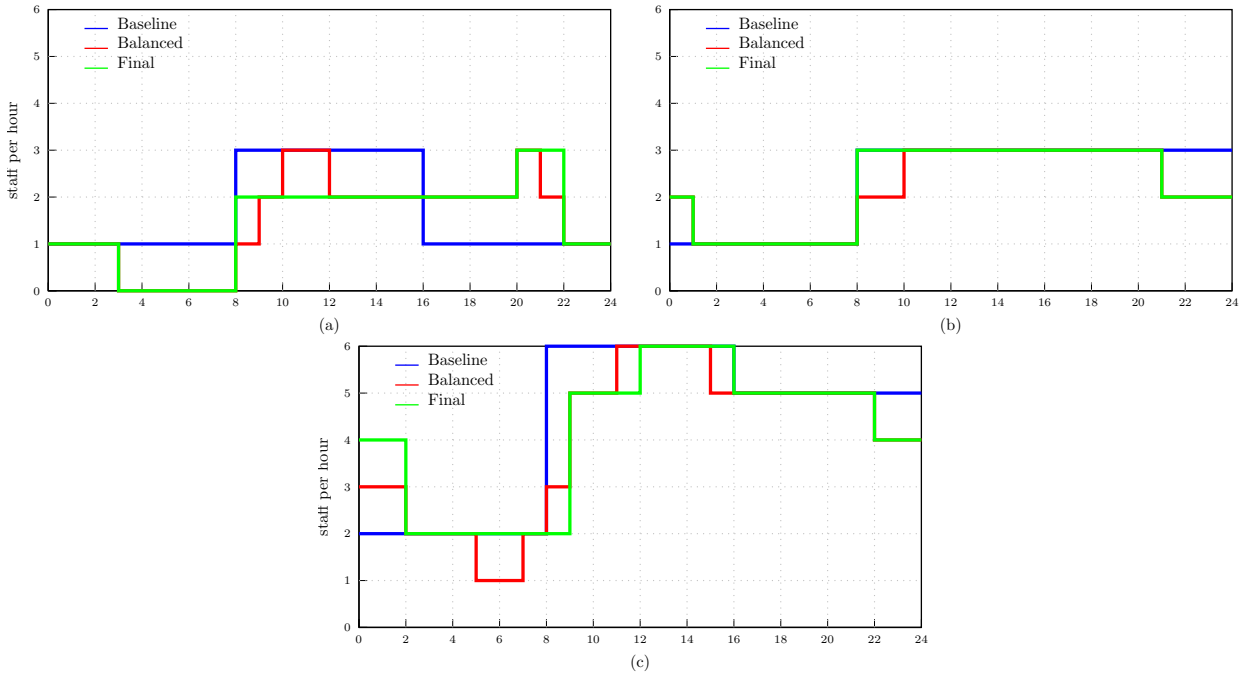

Figure 9: Baseline, balanced, and final staffing levels for : (a) ENPs, (b) radiologists, and (c) nurses.

the new predicted arrival rates to produce new profiles.

The feasibility of the staffing profiles and the resulting shift patterns is a major practical consideration. In fact, to implement the results of our method, one needs to convert shift schedules to employees rosters. This concerns assigning individual clinicians to scheduled shifts according to their preferences, working time directives, and hospital considerations (Buffa et al. 1976). Most hospitals curerently have some sort of manual or computeried rostering system which receives shift schedules and employees' preferecnes as the input and uses an exact or heuristic optimization routine to produce low-cost rosters that meet given constraints. Hence, the shifts scheduled by the method proposed here can be fed into these systems. If the system could not find any feasible solution, further changes need to be made in the staffing profiles and/or in the S-model parameters.

We note that our method does not consider forecasting uncertainty, i.e. uncertainty about the arrival rates and other elements of the model. Furthermore, we set the number of staff in response to projected loads, not adapteivly in response to observed loads.

7 Conclusions

We showed how queueing models equipped with simulation can be used to alleviate the congestion problem of emergency departments by modifying the staffing profiles. We focused on English A&E departments, in which a sojourn time target should be satisfied.

The proposed staffing algorithm relies on infinite server networks to compute the resources' time dependent workloads and highlights their ability in modeling complexities like multiple types of customers and resource sharing. We used the computed workloads in the square root staffing law, where a common value for the QoS parameter was applied for all resources at all times. Experiments confirm that this approach evens out the performance, as measured by various metrics like waiting times and sojourn times, over time.

Comparing the balanced staffing levels with baseline profiles, obtained by expected workload calculations, in a typical A&E departments shows that significant improvements can be made on the target without increase in total staff hours. The balanced profiles can be altered further to allow for practical considerations, as illustrated by a simple example. The balanced profiles can also be used in combination with a simple shift scheduling approach to produce feasible staffing levels, which also show significant improvement on the target whilst saving total staff hours. The

savings in the staff-hours might not be enought to reduce the number of clinicians working in the emergency department. However, in many hosptials it is possible to use medical staff in other departments when they are not needed in their original department.

In fact, we undertook staffing and scheduling routines in two consecutive steps. As mentioned above, employees' rosters need to be produced next. But this hierarchical approach to the problem may end up in sub-optimal solutions. For example, Ingolfsson et al. (2010) have already demonstrated inefficiencies arising from performing staffing and shift scheduling routines separately in single service systems. These inefficiencies are likely to amplify when rostering is to be performed next, and when networks of services are considered. They proposed a linear programming method, which iterates between a service quality evaluator and a schedule generator to produce shift patterns that satisfy service level targets. Extending their approach to networks of services, and to address rostering requirements are challenging yet important directions for future research.

References

Buffa, Elwood S., Michael J. Cosgrove, Bill J. Luce. 1976. An integrated work shift scheduling system. *Decision Sciences* **7**(4) 620–630.

Coats, T J, S Michalis. 2001. Mathematical modelling of patient flow through an accident and emergency department. *Emergency Medicine Journal* **18**(3) 190–192.

Cooke, M W, S Wilson, J Halsall, A Roalfe. 2004. Total time in english accident and emergency departments is related to bed occupancy. *Emergency Medicine Journal* **21**(5) 575–576.

Department of Health Annual Report. 2009. http://www.dh.gov.uk/dr_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_100819.pdf.

Department of Health Statistics. 2010. Total time spent in A&E. http://www.dh.gov.uk/en/Publicationsandstatistics/Statistics/Performancedataandstatistics/AccidentandEmergency/DH_079085.

Eick, Stephen G., William A. Massey, Ward Whitt. 1993. The physics of the $M_t/G/\infty$ queue. *Operations Research* **41**(4) 731–742.

Feldman, Zohar, Avishai Mandelbaum, William A. Massey, Ward Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Science* **54**(2) 324–338.

Fletcher, A., D. Halsall, S. Huxham, D. Worthington. 2006. The DH accident and emergency department

model: a national generic model used locally. *Jornal of the Operational Research Society* **58**(12) 1554–1562.

Green, Linda V., Peter J. Kolesar. 1997. The lagged psa for estimating peak congestion in multiserver markovian queues with periodic arrival rates. *Management Science* **43**(1) 80–87.

Green, Linda V., Peter J. Kolesar, Joao Soares. 2001. Improving the sipp approach for staffing service systems that have cyclic demands. *Operations Research* **49**(4) 549–564.

Green, Linda V., Peter J. Kolesar, Ward Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production And Operations Management* **16**(1) 13–29.

Green, Linda V., Jao Soares, James F. Giglio, Robert A. Green. 2006. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13**(1) 61–68.

Gunal, M M, M Pidd. 2009. Understanding target-driven action in emergency department performance using simulation. *Emergency Medicine Journal* **26**(10) 724–727.

Gunal, Murat M., Michael Pidd. 2006. Understanding accident and emergency department performance using simulation.

Halfin, Shlomo, Ward Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29**(3) 567–588.

Ingolfsson, Armann, Fernanda Campello, Xudong Wu, Edgar Cabral. 2010. Combining integer programming and the randomization method to schedule employees. *European Journal of Operational Research* **202**(1) 153–163.

Jennings, Otis B., Avishai Mandelbaum, William A. Massey, Ward Whitt. 1996. Server staffing to meet time-varying demand. *Management Science* **42**(10) 1383–1394.

Massey, William A., Ward Whitt. 1993. Networks of infinite-server queues with nonstationary poisson input. *Queueing Systems* **13**(1) 183–250.

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

Mayhew, L., D. Smith. 2008. Using queuing theory to analyse the Governments 4-h completion time target in accident and emergency departments. *Health Care Management Science* **11**(1) 11–21.

Mortimore, Andy, Simon Cooper. 2007. The "4-hour target": emergency nurses' views. *Emergency Medicine Journal* **24**(6) 402–404.

Munro, J, S Mason, J Nicholl. 2006. Effectiveness of measures to reduce emergency department waiting times: a natural experiment. *Emergency Medicine Journal* **23**(1) 35–39.

Sinreich, David, Ola Jabali. 2007. Staggered work shifts: a way to downsize and restructure an emergency

department workforce yet maintain current operational performance. *Health Care Management Science* **10**(3) 293–308.

Sinreich, David, Yariv Marmor. 2005. Emergency department operations: The basis for developing a simulation tool. *IIE Transactions* **37**(3) 233 – 245.

Vassilacopoulos, G. 1985. Allocating doctors to shifts in an accident and emergency department. *The Journal of the Operational Research Society* **36**(6) 517–523.

Wallace, Rodney B., Ward Whitt. 2005. A staffing algorithm for call centers with skill-based routing. *Manufacturing Service Operations Management* **7**(4) 276–294.

Whitt, Ward. 1992. Understanding the efficiency of multi-server service systems. *Management Science* **38**(5) 708–723.

Whitt, Ward. 2007. What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics* **54**(5) 476–484.