ROBUST CAPACITY PLANNING FOR ACCIDENT AND EMERGENCY SERVICES

Elvan Gökalp*

Abstract. Accident and emergency departments (A&E) are the first place of contact for urgent and complex patients. These departments are subject to uncertainties due to the unplanned patient arrivals. After arrival to an A&E, patients are categorized by a triage nurse based on the urgency. The performance of an A&E is measured based on the number of patients waiting for more than a certain time to be treated. Due to the uncertainties affecting the patient flow, finding the optimum staff capacities while ensuring the performance targets is a complex problem. This paper proposes a robust-optimization based approximation for the patient waiting times in an A&E. We also develop a simulation optimization heuristic to solve this capacity planning problem. The performance of the approximation approach is then compared with that of the simulation optimization heuristic. Finally, the impact of model parameters on the performances of two approaches is investigated. The experiments show that the proposed approximation results in good enough solutions.

1. Introduction

Healthcare is one of the largest sectors affecting millions of lives worldwide. The aging phenomenon, increased rates of long-term conditions and public access to the healthcare generated a dramatic growth in the demand [37]. On the other hand, the capacity has not risen sufficiently to match this growth due to the inflexibility and scarcity of resources. Inevitably, the healthcare managers are under a significant pressure to improve the existing capacity and resource allocation policies. The pressure for an efficient service is accompanied by the challenges inherited from the nature of the healthcare services. The most important one of these challenges is the variation in the demand and the time required to treat patients. These uncertainties result in long waiting times to receive the service in the peak demand sessions. The delays in the treatment risk the patient lives, and thus, should be avoided.

The service delays is especially critical on the patient outcomes in accident and emergency departments (A&E). This unit is the first point of contact for most of the complex and life-threating cases such as heart attack, stroke or loss of consciousness. Along with the overall increase in healthcare service demand, emergency unit arrivals have risen by 28% between 2002 and 2017 in England [44]. To improve the efficiency in A&E's,

the UK government has introduced the 4 h policy in 2004, requiring all patients in A&E to be treated within 4 h of their arrival. The target has been then reduced to 98% of patients in 2005 [42] and further to 95% in 2010 because the national levels fell below the previous targets [44]. In February 2018, the 4 h target has been suspended [45] after the national levels gradually fell to 83%. Although the financial sanctions are suspended, the hospitals are still obliged to report their A&E waiting times.

The UK government is planning to put alternative measures for emergency waiting times such as tighter waiting hour targets for serious cases [57]. In an A&E, the severity of cases are initially assessed by a triage nurse who also categorizes them. While urgent patients are prioritized to receive the main treatment, rest of the patients are served based on the first-come first-served (FCFS) policy. Although prioritization improves the patient outcomes, the levels of the human and physical resources have a significant effect on the waiting times. The physical resources such as the number of cubicles where the triage and treatment takes place in A&E cannot be changed much. On the other hand, the level of human resources *i.e.* triage nurses and doctors is a tactical decision and easier to adjust based on the long-term demand projections.

A significant challenge for A&E resource planning is the daily and seasonal variation in the demand. Besides, the time required to treat patients vary significantly; the urgent cases take significantly more time than the non-urgent ones. These variations result in very long waiting times in the high-demand sessions. Finding the optimum staffing capacity to reduce these *worst-case* waiting times is a difficult problem due to the uncertainties. Additionally, the uncertainty in arrival and waiting times may not follow a known distribution and therefore the underlying assumptions for classical (queuing) approaches are violated. This paper proposes a novel approach based on robust optimization and queuing theory to find the optimum staffing capacities in an A&E to keep the worst-case waiting times below a certain threshold. The proposed approach does not require any assumption regarding the distribution of uncertainties. The contributions of the paper are:

(1) The healthcare service in A&E is modelled using an approximation for the maximum waiting times combining the robust optimization and queuing theory, assuming that the arrival and service times can follow any distribution, and separating queues for urgent and non-urgent cases with prioritization. The results show that the resulting model can be solved to optimality.
(2) As a benchmark to the proposed optimization approach, we also implement a simulation optimization (SO) based heuristic to find the optimum staff capacities. The performances of the solutions obtained by the optimization model and the SO based heuristic are then compared. The impact of different model parameters on the solutions are also investigated.

The paper is organized as follows. The next section introduces the literature related to capacity planning problems in healthcare and accident and emergency department modelling. Section 3 provides the problem description and underlying assumptions along with the optimization model formulation. Section 4 presents the worst-case waiting time approximation and the structural analysis of the optimization model. Section 5 details the proposed SO heuristic. In Section 6, we introduce the design of experiments and results. Finally, Section 7 summarizes the study and provide several future directions.

2. Related literature

This section presents the related literature to capacity planning problem for A&E. First, we provide an overview of the capacity planning studies in healthcare. Then, we only focus on the A&E modelling and categorize the literature based on the modelling and solution techniques.

2.1. Capacity planning in healthcare

Queuing theory, a modelling approach to obtain performance measures in queuing systems, has been widely applied for the capacity planning of healthcare services; a related review can be found in [21]. Creemers *et al.* [16] use built-in queuing formulas to find the number of servers, *i.e.* capacity level, required to achieve a certain degree of performance. Hulshof *et al.* [32] use the queuing theory to model the elective patient admission process and

study the resource allocation problem for hospitals with uncertain treatment pathways. They consider different queues for different types of services with time-dependent capacity levels of resources. Similarly, Cochran *et al.* [13] apply the queuing theory to test various capacity design alternatives to be used in real time Hospital Emergency Departments when the capacity cannot meet the demand. Bretthauer *et al.* [10] consider the capacity planning problem for healthcare operations with blocking between different units. Castillo *et al.* [11] determine capacity and location of healthcare facilities using queuing models with exponential service times and Poisson arrivals. By considering time-varying demands in hospitals, Green *et al.* [26] analyze the staffing requirement in hospitals based on queuing analysis. Mingzhu [40] develops a queuing network analysis considering multiple patient types to find optimum number of servers in an outpatient clinic. The main drawback of queuing models comes from their intractability due to nonlinear formulations of performance metrics under certain distribution assumptions for arrival and service processes.

Simulation is an alternative approach to model the service systems when the queuing formulations are not useful due to their complexities. Harper *et al.* [30] introduce a discrete-event simulation model to analyze the operations management of an intensive care unit and use the data generated by the simulation approach to solve the stochastic optimization model which computes the optimum number of nurses required to achieve the service targets. De Angelis *et al.* [17] consider SO to determine the capacity of a transfusion centre under multiple objectives: cost minimization to achieve a fixed waiting time and minimization of waiting time under a limited budget. The queuing system is modelled with a discrete-event simulation and the objective functions are approximated by function fitting with data generated by the simulation model. Similarly, Alfonso *et al.* [2] model processes in a blood collection unit with a simulation-based approach. They evaluate possible blood-collection server configurations from a cost-effectiveness perspective. Although simulation is very useful to model complex systems, it can only provide approximate solutions that are affected by the bias of data generation.

Optimization models in healthcare capacity planning focus not only on single hospital or department but also the interconnection between departments and hospitals, which usually has significant effects on the overall performance. Several studies focus on this interconnection in different capacity planning problems modelled for networks of hospitals or departments [4, 5, 19, 24, 25, 28, 38, 50, 54–56]. Flessa [19] develops a model to allocate resources in the preventive and curative services in hospitals. Govind *et al.*, Gunes *et al.*, Santibanez *et al.* and Stummer *et al.* [25, 28, 54, 55] focus on the location and number of beds in hospitals within a network to minimize operation cost and maximize patient utility.

Pehlivan *et al.* [50] develop a mixed-integer optimization model to determine the capacity of maternity facilities in a network in view of uncertain patient arrivals and service times. The objective is to minimize the number of refused admissions which is formulated by using available queuing formulations. They assume the interarrival and service times are exponentially distributed. On the other hand, Asaduzzaman *et al.* [5] develop a queuing model to find the optimum capacities of neonatal centres to minimize refusal and overflow probabilities. They also assume exponential interarrival and service times. To the best of our knowledge, the proposed approximation in this paper has not been utilized before for a capacity planning problem.

2.2. Accident and emergency modelling

Simulation modelling and optimization

Mohiuddin *et al.* [41] identify and review 19 studies related to simulation modelling for emergency departments in the UK. Another comprehensive review of simulation modelling studies in emergency departments for normal and disaster conditions can be found in Gul *et al.* [27]. Among 106 reviewed papers, only few studies consider an optimization approach [1,18,22,51,59]. Fruggiero *et al.* [22] uses ant-colony optimization along with a simulation model to optimize the resources in an emergency department. Weng *et al.* [59], Ibrahim *et al.* [33] and Rico *et al.* [51] use OptQuest, a SO engine [49], to optimize nurse and physician numbers and nurse allocation in an influenza outbreak, respectively. Ghanes *et al.* [23] also uses SO to find the staffing levels in A&E with the objective of minimizing the patient length of stay. Chen and Wang [12] aim to find optimum number of staffing in A&E minimizing patient length of stay and the medical resources wasted by SO.

The most similar study to ours is Ahmed *et al.* [1] which presents a SO approach for capacity planning of an emergency department in Kuwait. They consider triage and a prioritized service queue. They consider stochastic constraints in a discrete SO problem maximizing the throughput. They also model and solve the optimization problem where the objective function is total cost of the staff and the constraints are the waiting times. This second problem description is very similar to ours. Their heuristic first identifies the feasible set of solutions and then finds the best solution among those based on random sampling. In each iteration, they compare the objective value of the new solution to the previous one and accept the new one if it supersedes in certain number of iterations. However, they have not provided any performance results for their heuristic results.

Mathematical modelling and queuing theory

Queuing theory has been mostly been applied into staff and bed optimization, ambulance deployment problems; for a detailed review, readers are referred to [36]. Emergency departments are also modelled by queuing theory [60]. Reviews of modelling and queuing theory studies for A&E can be found in Saghafian *et al.* [53], Costa *et al.* [15] and Hu *et al.* [31]. These papers have considered the average instead of maximum waiting times. Few papers model separate queues based on patients' severity [13]. For example, Cochran and Roche [13] develop a queuing network model for an A&E that uses separate queues for low and high acuity patients. Optimum capacity (either in terms of staff or waiting area limit) for each step of patient flow in A&E is computed. They use an approximate waiting time formulation [3] and target waiting times and utilization rates of each step to set up the capacities assuming that the arrival times are exponentially distributed.

Mayhew *et al.* [39] develop a queuing model for A&E department assuming that the arrivals and service times are exponentially distributed. They have divided the arrivals into two and added a triage step before the treatment. They have compared the predicted overall departure time from A&E with the real departure times in A&E's obtained from the NHS UK. They have used the model to test whether the 4 h target is achievable if part of the A&E service is carried in other units. They have not carried optimization or capacity planning.

3. Capacity planning model for A&E

This section first describes the underlying problem and our assumptions and then introduces the mathematical formulation.

3.1. Problem description and assumptions

We model the activities in a typical (major) A&E department in the UK [48] for a finite planning horizon. With small modifications, the model can be applied to any other emergency department. As a patient arrives to the A&E, s/he is put into an FCFS queue for triage. A triage nurse categorizes the patient as discharge, a type 1 (urgent) or type 2 (non-urgent) based on the medical assessment. Note that we do not model specific triage categories which may be more than 2. Instead, we only divide them based on whether the patient is urgent or not, as in Ahmed *et al.* [1]. Type 1 and 2 patients are placed into two separate FCFS queues for the treatment. Type 1 patients are given priority for treatment. After the patients are treated by an A&E doctor, they are either discharged, referred or admitted to the hospital. The time spent in the A&E from arrival until the disposal (discharge, referral or admission) should be less than 4 h for all patients. We assume that other medical activities required for the treatment such as laboratory tests are included in the treatment duration. Also note that we do not consider the single specialty cases such as opthalmology or dental. These patients go through a separate route than the other two categories in A&E [45] and constitute a small percentage (0.5%) of the 4 h breaches.

The uncertainties affecting the waiting times are the arrival times, the triage and treatment durations. Although arrival times can follow an exponential distribution, there is no consensus in the literature about the service time distributions in A&E; triangular [20], general [34], exponential [53] and uniform [1] distributions are used to model the A&E treatment duration.

The hospital management aims to find the staff capacities with minimum cost while satisfying the waiting time targets. One intuitive approach for this capacity planning problem would be to allocate resources proportionate to the demand rates for each service. However, that method would not match the actual workloads. The actual congestion peak times lag the arrival times as the number of service stations as in the A&E increases [34]. The approximation and heuristic approaches provided in this paper use queuing model to estimate the actual workloads, and thus provide a better performance than simple allocation of resources based on the demand rates [34].

Note that the arrival rates to an A&E can vary based on the time of the day. Here, we only consider a stable arrival rate. In other words, we approximate the varying arrival rate with its average value. The reason for this assumption is that the model can be easily extended to time-dependent arrival rates, by simply adding time indices to the arrival rates. In such a case, one would find the staff capacities for each time period. This can easily be done by following our approach seperately for each time period. In other words, extension to time-dependent arrival rates would not affect the model complexity or structure and thus the main objectives of this paper.

Another modelling choice is related to the "boarding". This term refers to the cases where after medical treatment is completed, the patient may need to wait for a bed in the hospital (if admitted). This may create additional delay on the patient's length of stay. However, since A&E beds are highly utilized and expensive resources, some hospitals put these admitted patients into "buffer" wards such as Critical Decision Units [43]. Besides, the "boarding" process would require us to model all the bed utilization in all wards of the hospital which is beyond the scope of this paper.

3.2. Problem formulation

This section provides a mathematical formulation for the capacity planning problem described above. The number of nurses in the triage is denoted by x_1. The fixed (unit) cost of triage nurses and doctors are shown with c_1 and c_2, respectively. We assume that the patients arrive to the A&E with the mean interarrival time $1/\lambda$ and standard deviation σ. After registration, they wait in the triage queue and assessed by a triage nurse under the FCFS rule. The mean triage time is $1/\mu$ with standard deviation σ_1. Maximum total time spent in the triage by any patient arrived during the planning horizon is $W_1(x_1)$. A certain percentage, θ, of the patients are discharged after triage. The others are categorized as type 1 or type 2 each of which has a separate queue for treatment. The rate of type 1 patients among whole arrivals is α. Patients wait in the treatment queues until they are seen by one of x_2 number of A&E doctors. Type 1 cases have the priority over type 2 and the queue is preemptive: the treatment of a type 2 case is stopped when a type 1 arrival occurs at the same time. The average treatment time for type 1 and type 2 are $1/\mu_{12}$ and $1/\mu_{22}$ with standard deviations σ_{12} and σ_{22}, respectively.

The hospital management aims to keep the total time spent in the A&E below $\overline{W}$. Due to the uncertainty in the arrival and service times, the waiting times experienced by the patients vary. Therefore, the model should be robust against the uncertainties in the arrival and service times. We denote the maximum time spent between triage and treatment as $W_{12}(x_2)$ and $W_{22}(x_2)$ for types 1 and 2 patients, respectively. An approximation for the maximum time spent in the A&E can be written as:

$$W_1(x_1) + W_{12}(x_2) + 1/\mu + 1/\mu_{12},$$

and

$$W_1(x_1) + W_{22}(x_2) + 1/\mu + 1/\mu_{22},$$

for types 1 and 2 patients, respectively. Figure 1 shows a summary of the A&E operations along with the notation used in the model.

For a stable queue, the utilization rate (traffic intensity) should be smaller than 1 [35]; $\dfrac{\lambda}{x_1\mu} < 1$ for the triage queue. In other words, the total service rate $(x_1\mu)$ should be larger than the total arrival rate (λ) such that the
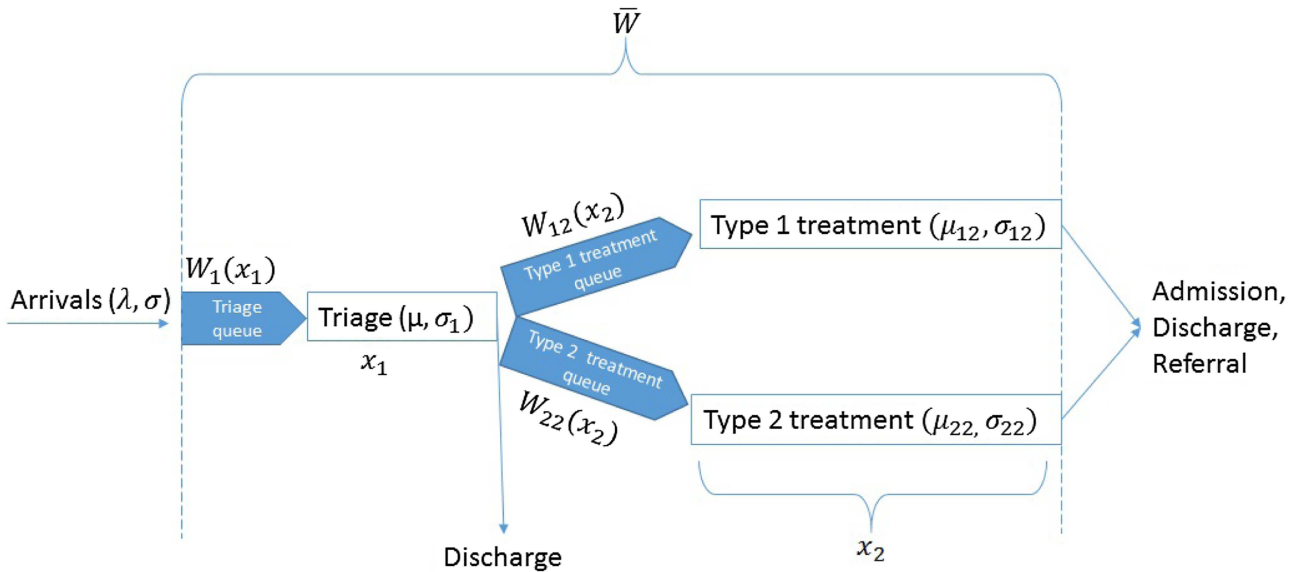
FIGURE 1. A description of the A&E service along with the notation used.

queue does not grow exponentially. This condition should be satisfied for the treatment queues as well,

$$x_2 > \lambda_{12}/\mu_{12}, \quad x_2 > \lambda_{22}/\mu_{22},$$

where λ_{12} and λ_{22} represent the arrival rates to the treatment queues of type 1 and type 2 patients, respectively.

The stochastic capacity planning model for the A&E can be formulated as follows:

$$\text{AE}: \quad \min c_1 x_1 + c_2 x_2, \tag{3.1}$$
$$\text{s.t. } \overline{W} \geq W_1(x_1) + W_{12}(x_2) + 1/\mu + 1/\mu_{12}, \tag{3.2}$$
$$\overline{W} \geq W_1(x_1) + W_{22}(x_2) + 1/\mu + 1/\mu_{22} \tag{3.3}$$
$$x_1 > \lambda/\mu, \tag{3.4}$$
$$x_2 > \lambda_{12}/\mu_{12}, \tag{3.5}$$
$$x_2 > (\lambda_{12} + \lambda_{22})/\mu_{22}, \tag{3.6}$$
$$x_1, x_2 \in Z^+. \tag{3.7}$$

As mentioned before, the government is planning to put waiting time targets for serious (type 1) patients only. Assuming that this policy is activated, the problem would be then modelled as:

$$\text{AE}^{\text{red}}: \quad \min c_1 x_1 + c_2 x_2,$$
$$\text{s.t. } (3.2), (3.4), (3.5), (3.7).$$

In the case of extension to a time-dependent arrival rate with $\lambda_{12}(t)$ and $\lambda_{22}(t)$, the model variables would be differentiated for each time period, *e.g.* $x_1(t)$ and $x_2(t)$, while the rest of the model would stay same. Since that extension does not affect the model, we continue with time-independent version in the rest of the paper.

In order to solve these models, we need to compute the waiting times for each patient arriving to the A&E. The exact computation of the waiting times in each scenario is difficult even with a fixed number of staff. The computational intractability due to combinatorial number of calculations has already been proven for a queuing

system of multiple servers with exponential arrivals and general service time distribution as in Tijms *et al.* [58]. The next section approximates the maximum waiting times by using robust optimization principles and provides a tractable approximate model.

4. APPROXIMATION WITH ROBUST OPTIMIZATION

This section presents an approximation for the maximum waiting times in the A&E. Different approaches exist to approximate the maximum waiting time in a queuing system; for instance, see Gupta *et al.* [29]. However, these approximations usually do not lead to realistic results when the arrival process follows a distribution different from Poisson [7]. As an alternative approach, Bandi and Bertsimas [7] proposed to approximate the maximum waiting time in an FCFS queue when the arrival and service times follow an unknown distribution. Their approach is based on developing *uncertainty sets* for the uncertain arrival and service durations based on historical data. They used the central limit theorem which asserts the asymptotic results for a large set of independent and identically distributed random variables. Readers are referred to [7] for more details regarding the approximation. According to Bandi and Bertsimas [7], the maximum waiting time in an FCFS queue with x_1 servers $\overline{W}(x_1)$, arrival and service rates λ and μ, arrival and service time variabilities Γ^a and Γ^s can be approximated as,

$$\overline{W}(x_1) = \frac{\lambda(\Gamma^a + \Gamma^s/\sqrt{x_1})^2}{4\big[1 - \lambda/(\mu x_1)\big]}. \tag{4.1}$$

The arrival and service time variabilities are set based on the desired conservativeness level. For example, they can be set as double or three times of the standard deviation of the corresponding uncertain parameter to cover most of the possible realizations.

We use (4.1) to approximate the maximum waiting times in the A&E. The variability in the arrival times to the A&E is denoted by Γ^a and the variabilities in triage duration and treatment duration for type 1 and type 2 patients are denoted by $\Gamma^s, \Gamma^s_{12}, \Gamma^s_{22}$, respectively. Based on [7], we set $\Gamma^a = k \cdot \sigma_a$ and $\Gamma^s = k \cdot \sigma_s$ with $k > 0$, where σ_a and σ_s are the standard deviations of the corresponding interarrival and service times, respectively. The parameter k is set based on the desired conservativeness level of the model: a larger k corresponds to a more conservative model against the uncertainties in the arrival and service times. We present the effect of different conservativeness levels on the results in Section 6.2. We approximate the maximum waiting times in the triage and type 1 treatment queues as follows:

$$\overline{W}_1(x_1) = \frac{\lambda(\Gamma^a + \Gamma^s/\sqrt{x_1})^2}{4\big[1 - \lambda/(\mu x_1)\big]}, \tag{4.2}$$

$$\overline{W}_{12}(x_2) = \frac{\lambda_{12}(\Gamma^a_{12} + \Gamma^s_{12}/\sqrt{x_2})^2}{4\big[1 - \lambda_{12}/(\mu_{12} x_2)\big]}. \tag{4.3}$$

Note that the computation of the waiting times of type 2 patients is more complicated. A type 2 patient in the treatment queue is always served after all existing type 1 patients are served. This would imply that type 2 patients always wait more than a type 1 patient. In the worst case, a type 2 patient would wait for the maximum waiting time for a type 1 patient in addition to the maximum waiting time in type 2 treatment queue. Then, an approximation for the maximum type 2 patient waiting time can be formulated as:

$$\overline{W}_{22}(x_2) = \frac{\lambda_{22}(\Gamma^a_{22} + \Gamma^s_{22}/\sqrt{x_2})^2}{4\big[1 - \lambda_{22})/(\mu_{22} x_2)\big]} + \overline{W}_{21}(x_2). \tag{4.4}$$

The model AE and AE$^{\text{red}}$, respectively, can be reformulated as:

$$\text{AE}_{\text{rob}} : \min c_1 x_1 + c_2 x_2,$$

$$\text{s.t. } \overline{W} \geq \frac{1}{\mu} + \frac{1}{\mu_{12}} + \overline{W}_1(x_1) + \overline{W}_{12}(x_2), \tag{4.5}$$

$$\overline{W} \geq \frac{1}{\mu} + \frac{1}{\mu_{22}} + \overline{W}_1(x_1) + \overline{W}_{22}(x_2), \tag{4.6}$$

$$(3.4), (3.5), (3.6), (3.7).$$

$$\text{AE}_{\text{rob}}^{\text{red}} : \min c_1 x_1 + c_2 x_2,$$

$$\text{s.t. } (4.5), (3.4), (3.5), (3.7)$$

which have linear objective functions and non-linear constraints. The next proposition states that the relaxed version of AE_{rob} has a convex feasible set. The same proposition and proof apply to $\text{AE}_{\text{rob}}^{\text{red}}$. Therefore, both models have global optimums [9].

Proposition 4.1. *For relaxed variables x_1, $x_2 \in R^+$, the model AE_{rob} has a convex feasible set.*

Proof. Let's first show the convexity of constraint (4.5). Let's define $f(x_1, x_2) = \overline{W}_1(x_1) + \overline{W}_{12}(x_2)$. For a multi-variate function to be convex, its Hessian matrix should be a positive semi-definite matrix. Let $H = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ denote the Hessian matrix of function $f(x_1, x_2)$, where its second order derivatives are denoted by a, b and c. For the function to be convex, all principal minors, a, c, $(ac - b^2)$, should be non-negative. Note that $b = \frac{\partial^2 f(x, \alpha)}{\partial x \partial \alpha}$ is always zero, because function $f(x_1, x_2)$ can be divided into two separate functions of variables x_1 and x_2. Therefore, it is enough to show that $a = \frac{\partial^2 f(x_1, x_2)}{\partial x_1^2}$ and $c = \frac{\partial^2 f(x_1, x_2)}{\partial x_2^2}$ are non-negative. These derivatives can be written as:

$$a = \frac{\partial^2 \overline{W}_1(x_1)}{\partial x_1^2}, \ c = \frac{\partial^2 \overline{W}_{12}(x_2)}{\partial x_2^2},$$

due to the separability of the maximum waiting time functions. We arrive the formulation of the second order derivative a after some intermediate calculations as:

$$a = \frac{\Gamma^a \Gamma^s \big(-m^2 + 6my^2 + 3y^4 \big) + 4y^3 \big(m(\Gamma^a)^2 + (\Gamma^s)^2 \big)}{2y^3(y^2 - m)^3},$$

where $y = \sqrt{x_1}$, and, $m = \lambda/\mu$. The denominator of a is always positive due to the traffic intensity condition $\mu x_1 > \lambda$. The absolute of the only negative term in the nominator, $-m^2 \Gamma^a \Gamma^s$, is always smaller than the second term of the nominator $6m\Gamma^a \Gamma^s y^2$ because $m < 1$ and $x_1 \geq 1$. Therefore, the second order derivative a is always positive. We omit the computations for c that follows the same structure as a. Since all principal minors of H are non-negative, $f(x_1, x_2)$ is convex. The second constraint (4.6) possesses the same structure, and is also convex. Note that all other constraints, (3.4)–(3.6) are also convex. Because all constraints are convex, the relaxed version of model AE_{rob} has a convex feasible set. $\square$

Since the relaxed version of AE_{rob} is a convex optimization problem with a global optimum, convex non-linear optimization solvers such as Bonmin [14] can find the global optimum solution efficiently following Bonami *et al.* [8].

5. SIMULATION OPTIMIZATION HEURISTIC

The models AE and AE^{red} are non-linear integer programming models that are very difficult to solve with traditional optimization techniques. A possible solution approach for these models is SO that is based on simulating alternative solutions and comparing the simulated objective function values. An intuitive method

for SO is to first enumerate all feasible solutions and simulate their performances. However, this method would require a long computation time, and therefore, more clever search strategies are needed. Our literature review shows that most of the SO studies employ built-in optimization packages within a commercial simulation modelling software such as OptQuest in Simul8.

To investigate the performance of the worst-case approximation and optimization approach presented in Section 4, we also design and implement an SO heuristic as a benchmark solution method. The performances of this heuristic and the optimization *via* a commercial non-linear integer solver are then compared.

Other than the stochastic waiting time constraints, our model has a set of deterministic constraints related to traffic intensity (3.7) and a monotonic objective function. Therefore, we do not need to search for the optimum solution randomly as in Ahmed *et al.* [1]; we can start the search from the smallest capacities satisfying the traffic intensity constraints and increase these capacities incrementally until the stochastic constraints are satisfied with a certain confidence rate.

For this purpose, we first develop a simulation model of the A&E operations described in Section 3.1 and implement it on Matlab. The planning period of the simulation is set to T minutes, while the time unit is one minute. One iteration of the simulation model comprises of j runs of n scenarios. In each iteration, the SO heuristic searches for a better solution based on the simulation outputs. Starting from the minimum possible levels of the capacity variables (satisfying the traffic intensity constraints), in each iteration, we increment the capacity variable that has the largest potential to decrease the maximum waiting times. To identify the variable with the largest potential improvement, we use the approximate maximum waiting time formulation (4.1). Note that the objective function (total cost) increases by c_1 and c_2 with one unit increase in x_1 and x_2, respectively. In other words, increasing x_2 by one would result in the same change in the objective value as increasing x_1 by c_2/c_1. Therefore, the potential improvements in the maximum waiting times should be computed for $x_1 + c$ and $x_2 + 1$, respectively, where $c = \lceil c_2/c_1 \rceil$ due to the integrality condition. The capacity variable with the largest potential improvement is incremented by one and the simulation model is run for another iteration. In each iteration, if $(1 - \epsilon)\%$ of the patients' total waiting time is lower than $\overline{W}$, where ϵ is the the desired confidence level, *e.g.* 1%, then the heuristic stops. Otherwise, the process is repeated again. Algorithm 1 presents the pseudo-code of the SO heuristic. The heuristic can be applied to AE^{red} by just removing the parameters related to type 2 treatment as presented in Algorithm 2.

Algorithm 1. SO Heuristic for AE.

Set ϵ, W^i to a very large number, $i = 0$ and compute $x_1^0 = \lambda/\mu$, $x_2^0 = \min\left\{\frac{\lambda_{12}}{\mu_{12}}, \frac{\lambda_{22}}{\mu_{22}}\right\}$, and $W^0 = \max\left\{\left(\overline{W}_1(x_1^0) + \overline{W}_{21}(x_2^0)\right), \left(\overline{W}_1(x_1^0) + \overline{W}_{22}(x_2^0)\right)\right\}$ using (4.2), (4.3), (4.4).

while $W^i \geq \overline{W}$, **do**

    Compute $\Delta W_1 = \overline{W}_1(x_1^i) - \overline{W}_1(x_1^i + c)$ and

    $\Delta W_2 = \max\left\{\left(\overline{W}_{12}(x_2^i) - \overline{W}_{12}(x_2^i + 1)\right), \left(\overline{W}_{22}(x_2^i) - \overline{W}_{22}(x_2^i + 1)\right)\right\}$.

    if $\Delta W_1 > \Delta W_2$ **then**

        $x_1^{i+1} = x_1^i + 1$, $x_2^{i+1} = x_2^i$.

    else

        $x_2^{i+1} = x_2^i + 1$, $x_1^{i+1} = x_1^i$.

    end if

    $i := i + 1$.

    Run simulation model for j runs and n scenarios with x_1^i and x_2^i. Set W^i to $(1 - \epsilon)\%$ of the waiting times obtained by the simulation model.

end while

return x_1^i and x_2^i.

Algorithm 2. SO Heuristic for AE^{red}.

Set ϵ, W^i to a very large number, $i = 0$ and compute $x_1^0 = \lambda/\mu$, $x_2^0 = \left\{ \frac{\lambda_{12}}{\mu_{12}} \right\}$, and $W^0 = \left(\overline{W}_1(x_1^0) + \overline{W}_{21}(x_2^0) \right)$ using (4.2), (4.3), (4.4).

while $W^i \geq \overline{W}$, **do**
    Compute $\Delta W_1 = \overline{W}_1(x_1^i) - \overline{W}_1(x_1^i + c)$ and
    $\Delta W_2 = \overline{W}_{12}(x_2^i) - \overline{W}_{12}(x_2^i + 1)$.
    if $\Delta W_1 > \Delta W_2$ **then**
        $x_1^{i+1} = x_1^i + 1$, $x_2^{i+1} = x_2^i$.
    else
        $x_2^{i+1} = x_2^i + 1$, $x_1^{i+1} = x_1^i$.
    end if
    $i := i + 1$.
    Run simulation model for j runs and n scenarios with x_1^i and x_2^i. Set W^i to $(1 - \epsilon)\%$ of the waiting times obtained by the simulation model.
end while
return x_1^i and x_2^i.

The computation time of the heuristic depends on the running time of the simulation model and therefore the levels of n, j and T. As the number of runs and scenarios increases, the robustness of the solution obtained by the heuristic increases as well.

6. Computational experiments

The computational experiments aim to illustrate the performances of the approximation approach and the SO heuristic as well as the impact of several model parameters on the results. For this purpose, we design two sets of computational experiments. The first set of experiments compares the performances of the solutions computed by the approximate optimization models and the SO heuristic. The second set of the experiments investigates the impact of model parameters on the solutions obtained by the approximation approach. All computational experiments are carried out on a PC with Windows 10 Enterprise operating system, CPU 4 GHz Intel Core i7 and 32 GB of RAM.

6.1. Input data

As all major A&E's in the UK follow the same service process, we use the arrival data of University Hospitals Coventry & Warwickshire (UHCW) provided in the online resources of the NHS UK [44]. The average treatment and triage times are obtained from Ahmed *et al.* [1]. The service times are assumed to follow an exponential distribution [53]. The effect of this assumption is investigated in the first set of computational experiments.

For the variability parameters, we first generate a dataset of arrival and service times by using the simulation model and the distribution information provided in Table 1. According to [7], the variability parameters (Γ^a and Γ^s) are then set such that most of the uncertain parameters are covered. The time spent in the A&E ($\overline{W}$) should be less than 4 h for all patients.

6.2. Comparison of SO heuristic and approximate optimization model

This section presents the results and the performances of two solution approaches for different (i) problem settings, (ii) conservativeness levels and (iii) service time distribution. The approach that solves the models, AE_{rob} and $\text{AE}_{\text{rob}}^{\text{red}}$, with a commercial solver (Gams/Bonmin) is referred as *Approximate Optimization (AO)* in the rest of this section.

TABLE 1. Input data for model parameters used in the numerical experiments.

Description of parameters	Value/Range	Source of data	Distribution
A&E arrival rates	0.57 patient/minute	[1, 44]	Exponential
Probability of discharge or diagnosis as Type 1 patient after triage, respectively	0.1 and 0.61	[44]	Binomial
Mean triage service duration	15 min	[1, 53]	Exponential
Mean treatment duration for type 1 patients	90 min	[1, 53]	Exponential
Mean treatment duration for type 2 patients	25 min	[1, 53]	Exponential
Cost of doctors with respect to nurses (c)	5	[47, 52]	–

TABLE 2. Base results of AO and the SO Heuristic when only type 1 patients are subject to waiting time limit.

Approach	AO	SO Heuristic					
Number of runs (j)	–	30		10		1	
Number of scenarios (n)	–	1000	150	1000	150	1000	150
Capacities (x_1, x_2)	9, 31	9, 30	9, 29	9, 30	9,28	9, 28	9, 28
Computation time (s)	2	3292	135	1161	198	52	4
Breaches	$10^{-5}\%$	$10^{-5}\%$	$10^{-5}\%$	$10^{-5}\%$	0.02%	0.02%	0.02%

Impact of problem setting. This section presents the results for the optimization models AE_{rob} and $\text{AE}_{\text{rob}}^{\text{red}}$ solved by a commercial solver Gams, by using the solver Bonmin. Similarly, we solve the reduced and full models, AE and AE^{red}, with the SO heuristic that is implemented in Matlab with $\epsilon = 0.0001$. We set the planning horizon of the problem to 1500 min that is found to be large enough to observe the queue dynamics. The SO heuristic is run for 40, 30, 10, and 1 runs to understand the impact of the number of runs on the heuristic's performance.

First, we solve the problem where only type 1 patients are subject to 4 h waiting time limit, *i.e.* AE^{red} and $\text{AE}_{\text{rob}}^{\text{red}}$. Table 2 shows the capacities found by AO and the SO heuristic for different number of runs and scenarios. The results of the SO heuristic with 40 runs is the same as those with 30 runs. The table also shows the computation times of these solution approaches in terms of seconds.

The number of doctors found by two methods is slightly different; AO is more conservative to the uncertainties in the arrival and service times. The computation time of the SO heuristic is significantly larger than that of AO especially as the number of iterations increases. The conservativeness of the solutions found by the SO heuristic also increases with a higher number of iterations.

The NHS sources [46] indicate that there are around 58 full-time-equivalent doctors in UHCW in March 2018. Assuming the doctors make 2.5 shifts per day, this would be equivalent to 23 doctors. The NHS statistics show the percentage of patients treated within 4 h in the A&E was 79.2% in that month [44]. Therefore, our results indicate that the performance can be improved by increasing the staff level from 23 to 31.

Next, we evaluate the performances of the solutions obtained by two approaches by giving these capacities to the simulation model as inputs. Figure 2 shows three frequency histograms for total waiting time in the A&E of (type 1) patients computed by the simulation model for 30 runs and 500 scenarios with the capacity levels found by AO and the SO heuristic for different number of runs.

The maximum waiting time in the A&E should not be larger than 135 min that is the difference between the limit on total time spent (240 min) and total average service time for triage and treatment (15 and 90 min, respectively). The graphs show that the frequency of patients waiting more than 135 min is significantly larger with the heuristic solutions, and the maximum waiting time computed by the Simulation model with AO
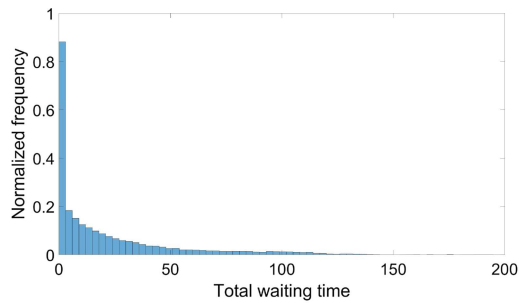
(A) Simulation outputs when staff capacities are obtained by AO
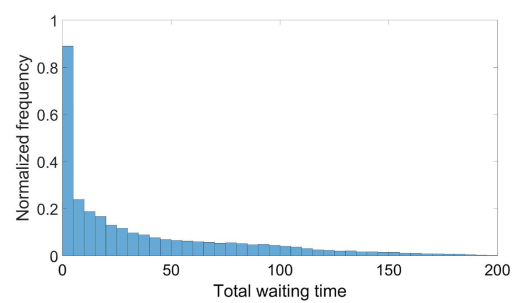
(B) Simulation outputs when staff capacities are obtained by the SO heuristic with 30 runs
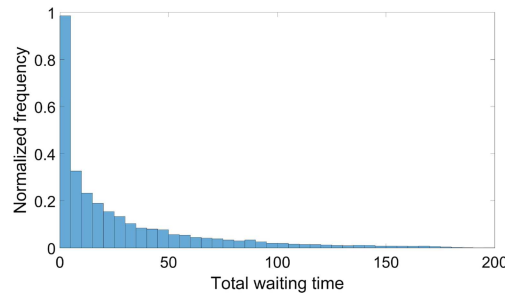
(C) Simulation outputs when staff capacities are obtained by the SO heuristic with 10 runs

FIGURE 2. Frequency histograms for total waiting time (of type 1 patients) in the A&E computed by the simulation model with the staff capacities obtained by different solution methods. (A) Simulation outputs when staff capacities are obtained by AO. (B) Simulation outputs when staff capacities are obtained by the SO heuristic with 30 runs. (C) Simulation outputs when staff capacities are obtained by the SO heuristic with 10 runs.

TABLE 3. Base results of AO and the SO heuristic when both types of patients are subject to 4 h.

Approach	AO	SO Heuristic	
Number of scenarios (n)	–	1000	100
Capacities (x_1, x_2)	12, 32	11, 32	11, 32
Computation time (s)	2	1003	210
Breaches	0.08%	0.08%	0.08 %

solutions is 130 min while that for the SO heuristic (with 30 iterations) is 150 min. This can be interpreted as the heuristic solutions perform worse than the AO solutions.

The results indicate that the proposed robust optimization based approximation is appropriate for the A&E capacity planning problem. They also suggest that AO performs better than the SO heuristic in terms of the computation time and the solution performance (Tab. 3).

Second, we assume that the health authorities replace the waiting time target of 4 h for both types of patients served in the A&E, *i.e.* AE and AE_{rob} are solved. Since the effect of number of runs is already shown, we set $j = 10$ and $n = 1000$ for the SO heuristic that produced the same solution with $j = 30$ and $n = 1000$.

TABLE 4. Capacities (x_1, x_2) found by the AO in different variability levels (Γ^s and Γ^a).

Variabilities	Service time	Arrival time
Doubled	18, 37	9, 31
Base	9, 30	9, 31
Halved	8, 28	9, 30

TABLE 5. Computation time and capacities (x_1, x_2) found by the SO heuristic in different conservativeness levels (ϵ).

ϵ	0.0001	0.001	0.05	0.1	0.2
Capacities (x_1, x_2)	9, 30	9, 26	9, 24	9, 23	9, 22
Computation time (s)	1161	64.18	38.03	25.3	12.5
Breaches	0	0.0016 %	0.021 %	0.061 %	0.1294 %

The computation times of both methods do not change significantly compared to the reduced problem setting. Note that type 2 patients have a lower priority and wait longer than type 1 patients. Therefore, the capacities are higher than those obtained for the reduced problem; they both increase by 2 units. We observe that the breaches have increased slightly for both methods compared to the previous problem setting. This may be due to increased complexity with the additional type 2 waiting time limit. These results indicate that the approximation of the waiting time for type 2 patients may not be as good as that for type 1 patients. In the rest of the experiments, we consider the reduced problem setting; only type 1 patients are subject to the 4 h waiting time limit. Also, the SO heuristic is always run for $j = 10$, $n = 1000$ unless stated otherwise.

Impact of conservativeness levels. In this section, we investigate the impact of conservativeness levels on the solutions obtained by two approaches. As explained before, the variability parameters, Γ^a and Γ^s, define the conservativeness levels of the AO; a larger variability corresponds to a more robust model against the uncertainties in the arrival and service times. To investigate the effect of the conservativeness, we solve the optimization model for two more variability levels: the double and half of the base variability levels used in the previous set of experiments. Table 4 shows the optimum number of triage nurses and doctors in the doubled, base and halved variabilities.

As different from the service time variability, the arrival time variability does not affect the solutions significantly. This is probably due to a lower variance in the arrival times leading to a lower arrival time variability.

For the SO heuristic, the conservativeness level is defined *via* parameter ϵ. Table 5 shows computation times and the solutions obtained by the SO heuristic in different (ϵ) levels. Note that the conservativeness level of the heuristic affects its computation time and the solutions significantly. Although a higher ϵ results in a shorter computation time, the quality of the solution drops significantly.

Note that the number of doctors found for the conservativeness level of 0.2 is almost equal to the that in UHCW in March 2018, *i.e.* 23 doctors [46], in which the hospital had breaching rate of 0.2 [44].

Impact of service time distribution. In this experiment, we investigate the impact of the distribution of treatment and triage durations on the results. For this purpose, the treatment and triage durations are assumed to be uniformly distributed, respectively, over [60, 120] and [10, 20] based on Ahmed *et al.* [1]. The solutions obtained by two approaches and the computation times are shown in Table 6. The rate of breaches is negligible for all solutions.

The computation times are not different to those with the exponential distribution assumption. However, the capacities obtained by the SO heuristic (with 1000 scenarios) and the AO are the same as different from

TABLE 6. Computation times and the solutions of AO and the SO heuristic when triage and treatment time distributions are uniform.

Approach	Aproximate optimization	SO Heuristic	
Number of scenarios (n)	–	1000	100
Capacities (x_1, x_2)	9 , 28	9, 28	9, 27
Computation time (sec.)	2	331	25

TABLE 7. The solutions obtained by the AO for different waiting time limits.

	$\overline{W_2}$		
$\overline{W_1}$	4 h	5 h	6 h
	x_1, x_2	x_1, x_2	x_1, x_2
3 h	12, 32	15, 31	15, 31
4 h	12, 32	10, 30	10, 30
5 h	12, 32	10, 30	9, 29

TABLE 8. Computation time and results of optimization model and SO heuristic when both categories of patients are subject to waiting time limits.

Ratio of type 1 patients	Discharge rate after triage	
	0.1*	0.2
after triage	x_1, x_2	x_1, x_2
0.45	16, 27	14, 27
0.61*	12, 32	12, 31
0.75	13, 34	–

the base case. This may indicate that the SO heuristic performs better when the service times follow a uniform distribution.

6.3. Sensitivity analysis

In this section, we test the impact of several model parameters on the solutions obtained by the AO.

Impact of waiting time limit. As the NHS looks for alternative performance monitoring policies, we investigate the impact of different waiting time limits on the solutions obtained by the AO. Table 7 shows these capacities for different waiting time limits for type 1 $(\overline{W_1})$ and type 2 $(\overline{W_2})$ patients.

The results indicate that for a fixed type 2 waiting time limit, the solutions are not affected significantly when the waiting time limit for type 1 patients is above 4 h. Similarly, when the waiting time limit is above 5 h for type 2 patients, the solutions do not change.

Impact of patient arrival rates. The NHS report [6] shows that the percentage of type 1 patients can vary in different hospitals. Therefore, this experiment investigates the effect of rates of type 1 patients and discharges after triage among all arrivals. Based on [6], we obtain the solutions by the AO for three rates for type 1 patients and two rates for discharge. Table 8 shows the solutions for different rates of discharge and type 1 patients. The base levels of these parameters are shown with * in the table. We have not conducted the experiments for the unrealistic case where 0.75 of all arrivals are type 1 and 0.2 of all arrivals are discharged after triage.

The results indicate that when the rate of type 1 patients decreases, it is beneficial to increase the number of triage nurses instead of doctors. On the other hand, when type 1 patients increase, the numbers of both triage nurses and doctors should be increased.

7. Conclusions

A&E's are the first point of contact for urgent and complex cases. The performance targets of A&E's have long been to reduce the patient waiting times below 4 h. As the hospitals have failed to satisfy this target, the government is planning to adopt alternative policies such as considering the waiting time targets for only serious cases. The staff planning in A&E's affect the performance levels critically. Due to the uncertainties involved in the A&E services, finding optimum capacities satisfying the performance targets is difficult with classical methods. This paper proposes to use a robust optimization based approximation for the computation of the worst-case waiting times in an A&E where patient are first triaged and then prioritized based on the urgency. We first model the problem with the approximation and then show that the model can be solved to optimality with the commercial solvers. We also develop an SO based heuristic where we use the approximation for the maximum waiting time in the search for a better feasible solution.

The computational experiments show that the AO outperforms the SO heuristic in terms of the computation time and the solution performance. The advantage of both the AO and the SO heuristic is their speed to provide an approximately good solution very quickly. As the problem complexity is increased with the waiting time limits for both patient types, the performance of the solutions obtained by the approximation drops when both patients are subject to waiting time limit. The experiments also indicate that the approximation approach still works well for different distribution assumptions for the treatment and triage durations. Another observation drawn from the experiment results is the non-linear effect of the waiting time limits on the solutions. The future studies may investigate the suitability of the approximation method for more complex A&E operations including more than two prioritization categories, the diagnostic test queues, etc.

References

[1] M.A. Ahmed and T.M. Alkhamis, Simulation optimization for an emergency department healthcare unit in Kuwait. *Eur. J. Oper. Res.* **198** (2009) 936–942.

[2] E. Alfonso, X. Xie, V. Augusto and O. Garraud, Modelling and simulation of blood collection systems: improvement of human resources allocation for better cost-effectiveness and reduction of candidate donor abandonment. *Vox Sang.* **104** (2013) 225–233.

[3] A.O. Allen, Probability, Statistics, and Queueing Theory. Academic Press (2014).

[4] M.H. Alrefaei and A. Diabat, Modelling and optimization of outpatient appointment scheduling. *RAIRO: RO* **49** (2015) 435–450.

[5] M. Asaduzzaman, T.J. Chaussalet and N.J. Robertson, A loss network model with overflow for capacity planning of a neonatal unit. *Ann. Oper. Res.* **178** (2010) 67–76.

[6] Audit General for Scotland: Emergency Departments. Tech. Rep., Audit General for Scotland (2010). Available from: http://www.audit-scotland.gov.uk/uploads/docs/report/2010/nr_100812_emergency_departments.pdf.

[7] C. Bandi and D. Bertsimas, Tractable stochastic analysis in high dimensions via robust optimization. *Math. Program.* **134** (2012) 23–70.

[8] P. Bonami, M. Kilinç and J. Linderoth, Algorithms and software for convex mixed integer nonlinear programs. In: Mix. integer nonlinear Program. Springer, New York (2012).

[9] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press (2004).

[10] K.M. Bretthauer, H.S. Heese, H. Pun and E. Coe, Blocking in healthcare operations: a new heuristic and an application. *Prod. Oper. Manag.* **20** (2011) 375–391.

[11] I. Castillo, A. Ingolfsson and T. Sim, Social optimal location of facilities with fixed servers, stochastic demand, and congestion. *Prod. Oper. Manag.* **18** (2009) 721–736.

[12] T.L. Chen and C.C. Wang, Multi-objective simulation optimization for medical capacity allocation in emergency department. *J. Simul.* **10** (2016) 50–68.

[13] J.K. Cochran and K.T. Roche, A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Comput. Oper. Res.* **36** (2009) 1497–1512.

[14] Computational Infrastructure for Operations Research: Bonmin (2018). Available from: https://www.coin-or.org/Bonmin/index.html.

[15] A.X. Costa, S.A. Ridley, A.K. Shahani, P.R. Harper, V. De Senna and M.S. Nielsen, Mathematical modelling and simulation for planning critical care capacity. *Anaesthesia* **58** (2003) 320–327.

[16] S. Creemers and M. Lambrecht, An advanced queueing model to analyze appointment-driven service systems. *Comput. Oper. Res.* **36** (2009) 2773–2785.

[17] V. De Angelis, G. Felici and P. Impelluso, Integrating simulation and optimisation in health care centre management. *Eur. J. Oper. Res.* **150** (2003) 101–114.

[18] H. Eskandari, M. Riyahifard, S. Khosravi and C.D. Geiger, Improving the emergency department performance using simulation and MCDM methods. In: *Proc. winter Simul. Conf. Winter Simulation Conference* (2011) 1211–1222.

[19] S. Flessa, Where efficiency saves lives: a linear programme for the optimal allocation of health care resources in developing countries. *Health Care Manag. Sci.* **3** (2000) 249–267.

[20] A. Fletcher, D. Halsall, S. Huxham and D. Worthington, The DH Accident and Emergency Department model: a national generic model used locally. *J. Oper. Res. Soc.* **58** (2007) 1554–1562.

[21] S. Fomundam and J.W. Herrmann, A Survey of Queuing Theory Applications in Healthcare. University of Maryland, College Park (2007).

[22] F. Fruggiero, A. Lambiase and D. Fallon, Computer simulation and swarm intelligence organisation into an emergency department: a balancing approach across ant colony optimisation. *Int. J. Serv. Oper. Inf.* **3** (2008) 142–161.

[23] K. Ghanes, M. Wargon, O. Jouini, Z. Jemai, A. Diakogiannis, R. Hellmann, V. Thomas and G. Koole, Simulation-based optimization of staffing levels in an emergency department. *Simulation* **91** (2015) 942–953.

[24] M. Gourgand, N. Grangeon and N. Klement, Activities planning and resources assignment on distinct places: a mathematical model. *RAIRO: OR* **49** (2015) 79–98.

[25] R. Govind, R. Chatterjee and V. Mittal, Timely access to health care: customer-focused resource allocation in a hospital network. *Int. J. Res. Mark.* **25** (2008) 294–300.

[26] L.V. Green, P.J. Kolesar and W. Whitt, Coping with time-varying demand when setting staffing requirements for a service system. *Prod. Oper. Manag.* **16** (2007) 13–39.

[27] S. Gul, B.T. Denton and J.W. Fowler, A progressive hedging approach for surgery planning under uncertainty. *Inf. J. Comput.* **27** (2015) 755–772.

[28] E.D. Güneş and H. Yaman, Health network mergers and hospital re-planning. *J. Oper. Res. Soc.* **61** (2010) 275–283.

[29] V. Gupta and T. Osogami, On Markov–Krein characterization of the mean waiting time in M/G/K and other queueing systems. *Queueing Syst.* **68** (2011) 339–352.

[30] P.R. Harper, N.H. Powell and J.E. Williams, Modelling the size and skill-mix of hospital nursing teams. *J. Oper. Res. Soc.* **61** (2010) 768–779.

[31] X. Hu, S. Barnes and B. Golden, Applying queueing theory to the study of emergency department operations: a survey and a discussion of comparable simulation studies. *Int. Trans. Oper. Res.* **25** (2018) 7–49.

[32] P.J.H. Hulshof, M.R.K. Mes, R.J. Boucherie and E.W. Hans, Tactical planning in healthcare using approximate dynamic programming. Tech. Rep. University of Twente (2013).

[33] I.M. Ibrahim, C.Y. Liong, S.A. Bakar, N. Ahmad and A.F. Najmuddin, Estimating optimal resource capacities in emergency department. *Indian J. Public Heal. Res. Dev.* **9** (2018).

[34] N. Izady and D. Worthington, Setting staffing requirements for time dependent queueing networks: the case of accident and emergency departments. *Eur. J. Oper. Res.* **219** (2012) 531–540.

[35] J. Köllerström, Heavy traffic theory for queues with several servers. I. *J. Appl. Probab.* **11** (1974) 544–552.

[36] C. Lakshmi and S.A. Iyer, Application of queueing theory in health care: a literature review. *Oper. Res. Heal. Care* **2** (2013) 25–39.

[37] B. Lunenfeld and P. Stratton, The clinical consequences of an ageing world and preventive strategies. *Best Pract. Res. Clin. Obstet. Gynaecol.* **27** (2013) 643–659.

[38] S. Mahar, K.M. Bretthauer and P.A. Salzarulo, Locating specialized service capacity in a multi-hospital network. *Eur. J. Oper. Res.* **212** (2011) 596–605.

[39] L. Mayhew and D. Smith, Using queuing theory to analyse the government's 4 h completion time target in accident and emergency departments. *Health Care Manag. Sci.* **11** (2008) 11–21.

[40] Z. Mingzhu and Q. Ershi, A multi-type queuing network analysis method for controlling server number in the outpatient. *Open Autom. Control Syst. J.* **8** (2016).

[41] S. Mohiuddin, J. Busby, J. Savović, A. Richards, K. Northstone, W. Hollingworth, J.L. Donovan and C. Vasilakis, Patient flow within UK emergency departments: a systematic review of the use of computer simulation modelling methods. *BMJ Open* **7** (2017) e015007.

[42] A. Mortimore and S. Cooper, The "4-hour target": emergency nurses' views. *Emerg. Med. J.* **24** (2007) 402–404.

[43] W. Munir, Critical analysis of the 4-hour A&E policy's impact on elderly patients. *Br. J. Nurs.* **17** (2008) 1188–1192.

[44] NHS England: AE Waiting times and activity. Tech. Rep. NHS England, London (2018). Available from: https://www.england.nhs.uk/statistics/statistical-work-areas/ae-waiting-times-and-activity/ae-attendances-and-emergency-admissions-2018-19/.

[45] NHS England: Commissioning Committee Report to Board. Tech. Rep. NHS England (2018). Available from: https://www.england.nhs.uk/wp-content/uploads/2018/11/13i-pb-28-11-18-commissioning-committeee-report-to-board-24-october.pdf.

[46] NHS UK: Hospital Accident and Emergency Activity, 2017–18, Provider Level Analysis. Tech. rep., National Health Services UK (2018). Available from: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-accident--emergency-activity/2017-18

[47] NHS UK: Pay for Doctors (2019). Available from: https://www.healthcareers.nhs.uk/explore-roles/doctors/pay-doctors.

[48] NHS UK: Urgent and Emergency Care: When to go to A&E (2019). Available from: https://www.nhs.uk/using-the-nhs/nhs-services/urgent-and-emergency-care/when-to-go-to-ae/.

[49] OptTek Systems: OptQuest (2019). Available from: https://www.opttek.com/products/optquest/.

[50] C. Pehlivan, V. Augusto, X. Xie and C. Crenn-Hebert, Multi-period capacity planning for maternity facilities in a perinatal network: a queuing and optimization approach. In: *Autom. Sci. Eng. (CASE). 2012 IEEE Int. Conf. IEEE* (2012) 137–142.

[51] F. Rico, E. Salari and G. Centeno, Emergency departments nurse allocation to face a pandemic influenza outbreak. In: *Proc. 39th Conf. Winter Simul. IEEE Press* (2007) 1292–1298.

[52] Royal College of Nursing: NHS pay scales 2017–18 (2019). Available from: https://www.rcn.org.uk/employment-and-pay/nhs-pay-scales-2017-18.

[53] S. Saghafian, G. Austin and S.J. Traub, Operations research/management contributions to emergency department patient flow optimization: review and research prospects. *IIE Trans. Healthc. Syst. Eng.* **5** (2015) 101–123.

[54] P. Santibáñez, G. Bekiou and K. Yip, Fraser Health uses mathematical programming to plan its inpatient hospital network. *Interfaces (Providence)* **39** (2009) 196–208.

[55] C. Stummer, K. Doerner, A. Focke and K. Heidenberger, Determining location and size of medical departments in a hospital network: a multiobjective decision support approach. *Health Care Manag. Sci.* **7** (2004) 63–71.

[56] S.S. Syam and M.J. Côté, A location allocation model for service providers with application to not-for-profit health care organizations. *Omega* **38** (2010) 157–166.

[57] The Telegraph: One in three hospital patients admitted as an emergency to be sent home without an overnight stay (2019). Available from: https://www.telegraph.co.uk/news/2019/03/08/one-three-hospital-patients-admitted-emergency-sent-home-without/.

[58] H.C. Tijms, M.H. Van Hoorn and A. Federgruen, Approximations for the steady-state probabilities in the M/G/c queue. *Adv. Appl. Probab.* **13** (1981) 186–206.

[59] S.J. Weng, B.C. Cheng, S.T. Kwong, L.M. Wang and C.Y. Chang, Simulation optimization for emergency department resources allocation. In: *Proc. 2011 Winter Simul. Conf. (WSC). IEEE* (2011) 1231–1238.

[60] J.L. Wiler, R.T. Griffey and T. Olsen, Review of modeling approaches for emergency department patient flow and crowding research. *Acad. Emerg. Med.* **18** (2011) 1371–1379.