DYNAMIC AND FLEXIBLE STAFF DEPLOYMENT IN ACCIDENT AND EMERGENCY DEPARTMENTS USING SIMULATION-BASED OPTIMIZATION

Elvan Gokalp

IDO, School of Management
University of Bath
Bath, UK
Corresponding author's e-mail: eg768@bath.ac.uk

Accident and emergency departments experience overcrowding due to staff shortages as well as to variations in patient arrivals and the time required to treat them. Several policies have been developed by hospitals to ensure that patients are not put at clinical risk during overcrowding. These policies suggest flexing nurses from different duties to the overcrowded section. However, the policies do not indicate the details of when exactly the flexing should be activated. We develop a mathematical model to find the optimum levels of triage and treatment queue lengths after which flexing should be activated. The performance indicators of the department are the waiting time targets and the disturbance due to nurse flexing. Because of the lack of closed-form formulations, we propose simulation optimization to solve the problem. By analyzing the model structure, we develop an efficient search procedure of the discrete solution space. We show the application of the proposed method using the data of a large hospital in the UK under different parameter settings. The results show that hospital management should focus on increasing the number of treatment nurses rather than flexing the nurses, and the queue of the service stream that requires tighter staffing should be controlled by an upper limit.

Keywords: Healthcare modelling, Simulation Optimization, Accident and Emergency Services, Discrete-event Simulation, Staff Planning

(Received on February 1, 2020; Accepted on July 3, 2021)

1. INTRODUCTION

Accident and emergency (A&E) department are the first point of contact for most of complex and life-threatening cases such as heart attack, stroke, or loss of consciousness. In an A&E department, the severity of a case is initially assessed by a triage nurse, so a senior nurse, who categorizes them as major, minor, and life-threatening (Gunal & Pidd, 2006). The last category of patients (life-threatening) are treated immediately. Following triage, the other two categories of patients wait in the waiting areas to be seen by one of the clinicians. Major cases are prioritized over minor ones, while both categories are served based on a first-come, first-served (FCFS) policy.

Commented [GML1]: What is major?

Commented [EG2R1]: Triage categorization, mentioned couple of lines ago.

Overcrowding is a significant problem in A&E departments (Proudlove, Gordon, & Boaden, 2003). Along with the overall increase in healthcare service demand, emergency unit arrivals rose by 28% between 2002 and 2017 in England (NHS England, 2018a). To improve the efficiency of A&E departments, the UK government introduced a 4-hour policy in 2004, requiring all patients in A&E to be treated within 4 hours of their arrival. In February 2018, the 4-hour target was suspended (NHS England, 2018b) but the hospitals are still obliged to report their A&E waiting times.

Two significant causes of A&E overcrowding are the variations in demand and the time required to treat patients. These variations may result in very long waiting times and put patients' health and well-being at risk. To deal with the periods of overcrowding, the National Institute of Health and Care Excellence (NICE), the leading guidance and regulatory agency of the healthcare system in the UK, has suggested developing ***escalation*** protocols (National Institute for Health and Care Excellence (NICE), 2016). These protocols outline action plans for the crowding in A&E that may lead to excessive waiting times and delay life-saving treatments. A common escalation action is to flex the nurses between different units, e.g. triage and treatment, to relieve the bottlenecks (Back et al., 2017). On the other hand, flexing may cause disruption on to the service that the staff are flexed from and therefore, should be planned carefully.

Although various possible actions are listed in the escalation plans, the triggers, i.e., when a specific action should be activated, is not provided with enough clarity and detail. Some hospitals define triggers based on total number of patients waiting in A&E or the number of arrivals (Portsmouth Hospitals NHS Trust, 2018). However, an aggregate level of utilization may not be an efficient indicator of a bottleneck located in a specific stream such as triage. Besides, the triggers are, in practice, selected in a rather ad-hoc manner. Yet, the selection of the triggers is actually a complex decision-making problem involving uncertainties in both demand and treatment time. Besides, such decisions should also take the interdependencies between the service streams into account.

This paper proposes a simulation-based optimization approach to find the best trigger levels in different service streams of A&E. We specifically consider two performance measures to evaluate a solution: (i) the percentage of patients waiting more than a certain threshold and (ii) the possible disturbance due to flexing. We develop a simulation model of A&E services including a triage queue followed by a treatment queue. Due to the lack of closed form formulations for this type of queues, we solve the problem with a simulation optimization (SO) approach. We assume a time-dependent arrival rate to A&E.

To the best of our knowledge, this problem has not been studied yet. The contributions of the paper can be summarized as developing an efficient solution approach to find the best escalation trigger levels in an A&E. More specifically, the solution approach can find a good enough solution within a minute. We also conduct several computational experiments using the data of University Hospitals Coventry and Warwickshire in the UK to show the performance of the proposed approach and to investigate the impact of the model parameters. The proposed approach can be applied into any service with two successive queues and waiting time targets, such as bank branches. The paper is organized as follows. Section 2 provides the problem description and underlying assumptions along with the optimization model formulation. Section 3 details the proposed SO algorithm. In Section 4, we introduce the design of experiments and results. Finally, Section 5 provides a summary of the study as well as the future research directions.

2. RELATED LITERATURE

This section provides an overview of the related literature on the modelling of A&E department using simulation and optimization. SO utilizes simulation to find the performance of a solution and is especially useful for complex problems with no analytical formulation. The optimization within SO can be conducted through several algorithms such as random search (Andradóttir, 2006), response surface methodology (Barton, 2013), and heuristics like genetic algorithm (Yeh & Lin, 2007), tabu search (Henderson & Nelson, 2006).

Mohiuddin et al. (2017) 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 & Guneri (2015). Among 106 reviewed papers, only few studies consider an optimization approach (Ahmed & Alkhamis, 2009; Eskandari, Riyahifard, Khosravi, & Geiger, 2011; Feng, Wu, & Chen, 2017; Fruggiero, Lambiase, & Fallon, 2008; Ko, Song, Morrison, & Hwang, 2014; Rico, Salari, & Centeno, 2007; Weng, Cheng, Kwong, Wang, & Chang, 2011; Yeh & Lin, 2007). A common theme of the studies is to find the optimum number of staffing in A&E to minimize patient waiting times (Feng et al., 2017; Ghanes et al., 2015) and medical resources wasted (T. L. Chen & Wang, 2016). Several authors (Ibrahim, Liong, Bakar, Ahmad, & Najmuddin, 2018; Rico et al., 2007; Weng et al., 2011) use OptQuest, an SO engine, to find the optimum staffing levels. Other simulation software utilized is IDS Scheer ARIS™ and Rockwell Arena (Wang, Guinet, Belaidi, & Besombes, 2009). Fruggiero et al., (2008) use ant-colony optimization along with a simulation model to optimize the resources in an emergency department. Distinct from the others, as well as modelling an A&E, Rabbani, Farshbaf-Geranmayeh, & Yazdanparast (2018) consider the departments interacting with the A&E such as radiology and pharmacy. They have used data envelopment analysis, a multi-layer perceptron artificial neural network, and radial basis functions for purposes of optimization. An emergency department is modelled by discrete-event simulation in Gul & Guneri (2012). By utilizing the DES model, they have compared the performance of various resource allocation scenarios. DES has also been used by Joshi & Rys (2011) to study the impact of arrival patterns on the performance of an A&E.

Although there are plenty of studies utilizing SO for unconstrained models of A&E planning, the literature on the stochastic constrained problems is limited (Ahmed & Alkhamis, 2009; W. Chen, Guo, & Tsui, 2019; Diefenbach & Kozan, 2011; Guo, Gao, Tsui, & Niu, 2017; Zeinali, Mahootchi, & Sepehri, 2015). Guo et al. (2017) propose a search method based on computing the objective values of all possible solutions. This method would be very time-consuming on the problems where the objective function value can only be estimated through simulation models. W. Chen et al. (2019) present an SO method for optimization of medical staff allocation in an A&E. The objective of the model is to meet the target performance measures regarding the response times to critical patients. They utilize simulated annealing to solve the stochastic constrained, discrete optimization problem. Zeinali et al. (2015) develop a simulation model of an A&E to minimize the average waiting time for patients subject to budget and capacity constraints. They use meta-modelling to replace the simulation model with an efficient metamodel to find the best medical staff configuration. Diefenbach & Kozan (2011) use a software, Extend and its optimization toolbox, to model the operations in an A&E and to optimize the bed configurations taking into account waiting time targets.

Ahmed & Alkhamis (2009) present an SO approach for capacity planning of an emergency department in Kuwait. They consider stochastic constraints in a problem with a discrete solution space. They also model and solve the optimization problem where the objective function is total cost of the staffing, and the constraints are the waiting times. Their algorithm first identifies a 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 surpasses a certain number of iterations. However, they do not provide any performance results for the proposed algorithm. Also, their method can be trapped into a local optimum.

Instead of performing a random search as in Ahmed & Alkhamis, (2009), we explore the structural properties of the model to find the optimum solution efficiently. The heuristic approaches such as genetic or tabu algorithm are not effective to solve the problem presented in this paper because of its structural properties (i.e., monotonicity) investigated in Section 5. Finally, to the best of our knowledge, there is no study focusing on the flexing rules in overcrowded A&Es.

3. PROBLEM DESCRIPTION AND MATHEMATICAL MODEL

This section first describes the underlying problem as well as our assumptions and then introduces the mathematical formulation. We model the activities in a typical (major) A&E department in the UK (NHS UK, 2019) for a finite planning horizon. Note that we do not consider the single specialty cases such as ophthalmology or dentistry which go through a separate route (NHS England, 2018b). 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 major, minor, or life-threatening which group receives immediate care (Gunal & Pidd, 2006). Major and minor categories have separate queues for treatment while the major ones have priority over the minor. There are only certain number of treatment cubicles that contain special beds for major patients. When all treatment cubicles are full, the patient is kept waiting in some other place, e.g., a buffer ward or the triage waiting area. After the patients are treated, they are either discharged, referred, or admitted to the hospital.

When the A&E is overloaded, an escalation plan is activated that involves flexing nurses between treatment and triage areas. However, the nurses who can be flexed, denoted with S, are the senior ones with multi-tasking that exist in limited numbers in A&Es, i.e. around 46% of all (Harper, Powell, & Williams, 2010). The escalation plan is activated when the queue for triage or treatment goes above certain levels. These levels are denoted with α and θ for the triage and the treatment queues, respectively. When the triage queue length is above α, one of the flexible nurses $s \in 1, \ldots, S$ in the treatment is called to triage. The flexed nurse (in the treatment) that finishes her current task earliest answers the call and placed into triage as a new server. If the queue length is still above α, $\bar{t}$ minutes after the flexible nurse started triage, another flexible nurse is called. Similarly, when the number of major patients waiting for the treatment in the buffer area goes above θ, one of the flexible nurses currently in triage is deployed to the treatment as a new server. If there are no flexible nurses left in the other service, then the call cannot be answered. When the length of the respective queue drops to zero, the flexed nurse(s) is deployed back to his/her original stream. Figure 1 shows the flexing rules for the triage queue, where Q_t represents the length of triage queue at period t. Same rules apply to the treatment queue.

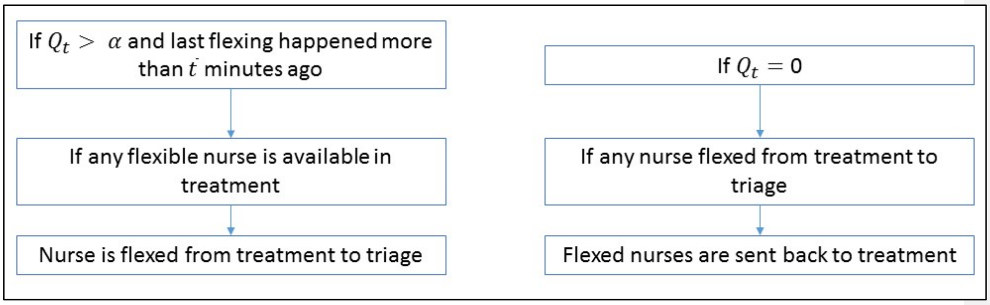

If $Q_t >$ α and last flexing happened more than $\bar{t}$ minutes ago
If any flexible nurse is available in treatment
Nurse is flexed from treatment to triage
If $Q_t = 0$
If any nurse flexed from treatment to triage
Flexed nurses are sent back to treatment

Figure 1. Flexing rules for triage queue

Each time a flexible nurse is deployed to the other service, a disruption may occur as that nurse may be required to give some clinical information regarding the patients seen previously by him/her (Back et al., 2017). Therefore, the flexing should be minimized. Average number of times that a nurse is flexed during the planning period is denoted with n.

As mentioned earlier, an important performance measure is the percentage of patients staying more than 4 hours in the A&E. Let us W denote the 95% upper confidence limit of the patients' length of stay arrived in the A&E within the planning horizon. The uncertainties affecting the waiting times are the arrival times, and the duration of both triage and treatments. Although arrival times can follow a time-dependent exponential distribution (Ahmed & Alkhamis, 2009), there is no consensus in the literature about the triage and treatment time distributions in A&E; triangular (Fletcher, Halsall, Huxham, & Worthington, 2007), general (Izady & Worthington, 2012), exponential (Saghafian, Austin, & Traub, 2015) and uniform (Ahmed & Alkhamis, 2009) distributions are used to model the A&E triage and treatment durations.

The number of nurses involved in the triage and treatment during the planning horizon are denoted with x_1 and x_2, respectively. We assume that the staffing is fixed during the planning horizon, i.e., a single shift. The model can easily be extended to varying staff levels, by adding time-dependency to x_1 and x_2. Also note that we do not model the doctor levels in the A&E since they are not flexed between triage and treatment. Besides, the recent figures show that NHS has a bigger shortage of nurses than doctors (The Health Foundation, 2019). Therefore, possibly, the number of nurses is a bottleneck and defines the waiting times. If the number of doctors is not enough, that may create a bottleneck in the treatments. In such cases, doctors from other departments may be called on duty. However, this escalation activity is not the scope of this paper.

A certain percentage, η of patient arrivals are categorized as major. The number of major treatment cubicles is denoted with N. We assume that the patients arrive to the A&E with the mean inter-arrival rate λ_t and standard deviation σ_t at time period t. After registration, they wait in the triage queue and are assessed by a triage nurse under the FCFS rule. The mean triage time is $1/\mu$ with standard deviation σ_1. The average treatment time for major cases is $1/\mu_2$ with standard deviation σ_2. We assume that other medical activities required for the treatment such as laboratory tests are included in the treatment duration. Finally, the target length of stay, i.e., 4 hours, is denoted with $\overline{W}$. Below, we summarize several modelling assumptions.

Assumptions:

- A modelling choice is related to `boarding' which refers to the cases where an admitted patient is kept waiting in A&E due to the lack of available bed on the other hospital wards. 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 (Munir, 2008). 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.
- The only escalation action considered in the model is the flexing of nurses. There may be other escalation actions such as discharging patients from beds or calling extra staff from other departments within hospital. However, these actions have serious disadvantages and are usually used if the flexing cannot handle the overcrowding (Back et al., 2017).
- Another assumption is related to the patient categories. The major cases have priority over minors, and they may share the clinicians. Since the majors always have priority, take significantly longer to treat (Ahmed & Alkhamis, 2009) and breach the waiting time targets (Gunal & Pidd, 2006), we only model the waiting times of the major cases. In other words, even if we would model the queue for minor cases, it would not affect the waiting time performance significantly. This assumption is tested in Section 5.1. Besides, the NHS plans to impose the waiting time limits only for major instead of all patient groups (NHS, 2019).
- There may be more complicated flexing rules such as re-deploying the nurses when queue lengths go below a certain level rather than zero. However, the escalation plans do not specify when the flexing should be stopped.

3.1. Mathematical Model

The A&E managers have two main objectives: (1) decrease the average frequency of the nurse flexing, $n(\alpha,\theta)$, (2) maintain a reasonable performance measured by the waiting times, especially by $W(\alpha,\theta)$. These two objectives conflict each other, i.e., if the nurses are flexed less, the waiting times will increase. The factors that the managers can vary to achieve these objectives are the queue limits at which flexing is activated; $\alpha \epsilon \alpha^b,\ldots,\alpha^f$ and $\theta \epsilon \theta^b,\ldots,\theta^f$, where α^b, α^f and θ^b,θ^f represent the minimum and maximum possible levels for α and θ, respectively. These boundaries can be set based on the historical data of the queue lengths. The stochastic programming model can be defined as:

$$\begin{aligned} &\min n(\alpha,\theta)\,, \\ &s.t. \overline{W} \geq W(\alpha,\theta), \\ &\alpha \,\epsilon\, \alpha^b,\ldots,\alpha^f, \theta \,\epsilon\, \theta^b,\ldots,\theta^f. \end{aligned} \qquad (1)$$

In this model, the waiting times are limited with an upper bound, $\overline{W}$, that can be set based on the waiting time targets of National Health Service (NHS) in the UK, such as 95% of patients should spend less than 4 hours in A&E. Note that both waiting times and the number of flexing are stochastic variables. In other words, the problem is a constrained problem with a stochastic objective function. 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 is difficult even with fixed queue limits, α and θ. The computational intractability is already shown for a queue of multiple servers with exponential arrivals and general service time distribution (Tijms, Van Hoorn, & Federgruen, 1981). An alternative approach to compute the waiting times and the number of flexing is to use simulation modelling. The details of this approach are provided in the next section.

4. SIMULATION OPTIMIZATION

Our literature review shows that most of the SO studies employ built-in optimization packages within a commercial simulation modelling software program such as OptQuest in Simul8. These built-in packages are not very flexible and therefore, we design and implement an SO algorithm to solve the model proposed in the previous section. For this purpose, we first develop a simulation model of the A&E operations described in Section 3 and implement it on Matlab. Specifically, we model an FCFS queue for triage and move η% of the triaged patients to the treatment area which holds a limited number of patients N. If the treatment area is full, then the patients are put into another FCFS queue, defined as treatment queue. 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 z scenarios in which uncertain parameters are randomly generated from the corresponding distributions. Below we explain the details of the SO approach.

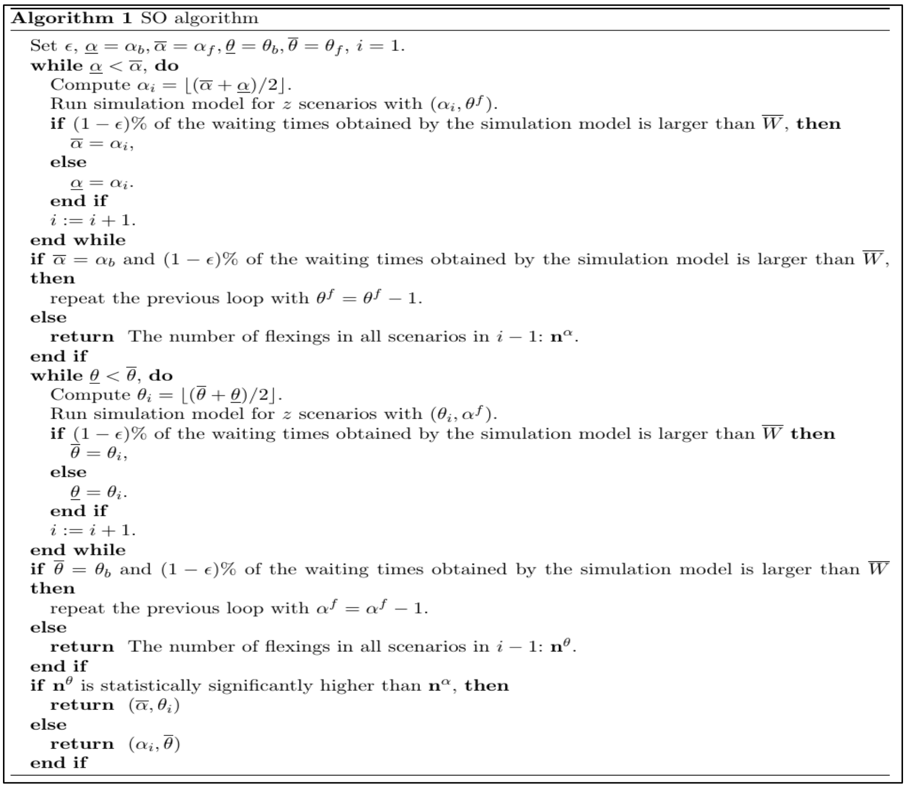
Note that the objective function (1) is monotonic with respect to each variable. To see the monotonicity, consider the case with a fixed θ. As α decreases, the amount of flexing would expected to increase. On the other hand, the length of the treatment queue may increase as nurses are flexed more often, which may result in even more flexing depending on θ. Therefore, the number of flexing always increases where there is a decreasing of α. The same logic applies to the case where α is fixed instead of θ. Due to the monotonocity, we do not need to search for the optimum solution randomly as in Ahmed & Alkhamis, (2009); we can use line search methods. Let us show the solution space with a matrix,

$$H = \begin{bmatrix} \alpha^b, \theta^b & \cdots & \alpha^b, \theta^f \\ \vdots & \ddots & \vdots \\ \alpha^f, \theta^b & \cdots & \alpha^f, \theta^f \end{bmatrix}.$$

Note that since the objective function is monotonic in the model, for a fixed row H, the optimum solution is always in the last column for an unconstrained version. Similarly, for a fixed column, the optimum solution is always in the last row for an unconstrained version. Therefore, the search for the optimum solution should always be in the last row and the last column of the matrix H. However, if none of the solutions in these edges are feasible, then the search should move into the next column or row.

Since the objective function is monotonic, we can also use section search methods with respect to each variable. Once the optimum solution is found for one variable (while the other is fixed) and vice versa, we can compare the optimum objective values of each and choose the solution that gives the minimum of two objective values. For a fixed decision variable, we conduct a bisection search on the feasible range of the other variable to find the minimum objective function value while satisfying the waiting time constraint (1). Note that since the decision variables are integer, the smallest value they can take is 1; $\alpha^b = 1$. The largest value, α^f can be determined by the decision-makers as a large enough value based on the historical data where queue lengths are recorded. A bisection search divides the variable range into two halves at each iteration and the feasibility of the new boundaries are then checked.

For the sake of clarity, let us consider finding the optimum α level. Note that θ is first fixed to its maximum value, θ^f since we know that the optimum solution is on the edges of the feasible solution space. The aim is to find α level that satisfies (1) with minimum n. First, the infeasible and feasible bounds are set to $\bar{\alpha} = \alpha^f$ and $\underline{\alpha} = \alpha^b$. The number of iterations, denoted with I is set to a large enough number based on the ranges of variables, thus, the possible number of iterations. In each iteration $i = 1, \dots, I$ of the algorithm, first α_i is set to $\lfloor (\underline{\alpha} + \bar{\alpha})/2 \rfloor$ and $W(\alpha_i)$ and $n(\alpha_i)$ are found with the simulation. If $(1 - \varepsilon)$ % of the patients' total waiting time is lower than $\bar{W}$, then $\underline{\alpha} = \alpha_i$, otherwise, $\bar{\alpha} = \alpha_i$. Therefore, the solution range is divided into two at each iteration. The search stops when $\bar{\alpha} = \underline{\alpha}$ and the resulting objective function values for all scenarios $\boldsymbol{n}^{\alpha}$ are reported. If there is no feasible solution in this range, then the same procedure is repeated for a fixed $\theta^f = \theta^f - 1$.



```
Algorithm 1 SO algorithm
  Set $\epsilon$, $\underline{\alpha} = \alpha_b, \overline{\alpha} = \alpha_f, \underline{\theta} = \theta_b, \overline{\theta} = \theta_f$, $i = 1$.
  while $\underline{\alpha} < \overline{\alpha}$, do
    Compute $\alpha_i = \lfloor(\overline{\alpha} + \underline{\alpha})/2\rfloor$.
    Run simulation model for $z$ scenarios with $(\alpha_i, \theta^f)$.
    if $(1-\epsilon)\%$ of the waiting times obtained by the simulation model is larger than $\overline{W}$, then
      $\overline{\alpha} = \alpha_i$,
    else
      $\underline{\alpha} = \alpha_i$.
    end if
    $i := i + 1$.
  end while
  if $\overline{\alpha} = \alpha_b$ and $(1-\epsilon)\%$ of the waiting times obtained by the simulation model is larger than $\overline{W}$,
  then
    repeat the previous loop with $\theta^f = \theta^f - 1$.
  else
    return The number of flexings in all scenarios in $i-1$: $\mathbf{n}^\alpha$.
  end if
  while $\underline{\theta} < \overline{\theta}$, do
    Compute $\theta_i = \lfloor(\overline{\theta} + \underline{\theta})/2\rfloor$.
    Run simulation model for $z$ scenarios with $(\theta_i, \alpha^f)$.
    if $(1-\epsilon)\%$ of the waiting times obtained by the simulation model is larger than $\overline{W}$ then
      $\overline{\theta} = \theta_i$,
    else
      $\underline{\theta} = \theta_i$.
    end if
    $i := i + 1$.
  end while
  if $\overline{\theta} = \theta_b$ and $(1-\epsilon)\%$ of the waiting times obtained by the simulation model is larger than $\overline{W}$
  then
    repeat the previous loop with $\alpha^f = \alpha^f - 1$.
  else
    return The number of flexings in all scenarios in $i-1$: $\mathbf{n}^\theta$.
  end if
  if $\mathbf{n}^\theta$ is statistically significantly higher than $\mathbf{n}^\alpha$, then
    return $(\overline{\alpha}, \theta_i)$
  else
    return $(\alpha_i, \overline{\theta})$
  end if
```

Figure 2. The pseudo-code for Simulation Optimization algorithm

In the next phase, the model is solved for θ with a fixed α^f following the same steps. After the optimum θ is found, the objective function values $\boldsymbol{n}^\alpha$ and $\boldsymbol{n}^\theta$ are compared with a statistical significance test. If $\boldsymbol{n}^\theta$ is statistically significantly higher than $\boldsymbol{n}^\alpha$, then the optimum solution is (α^f, θ_l); otherwise, it is (α_l, θ^f). Algorithm 1 presents the pseudo-code of the SO algorithm for the model.

The computational time of the algorithm depends on the running time of the simulation model, and thus, the number of scenarios in the simulation and the length of the planning period T. As the number of scenarios increases, the robustness of the solution obtained by the algorithm increases as well. For a reasonable size of variable ranges of 40, the number of iterations would be 6 for each decision variable, leading to 12 runs of the simulation model, where each run has a fixed number of scenarios, z. On the other hand, the enumeration of all possible solutions would require running the simulation model 1600 times. Therefore, the algorithm is expected to significantly decrease the computational time required.

5. COMPUTATIONAL EXPERIMENTS

The computational experiments aim to illustrate the performance of the SO algorithm 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 provides the results of the model under different parameter settings. The second set of experiments analyses the performance of a particular solution with respect to queue dynamics. All computational experiments are carried out on a PC with a Windows 10 Enterprise operating system, CPU 4GHz Intel Core i7 and 32GB of RAM.

5.1. Input Data

We use the hourly arrival data of University Hospitals Coventry & Warwickshire provided in the online resources of the NHS UK (NHS UK, 2018) and presented in Appendix **Error! Reference source not found.**. The rest of the data are presented in Table 1. The distribution of inter-arrival times is assumed to be exponential for each hour (Ahmed & Alkhamis, 2009). We have considered the planning horizon as the day shift, i.e., 8 am. - 6.30 pm. which has the same number of nurses throughout. The average treatment and triage times are obtained from (Ahmed & Alkhamis, 2009) and are assumed to follow an exponential distribution (Saghafian et al., 2015). To ensure validity, we have also examined the results obtained by assuming a uniform distribution for treatment and triage time and observed no significant impact on the distribution on the results.

Some of the data specific to the hospital are collected through mining expert opinion and in person visits to the hospital. Other parameters such as number of triage, treatment and flexible nurses vary in different seasons of the year, and thus, we conduct the experiments for a plausible range of these parameters estimated by experts. The number of scenarios in each simulation run is 500. Finally, ε is set to 0.05 throughout the experiments.

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

Description of Parameter	**Value**	**Source of Data**
A&E arrival rates	Time-dependent	(NHS UK, 2018)
Mean triage duration	15 min.	(Ahmed & Alkhamis, 2009)
Mean treatment duration	90 min.	(Ahmed & Alkhamis, 2009)
Number of treatment beds	12	Expert opinion
Buffer period between consecutive flexing	15 min.	Expert opinion
Waiting time limit	4 hours	(NHS England, 2018b)
Percentage of major patients	61%	(Ahmed & Alkhamis, 2009)
Ranges for variables ($[\alpha^f, \alpha^b]$, $[\theta^f, \theta^b]$)	[4, 40], [4, 40]	Expert opinion

5.2. Validation

The outputs of the simulation model are compared with the real data to ensure that the simulation model reflects the actual process. Specifically, we obtain the simulation results of *percentage of patients staying less than 4 hours*, *average length of stay*, and *average time to triage from arrival*. The real values of those parameters in 2017/2018 are collected from NHS England (2019), NHS UK (2018), and Ashraf (2019), respectively. Since these data sources have not differentiated the performance measures with respect to major and minor patient groups, we also obtain the results for these two patient groups combined. The number of triage, treatment and flexible nurses are assumed to be 6, 24, and 2, respectively. The real data and the simulation outputs (the mean and ± standard deviation over the scenarios) are presented in Table 2~~Table 2~~. The comparison of two sets of outputs indicates that the model is a good representation of the real-life setting.

Table 2. Comparison of real data and the simulation results for model validation

	% staying < 4 hours	**Mean length of stay (min.)**	**Mean time to triage (min.)**
Real	79%	164	9
Simulated	80% ± 0.44	168 ± 2.59	9 ± 0.19

5.3. Results

The initial available nurse levels, $x_1$and x_2, as well as the number of overall flexible nurses, S, may affect the solution significantly. Table 3~~Table 3~~ presents the solutions and the objective function value n along with the relevant ± standard deviation obtained by solving the model using SO for various levels of these parameters. We can report that the computational time of the algorithm is less than 1 minute in all parameter settings.

Table 3. Sensitivity analysis with respect to the numbers of initial and flexible nurses

		S					
Scenario	x_1, x_2	**4**		**6**		**8**	
1	6, 24	$\alpha = 7$	21.5 ± 3.16	$\alpha = 7$	22 ± 2.84	$\alpha = 10$	20 ± 2.52
2	7, 23	$\alpha = 7$	16.1 ± 1.02	$\theta = 7$	23.5 ± 1.89	$\theta = 15$	25 ± 3.05

3	8, 22	$\alpha = 7$	18 ± 2.21	$\theta = 7$	20 ± 1.89	$\theta = 8$	24 ± 1.65
		$\boldsymbol{\theta}$	$\boldsymbol{n}$	$\boldsymbol{\theta}$	$\boldsymbol{n}$	$\boldsymbol{\theta}$	$\boldsymbol{n}$
4	10, 20	-	-	-	-	8	11 ± 4.58
5	10, 21	8	2.61 ± 0.2	15	2.28 ± 0.18	18	2.89 ± 0.24
6	10, 22	7	2.2 ± 0.2	14	2.13 ± 0.19	17	2.95 ± 0.21
		$\boldsymbol{\alpha}$	$\boldsymbol{n}$	$\boldsymbol{\alpha}$	$\boldsymbol{n}$	$\boldsymbol{\alpha}$	$\boldsymbol{n}$
7	5, 26	7	5.5 ± 1.01	18	10 ± 0.8	20	10 ± 0.79
8	6, 26	7	2.4 ± 0.02	37	2.5 ± 0.01	37	2.9 ± 0.09

The flexible nurses are distributed equally between the treatment and triage initially, i.e., if there are four flexible nurses, 2 of them are included in $x_1$and x_2. The first three scenarios in Table 3~~Table 3~~ show the cases where total number of nurses is kept at 30, whereas scenarios 4 to 6 are for the cases with varying x_2 levels. Finally, scenarios 7 and 8 only differ in x_1. Due to the structure of the solutions, only one variable attains a varying level (between its maximum and minimum) while the other is at its maximum. For scenarios 4 to 6, the optimum solution was always defined with θ and α^fwhile in scenarios 7 and 8, it was defined by only α and θ^f. In scenario 4, the model was infeasible for $S = 4, 6$. Below, we analyze the results based on the effect of the initial and flexible nurse levels.

5.3.1. Effect of initial nurse levels

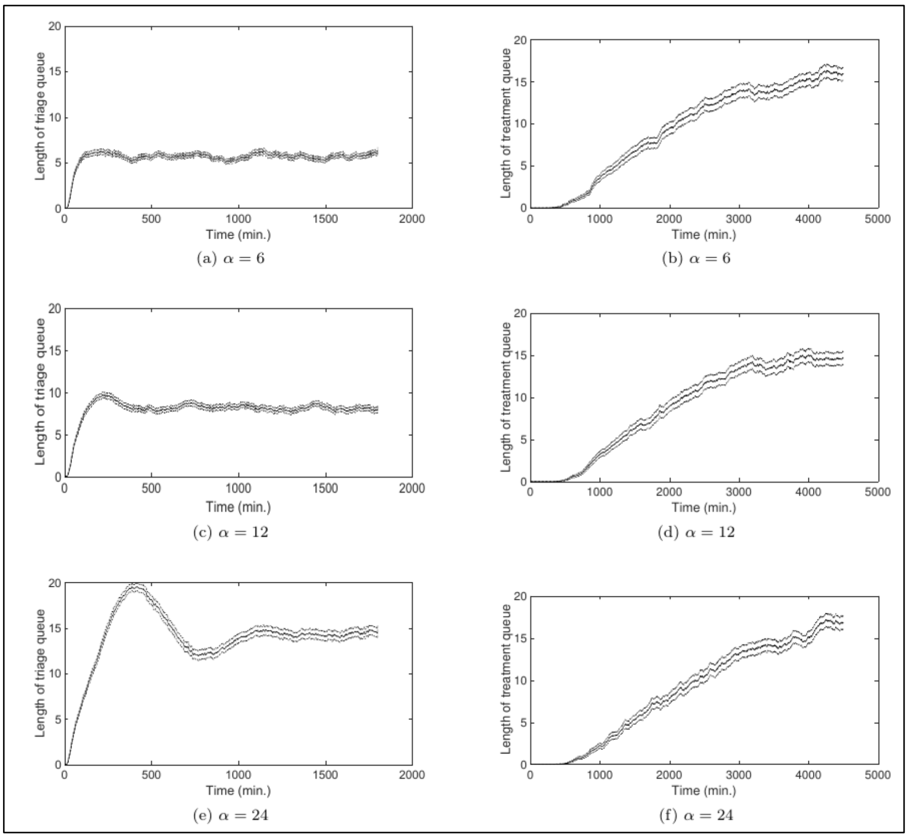
- Due to the high $x_1$and limited x_2 in scenarios 4 to 6, the solution depends only on θ. We observe a similar phenomenon in other experiments, the queue of the service stream that has more limited staff is kept under control to reach the waiting time target.
- Scenarios 7 and 8 show that a change in x_1 has a very significant effect on the solution, as different from a change in x_2.
- As the total number of nurses is increased, the objective value decreases. On the other hand, for the same number of total nurses (scenarios 1 to 3) and for fixed x_1 levels (scenarios 4 to 7), we do not see a significant difference in the objective function level for the same S levels. However, this does not hold for scenarios 7 and 8, where the objective value significantly decreases as x_1 increases. This again shows the significant effect of the level of x_1.

5.3.2. Effect of number of flexible nurses

- The trend with respect to S exhibits a nonlinear fashion and depends on initial staff levels. When the triage nurse level is limited, as in scenario 1, the increase in S from 4 to 6 does not affect the solution, whereas increase from 6 to 8 affects. Finally, when the treatment nurse is more limited (scenario 3), the change in S does not affect the solution significantly, probably because all the flexible nurses are used in all settings in this scenario. In scenarios 5 & 6, we see that as S is increased, θ also increases, indicating that a more flexible control is enough when there are more flexible nurses.
- Scenarios 5 and 6 have very close solutions for a fixed S. This indicates that increasing the treatment nurses from 21 to 22 does not affect the solutions significantly, possibly a larger increase is needed to see any significant effect.
- The objective function values are affected by S, the initial nurse levels and the solution. For example, the value of the objective function does not change for different S in scenarios 1 to 3, even though the solutions differ. On the other hand, in scenarios 7 and 8, the objective function value depends significantly on the solution. In scenarios 4 to 6, the increase in S from 6 to 8 results in a slightly higher objective function value. The main reason is that as there are more flexible nurses, then further flexing may not be necessary as the queue may start decreasing soon after. Note that the rule on re-flexing, i.e., after the queue goes to zero, has an impact on this observation.

In summary, the results indicate that A&E managers should put an upper limit on the queue of the section that has more limited staff. Another suggestion would be focusing on increasing the level of triage or treatment nurses instead of having more flexible staff. The results presented in Table 3~~Table 3~~ can be utilized by the A&E managers in several ways. First, they can compare the cost of increasing the number of triage or treatment nurses with the improvement which such a change would bring. Alternatively, they can set a certain upper threshold on the number of changes, n, and find the best staffing levels and queue limits to satisfy this threshold.

Next, we evaluate the performances of several solutions by running the simulation model for 1000 scenarios with the corresponding solution. Figure 3~~Figure 3~~ and Figure 4~~Figure 4~~ show the lengths of two queues along with the planning horizon for different α and θ levels (while the other is at its maximum level), for scenarios 1 and 6, respectively, and with $S = 6$. Note that the purpose of the second set of experiments is to present the queuing dynamics of A&E instead of comparing the results obtained in different scenarios. Since we would like to show the impact of changing both variables (α and θ) on the queue dynamics, we have chosen two scenarios where α and θ appear in the optimum solution, respectively. Another reason for choosing Scenario 1 is because its nurse staffing levels is the closest match with the actual situation. The solid and dashed lines respectively, show the average and ± standard errors of the queue lengths over all scenarios. The time horizon of these simulation runs is kept longer than the previous runs to be able to show the stable trend in queue lengths.



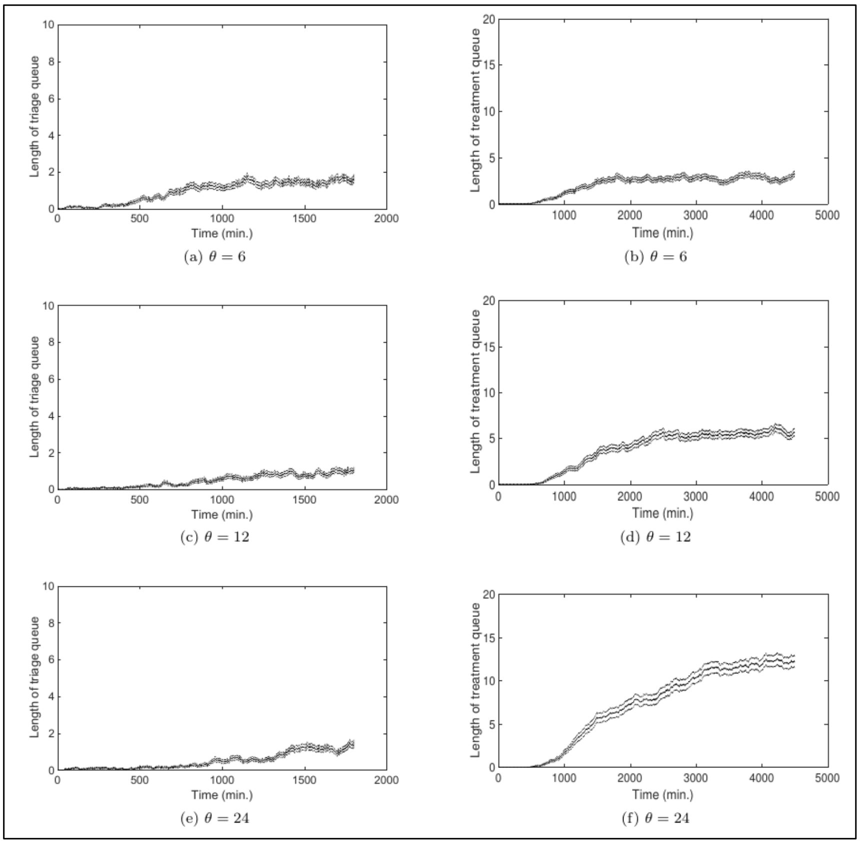
(a) $\alpha = 6$ (b) $\alpha = 6$

(c) $\alpha = 12$ (d) $\alpha = 12$

(e) $\alpha = 24$ (f) $\alpha = 24$

Figure 3. Simulation outputs (average and confidence interval) of the length of two queues for different α levels

Figure 3 shows that the triage queue length stabilizes around the threshold level α, except where α is large. The treatment queue increases slightly but stabilizes after a while. However, treatment queue on average is higher for a larger α. This result seems counterintuitive but is due to the flexing rules. For a larger α, flexing is made when the triage queue is already longer. Therefore, after the flexing, the time to send back the flexed nurse is longer which results in a slightly larger queue on the treatment service.



(a) $\theta = 6$

(b) $\theta = 6$

(c) $\theta = 12$

(d) $\theta = 12$

(e) $\theta = 24$

(f) $\theta = 24$

Figure 4. Simulation outputs (average and confidence interval) of the length of two queues for different θ levels

Similar to Figure 3, Figure 4 shows that treatment queue length is stabilized around θ level, except where θ is large. As different from the previous figure, triage queue shows a stable behavior in different θ levels. This is probably due to the high level of x_1 that is not significantly affected by the flexing.

6. CONCLUSION

The crowding in A&Es is a major problem in the UK. Hospitals have developed escalation policies to resolve overcrowding. A common escalation action is to move nurses to a crowded section. However, the levels of crowding after which the flexing needs to be introduced is a complex decision-making problem. This paper proposes a simulation-based optimization approach to find what length of queue provides a trigger event. By exploiting the model structure, we develop an efficient solution algorithm that can solve the model within a minute, i.e., the simulation-optimization results are obtained within a minute.

We conduct experiments for several scenarios in which the initial and flexible nurse levels are varied. The results show that the queue of the service that has a more limited staffing level should be controlled more tightly. We also see that the impact of changes in the triage nurse levels is more significant compared with that of the treatment nurses. Also, the increase in the number of flexible nurses is not always better and may require optimizing the re-flexing rules, i.e., when to send back the flexed nurse. The proposed method can be applied to any A&E with small amendments in the simulation model if necessary. The input dataset for the simulation model and the algorithm can be extracted from the historical data of the corresponding A&E. Then, the A&E's Operational Manager can enter the data and run the algorithm (directly in Matlab or using another interface designed for the purpose) when the nurse assignments for that day are ready and finalized, i.e., one or two days before the actual day of implementation. The algorithm would then provide the optimum queue trigger levels within a minute. When the respective day comes, these levels are used for the escalation rules presented in Figure 1. Due to the short computational time of the algorithm, even last-minute changes on the nurse schedule can be accommodated; the algorithm being then re-run to obtain the new solutions based on the most recent change in the nurse schedule.

Future studies may aim to find the optimum re-flexing rules, as well as modelling more escalation policies such as speeding the discharge process and comparing these alternative policies. Another line of research would be to incorporate more complexities into the treatment phase such as laboratory testing or the number of physicians. Finally, the approach, i.e., using simulation-optimization to find best queue trigger limits, can actually be applied to even more than two queues by simulating the additional queues as well. However, the structural properties (e.g., monotonicity) of this new problem need to be analyzed to see whether simulating a limited number of solutions (as in the two-queue case) would still be enough or not. This case can be the focus of the further studies.

REFERENCES

Ahmed, M. A., & Alkhamis, T. M. (2009). Simulation optimization for an emergency department healthcare unit in Kuwait. *European Journal of Operational Research*, *198*(3):936–942.

Andradóttir, S. (2006). An overview of simulation optimization via random search. *Handbooks in Operations Research and Management Science*, *13*:617–631.

Ashraf, S. (2019). *Exploration of Length of Stay Trends in Emergency Departments for Operational Efficiency*. The University of Warwick.

Back, J., Ross, A. J., Duncan, M. D., Jaye, P., Henderson, K., & Anderson, J. E. (2017). Emergency department escalation in theory and practice: a mixed-methods study using a model of organizational resilience. *Annals of Emergency Medicine*, *70*(5):659–671.

Barton, R. R. (2013). Response surface methodology. *Encyclopedia of Operations Research and Management Science*, 1307–1313.

Chen, T. L., & Wang, C. C. (2016). Multi-objective simulation optimization for medical capacity allocation in emergency department. *Journal of Simulation*, *10*(1):50–68.

Chen, W., Guo, H., & Tsui, K.-L. (2019). A new medical staff allocation via simulation optimisation for an emergency department in Hong Kong. *International Journal of Production Research*, 1–20.

Diefenbach, M., & Kozan, E. (2011). Effects of bed configurations at a hospital emergency department. *Journal of Simulation*, *5*(1):44–57.

Eskandari, H., Riyahifard, M., Khosravi, S., & Geiger, C. D. (2011). Improving the emergency department performance using simulation and MCDM methods. In *Proceedings of the winter simulation conference* (pp. 1211–1222). Winter Simulation Conference.

Feng, Y.-Y., Wu, I.-C., & Chen, T.-L. (2017). Stochastic resource allocation in emergency departments with a multi-objective simulation optimization algorithm. *Health Care Management Science*, *20*(1):55–75.

Fletcher, A., Halsall, D., Huxham, S., & Worthington, D. (2007). The DH Accident and Emergency Department model: a national generic model used locally. *Journal of the Operational Research Society*, *58*(12):1554–1562. https://doi.org/10.1057/palgrave.jors.2602344

Fruggiero, F., Lambiase, A., & Fallon, D. (2008). Computer simulation and swarm intelligence organisation into an emergency department: a balancing approach across ant colony optimisation. *International Journal of Services Operations and Informatics*, *3*(2):142–161.

Ghanes, K., Wargon, M., Jouini, O., Jemai, Z., Diakogiannis, A., Hellmann, R., … Koole, G. (2015). Simulation-based optimization of staffing levels in an emergency department. *Simulation*, *91*(10):942–953.

Gul, M., & Guneri, A. F. (2012). A computer simulation model to reduce patient length of stay and to improve resource utilization rate in an emergency department service system. *International Journal of Industrial Engineering*, *19*(5):221–231.

Gul, M., & Guneri, A. F. (2015). A comprehensive review of emergency department simulation applications for normal and disaster conditions. *Computers & Industrial Engineering*, *83*:327–344.

Gunal, M. M., & Pidd, M. (2006). Understanding accident and emergency department performance using simulation. In *Proceedings of the 38th conference on Winter simulation* (pp. 446–452). Winter Simulation Conference.

Guo, H., Gao, S., Tsui, K.-L., & Niu, T. (2017). Simulation optimization for medical staff configuration at emergency department in Hong Kong. *IEEE Transactions on Automation Science and Engineering*, *14*(4):1655–1665.

Harper, P. R., Powell, N. H., & Williams, J. E. (2010). Modelling the size and skill-mix of hospital nursing teams. *Journal of the Operational Research Society*, *61*(5):768–779.

Henderson, S. G., & Nelson, B. L. (2006). *Handbooks in operations research and management science: simulation* (Vol. 13). Elsevier.

Ibrahim, I. M., Liong, C.-Y., Bakar, S. A., Ahmad, N., & Najmuddin, A. F. (2018). Estimating Optimal Resource Capacities in Emergency Department. *Indian Journal of Public Health Research & Development*, *9*(11).

Izady, N., & Worthington, D. (2012). Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments. *European Journal of Operational Research*, *219*(3):531–540.

Joshi, A. J., & Rys, M. J. (2011). STUDY ON THE EFFECT OF DIFFERENT ARRIVAL PATTERNS ON AN EMERGENCY DEPARTMENT'S CAPACITY USING DISCRETE EVENT SIMULATION. *International Journal of Industrial Engineering*, *18*(1).

Ko, Y. D., Song, B. D., Morrison, J. R., & Hwang, H. (2014). LOCATION DESIGN FOR EMERGENCY MEDICAL CENTERS BASED ON CATEGORY OF TREATABLE MEDICAL DISEASES AND CENTER CAPABILITY. *International Journal of Industrial Engineering*, *21*(3).

Mohiuddin, S., Busby, J., Savović, J., Richards, A., Northstone, K., Hollingworth, W., … Vasilakis, C. (2017). Patient flow within UK emergency departments: a systematic review of the use of computer simulation modelling methods. *BMJ Open*, *7*(5), e015007.

Munir, W. (2008). Critical analysis of the 4-hour A&E policy's impact on elderly patients. *British Journal of Nursing*, *17*(18).

National Institute for Health and Care Excellence (NICE). (2016). *Safe staffing for nursing in A&E departments: the full NICE safe staffing guideline*. London. Retrieved from https://www.nice.org.uk/guidance/gid-sgwave0762/documents/accident-and-emergency-departments-appendix-1-evidence-to-recommendations-tables2

NHS. (2019). Clinically-led Review of NHS Access Standards - Interim Report from the NHS National Medical Director, 1–40. Retrieved from https://improvement.nhs.uk/improvement-hub/quality-improvement/%0Ahttps://www.england.nhs.uk/clinically-led-review-nhs-access-standards/

NHS England. (2018a). *AE Waiting times and activity*. London. Retrieved from https://www.england.nhs.uk/statistics/statistical-work-areas/ae-waiting-times-and-activity/ae-attendances-and-emergency-admissions-2018-19/

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

NHS England. (2019). *A&E Attendances and Emergency Admissions 2017-2018*. London. Retrieved from https://www.england.nhs.uk/statistics/statistical-work-areas/ae-waiting-times-and-activity/ae-attendances-and-emergency-admissions-2017-18/

NHS UK. (2018). *Hospital Accident Emergency Activity, 2017-18*. London. Retrieved from https://digital.nhs.uk/data-and-information/publications/statistical/hospital-accident--emergency-activity/2017-18

NHS UK. (2019). Urgent and Emergency Care: When to go to A&E. Retrieved May 10, 2019, from https://www.nhs.uk/using-the-nhs/nhs-services/urgent-and-emergency-care/when-to-go-to-ae/

Portsmouth Hospitals NHS Trust. (2018). *TRUST CAPACITY MANAGEMENT POLICY*. Retrieved from https://www.porthosp.nhs.uk/about-us/policies-and-guidelines/management-policies.htm

Proudlove, N. C., Gordon, K., & Boaden, R. (2003). Can good bed management solve the overcrowding in accident and emergency departments? *Emergency Medicine Journal*, *20*(2):149–155.

Rabbani, M., Farshbaf-Geranmayeh, A., & Yazdanparast, R. (2018). A simulation optimization approach for integrated resource allocation in an emergency department, pharmacy, and lab. *Intelligent Decision Technologies*, *12*(2), 187–212.
Rico, F., Salari, E., & Centeno, G. (2007). Emergency departments nurse allocation to face a pandemic influenza outbreak. In *Proceedings of the 39th conference on Winter simulation* (pp. 1292–1298). IEEE Press.

Saghafian, S., Austin, G., & Traub, S. J. (2015). Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering*, *5*(2):101–123.

The Health Foundation. (2019). *Falling short: NHS workforce challenge*. London. Retrieved from https://www.health.org.uk/publications/reports/falling-short-the-nhs-workforce-challenge

Tijms, H. C., Van Hoorn, M. H., & Federgruen, A. (1981). Approximations for the steady-state probabilities in the M/G/c queue. *Advances in Applied Probability*, 186–206.

Wang, T., Guinet, A., Belaidi, A., & Besombes, B. (2009). Modelling and simulation of emergency services with ARIS and Arena. Case study: the emergency department of Saint Joseph and Saint Luc Hospital. *Production Planning & Control*, *20*(6), 484–495. https://doi.org/10.1080/09537280902938605

Weng, S.-J., Cheng, B.-C., Kwong, S. T., Wang, L.-M., & Chang, C.-Y. (2011). Simulation optimization for emergency department resources allocation. In *Proceedings of the 2011 Winter Simulation Conference (WSC),* (pp. 1231–1238). IEEE.

Yeh, J.-Y., & Lin, W.-S. (2007). Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department. *Expert Systems with Applications*, *32*(4):1073–1083.

Zeinali, F., Mahootchi, M., & Sepehri, M. M. (2015). Resource planning in the emergency departments: A simulation-based metamodeling approach. *Simulation Modelling Practice and Theory*, *53*:123–138.

APPENDIX

Table 4. Average number of total arrivals to UHCW in 2017/18

Hour	Arrivals
9 am.	12370
10 am.	13390
11 am.	13815
12 pm.	13225
1 pm.	11965
2 pm.	11655
3 pm.	11845
4 pm.	11065
5 pm.	10865
6 pm.	11095