Staff optimization for time-dependent acute patient flow

Andersen, Anders Reenberg; Nielsen, Bo Friis; Reinhardt, Line Blander; Stidsen, Thomas Riis

Published in:
European Journal of Operational Research

Link to article, DOI:
10.1016/j.ejor.2018.06.015

Publication date:
2019

Document Version
Peer reviewed version

Staff Optimization for Time-Dependent Acute Patient Flow

Andersen, Anders Reenberg *
Technical University of Denmark
Department of Engineering Management
Anker Engelunds Vej 1, 2800 Kgs. Lyngby, Denmark

Nielsen, Bo Friis
Technical University of Denmark
Department of Applied Mathematics and Computer Science
Anker Engelunds Vej 1, 2800 Kgs. Lyngby, Denmark

Reinhardt, Line Blander
Aalborg University
Department of Mechanical and Manufacturing Engineering
A.C. Meyers Vænge 15, 2450 Copenhagen, Denmark

Stidsen, Thomas Riis
Technical University of Denmark
Department of Engineering Management
Anker Engelunds Vej 1, 2800 Kgs. Lyngby, Denmark

Abstract

The emergency department is a key element of acute patient flow, but due to high demand and an alternating rate of arriving patients, the department is often challenged by insufficient capacity. Proper allocation of resources to match demand is, therefore, a vital task for many emergency departments.

Constrained by targets on patient waiting time, we consider the problem of minimizing the total amount of staff-resources allocated to an emergency department. We test a matheuristic approach to this problem, accounting for both patient flow and staff scheduling restrictions. Using a continuous-time Markov chain, patient flow is modeled as a time-dependent queueing network where inhomogeneous behavior is evaluated using the uniformization method. Based on this modeling approach, we recursively evaluate and allocate staff to the system using integer linear programming until the waiting time targets are respected in all queues of the network. By comparing to discrete-event simulations of the associated system, we show that this approach is adequate for both modeling and optimizing the patient flow. In addition, we demonstrate robustness to the service time distribution and the associated system with multiple classes of patients.

Keywords— OR in health services, Queueing, Markov chain, Heuristics, Stochastic optimization

1 Introduction

In this study, we consider the well-known problem of optimizing the patient flow for an Emergency Department (ED). With many hospitalizations on a daily basis, the ED is often considered a vital element to the hospital compared to other hospital departments. ED hospitalizations are further characterized by a large variety of different diagnoses, requiring staff from a range of different specializations around the clock.
A report from the Danish Ministry of Health [1] places Denmark below the average lifespan for countries in the Organization for Economic Co-operation and Development (OECD), but above the average on fraction

*Corresponding author: E-mail: arean@dtu.dk
This research was funded and supported by Region Sjælland.

of Gross Domestic Product (GDP) used on public health care; hence suggesting that a general increase in the utilization of resources is required. In this study, we address this issue by providing hospital management with a method for deriving the minimum required staff for an ED constrained by targets on patient waiting time. Such method is especially relevant for hospitals that are governed by their efficiency, and therefore seek to rearrange the excess resources for instance by validating the difference between the minimum required and currently available resources.

Operations Research literature related to Emergency Department (ED) planning and dimensioning is relatively unexplored as regards analytical modeling of acute patient flow combined with optimization.
Lim et al., 2012 [21] conducted an elaborate survey on the use of mathematical modeling of ED patient waiting times and found 29 relevant studies. From these, four overall modeling techniques were uncovered: (1) Queueing Theoretic (QT) models covered a total of four different studies, (2) Discrete Event Simulation (DES) covered 22 different studies, (3) System Dynamics (SD) covered two studies and (4) Agent-Based Modeling (ABM) covered two studies likewise. Substantial weight is obviously given to the three simulation-related approaches as only four studies were conducted using QT modeling.
Lim et al., 2012 further found that a recurrent objective is to use the model to test one or more scenarios and rarely to optimize the system. Examples in QT modeling are Cochran & Roche, 2009 [8] and Mayhew & Smith, 2008 [22], where open queueing network models are developed with a view to investigate how to increase patient throughput. In the area of DES, Medeiros et al., 2008 [23] tested an approach named Provider Directed Queueing for improving ED performance. Additionally, Khadem et al., 2008 [15] assessed a new layout for an ED and found the new layout to reduce patient waiting time by a substantial amount. In SD modelling, Storrow et al., 2008 [32] assessed the effect of decreasing lab turnaround times, focusing on emergency medical services, patient throughput and length of stay. Lane et al., 2000 [18] assessed changes in waiting times as bed capacity is changed. Further, in the area of ABM, Wang, 2009 [35] evaluated different settings of triage and radiology procedures. Lastly, some studies combine different modeling approaches to attain their objective. Laskowski et al., 2009 [19] evaluated patient flow using two different models. One based on ABM and the second based on queueing theory. In their study, the two models are applied and compared by using a number of relatively simple scenarios.
Getting an understanding of acute patient flow based on simulation seems well explored. However, Lim et al., 2012 only obtained two studies that use modeling of patient flow in an actual optimization scheme. The first is Yeh & Lin, 2007 [38] where schedules are adjusted for a fixed amount of nurses by using a combination of DES and a Genetic Algorithm (GA). The aim was to find the configuration of schedules that minimizes patient waiting time. Secondly, Ahmed & Alkhamis, 2009 [2] combined DES with a local search heuristic by applying statistical hypothesis testing. The goal was to determine the optimal number of different staff types by maximizing the throughput of patients constrained by department budgets.

Besides the studies in Lim et al., 2012 we were able to identify four studies where optimization is conducted in the context of acute patient flow. Firstly, Sinreich et al., 2012 [29] use a DES model together with Mixed Integer Linear Programming (MILP) to derive two different heuristics with the aim of determining efficient work-shift schedules that minimize patient waiting time. Further, Daldoul et al., 2015 [9] determined the optimal amount of staff and equipment by using a MILP model. Interestingly, system stochasticity was not incorporated in this model. In addition, Cabrera et al., 2012 [6] used ABM and exhaustive search to optimize the configuration of different staff types, and lastly, Wang, 2013 [36] used a modeling approach known as Separated Continuous Linear Programming to determine the level of staffing that would minimize the overall cost of the ED.
Now, when we consider studies that focus only on queueing theoretic modeling, then queues with non-homogeneous Poisson arrivals or even processes with more general time dependent arrivals has received substantial interest. See e.g. Schwarz et al., 2016 [26] and Defreye & Inneke, 2016 [10] for two recent review papers. The literature on time dependent queuing networks specifically is less abundant, but see Armony et

al., 2015 [4] for a data-based analysis of ED's viewed through the lens of a queueing scientist.
Moving to different application areas, in manufacturing Bitran & Morabito, 1994 [5] conducted a survey on stationary open queueing networks, presenting both exact and approximate solutions to a range of different problem structures. Related to our study, the problem of minimizing cost by allocating machines, constrained by an upper bound on a Work-In-Progress (WIP) level, may be solved approximately by a heuristic. However, if the number of machines is fixed, and the objective is to minimize the WIP level, then an exact solution can be derived.
Further, on optimizing stationary queueing networks, Smith et al., 2010 [30] present an exact solution to the machine allocation problem for a finite queueing network by using Powell's algorithm. For a general open queueing network, Giloni, 2001 [12] derives conditions under which the problem is reduced to solving a concave or convex problem. Additionally, Seshadri & Pinedo, 1999 [27] exploit an approach where a heuristic is used to minimize the WIP level. Lastly, Yoneda et al., 1992 [39] apply simulated annealing to optimize their system.

In the area of call center staffing, several studies have been conducted considering both time-varying arrivals and staff optimization. For single queues, Feldman et al., 2004 [11] investigate three different methods for deriving the minimal time-dependent staffing level, s_t, to maintain time-stable performance. The study proposes a simulation-based algorithm, along with an extension of the square-root-staffing formula [14]. Lastly, for queues with customer abandonments, $M_t/M/s_t + M$, they show for a certain setting that staffing can be adjusted to match the expected load in the associated infinite-server system. Related hereto, Whitt, 2006 [37] maximizes the revenue of an $M/GI/s+GI$ queue by firstly modeling the system as a deterministic fluid model, and secondly as the associated $M/M/s + M(n)$ model. The optimization is conducted by adjusting the number of servers in the system. Further, Sze, 1984 [33] focuses on choosing an adequate $M/G/s$ model for staffing purposes, taking arrival variability into account.
Turning to queueing networks in call center staffing, Tipper & Sundareshan, 1990, [34] consider a network of single-server queues with time-varying arrival rate, using two models. The first is based on Chapman-Kolmogorov differential equations, and the next on non-linear differential equations modeling the mean queue lengths in the network. We have noticed that neither in this study nor the proceeding three studies on single queues is emphasis put on incorporating staff in more complex shift structures. Liao et al., 2012 [20] derives the optimal staffing level of a single queue, incorporating back-office jobs, by using both stochastic and robust programming, respectively, but assumes a single-shift structure.

We acknowledge that modeling of queueing networks is an extensive field covering many other applications as have not been mentioned above. In this review, we mainly focus on the literature covering optimization of acute patient flow, and two related application areas. In the area of manufacturing, non-stationary cases seem to be rarely considered, whereas for call center modeling and staffing, limited emphasis is put on optimizing staff with more elaborate shift structures. In the area of modeling flow for acute patients queueing theory combined with optimization is in general an uncommon approach.
In our study, we present an approach based on a continuous-time Markov chain (CTMC) for modeling the time-dependent behavior of acute patient waiting time, and the interaction of this approach with an Integer Linear Programming (ILP) model. The ILP will serve as the method we use to efficiently allocate staff to specific working-patterns, as has been proven adequate by other studies [7]. We combine the CTMC and ILP in a matheuristic search procedure with the objective of minimizing the total amount of staff that is allocated to the ED. This heuristic procedure is further divided into two variations, yielding two models for our numeric experiments. Further, we have used a Danish ED as the basis for our study and have constructed a representation of patient flow as well as conducted tests based on data from this ED.

Specifically, our contribution to the area of acute patient modeling and optimization is:

- Applying an analytical approach for modeling time-dependent flow for acute patients, going from triage

to specialized treatment. Specifically, we employ a numerical method for modeling the system as an open queueing network.

- Combining the queueing network with an ILP model in a simple and generalizable matheuristic procedure for minimizing staff, taking constraints on patient waiting time, as well as staff working-patterns into account.

In Section 2 we elaborate on the specific problem and data at hand. In Section 3, we present the CTMC model that is used to evaluate patient waiting time and the structure of the matheuristic incorporating both CTMC and ILP modeling. In Section 4 we evaluate our CTMC approach, present how the tuning of parameters is conducted, and demonstrate the performance of our matheuristic. Lastly, we present our conclusion in Section 5.

2 Problem Description

Any patient who is admitted to an Emergency Department (ED) will be dependent on a range of different resources. Upon arrival, the patient is firstly triaged to determine the severity of the patient's condition. Next, an examination is conducted by a physician to determine whether a more in-depth treatment is required. If there is no need for this, the patient can be discharged immediately; otherwise, a specialized physician is further required.
Obviously, the admission of a patient involves the use of a range of many different resources. In case one or more of these resources are absent, the treatment quality will decrease accordingly. Still, like any other organization the ED is subject to a limited capacity and is thus faced with the problem of balancing quality of care against the department expenditures. Being able to make clear objectives and utilize resources accordingly is, therefore, a core responsibility of the department.

Our objective is to contribute to the methodology related to balancing ED capacity against service, by minimizing the total amount of staff allocated to the department taking the resulting effect on treatment quality into account. As waiting time has been shown to directly influence the treatment quality of acute patients [13, 24, 25], the total amount of allocated staff will be constrained by targets on patient waiting time. Due to union settlements, we further consider that staff resources are constrained by a number of fixed working patterns.

2.1 System and Data Description

We consider patients of a single class arriving according to a process with time-varying intensity to an ED. Upon arrival, the patients are physically admitted to a bed, where they will stay until discharged. During this time, however, the patients require attention from a range of different staff types depending on their diagnosis. Each staff type is drawn from a "pool" of limited capacity and will attend the patients for a random amount of time. Thus, we assume that the stay of a patient can be modeled as an open queueing network, where the change in required attention between staff types corresponds to moving from one network node to the next. Due to this approach, we assume that a patient can only be treated by a single care-provider at a time.
In case a patient requires attention from a pool of staff where all members are occupied, a queue is created, and the patient will receive attention according to a *first-come first-served* (FCFS) discipline. To represent the diversity of diagnoses and need, the routing of patients from one queue to the next occurs randomly, but with known probability. Additionally, in case a patient requires attention from the same care-provider more than once, the patient is looped back to the same node. Specifically, we interpret the patient flow as the queueing network presented in Figure 1. Here, each queue of the network represents the following five

staff types (by queue number):

1. Triage Nurses
2. Basic Physicians
3. Specialized Medical Physicians
4. Organ Surgeons
5. Orthopedic Surgeons

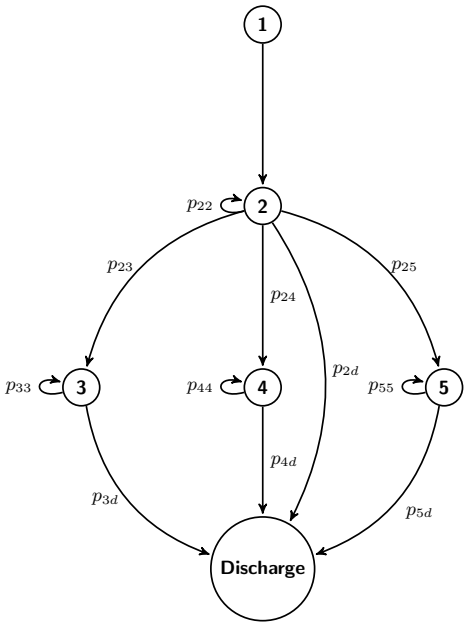

Figure 1: The stay of patients modeled as a network of queues. Each node represents a single queue with servers of only one staff type. Staff types by queue number are: (1) Triage Nurses, (2) Basic Physicians, (3) Specialized Medical Physicians, (4) Organ Surgeons and (5) Orthopedic Surgeons. The parameter, p_{ij}, defines the routing probability.

2.1.1 Patient Data

We obtained one year of patient data from a Danish ED, showing the exact arrival time and triage level of each patient. Our case-ED uses four triage levels between which patients are initially distributed (in ascending priority) with 9% on level 1, 63% on level 2, 25% on level 3, and 3% on level 4. Furthermore, we naturally found that all patients were triaged on arrival, but then have their priority level adjusted after the first examination by a physician. That is, after the examination about 72% of the patients on level 3 were re-evaluated to level 2, essentially changing the distribution to 81% of the patients on level 2 and only 7% on level 3 for the remaining queues in the network.

Based on the ED data, we further investigated the patient inter-arrival time by modeling the arrival rate as the Poisson regression, shown in (1),

$$log(\lambda_{ij}(u)) = \alpha + \beta u + \theta u^2 + \gamma_j + \delta_i + \phi_j u + \zeta_j u^2 + \psi_i u + \xi_i u^2 + \rho_{ij} + \eta_{ij} u + \omega_{ij} u^2 \tag{1}$$

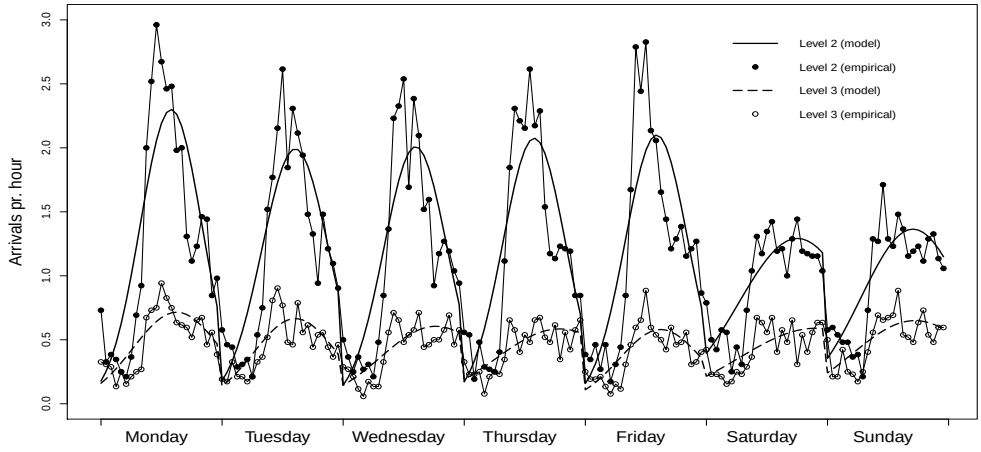

Figure 2: The modeled, according to (1), and empirical time-varying arrival rate for patients of triage level 2 and 3 respectively.

where $\lambda_{ij}(u)$ is the expected number of arrivals on hour of the day $\{u \in \mathbb{R} | 0 \leq u \leq 24\}$, on the day of the week $j \in \{Monday, Tuesday, \ldots, Sunday\}$ for patients of triage priority $i \in \{1, 2, 3, 4\}$. We used explanatory variables to model the effect of day of the week, j, and triage priority i. Due to the limited amount of data obtained, we modeled the effect from time of the day, u, as a second order polynomial. The resulting modeled arrival rate is demonstrated in Figure 2, showing both the modeled and empirical rates for patients of triage level 2 and 3, respectively.

We evaluated our model by examining the distribution of $\epsilon = (y - \hat{\lambda})/\sqrt{\hat{\lambda}}$, where $\hat{\lambda}$ is the model fit and y the observations. In addition, we conducted a graphical test where the model was fitted to the first six months of data and then compared to the last six months. Lastly, we estimated the dispersion parameter at $\hat{\phi} = 0.82$, and conducted a Pearson's goodness-of-fit based on a model deviance of 41346 with 50274 degrees of freedom, yielding a right-tailed probability of $p = 1$. Thus, a very large p-value. From these both graphical and quantifiable measures we have found Poisson behavior to fit the data well.

As patients are admitted to the ED, they will require attention from a range of different staff types, which we interpret as the service times of the queueing network. For the case ED, we were not able to obtain reliable data on time spent on patients. Therefore, in our subsequent modeling we will assume for convenience that service times are exponentially distributed, even though we acknowledge that such distribution might not fit real-life inter-service times of an ED. Later, in Section 4.1, we elaborate more on the robustness of this assumption. The specific parameters that we have used for our service time distributions are presented in Appendix A, Table 6. These were estimated based on interviews with hospital staff.
Lastly, regarding the use of prioritizes, recall that a fairly large fraction of patients are prioritized on triage level 2, especially after the examination by a physician. Therefore, to ensure computational tractability of our modeling approach, we will only be considering a single class of patients with arrival rate corresponding to the sum of triage level 2 and 3. Since we only consider this single class, we may drop the index on triage level, such that the arrival rate simplifies to $\lambda_j(u)$. Later, in Section 3.1, we will refer to the arrival rate as $\lambda(\xi)$, where ξ is any continuous point in time within one week. We elaborate more on the implications of this assumption in Section 4.2.2.
Besides the arrival rates derived from (1), we test two additional arrival patterns for our optimization experiments, $g_j(u) = \lambda_j(u) \cdot 0.9$ and $h_j(u) = \lambda_j(u) \cdot 0.9^2$, depicted in Figure 3.

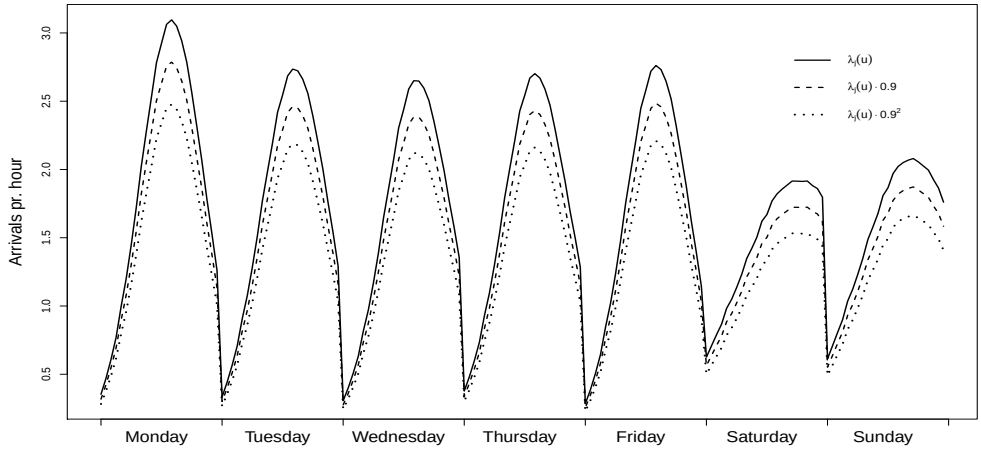

Figure 3: The three arrival rate patterns used to test the performance of our matheuristic approach.

2.1.2 Routing

Depending on condition and diagnosis, any patient arriving at the ED will have a unique need for care, and as a result, there is a large range of different combinations of services to account for in the patient flow. From the perspective of our queueing network, the necessary care will be reflected in the patient being routed to either a specialized physician, looped to the current care provider, or discharged. The question is whether such routing occurs randomly or depends on some underlying policy. For instance, in situations of overcrowding at the optimal care resource, it might be worth to consider an alternative option for the patient. However, our interviews with the ED staff indicated that the patients are provided with the care optimal to each patient, in which case routing occurs randomly in accordance with the random occurrence of condition and diagnosis. Furthermore, we found that patients require care resources according to some distribution. Let p_{ij} define the probability of being served by staff type $j \in C$ successive to $i \in C$, and let p_{id} define the probability of being discharged and leaving the system upon completion of $i \in C$, as shown in Figure 1. To derive the value of these, we obtained patient data showing the specific staff resources that were required during each patient's stay. The result is presented in Appendix A, Table 7.

3 Modeling & Solution Approach

In this section we present an approach for the problem of minimizing the total amount of staff allocated to the ED, constrained by targets on the patient waiting time. We consider a set of staff types, C, subject to a limited set of working-patterns, J, and that patient waiting time is a non-linear function of the available capacity in the department. This leads to the master problem shown in equation (2a)-(2c),

$$\begin{aligned}
\text{min.} \quad & \textstyle\sum_{c\in C}\sum_{j\in J} x_{cj} && \text{(2a)}\\
\text{s.t.} \quad & \\
& L_{ct}(\mathbf{Z}) \geq \tau \qquad \forall t \in T,\, c \in C, \text{ where } z_{ct} = \textstyle\sum_{j\in J} a_{cjt}x_{cj} && \text{(2b)}\\
& \textstyle\sum_{j\in J} a_{cjt}x_{cj} \geq \beta_{ct} \qquad \forall t \in T,\, c \in C && \text{(2c)}\\
& x_{cj} \in \mathbb{N}_0 \qquad \forall j \in J,\, c \in C
\end{aligned}$$

where x_{cj} is the amount of staff type $c \in C$ assigned to working-pattern $j \in J$. Thus, (2a) is the total

amount of staff allocated to the department. Furthermore, $L_{ct}(\mathbf{Z})$ is the fraction of patients waiting for staff type $c \in C$ below a predefined time, in time period $t \in T$, where T is a discrete set of the weekly hours $T = \{1, 2, \dots, 168\}$. Here, $\mathbf{Z}$ is a $|T| \times |C|$ matrix defining the resulting allocation of each staff type for each time period in the entire planning period. Let z_{ct} be an element of $\mathbf{Z}$, and let $a_{cjt} \in \{0, 1\}$ be equal to 1 if working-pattern $j \in J$ assigns staff type $c \in C$ to time period $t \in T$; otherwise 0. Then, $z_{ct} = \sum_{j \in J} a_{cjt} x_{cj}$, is the amount of staff type $c \in C$ allocated to time period $t \in T$. Since $\{\tau \in \mathbb{R} | 0 < \tau < 1\}$, (2b) constraints the fraction of patients with a waiting time below a predefined amount of time.
Lastly, we assume that the system has a limit for each staff type $c \in C$ and time period $t \in T$, β_{ct}, after which the system is no longer operative. Constraints (2c) is introduced to ensure that the staff limit is never violated.
We evaluate $L_{ct}(\mathbf{Z})$ by using a continuous-time Markov chain, presented in Section 3.1. Due to the non-linear and complex structure of $L_{ct}(\mathbf{Z})$ there exists, to our knowledge, no standard approach to solve (2a)-(2c). We present a heuristic approach in Section 3.2.

3.1 Modeling Patient Waiting Time

As previously mentioned, we consider five staff types, $C = \{1, 2, \dots, 5\}$ interpreted as five different nodes in a queueing network. To model the occupancy and flow between these queues, we introduce a continuous-time Markov chain (CTMC) with state definition $s = \{k_1, k_2, \dots, k_5\}$, where k_i is the number of patients waiting for *or* in service by staff type $i \in C$. We further consider a truncation of the patient capacity, $M_i \geq k_i$, and choose this so the probability of having M_i patients waiting for or being served by staff $i \in C$ has negligible effect on the behavior of the system. Then the CTMC has state space $S = \{0, \dots, M_1\} \times \{0, \dots, M_2\} \times \cdots \times \{0, \dots, M_5\}$ of size $|S| = \prod_{i \in C}(M_i + 1)$.
Furthermore, let $\lambda(\xi)$ define the arrival rate of a single class of patients at time ξ. In addition, let μ_i define the service rate of staff type $i \in C$. Moreover, let w_i define the number of servers of staff type $i \in C$, and assume that $w_i < M_i$.
Let Q define the transition rate matrix of the CTMC, with q_{ss^*} the transition rate from the current state $s \in S$ to a new state $s^* \in S$. Then we have,

$$q_{ss^*} = \begin{cases} \lambda(\xi) & \text{if } s^* = (k_1 + 1, k_2, \dots, k_5) \text{ and } k_1 < M_1 \\ \mu_1 k_1 & \text{if } s^* = (k_1 - 1, k_2 + 1, \dots, k_5) \text{ and } k_1 > 0, k_2 < M_2, k_1 \leq w_1 \\ \mu_1 w_1 & \text{if } s^* = (k_1 - 1, k_2 + 1, \dots, k_5) \text{ and } k_1 > 0, k_2 < M_2, k_1 \geq w_1 \\ \mu_2 k_2 p_{2j} & \text{if } s^* = (k_1, k_2 - 1, \dots, k_j + 1, \dots) \text{ and } k_2 > 0, k_j < M_j, k_2 \leq w_2 \quad \forall j \in C \setminus \{1, 2\} \\ \mu_2 w_2 p_{2j} & \text{if } s^* = (k_1, k_2 - 1, \dots, k_j + 1, \dots) \text{ and } k_2 > 0, k_j < M_j, k_2 \geq w_2 \quad \forall j \in C \setminus \{1, 2\} \\ \mu_i k_i p_{id} & \text{if } s^* = (k_1, \dots, k_i - 1, \dots) \text{ and } k_i > 0, k_i \leq w_i \quad \forall i \in C \setminus \{1\} \\ \mu_i w_i p_{id} & \text{if } s^* = (k_1, \dots, k_i - 1, \dots) \text{ and } k_i > 0, k_i \geq w_i \quad \forall i \in C \setminus \{1\} \end{cases}$$

where all other transition rates, q_{ss*}, are 0.
All patients arrive at the first node of the network, and therefore only k_1 is subject to increase by a rate of $\lambda(\xi)$. Consider a case where $M_1 = 10$. Then the transition $s = (k_1, k_2, k_3, k_4, k_5) = (5, 10, 2, 3, 2) \rightarrow s^* = (\mathbf{6}, 10, 2, 3, 2)$ occurs with a rate of $\lambda(t)$. Internal flows of the network occurs from either node 1 or 2, and all discharges from either node 2, 3, 4 or 5. These are all dependent on both service rates, assigned staff and the routing probabilities of the node where the patient has just completed service. Thus if $M_3 = 5$ and $w_2 = 2$, $s = (6, 10, 2, 3, 2) \rightarrow s^* = (6, \mathbf{9}, \mathbf{3}, 3, 2)$ occurs with a rate of $\mu_2 w_2 p_{23}$, and if $w_3 = 4$, $s = (6, 9, 3, 3, 2) \rightarrow s^* = (6, 9, \mathbf{2}, 3, 2)$ occurs with a rate of $\mu_3 k_3 p_{3d}$.

3.1.1 Time-Dependent Behavior

To derive the waiting times from the queueing network, we do not only have to take the assigned staff into account, but also the effect of the time-varying arrival rate, as was defined in Section 2. The approach

we follow could be classified as a piecewise transient model according to Schwarz et al., 2016 [26] and the solution method we apply is uniformization [31], also denoted randomization.
Notice that, as the arrival rate is weekly cyclical and *if* the working-patterns are weekly cyclical as well, then the process eventually stabilizes with the distribution given as a weekly-periodic vector function, $f(\xi)$. We make a numerical approximation to this distribution by first assuming that the change in arrival rate is negligible within some limited time interval, for instance, one hour. We denote the length of these time intervals by δ such that the length of the period, of one week, τ^{week} is an integer multiple of δ. Now, let $\lambda(\xi)$ define the arrival rate of patients at time ξ, and assume there exists a negligible change $|\lambda(\xi) - \lambda(\xi + \delta)|$, so $\lambda(\xi)$ can be discretized into a vector $\boldsymbol{\lambda}$ of size $\tau^{week}/\delta \in \mathbb{N}^+$.
Let $\pi_i(t)$ define the i'th segment of the stabilized distribution a function of t with $\{t \in \mathbb{R} | 0 \leq t \leq \delta\}$. Furthermore, let $\Upsilon = \{\xi \in \mathbb{R} | 0 \leq \xi \leq \tau^{week}\}$ and $t = 0$ represent the beginning of a segment, i, on the time line of Υ, and $t > 0$ the duration of time spent in such a segment, so $\pi_i(t)$ can be determined on any $\xi \in \Upsilon$ using both i and t, as illustrated in Figure 4.
Then, (3) represents the time-dependent state distribution of the process for any point in time of the week, $\xi \in \Upsilon$, where all entries in the vector function $f(\xi)$ are piecewise constant.

$$f(\xi) = \pi_i(t), \quad i = \lceil \xi/\delta \rceil \wedge t = \xi - (i-1) \cdot \delta \tag{3}$$

where $\pi_{i-1}(\delta) = \pi_i(0)$.

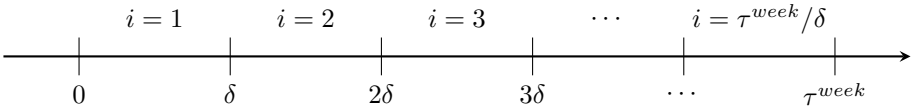

Figure 4: The segmented time-line of length τ^{weeks}. Applied in the modelling of the time-dependent state distribution, $f(\xi)$.

For each of the τ^{week}/δ time-intervals, illustrated in Figure 4, we require a method for deriving the time-inhomogeneous state distributions, $\pi_i(t)$, from which we derive the entire weekly behavior of the process.
From standard theory we have,

$$\pi_i(t) = \pi_i(0) e^{Q_i t} \tag{4}$$

where Q_i is the transition rate matrix containing the element λ_i from $\boldsymbol{\lambda}$ corresponding to the i'th segment. We use *uniformization* as presented below to calculate the matrix exponential.
Let γ_i be at least as large numerically as the largest diagonal element of Q_i. We then write,

$$P_i = Q_i/\gamma_i + I \tag{5}$$

where P_i is a transition probability matrix, with each element defining the probability of going from a state $s \in S$ to a new state $s^* \in S$, and I the identity matrix. Then, from an initial distribution $\pi_i(0)$, the distribution at time t, in segment i, $\pi_i(t)$, can be derived using (6) [31].

$$\pi_i(t) = \sum_{k=0}^{\infty} \pi_i(0) P^k \frac{(\gamma_i t)^k}{k!} e^{-\gamma_i t} \tag{6}$$

The transformation of Q_i into P_i, allows us to interpret our CTMC as an embedded Markov chain with a random number of transitions. The number of state changes in the embedded Markov chain, P_i, has probability $e^{-\gamma_i t}(\gamma_i t)^k/k!$ according to a Poisson distribution with parameter $\gamma_i t$, and thus depends on the time t. From $\pi_i(0)P^k$, the state distribution after exactly k changes is determined, so by using $e^{-\gamma_i t}(\gamma_i t)^k/k!$ we may weigh each of these distributions, according to time t and form the resulting state distribution $\pi_i(t)$. In implementing (6), we computationally use a recursive formulation to approach $\pi_i(t)$ until convergence.

Let K be the minimum number of terms in (6) required to attain an accuracy of ϵ, then the following statement has to be satisfied [31]:

$$\sigma_K = \sum_{k=0}^{K} \frac{(\gamma t)^k}{k!} \geq (1-\epsilon)e^{\gamma t} \tag{7}$$

From (7), a recursive formulation to determine K can be established – presented below:

1. Initialize setting $\zeta \leftarrow \sigma \leftarrow 1$ and $K \leftarrow 0$.
2. If $\sigma \geq (1-\epsilon)e^{\gamma_i t}$, then **stop**; otherwise continue.
3. Set $\zeta \leftarrow \zeta \frac{\gamma_i t}{K+1}$ and $\sigma \leftarrow \sigma + \zeta$.
4. Set $K \leftarrow K+1$ and go to **2**.

The stabilized distribution of the segment i, $\pi_i(t)$, can then be computed by using the recursion:

1. Initialize setting $y \leftarrow \pi \leftarrow \pi_i(0)$, and $k \leftarrow 0$.
2. If $k = K$, then **stop** as $\pi_i(t) \approx e^{-\gamma_i t}\pi$ with an accuracy of ϵ. Otherwise set $k \leftarrow k+1$ and continue.
3. Set $y \leftarrow y\left(P_i \frac{\gamma_i t}{k}\right)$, $\pi = \pi + y$ and go to **2**.

This concludes the approach we use to derive the state distribution at time t in segment i, $\pi_i(t)$. Our implementation of the step function, $f(\xi)$, containing the state distribution at every point in time of the week, is derived recursively by using Algorithm 1, as is presented in the following Section 3.1.2.

3.1.2 Waiting Times

In the above, we have presented an approach to obtain the occupancy distribution for a duration of time t in a segment i, denoted $\pi_i(t)$, given the initial value $\pi_i(0)$. The entries of the transition rate matrix in each segment, Q_i, is determined by $\lambda(\xi)$, along with the allocation of staff to each queue of the network. Through (3), $\pi_i(t)$ is used to determine the state distribution for the entire week, $f(\xi)$. Though $f(\xi)$ specifies how many patients are expected to be present in each queue, the measure does not directly reflect the resulting waiting times.

Let W denote the waiting time at a queue with w servers, and let k define the number of patients present at the queue at the time of arrival. Then $W = 0$ if $k \leq w-1$. For exponential service times and $k \geq w$ we have $W = \sum_{i=w}^{k} Z_i$, where Z_i are independent exponential random variables of rate w times the service rate of each server. Our aim is to derive the fraction of patients waiting below a specific target as function of time of the week, ξ. That is, $L(\xi) = Prob\{W(\xi) \leq \nu\}$, where ν is the target upper waiting time and $W(\xi)$ the time-dependent waiting time distribution. Letting $K(\xi)$ define the random number of patients present at the queue at time ξ, we assume the time-inhomogeneous behavior within a segment is negligible so,

$$Prob\{W(\xi) \leq \nu\} = \sum_{k=0}^{M_c} Prob\{W(\xi) \leq \nu | K(\xi) = k\} \cdot Prob\{K(\xi) = k\} \tag{8}$$

and therefore, by letting $f_{ci}(\xi) = \sum_{j \in J} Prob\{s = (\dots, k_c = i, \dots)\}$ define the marginal time-dependent state distribution obtained in Section 3.1.1, where $J = S \setminus \{s = (\dots, k_c \neq i, \dots)$ — that is, the probability that queue $c \in C$ is occupied by i patients, we get

$$Prob\{W_c(\xi) \leq \nu_c\} = \sum_{i=0}^{w_c-1} f_{ci}(\xi) + \sum_{k=w_c}^{M_c} f_{ck}(\xi) \cdot Prob\left(\sum_{i=1}^{k} z_i \leq \nu_c\right) \quad \forall c \in C \tag{9}$$

from [17]. The first term of (9) accounts for the probability that there is no waiting time on arrival – namely when at least one of the servers is free. The second term contains the probability that the queue is occupied by w_c or more patients, and the probability that the sum of service times for these patients is equal to or less than the queue dependent target ν_c. Furthermore, as

$$Prob\left(\sum_{i=1}^{k} z_i \leq \nu_c\right) = \int_0^{\mu_c w_c \nu_c} \frac{u^{k-1}}{(k-1)!} \cdot e^{-u} du = 1 - \sum_{j=0}^{k-1} \frac{(\mu_c w_c \nu_c)^j}{j!} \cdot e^{-\mu_c w_c \nu_c} \tag{10}$$

this allows us to write (9) on the form,

$$L_c(\xi) = \sum_{i=0}^{w_c-1} f_{ci}(\xi) + \sum_{k=w}^{M_c} f_{ck}(\xi) \cdot \left(1 - \sum_{j=0}^{k-1} \frac{(\mu_c w_c \nu_c)^j}{j!} \cdot e^{-\mu_c w_c \nu_c}\right) \quad \forall c \in C \tag{11}$$

where $L_c(\xi)$ is the fraction of patients waiting for staff type $c \in C$ below the target ν_c at time ξ. Let t define an hour of the week in the set $T = \{1, 2, \ldots, 168\}$, then for the remaining of this study, we refer to (11) as the function $L_{tc}(\mathbf{Z})$, presented in the master problem (2a)-(2c). The time targets for staff type $c \in C$, ν_c, are presented in Appendix A, Table 8. Finally, we apply (11) based on the time-dependent distribution, $f(\xi)$, using Algorithm 1.

Algorithm 1 Algorithm for evaluating the system over a full week.

```
π_0 ← (1, 0, 0, ..., 0)^T
L_0 ← WAITINGTIME(π_0)
while d > tol do                          ▷ Run until tolerance is satisfied
    i ← 1
    while i < 169 do
        π_i ← UNIFORMIZE(π_{i−1})         ▷ Uniformize at the end of the i'th hour
        L_i ← WAITINGTIME(π_i)            ▷ Evaluate waiting times in network using (11)
        i ← i + 1
    end while
    d ← RELATIVETOL(L_168, L_0)           ▷ d = max_{c∈C}(L_{168,c} − L_{0,c}))/L_{0,c}
    π_0 ← π_168
    L_0 ← L_168
end while
    return L
```

Notice, as we are only concerned with the instance where $t = 0$, we suppress the dependency on t, and let $\pi_i(0) = \pi_i$. Further, for convenience in Algorithm 1 we let L_i define a vector of the elements $L_{tc}(\mathbf{Z})$ for all $c \in C$ with time index t corresponding to the i'th segment.
We initialize the algorithm by an empty system, setting $\pi_0 \leftarrow (1, 0, 0, \ldots, 0)^T$. We then discretize to form $\tau^{week}/\delta = 168$ time-intervals – one for each hour of the week. Next, we evaluate the system in each time interval by uniformizing the process at the end of the hour, using the preceding hour as input. After all hours have been evaluated, the maximum relative difference in waiting time from the beginning, L_0, to the end of the week, L_{168}, is used as stopping criteria. Notice, for a slightly faster algorithm, the evaluation of L_i for the remaining segments of the week can be postponed until $d \leq tol$.

3.2 Optimization Heuristic

In Section 3.1 we have presented how to model ED patient flow using a continuous-time Markov chain (CTMC) and derive the time-dependent behavior of the system by recursively uniformize the model until

convergence. This approach yields a complex non-linear relation between assigning staff and the resulting patient waiting time.
Now, consider again the master problem (2a)-(2c). Let b_{ct} define a lower bound on staff of type $c \in C$ in time period $t \in T$ which is required to respect (2b) and (2c). Then (2a)-(2c) may be re-written in the form of an Integer Linear Programming (ILP) problem presented in (12a)-(12b) – which we can solve in reasonable time by applying a standard commercial solver software.

$$\begin{aligned} \text{min.} \quad & \textstyle\sum_{c\in C}\sum_{j\in J} x_{cj} && \text{(12a)} \\ \text{s.t.} \quad & \\ & \textstyle\sum_{j\in J} a_{cjt}x_{cj} \geq b_{ct} \qquad \forall t \in T,\, c \in C && \text{(12b)} \\ & x_{cj} \in \mathbb{N}_0 \qquad \forall j \in J,\, c \in C \end{aligned}$$

Still, we are faced with the problem of deriving b_{ct} in (12b), in order to respect the master problem. Sinreich et al. 2012, [29] presented two recursive heuristic algorithms, combining both simulation and mixed integer programming. Their simulation model, which is developed by Sinreich & Marmor, 2005 [28], accounts for many different patient pathways upon which their heuristics have been developed. Their overall approach is to minimize the length of stay for patients by recursively identifying and removing bottlenecks in the system. In our study, the representation of the ED is more simple, as we consider only a single class of patients and queues consisting only of one staff type. Our matheuristic approach reflects this representation by recursively assigning staff to the queues of the network constrained by a fixed set of working-patterns and the waiting time distributions evaluated by (11). We elaborate more on this matheuristic in the following section.

3.2.1 Recursive Bound Adaptation

In this section we present a matheuristic search procedure, where working-patterns and constraints on waiting time are incorporated in a recursive manner until a solution is derived. The heuristic has two stages: First a solution is constructed by recursively solving (12a)-(12b) and evaluating the resulting solution through the CTMC. The progressing of this first step constructs a feasible solution to x^*_{cj} in a greedy manner. Next, in the second stage, the meta-heuristic approach known as *Tabu Search* (TS) is used to search for an improved solution by further minimizing $\sum_{c\in C}\sum_{j\in J} x_{cj}$. We refer to this optimization strategy as Recursive Bound Adaptation (RBA).

The first stage consists of two parts:

1. **Optimization**. Let b^k_{ct} be a lower bound on required staff of type $c \in C$ in time period $t \in T$ for iteration k. Initializing with $b^0_{ct} = \beta_{ct}$, solve the ILP problem (12a)-(12b).

2. **Evaluation**. Starting from the solution, x^*_{cj}, derive the resulting allocation of staff $c \in C$ in time period $t \in T$, through $z_{ct} = \sum_{j\in J} a_{cjt}x^*_{cj}$. Then, evaluate the waiting times $L_{ct}(\mathbf{Z}) \quad \forall t \in T, c \in C$ by using the CTMC. Let $U_c \subseteq T$ be the set of time periods for staff type $c \in C$ for which, (2b), the waiting time constraint is violated. That is, $L_{ct}(\mathbf{Z}) < \tau$. Then, if $U_c \neq \emptyset$, make the adjustment: $b^{k+1}_{ct} = 1 + \sum_{j\in J} a_{cjt}x^*_{cj} \quad \forall t \in U_c, c \in C$, and $b^{k+1}_{ct} = b^k_{ct} \quad \forall t \in T \setminus U_c, c \in C$. Then, go to step 1 to generate a new allocation of staff, using b^{k+1}_{ct} as the new lower bound for (12b). Otherwise, if $U_c = \emptyset$, **stop**.

This recursive procedure ensures to not only derive a feasible solution to the master problem, but additionally as $b^0_{ct} = \beta_{ct}$, and b^k_{ct} is subsequently increased by 1, only in the time periods where $L_{ct}(\mathbf{Z}) < \tau$, ensures that x^*_{cj} is derived based on a tight lower bound. This solution should, therefore, serve as a promising input for the second stage. Here, we use a classic TS heuristic structure consisting of a neighborhood adjacent to the

current solution, $N(x_{cj})$, the admissible subset of the neighborhood, $\tilde{N}(x_{cj})$, as well as a "tabu list" L of length $|L| = l$.

Furthermore, we consider two variations of the neighborhood definition. In the first, a probabilistic set of pattern-staff pairs is chosen from the total set $J \times C$. Here, a fraction, p_f, of the pairs are already used in the solution x_{cj}. For each of the chosen pairs, a random number $r \in \{-1, 1\}$ is generated, so that the neighborhood to be tested is $x_{cj} + r$. In the remaining of this paper, we refer to this definition as *add-remove*. In the second variation, a probabilistic set of pattern-staff pairs are chosen from the set $J \times C$ again. However, all pairs must be used in the solution x_{cj}. Then, instead of adding additional staff to the solution, we consider that staff can be moved to another pattern that may, or may not, be used by x_{cj} already. Thus, a *move* is defined by the change $x_{cj} - 1$ followed by $x_{ci} + 1$, where $j \in Z_c$ is the set of patterns that *is* used by staff type $c \in C$, and $i \neq j \in J_c$, where J_c is the set of all patterns that *can* be used by staff type $c \in C$. To make sure that $\sum_{c \in C} \sum_{j \in J} x_{cj}$ is minimized, *moves* are a fraction of size p_f of all elements in the neighborhood, where the rest are pure removals, as in the first neighborhood definition. Thus, we refer to this definition as *move-remove*.

Lastly, in order to evaluate the elements of the neighborhood, let $y \in \mathbb{R}_+$ be a "large" number which defines the penalty of violating (2b), so that the total penalty a solution generates is $\sum_{c \in C} \sum_{t \in U_c} y(\tau - L_{ct}(\mathbf{Z}))$. The function from which we evaluate each solution in the neighborhood is then $\sum_{j \in J} \sum_{c \in C} x_{cj} + \sum_{c \in C} \sum_{t \in U_c} y(\tau - L_{ct}(\mathbf{Z}))$.

Our TS heuristic is presented in Appendix B. The overall structure of the RBA heuristic is presented in Algorithm 2.

Algorithm 2 The overall structure of the Recursive Bound Adaptation heuristic.

$b_{ct} \leftarrow \beta_{ct}$ ▷ Initialize
$x_{cj} \leftarrow SOLVE(b_{ct})$
$U_c \leftarrow EVALUATE(x_{cj})$
while $U_c \neq \emptyset \quad \forall c \in C$ **do** ▷ Adjust bound b_{ct} until x_{cj} is feasible cf. (2b)
    $b_{ct} \leftarrow 1 + \sum_{j \in J} a_{cjt} x_{cj} \quad \forall t \in U_c, c \in C$
    $x_{cj} \leftarrow SOLVE(b_{ct})$
    $U_c \leftarrow EVALUATE(x_{cj})$
end while
$x^*_{cj} \leftarrow x_{cj}$
while *elapsedtime* $<$ *maxtime* **do** ▷ Attempt to improve the solution by using tabu search
    $x^*_{cj} \leftarrow TABUSEARCH(x^*_{cj})$
end while
return x^*_{cj}

4 Results

In this section, we test and apply the continuous-time Markov Chain (CTMC), as well as the Recursive Bound Adaptation (RBA) matheuristic presented in Section 3. Firstly, we derive the truncation of the CTMC and evaluate the model by comparing to a simulation of the associated system. This is presented in Section 4.1. In Section 4.2 we demonstrate the RBA matheuristic by firstly tuning the parameter setting, and subsequently conducting optimizations experiments for a number of different input datasets. We then evaluate our approach by comparing to a simulation, taking all patient classes into account.

4.1 Evaluation of the CTMC Model

Recall from Section 3.1 that our modeling approach assumes a finite upper bound, $M_i \ \forall i \in C$, limiting the number of patients that can be contained in the system at each queue in the network. To decide on a setting of these $|C|$ parameters, we conduct a number of tests, where we gradually increase each parameter in sequence until the maximum probability of attaining the bound respects a predetermined tolerance. In increasing M_i, we choose a sequence going downstream the network, such that the parameter for all potential upstream queues are determined. Furthermore, we realize that as the marginal state probabilities depend on the load of the system, so does the appropriate setting of M_i. For this reason, we conduct our tests using the arrival rate $\boldsymbol{\lambda}$ and the lower bound β_{tc}, cf. the optimization problem (2a)-(2c), yielding the largest load that will ever be encountered by the system. For the remaining of this study, we choose β_{tc} such that $\beta_{tc} = \left\lceil \frac{\lambda_t p_{ic}}{\mu_c(1-p_{cc})} \right\rceil \ \forall t \in T, c \in C$, where p_{ic} defines the probability of a patient going to queue c from the predecessor i. Notice that this definition ensures the minimum number of servers that prevents an over-utilized system for each segment of the time-line separately.
We conduct our tests using three different tolerance levels. The resulting setting of each $M_i \ \forall i \in C$ and the runtimes associated with evaluating the system, is presented in Table 1. Here we notice that both the required state space, as well as the associated runtime, increase excessively, despite the fairly limited size of the state space. This would indicate that the system can become computationally intractable if the arrival rate increases, or a small tolerance is required to attain sufficient accuracy.

M_1	M_2	M_3	M_4	M_5	Tolerance	Runtime (s)
27	62	22	10	4	$5 \cdot 10^{-2}$	1254.7
63	104	35	14	6	$1 \cdot 10^{-2}$	32716.8
63	104	35	15	6	$5 \cdot 10^{-3}$	34992.8

Table 1: Results from adjusting the limit $M_i \ \forall i \in C$. Shows the resulting parameter setting, probability tolerance used in each test, and the runtime associated evaluating the system.

	Service Time		Servers				
#	Distribution	Standard Dev.	w_1	w_2	w_3	w_4	w_5
1	Exponential	$\sigma_c = 1/\mu_c$	1	3	3	2	1
2	Log-normal	$\sigma_c = 1/\mu_c$	1	3	3	2	1
3	Log-normal	$\sigma_c = 2/\mu_c$	1	3	3	2	1
4	Exponential	$\sigma_c = 1/\mu_c$	2	4	3	2	2
5	Log-normal	$\sigma_c = 1/\mu_c$	2	4	3	2	2
6	Log-normal	$\sigma_c = 2/\mu_c$	2	4	3	2	2

Table 2: Overview of the simulation experiments used to assess the CTMC model adequacy. Shows the service time distribution and the number of servers used in each run.

Now, in our subsequent experiments we demonstrate that using the setting $M_1 = 27$, $M_2 = 62$, $M_3 = 22$, $M_4 = 10$ and $M_5 = 4$ is adequate. We conduct these experiments by comparing the marginal state distributions, and waiting times as they were defined in (11) to a discrete-event simulation of the CTMC behavior. We conduct these experiments using two different staff profiles for which we fix the number of servers over the entire week in each queue. Furthermore, we assess the model sensitivity to the assumption that service times are exponentially distributed by comparing to simulations where service times follow a log-normal distribution.
Our simulation model was implemented using the modeling language Matlab, and all experiments were conducted for a simulation time of 416 weeks (8 years) including 8 weeks of burn-in. An overview of all

simulation experiments are presented in Table 2, showing the service time distribution and the staff profile used in each run.
The results were evaluated by graphically comparing the two measures. We assessed the marginal state distributions on the expected state, according to $f_{ci}(\xi)$, by sampling the system state at the beginning of each hour in the simulation period. Further, the waiting time service level was evaluated by sampling waiting times on arrival to the respective queues, and then deriving the fraction corresponding to $L_c(\xi)$ from the resulting distributions. Examples of the experiments from Table 2 are presented in Figure 5, showing the waiting time service level in experiment 1, 2 and 3 for queue 1 (triage) and queue 5 (orthopedic surgeons), respectively.

For the experiments where simulation is compared directly to the CTMC (1 and 4, cf. Table 2), we find that the difference in expected state, as well as waiting time service level, is fairly negligible in all cases. Moreover, when we adjust the service time distribution to log-normal, the change is only distinct in the cases where standard deviation is twice the expected service time. Furthermore, Figure 5 demonstrates that the system is dependent on the service time distribution, but that the sensitivity depends greatly on the queue in focus.

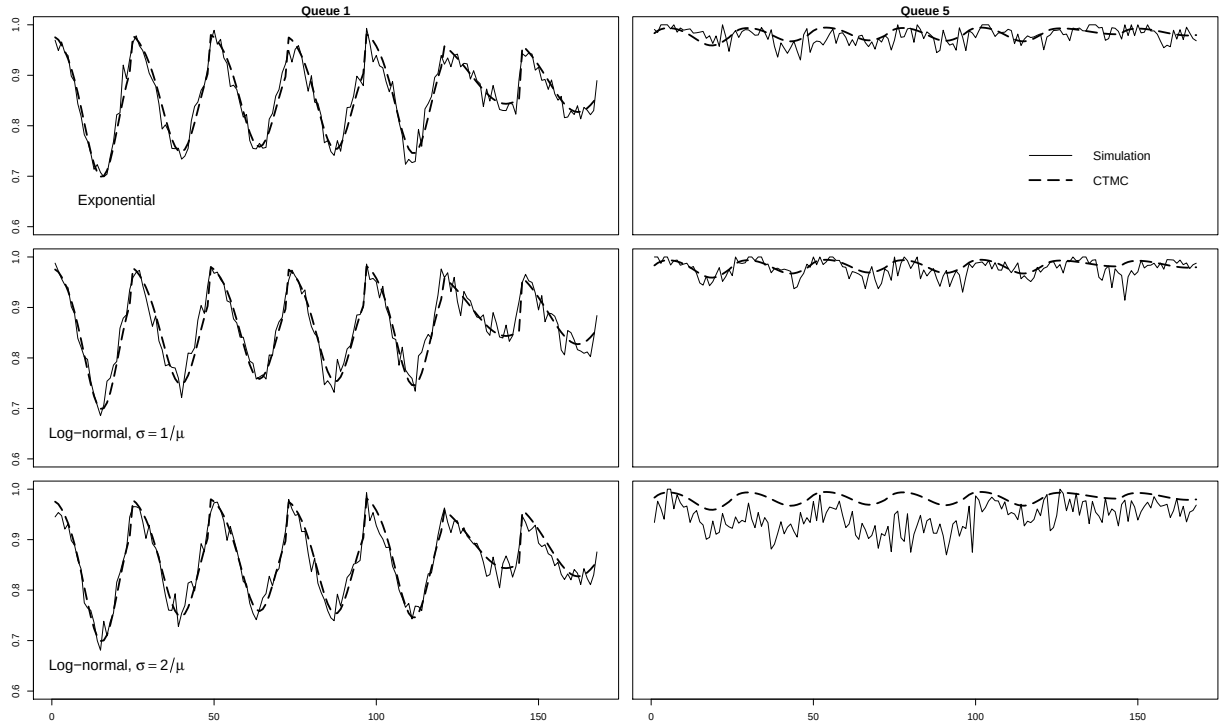

Figure 5: The waiting time service level as function of week-hour. Compares the CTMC model and simulation, on experiment 1, 2 and 3 (cf. Table 2), and queue 1 and 5, respectively.

4.2 Evaluation of the RBA Heuristic

In order to demonstrate our heuristic search procedure we have defined a range of datasets consisting of the interaction between the three arrival patterns, $\Lambda = \{\lambda(\xi) \cdot 0.9^2, \lambda(\xi) \cdot 0.9, \lambda(\xi)\}$ (cf. Section 2.1.1), along with three waiting time service levels, τ, $\mathcal{T} = \{0.7, 0.8, 0.9\}$, cf. (2b). Together, these make up nine different datasets, presented in Table 3.

Reference	Low70	Medium70	High70	Low80	Medium80	High80	Low90	Medium90	High90
Λ	$\lambda(\xi) \cdot 0.9^2$	$\lambda(\xi) \cdot 0.9$	$\lambda(\xi)$	$\lambda(\xi) \cdot 0.9^2$	$\lambda(\xi) \cdot 0.9$	$\lambda(\xi)$	$\lambda(\xi) \cdot 0.9^2$	$\lambda(\xi) \cdot 0.9$	$\lambda(\xi)$
$\mathcal{T}$	0.7	0.7	0.7	0.8	0.8	0.8	0.9	0.9	0.9
Used in	Tuning	Testing	Testing	Testing	Tuning	Testing	Testing	Testing	Tuning

Table 3: Datasets used in parameter tuning and testing of our two heuristic approaches.

To determine the appropriate parameter setting for our subsequent optimization experiments, we apply the RBA heuristic to the following datasets: Low70, Medium80, High90 which essentially represent three different levels of load to the system.
Let z_d^* define the best known solution for dataset $d \in D$, where $D = \{\texttt{Low70}, \texttt{Medium80}, \texttt{High80}\}$. Then, the performance of each specific parameter setting is evaluated for dataset d, by using the average percentage gap, E_d, and variance, σ_d^2, presented in (13a)-(13b),

$$E_d = \frac{1}{N} \sum_{i=1}^{N} \frac{z_i - z_d^*}{z_d^*} \cdot 100\% \qquad \sigma_d^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - E_d)^2 \qquad \text{(13a, 13b)}$$

where z_i and $x_i = (z_i - z_d^*)/z_d^*$ are the resulting fitness and percentage gap of replication $i \in \{1, 2, 3\}$, respectively. Further, to determine the overall performance of each of the tested parameter settings we let E_{tot} define the overall average percentage gap, and σ the pooled standard deviation, presented in (14a)-(14b).

$$E_{tot} = \frac{1}{n_D} \sum_{d=1}^{n_D} E_d \qquad \sigma = \sqrt{\frac{\sum_{d=1}^{n_D} (N-1)\sigma_d^2}{n_D(N-1)}} \qquad \text{(14a, 14b)}$$

We conducted a full interaction test adjusting the penalty y on the levels 40 and 10000, and the fraction p_f on the levels 0.25 and 0.75, respectively. The remaining parameters were fixed based on preliminary testing. The experiments were again conducted on the two variations of the heuristic. That is, the *add-remove* and *move-remove*. Each setting was replicated twice for each of the three datasets with a time-limit of 10 hours. Running these experiments, none of the settings were able to improve the solution subsequent to the first recursive stage of the heuristic. For this reason, we have chosen an arbitrary parameter setting, presented in Appendix A, for our later optimization experiments.

4.2.1 Optimization Experiments

The optimization experiments were conducted on the remaining six datasets: Medium70, High70, Low80, High80, Low90 and Medium90. Each run of the RBA heuristic was replicated three times using af fixed setting of **24 hours of runtime** in the second stage. The ILP problem in (12a)-(12b) was solved by using the IBM ILOG CPLEX Optimizer.

The results for each dataset and variation of the heuristic are presented in Table 4, showing the total amount of staff that is initially derived by the first recursive stage, and subsequently by the second TS stage. The latter is presented in three columns which contains the results obtained in each replication of the heuristic, whereas the first stage is presented in a single column due to its deterministic progression.
Now, our experiments show that the TS variations produce similar and quite consistent results. Regarding the difference between the different datasets, the amount of allocated staff is clearly sensitive to the arrival rate and the specified service level targets. The ILP problem in (12a)-(12b) was solved in less than 10 seconds for all cases, and with 2-6 iterations in the first stage. Moreover, the input for the second TS stage turns out to improve in only a single case, indicating that the first stage returns solutions that are close or exact optimums, *or* since the optimal solution is unknown it may also be the case that our second TS stage implementation is inefficient.

	First Stage			Second Stage					
				TS add-remove			TS move-remove		
Dataset	Allocated Staff	Iterations	Runtime (s)	1	2	3	1	2	3
Medium70	33	2	3017	33	33	33	33	33	33
High70	35	3	4496	35	35	35	35	35	35
Low80	33	2	2839	33	33	33	**32**	33	33
High80	39	3	4571	39	39	39	39	39	39
Low90	40	5	8079	40	40	40	40	40	40
Medium90	42	6	9826	42	42	42	42	42	42

Table 4: Results from testing the RBA heuristic on the remaining datasets. Shows both the solution that is derived in the first stage of the heuristic, and the subsequent (replicated) TS solution.

4.2.2 Solution Evaluation

Now recall from Section 2 that the ED is in fact subject to four different patient classes determining the order in which patients are prioritized. To assess whether the solutions derived in the preceding section has any implications for a system incorporating all four classes, we applied each solution to our discrete-event simulation model from Section 4.1 by distinguishing between patient classes, as well as including the possibility of changing priority subsequent to the second queue.
Once again, our simulation experiments were conducted using a simulation time of 416 weeks including 8 weeks of burn-in. In order to depict the general implications for each respective patient class, we have derived the waiting time service level as an average, *weighted* according to the number of patients arriving at each queue over time.

The results are presented in Table 5 showing each patient class and dataset, respectively. As expected, the service level is increasing in accordance with both the priority of each class and the target of each dataset. Furthermore, we find that the service level is always attained above the target for patients of level 3 and 4, but slightly violated in a single case on level 2 and half of the cases on level 1. In this regard, note that for some ED cases is the service level dependent on the triage level, often yielding a less ambitious waiting time target for patients of lower priority.

Triage / Datasets	Level 1	Level 2	Level 3	Level 4
Medium70	0.801	0.839	0.907	0.970
High70	0.814	0.861	0.917	0.973
Low80	0.787	0.832	0.900	0.967
High80	0.880	0.916	0.968	0.986
Low90	0.876	0.906	0.984	0.994
Medium90	0.855	0.889	0.959	0.984

Table 5: The simulated waiting time service level for each patient class. Presented as an average, weighted according to the amount of patients arriving at each queue over time.

4.3 Discussion

Through Section 3 and 4, we have presented and tested an approach for modeling ED patient flow based on a CTMC, which accounts for the time-dependency in the system resulting from a realistic time-varying arrival rate of patients, and presence of staff. Even though we capture many of the essential elements of an ED, is our model based on a few simplifications such as assuming that service times are exponentially

distributed. We have further noted that our flow system does not incorporate the extensive structure as has been considered by related simulation studies [19, 28]. However, simulation experiments have indicated that our CTMC is robust to the service time distribution, and derives an accurate state distribution faster than the associated simulations. We have further found that our approach, considering only a single merged class of patients, is adequate in evaluating the performance of the associated multi-class system.

For the optimization of the system, we have been investigating a matheuristic approach on a number of different input datasets. The approach consists of firstly a recursive stage, where the lower bound on staff is greedily increased until the constraint on waiting time is respected. As allocating staff to one period affects the service level in all other time periods, optimality cannot be guaranteed by using this procedure. A TS heuristic is, therefore, added to search for any "excess" staff.
The advantage of this procedure is the greedy adaptation of a lower bound, initialized at its lowest possible level, and therefore inclined to produce a promising solution. On the other hand, the approach faces the problem of repeatedly evaluating both the CTMC and ILP problem, which can be computationally expensive for more complex cases. For our case the problem of assigning staff to a limited set of working-patterns can be solved in below 10 seconds, and for this reason the RBA heuristic is able to derive a feasible solution within a reasonable amount of time from the first stage of the heuristic alone.
Lastly, an important question remains as to how far the obtained solutions are from the true optimum. A time-dependent queueing network makes up a range of dependencies, such as the load-dependency between queues in the network, and the effect that one time period has on all other time periods due to the weekly cyclic behavior. There is to our knowledge no standard method of deriving an optimality gap in a system that comprises these relations that does not involve an exhaustive evaluation of all permutations for a fixed sum $\sum_{c\in C}\sum_{j\in J} x_{cj}$, which can be quite computationally expensive, as we have demonstrated earlier. However, recall that results from the second stage in Table 4 could indicate that our solutions are near-optimal, since there is only improvement in a single case.

5 Conclusion & Future Work

In this study, we have aimed at providing a continuous-time Markov chain (CTMC) approach for the modeling of time-varying behavior of patient waiting time, and the interaction of this approach with an Integer Linear Programming (ILP) model. We have tested a matheuristic approach to the problem of allocating staff to an Emergency Department (ED) which we refer to as Recursive Bound Adaptation (RBA).
In the literature, we have found that a range of different methods is used in patient flow modeling, but only a few of these studies considered optimizing the system. Even fewer studies have explored modeling and optimizing the ED based on a time-dependent queueing network. In our study, we have modeled time-dependency by discretization of the patient arrival rate and defining a step function of consecutive uniformizations of the CTMC. By conducting numerous simulation experiments, we have found that this approach is adequate for modeling the system occupancy, as well as waiting time, and is fairly robust to adjustments in the service time distribution.
By applying the CTMC to our matheuristic approach, we provide solutions that satisfy targets on patient waiting time, when we reduce the system to that of only a single class of patients. Further simulation experiments have shown show that these solutions perform well in an associated multi-class system, with only slight violations for the least prioritized patients.

Our model approach has been based on the essential elements of acute patient flow, which might be insufficient for other hospital cases. However, with this study, we have provided a method that adequately evaluates patient waiting times, which do not rely on sampling, and is therefore suitable in the context of optimization. Moreover, we deem that the approach presented in this study, may serve as a basis for further exploration within the area of ED optimization. Finally, the reader should notice that our matheuristic is not only

limited to the specific system nor ILP that has been tested in this study, but can be used for other similar cases.

5.1 Future Work

The approach that was presented in this study provides a range of different aspects to consider in future work. The recursive first stage of the RBA heuristic could be diversified by the use of a restricted candidate list, as in the well-known Greedy Randomized Adaptive Search Procedure (GRASP). Furthermore, over-allocation of staff may be avoided by adjusting time periods of the bound sequentially.
Our study did not include any lower bounds to the master problem presented in (2a)-(2c) which is otherwise necessary to conduct a proper assessment of the solutions obtained by our matheuristic. For this reason, we deem that further work into exact solutions of the relaxed master problem should be considered.
Lastly, the CTMC have provided an analytical approach for evaluating time-dependent patient waiting time in an ED. Further patient data should be obtained to evaluate this modeling approach — for instance on patient waiting time and the service time distributions of each staff type. Moreover, analysis into larger flow systems should be studied to approach ED cases of more complex structure.

Acknowledgements

This research was supported and funded by Region Sjælland. The managing organization of seven public hospitals. Particularly, we thank the department of Production, Research and Innovation (Produktion, Forskning og Innovation) for their support in providing data and insight into the operations of the Danish hospitals, and Associate Prof. Anders Stockmarr for statistical advice.

References

[1] Status på sundhedsområdet. Ministry of Health, 2015.

[2] Mohamed A. Ahmed and Talal M. Alkhamis. Simulation optimization for an emergency department healthcare unit in kuwait. *European Journal of Operational Research*, 198(3):936–942, 2009.

[3] Anders R. Andersen. Assessment of capacity and waiting time in emergency departmentst. Master's thesis, Technical University of Denmark, Department of Engineering Management, Anker Engelunds Vej 1, 2800 Kgs. Lyngby, Denmark, 2014.

[4] Mor Armony, Shlomo Israelit, Avishai Mandelbaum, Yariv N. Marmor, Yulia Tseytlin, and Galit B. Yom-Tov. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, 5(1):146–194, 2015.

[5] Gabriel R. Bitran and Reinaldo Morabito. Open queueing networks: Optimization and performance evaluation models for discrete manufacturing systems.

[6] Eduardo Cabrera, Emilio Luque, Manel Taboada, Francisco Epelde, and Ma Luisa Iglesias. Abms optimization for emergency departments. *Proceedings - Winter Simulation Conference*, page 6465116, 2012.

[7] B Cheang, H Li, A Lim, and B Rodrigues. Nurse rostering problems - a bibliographic survey. *European Journal of Operational Research*, 151(3):447–460, 2003.

[8] Jeffery K. Cochran and Kevin T. Roche. A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Computers and Operations Research*, 36(5):1497–1512, 2009.

[9] Dorsaf Daldoul, Issam Nouaouri, Hanen Bouchriha, and Hamid Allaoui. Optimization on human and material resources in emergency department. pages 633–638, 2015.

[10] Mieke Defraeye and Inneke Van Nieuwenhuyse. Staffing and scheduling under nonstationary demand for service: A literature review. *Omega-international Journal of Management Science*, 58:4–25, 2016.

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

[12] Avi Giloni and Sridhar Seshadri. Optimal configurations of general job shops. *Queueing Systems*, 39(2-3):137–155, 2001.

[13] E.C. Jauch, J.L. Saver, H.P. Adams, A. Bruno, J.J.B. Connors, B.M. Demaerschalk, P. Khatri, P.W. McMullan Jr., A.I. Qureshi, K. Rosenfield, P.A. Scott, D.R. Summers, D.Z. Wang, M. Wintermark, and H. Yonas. Guidelines for the early management of patients with acute ischemic stroke: A guideline for healthcare professionals from the american heart association/american stroke association. *Stroke*, 44(3):870–947, 2013.

[14] O.B. Jennings, A. Mandelbaum, W.A. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394, 1996. cited By 131.

[15] Mohammed Khadem, Hamdi A. Bashir, Yasin Al-Lawati, and Fatma Al-Azri. Evaluating the layout of the emergency department of a public hospital using computer simulation modeling: A case study. *I C Indus E*, pages 1709–1713, 2008.

[16] Ketki Kulkarni and Jayendran Venkateswaran. Iterative simulation and optimization approach for job shop scheduling. *Proceedings - Winter Simulation Conference*, 2015-:7020013, 1620–1631, 2015.

[17] Laszlo Lakatos, Laszlo Szeidl, and Miklos Telek. *Introduction to Queueing Systems with Telecommunication Applications*. Springer, 2013.

[18] D C Lane, C Monefeldt, and J V Rosenhead. Looking in the wrong place for healthcare improvements: A system dynamics study of an accident and emergency department. *Journal of the Operational Research Society*, 51(5):518–531, 2000.

[19] Marek Laskowski, Robert D. McLeod, Marcia R. Friesen, Blake W. Podaima, and Attahiru S. Alfa. Models of emergency departments for reducing patient waiting times. *PLoS One*, 4(7):Article No.: e6127, 2009.

[20] S. Liao, G. Koole, C. van Delft, and O. Jouini. Staffing a call center with uncertain non-stationary arrival rate and flexibility. *OR Spectrum*, 34(3):691–721, 2012.

[21] Morgan E. Lim, Tim Nye, James M. Bowen, Jerry Hurley, Ron Goeree, and Jean-Eric Tarride. Mathematical modeling: The case of emergency department waiting times. *International Journal of Technology Assessment in Health Care*, 28(2):93–109, 2012.

[22] 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 Management Science*, 11(1):11–21, 2008.

[23] D. J. Medeiros, Eric Swenson, and Christopher DeFlitch. Improving patient flow in a hospital emergency department. *2008 Winter Simulation Conference, Vols 1-5*, pages 1526–1531, 2008.

[24] Richard J. Mullins and N. Clay Mann. Population-based research assessing the effectiveness of trauma systems. *Journal of Trauma: Injury, Infection, and Critical Care*, 47(SUPPLEMENT):S59–S66, 1999.

[25] Ronny M. Otero, H. Bryant Nguyen, David T. Huang, David F. Gaieski, Munish Goyal, Kyle J. Gunnerson, Stephen Trzeciak, Robert Sherwin, Christopher V. Holthaus, Tiffany Osborn, and Emanuel P. Rivers. Early goal-directed therapy in severe sepsis and septic shock revisited - concepts, controversies, and contemporary findings. *Chest*, 130(5):1579–1595, 2006.

[26] Justus Arne Schwarz, Gregor Selinka, and Raik Stolletz. Performance analysis of time-dependent queueing systems: Survey and classification. *Omega-international Journal of Management Science*, 63:170–189, 2016.

[27] Sridhar Seshradi and Michael Pinedo. Optimal allocation of resources in a job shop environment. *Iie Transactions Industrial Engineering Research and Development*, 31(3):195–206, 1999.

[28] D Sinreich and Y Marmor. Emergency department operations: The basis for developing a simulation tool. *Iie Transactions*, 37(3):233–245, 2005.

[29] David Sinreich, Ola Jabali, and Nico P. Dellaert. Reducing emergency department waiting times by adjusting work shifts considering patient visits to multiple care providers. *Iie Transactions Industrial Engineering Research and Development*, 44(3):163–180, 2012.

[30] J.M. Smith, F.R.B. Cruz, and T. Van Woensel. Optimal server allocation in general, finite, multi-server queueing networks. *Applied Stochastic Models in Business and Industry*, 26(6):705–736, 2010.

[31] William J. Stewart. *Probability, Markov Chains, Queues, and Simulation - The Mathematical Basis of Performance Modeling*. Princeton University Press, 1 edition, 2009.

[32] Alan B. Storrow, Chuan Zhou, Gary Gaddis, Jin H. Han, Karen Miller, David Klubert, Andy Laidig, and Dominik Aronsky. Decreasing lab turnaround time improves emergency department throughput and decreases emergency medical services diversion: A simulation model. *Academic Emergency Medicine*, 15(11):1130–1135, 2008.

[33] David Y. Sze. Queueing model for telephone operator staffing. *Operations Research*, 32(2):229–249, 1984.

[34] D. Tipper and M.K. Sundareshan. Numerical methods for modeling computer networks under nonstationary conditions. *IEEE Journal on Selected Areas in Communications*, 8(9):1682–1695, 1990.

[35] Lu Wang. An agent-based simulation for workflow in emergency department. pages 19–23, 2009.

[36] X. Wang. Emergency department staffing: A separated continuous linear programming approach. *Mathematical Problems in Engineering*, 2013, 2013.

[37] W. Whitt. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, 15(1):88–102, 2006.

[38] JY Yeh and WS Lin. Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department. *Expert Systems With Applications an International Journal*, 32(4):1073–1083, 2007.

[39] K. Yoneda, I. Wada, and K. Haruki. Job shop configuration with queueing networks and simulated annealing. pages 407–410, 1992.

Appendices

A Parameters

Staff Types	Service Times (h)
Triage Nurse	1/6
Basic Physicians	1/3
Specialized Medical Physicians	3/4
Organ Surgeons	3/4
Orthopedic Surgeons	3/4

Table 6: Assumed average service times for each of the five staff types.

From \ To	1	2	3	4	5	Discharge
1		1.00				
2		0.10	0.53	0.25	0.11	0.01
3			0.50			0.50
4				0.50		0.50
5					0.50	0.50

Table 7: Routing probabilities for the queueing network presented in Figure 1.

Staff Type	Waiting Time Target (h)
Triage Nurse	1/6
Basic Physician	1
Specialized Medical Physician	3
Organ Surgeons	3
Orthopedic Surgeons	3

Table 8: Waiting time targets, ν_c, used to evaluate the performance of the ED.

Variation	l	a	p_f	y
add-remove	15	5	0.75	40
move-remove	15	5	0.75	10,000

Table 9: Parameters used in the Recursive Bound Adaptation tests.

B Algorithms

Algorithm 3 The tabu search heuristic.

$x_{cj} \leftarrow INITIALIZE(), L \leftarrow \emptyset$ ▷ Initialize solution x_{cj} and tabu list L
$x^*_{cj} \leftarrow x_{cj}, f^* \leftarrow EVALUATE(x^*_{cj})$
while $elapsedtime < maxtime$ **do**
    $N \leftarrow CREATE(x_{cj}, p_f, a)$ ▷ Create neighborhood of size a, using solution x_{cj} and fraction p_f
    $j \leftarrow 1, b \leftarrow N[j]$
    for $i = 2$ to $|N|$ **do** ▷ Find the best solution in the neighborhood
        $f \leftarrow EVALUATE(N[i])$
        if $f < b$ and ($N[i] \notin L$ or $f < f^*$) **then**
            $b \leftarrow f, j \leftarrow i$
        end if
    end for
    $x_{cj} \leftarrow N[j], L \leftarrow UPDATE(N[j], l)$ ▷ Move to the best permissible solution and update the tabu list
    if $f < f^*$ **then**
        $f^* \leftarrow f, x^*_{cj} \leftarrow x_{cj}$ ▷ Save the best known solution
    end if
end while
    return x^*_{cj}