Operations Research for Health Care 8 (2016) 62–70

Contents lists available at ScienceDirect

Operations Research for Health Care

journal homepage: www.elsevier.com/locate/orhc

Assigning treatment rooms at the Emergency Department

Maartje van de Vrugt[a,b,c,*], Richard J. Boucherie[a,b]

[a] Centre for Healthcare Operations Improvement and Research (CHOIR), University of Twente, Enschede, The Netherlands

[b] Stochastic Operations Research, Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

[c] Department of Quality and Safety, Jeroen Bosch Hospital, 's Hertogenbosch, The Netherlands

ARTICLE INFO

Article history:
Received 2 November 2014
Accepted 29 June 2015
Available online 7 August 2015

Keywords:
Queueing theory
Emergency department
Discrete event simulation

ABSTRACT

Increasing efficiency at the Emergency Department (ED) reduces overcrowding. At the ED in typical Dutch Hospitals treatment rooms are mostly shared by two residents of different specialties: a Surgeon and an Internist. Each resident uses multiple rooms in parallel; while one patient awaits test results in a treatment room, the resident visits other patients. The assignment of rooms among the residents is often unbalanced, which affects the blocking probability and waiting and sojourn times of patients. Invoking a queueing model in a random environment, we analytically investigate expected sojourn times of (semi-urgent) patients for both types of residents for different room assignment policies and working routines of the residents. We determine the Pareto efficient policies and working routines for all performance measures. We conduct a Discrete Event Simulation to validate our model and present numerical results for a large Dutch teaching hospital and other illustrative cases.

1. Introduction

Emergency department (ED) overcrowding has many negative effects, for both patients and ED staff. ED overcrowding leads for example to an increase in the number of medical errors, long waiting times and high levels of stress for ED staff (see e.g. [1–3]). Therefore, governments have set legal norms on the waiting and sojourn times of patients visiting the ED. Meeting these norms is often difficult due to an increasing demand for acute care and the closure of many EDs and hospital beds [4].

The waiting and sojourn times of ED patients are influenced by many factors, among others the patient's urgency class and type. In the Netherlands, patients are both triaged and assigned to a specialty upon arrival at the ED. The triage category of a patient determines the order in which patients are treated. Treatment can start when there is a treatment room and the patient's physician is available. The specialties with the highest number of patients are represented at the ED at all times, either by a specialist or a resident (together with physicians collectively called 'doctors' in this paper). Other specialties have to be called for assistance, implying that patients at least have to wait for the travel time of the doctor.

Typically at an ED, one doctor occupies multiple rooms in parallel; when a patient requires diagnostic tests, the doctor visits other patients. Each specialty/doctor has a different working routine: Emergency Physicians are for example specially trained for only providing urgent care, while other doctors tend to provide care more like they are used to outside the ED [4]. As a consequence, sojourn and waiting times of different patient types differ significantly, and the room utilization can be unbalanced over the doctor types. The unbalanced room utilization possibly also affects the sojourn and waiting times of the patients.

Besides the sojourn and waiting times, the probability that the ED is 'full' (*blocking probability*) is an important performance measure. When the number of patients at the ED exceeds a certain level, the ED staff will call in an extra doctor and/or defer new patients. The first will result in additional costs, while the latter negatively affects the image of the hospital.

In this paper we investigate several room assignment policies and doctor working routines at an ED and determine the Pareto-efficient combinations with respect to the performance measures mentioned above. To this end, we invoke a continuous time queueing model in which patients may require diagnostic tests, implying that doctors visit their patients a random number of times and interact in sharing the available treatment rooms. We investigate a case study of a large Dutch teaching hospital (the Jeroen Bosch Hospital, JBH) and several additional illustrative cases. By means of a Discrete Event Simulation, we investigate how the obtained policies perform in a more realistic setting.

* Correspondence to: Centre for Healthcare Operations Improvement and Research (CHOIR), University of Twente, Postbox 217, 7500 AE, Enschede, The Netherlands.

E-mail address: n.m.vandevrugt@utwente.nl (M. van de Vrugt).

http://dx.doi.org/10.1016/j.orhc.2015.06.004

We also study the sojourn time of an accepted patient conditioned on the state found in the ED upon arrival. This conditional waiting and sojourn times information may be provided to the patients to inform them about their expected length of stay at the ED, which adds to patient satisfaction [5].

In the next section we will provide an overview of related literature. The model is described in Section 3. Section 4 provides the analytical results, Section 5 the numerical results, and Section 6 the conclusion and discussion.

2. Related literature

The literature that is related to this paper consists of papers on EDs and of queueing models that are similar to the one considered in this paper. Both types of literature will be summarized below.

EDs are modeled most often by means of queueing theory and computer simulation. Typical focus areas of these works are capacity dimensioning and patient routing, with the objectives to minimize patient waiting time, maximize patient throughput or increase staff utilization [6]. In patient routing the effects of for example parallelization of tasks, a 'fast-track' system, and letting patients wait outside the treatment room are studied.

There are many papers that apply queueing models to an ED, but only a few papers are closely related to this paper as they consider patients who are visited more than once by the doctor, cf. [7–9]. These studies consider a multi-server queueing network in which customers can re-enter a queue, and the authors aim to determine appropriate staffing levels to achieve certain performance targets. Differences between the models in these references (and the references therein) and the model considered here, are that patients do not return to their own doctor and that it is assumed that treatment can start once the doctor is available. This implies that either there always is a sufficient number of treatment rooms, or patients await their test results outside the room.

In a recent review, Wiler et al. [3] state that most queueing approaches to modeling an ED assumed that: (1) arrival rates are constant, (2) patients do not deteriorate, and (3) patients are not seen by same doctor again. Here we also assume (1) and (2), but allow for patients to be seen by the same doctor more than once. For more literature on modeling EDs, the reader is referred to recent literature reviews [2,3,6] and the references therein.

A different but related topic is the optimal assignment of servers to two types of calls in a callcenter, studied by e.g. Bhulai and Koole [10]. Here the question is when to assign servers to outbound calls in such a way that inbound calls still achieve high service levels. Differences with our work are that in [10] each call visits the server only once, and each server has its own line (opposed to doctors sharing multiple treatment rooms).

Our contribution to both fields of literature is that we incorporate the ordering of diagnostic tests and doctors revisiting their patients in a queueing model for an ED. We consider a model with two resource types (doctors and rooms) in which a customer can only get service when both resource types are available at the same time. To the best of our knowledge, this type of model has not been analyzed before. This model is applied to an ED, but could easily be adapted to, for example, an outpatient clinic in which multiple physicians share the available treatment rooms.

The method used in this paper has been applied before to telecommunication systems, see for the most related papers [11–14]. These systems typically have two call types: inelastic (speech) calls, and elastic (data) calls. The inelastic calls require one unit of capacity (bandwidth) throughout their entire service time. The elastic calls share the remaining capacity in a processor sharing fashion. Important performance measures are the (average) throughput, sojourn time and blocking probability. At an ED, multiple types of patients also have to share the available capacity (treatment rooms). Although the processor sharing service discipline is not applicable for an ED, it will be shown that the methods applied to telecommunication systems can still be applied.

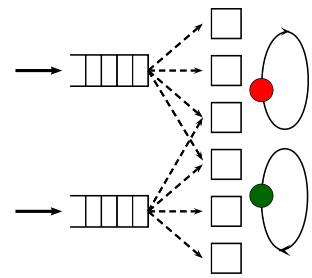

Fig. 1. Schematic representation of the queueing network.

3. The model

In this paper, we focus on a system with two doctors, since in many Dutch hospitals such as the JBH most of the non-urgent patients require a Surgeon or an Internist and both specialties are represented at the ED at all times. Patients that require other specialties are either treated by an Emergency Physician (working at the ED 7:00–23:00 h) or by a specialist doctor who is called for assistance.

Upon arrival at the ED, patients are triaged, assigned to a doctor, and join a waiting area. We only consider patients that are triaged as non-urgent, and are therefore treated on a first come, first serve (FCFS) basis. The triage is typically not the bottleneck at an ED, and therefore we only focus on the process after triage.

To model this ED in which two doctors share multiple treatment rooms to treat two types of patients with equal priority and each type of patient has its own queue and doctor, we consider a queueing system as depicted in Fig. 1. There are more treatment rooms than doctors, depicted by squares and circles respectively. The number of patients in the system is finite for both patient types and patients are assumed to arrive at their assigned queue according to a random process with constant arrival rate (we numerically investigate this assumption in Section 5). When the required doctor is inactive upon arrival of a patient and a treatment room is available, the treatment commences immediately. Otherwise, the patient is either blocked (if the maximum number of patients of this type is reached) or placed in a waiting area.

The treatment of a patient consists of at least one visit by the doctor; between two visits of the doctor, patients take diagnostic tests. At the instant the treatment of a patient begins, he is assigned a treatment room. This room will be assigned to this patient during the entire treatment time, even if the patient is not always physically in the room, which is the common policy at Dutch hospitals. We call each time a doctor visits a patient a 'phase' and the 'consultation time' of a patient denotes the total time a doctor has to spend with this patient in a room (so the sum of all phases). The treatment time thus includes the consultation time and possibly time between two doctor visits and time for taking diagnostic tests. To enhance tractability, we assume that both types of doctors visit their patients in random order. The sojourn time equals the waiting time before a room is assigned plus the treatment time.

During each visit the doctor can decide that the patient requires (further) diagnostic tests. Therefore, after each phase completion a patient stays in the system with a certain probability. When a patient's phase is completed but this patient does not leave the ED, the doctor can consult either a patient from the waiting area ('new patient') or in another treatment room ('existing patient'). The first option requires an extra room. The decision which of the two options to choose, depends on the working routine of the doctor and will be varied. If at phase completion the patient cannot leave the system and there is no other patient in the system, the doctor will become 'inactive', e.g. perform administrative tasks, for a random time. If a new patient arrives during an inactive period and there is a room available, the new patient is consulted immediately.

An interesting feature of this model is that in order to be treated, each patient requires two resources at the same time: a room and a doctor. Rooms 'serve' patients in a first come first serve fashion, while doctors visit their patients in random order. A patient occupies a room throughout its entire treatment time, while the doctor will only be in the room with the patient occasionally. Without the doctor present in the room, the treatment of the patient is 'paused'.

The system under consideration can be modeled as a continuous time Markov process with two types of patients and one doctor for each patient type. This system has two independent Poisson arrival streams with rates λ_1 and λ_2. The state of the system is denoted by $\mathbf{n} = (n_1, r_1, i_1; n_2, r_2, i_2)$, with: n_j the number of patients in system; r_j the number of rooms occupied; $i_j \in \{0, 1\}$ an indicator that equals 1 if the doctor is with a patient and 0 otherwise; and $j = 1, 2$ denotes the different patient and doctor types. The indicator is used when there is a phase completion of a patient that is the only patient present of his type. At such phase completions, the doctor will be inactive until either the patient returns from his tests, or a new patient of his type arrives. The time a doctor is inactive is therefore the minimum of two exponential times, with parameters ν_j and λ_j respectively.

Each phase of consultation time is exponentially distributed with rate μ_j for type j. Define p_j the probability that the patient's consultation time ends after completion of a phase. Then, the consultation time consists of a geometric number of exponential phases, and thus it is readily derived that its probability mass function is given by $f_j(x) = p_j\mu_j \exp(-p_j\mu_j x)$. Furthermore, let M_j be the maximum number of patients in the system for type j. The generator $\mathbf{Q}$ for this process is developed in Appendix A.

4. Analytical results

Since for both patient types the analysis is the same, we will omit most patient type subscripts in the development which follows, using the subscript j to indicate one of the types where necessary, and refer to 'patient' instead of mentioning the types. The analysis follows the lines of e.g. [12,14].

4.1. Basic performance measures

We obtain the stationary distribution $\boldsymbol{\pi}$ of this system by solving $\boldsymbol{\pi}\mathbf{Q} = 0$. Using π, we can obtain average performance measures like the expected number of patients in the queue (N^q) and system (N), and the blocking probability P^b for both patient types:

$$N_j^q = \sum_{\mathbf{n}|n_j \geq r_j} (n_j - r_j) \cdot \pi_{\mathbf{n}}, \qquad N_j = \sum_{\mathbf{n}} n_j \cdot \pi_{\mathbf{n}},$$
$$P_j^b = \sum_{\mathbf{n}|n_j = M_j} \pi_{\mathbf{n}}. \tag{1}$$

Using Little's formula, we can obtain the expected waiting time W_j and sojourn time S_j of an accepted patient by:

$$W_j = \frac{N_j^q}{\lambda_j\left(1 - P_j^b\right)}, \qquad S_j = \frac{N_j}{\lambda_j\left(1 - P_j^b\right)}. \tag{2}$$

4.2. The conditional expected sojourn time

The sojourn time of a patient consists of two stages: the time from triage until entering the treatment room (waiting time) and the treatment time. We determine the conditional expected sojourn time by tagging an arriving patient and running the Markov chain from the state the patient encountered upon arrival until the patient leaves the system. The conditional expected waiting time $\boldsymbol{\alpha}_{\mathbf{n},j}$ depends on the number of patients in the system upon arrival of the tagged patient. The treatment time of the tagged patient depends both on his consultation time, and the (possibly varying) number of rooms his doctor uses during his treatment time. The conditional expected treatment time $\boldsymbol{\tau}_{\mathbf{n}',j}(x)$ of the tagged patient is conditioned on his consultation time x and the state of the system in which the treatment of this patient commences. This state is not necessarily the state which the tagged patient encountered upon arrival. Therefore, we also need to determine the conditional transition probabilities $\boldsymbol{\psi}_j(\mathbf{n}, \mathbf{n}')$: the probability that a patient arriving in state $\mathbf{n}$ commences his service in state $\mathbf{n}'$. The expected sojourn time of an admitted patient, conditioned on the state of the system at the instant the patient has arrived, is then expressed by:

$$\mathbf{S}_{\mathbf{n},j} = \int_0^\infty \sigma_{\mathbf{n},j}(x) f_j(x) \mathrm{d}x,$$
$$\text{with } \boldsymbol{\sigma}_{\mathbf{n},j}(x) = \boldsymbol{\alpha}_{\mathbf{n},j} + \sum_{\mathbf{n}'} \boldsymbol{\tau}_{\mathbf{n}',j}(x)\boldsymbol{\psi}_j(\mathbf{n}, \mathbf{n}'). \tag{3}$$

Here, $\boldsymbol{\sigma}_{\mathbf{n},j}(x)$ denotes the conditional sojourn time of a patient of type j arriving in state $\mathbf{n}$ (including the new patient) with consultation time x. Note that the sojourn time does not exist for certain states, for example states with $n_j = 0$. From the conditional expected sojourn time, the expected sojourn time can be obtained by:

$$S_j = \sum_{\mathbf{n}|n_j>0} \frac{\pi_{\mathbf{n}-\mathbf{e}_j}}{\sum\limits_{\mathbf{n}|n_j>0} \pi_{\mathbf{n}-\mathbf{e}_j}} S_{\mathbf{n},j} = \sum_{\mathbf{n}|n_j>0} \frac{\pi_{\mathbf{n}-\mathbf{e}_j}}{1 - P_j^b} \cdot S_{\mathbf{n},j},$$

with $\mathbf{n} - \mathbf{e}_j$ denoting a system with one patient of type j less. Eq. (3) consists of multiple parts that can be obtained by the steps summarized in Appendix B, and in more detail in [13,14]. We implemented Eq. (3) and all calculation steps in MATLAB® [15] in order to perform a numerical study.

5. Numerical results

In this section we first describe the JBH case study and its results. Hereafter, we provide numerical results for several illustrative cases. The policies we investigate for this queueing model are all of threshold-type: for both specialties we define a maximum number of rooms δ_j that specialty j can use in parallel. The room assignment policies we investigate only include policies in which all rooms could theoretically be used; for example for a case with five rooms available the policy $(\delta_1, \delta_2) = (3, 1)$ was not evaluated since the maximum number of rooms in use is four here. We further distinguish two different working routines per doctor: either he visits a new patient first at phase completion (called NF), or he returns to an existing patient first (called EF). The working routines are therefore denoted by: EF–EF, EF–NF, NF–EF and NF–NF (in which the working routine of doctor 1 is mentioned first). For all performance measures, we determine the Pareto efficient 'working routine + room assignment policy' combinations (RPCs), which we denote by (*working routine doctor 1*, δ_1, *working routine doctor 2*, δ_2).

In addition to the analytical results, we conduct a discrete event simulation study to investigate the policies in a more realistic setting: the arrival rates are time-dependent and obtained from data of the JBH, see Fig. 2. We simulate 5000 days for 20 runs such that the relative precision of the waiting and sojourn times is always below 1%. We use common random numbers [16].

5.1. JBH case study

Data. We only consider Surgical (type 1) and Internal Medicine (type 2) patients in our case study, which represents 66% of the

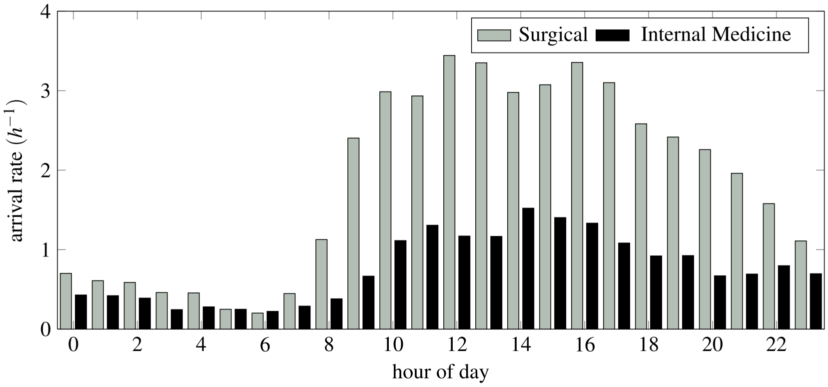

Fig. 2. Time-dependent arrival rates obtained from JBH data.

Table 1
Input parameters for the JBH case study.

	λ (h^{-1})	μ (h^{-1})	ν (h^{-1})	p	M
Surgical	3.15	8.57*	1.93*	0.45*	15
Internal medicine	1.26	7.09	1.93	0.30*	10

* Parameter is estimated.

non-urgent ED-patients at the JBH. Typically at an ED, arrival rates are not constant during the day. The problem of dividing rooms among the doctors is most prevalent during the peak-hours. Therefore, we take the average arrival rate between 10 a.m. and 5 p.m. as input for the model, and refer to this rate as the 'peak arrival rate'. All input parameters are obtained from JBH data or estimated when not available, and are displayed in Table 1.

In the current working routine of the JBH, there is no policy for the assignment of rooms and both doctors visit new patients first. There are eight available treatment rooms at the ED. In our study, since we only consider 66% of the patients, we assume that there are only five treatment rooms. For five treatment rooms, there exist 19 allowed room assignment policies: (5, 5), (5, 4), (5, 3), (5, 2), (5, 1), (4, 5), (4, 4), (4, 3), (4, 2), (4, 1), (3, 5), (3, 4), (3, 3), (3, 2), (2, 5), (2, 4), (2, 3), (1, 5), (1, 4). There are two possible working routines for each doctor, thus the total number of RPCs investigated is $19 \times 4 = 76$.

Basic performance measures. We refer to the RPC (NF, 5, NF, 5) as the current situation; both doctors visit new patients first and are allowed to use all treatment rooms in parallel. From the analytical model we obtain that for the current situation the expected sojourn time equals 3.4 h for Surgical patients, and 3.9 h for Internal Medicine patients. These values are significantly higher than the measured average sojourn times, as was expected since the peak-arrival rate is used in the analytical model. In order to determine the Pareto efficient RPCs, we consider six objectives: three performance measures for both patient types. For each performance measure we therefore determine the Pareto front from all RPCs; if an RPC is Pareto efficient it is impossible to establish an improvement for one patient type without increasing the blocking probability, the expected waiting time and/or the expected sojourn time for the other type. All RPCs that are not Pareto efficient are said to be dominated. There exist RPCs for which all three performance measures are Pareto efficient, which we denote by '*efficient*' RPCs.

The Pareto front for the expected waiting time is depicted in Fig. 3. Here all markers indicate one or multiple RPCs and the different marker-types indicate the different working routines. One marker can indicate multiple RPCs; for example for the routine where both doctors treat existing patients first (EF–EF, marker square), both doctors will by assumption never occupy more than two rooms in parallel, resulting in the same expected waiting time for all policies with $\delta_j \geq 2$.

Both for the blocking probability and the expected sojourn time the Pareto-front consists of a single (P_1^b, P_2^b)- and (S_1, S_2)-coordinate respectively, and this optimum is attained in the policies $(\delta_1, \delta_2) \in \{(3, 3); (3, 2); (2, 3)\}$ for all working routines.

For the JBH case study, the *efficient* RPCs and the resulting performance measures are given in Table 2. It appears that the improvements when *efficient* policies are used are significant, especially for type 1 patients. The *efficient* policies all require both doctors to visit new patients first when a patient finishes (a phase of) his treatment. This policy implies that both doctors occupy as many treatment rooms in parallel as possible, which is beneficial for the waiting time of patients. The optimal room assignment policies indicate in all cases that one doctor should never use less than two rooms in parallel, which ensures that the doctor does not have to be inactive while waiting for test results. The room assignment policies (3, 2) and (2, 3) ensure that each doctor has his own treatment rooms, and no rooms are used by both doctors. For this case study sharing all the available rooms, which is the current room assignment policy at the JBH, appears to result in higher blocking probability, waiting time and/or sojourn time for patients.

Simulation results. From the simulation study, as expected, it appears that the value of all performance indicators decreases when time-varying arrival rates are incorporated. For example, for the current policy the average sojourn time in the simulation is 1.1 h for type 1 patients and 1.6 h for type 2 patients, and the average sojourn times with the Pareto efficient RPCs are 1.0 h and 1.2 h respectively. The improvement achieved with the Pareto efficient RPCs is less than in the analytical results: −6% and −26% respectively.

Similar to the results of the analytical model, in the simulation results a single (P_1^b, P_2^b)-coordinate forms the Pareto front for the blocking probability. For the average sojourn time, three Pareto efficient coordinates exist; of these coordinates two are uniquely attained and one is attained in 30 RPCs.

For 94% of the (3 indicators $\times$ 76 RPCs $=$) 228 performance indicators, the conclusion whether the indicator is Pareto efficient is the same in the simulation results and analytical results. From the analytical model three RPCs appear to be *efficient* (see Table 2), but only (NF, 3, NF, 3) is *efficient* in the simulation results.

Conditional performance measures. We study the conditional expected sojourn time when the doctors both visit a new patient first at phase completion, since the JBH already works according to this routine, and from the analytical model it appears to be the only routine with Pareto efficient RPCs. Figs. 4 and 5 display the conditional expected sojourn time of both patient types for different system states with $n_2 = 5,\ r_2 = 1,\ i_2 = 1$ (a typical example) when, respectively, room assignment policies (3, 3) and (4, 2) would be implemented and both doctors see new patients first.

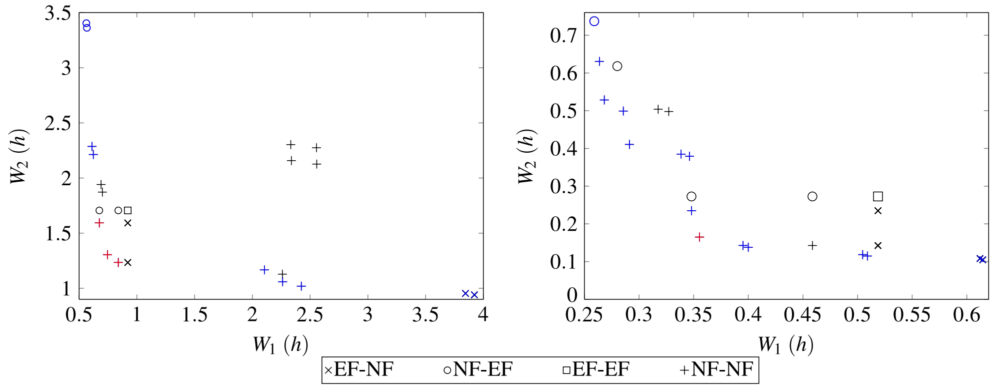

Fig. 3. Analytical (left) and simulated (right) waiting times for different policies for the JBH case study, non-dominated marked blue, *efficient* marked red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

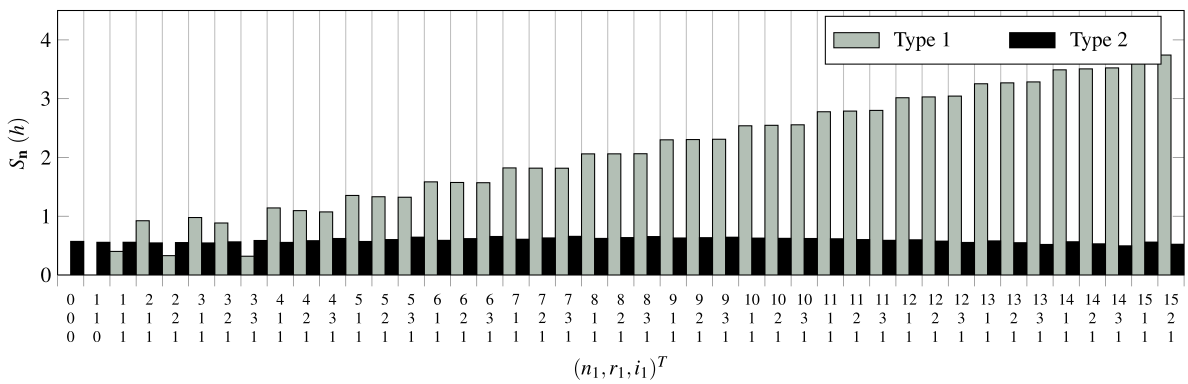

Fig. 4. Conditional sojourn time for states with $n_2 = 5,\ r_2 = 1,\ i_2 = 1$, RPC (NF, 3, NF, 3).

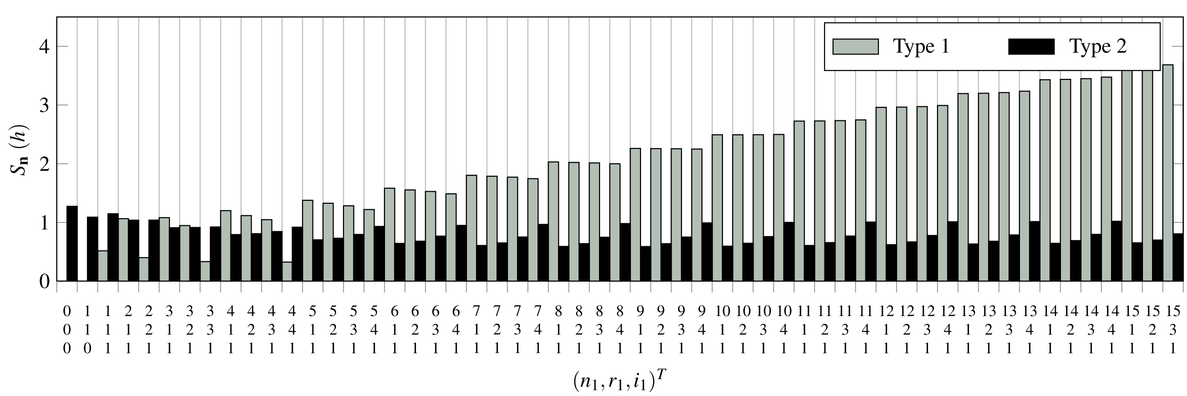

Fig. 5. Conditional sojourn time for states with $n_2 = 5,\ r_2 = 1,\ i_2 = 1$, RPC (NF, 4, NF, 2).

Table 2
Waiting times for the current situation (1st line) and the *efficient* RPCs, all with routine NF–NF.

δ_1	δ_2	W_1	W_2	S_1	S_2	P_1^b	P_2^b
5	5	2.56	2.27	3.39	3.89	0.18	0.19
3	3	0.75 (−71%)	1.30 (−43%)	1.34 (−60%)	2.53 (−35%)	0.01 (−94%)	0.07 (−64%)
3	2	0.67 (−74%)	1.59 (−30%)	1.34 (−60%)	2.53 (−35%)	0.01 (−94%)	0.07 (−64%)
2	3	0.84 (−67%)	1.23 (−46%)	1.34 (−60%)	2.53 (−35%)	0.01 (−94%)	0.07 (−64%)

The conditional expected sojourn time for type 1 patients increases with the number of patients present upon arrival, and decreases with the number of rooms that the type 1 doctor occupies for states with relatively few type 1 patients. The conditional expected sojourn time of a type 2 patient is almost equal for all system states in Fig. 4. However, for policies that are not *efficient*, there is a difference for states with relatively many rooms; compare for example states with $n_1 = 6$ and $r_1 = 3, 4$ in Fig. 5. From this figure it also appears that the expected sojourn time of type 2 patients is higher when there are few type 1 patients at arrival, because for these states the type 2 doctor can take an additional room and as a consequence the sojourn time of all existing patients increases.

Table 3
Input parameters for cases with eight treatment rooms.

Case	λ_1	μ_1	ν_1	p_1	M_1	λ_2	μ_2	ν_2	p_2	M_2
1	6.00	12.00	2.00	0.67	15	2.00	7.50	2.00	0.33	10
2	6.25	9.26	2.00	0.90	15	2.00	7.50	2.00	0.33	10
3	2.00	7.50	2.00	0.33	10	2.00	7.50	2.00	0.33	10
4	6.00	12.00	2.00	0.67	15	2.00	7.50	1.00	0.33	10

Table 4
Results from the analytical model for eight treatment rooms.

		W_1 (h)	W_2 (h)	S_1 (h)	S_2 (h)	P_1^b (h)	P_2^b (h)
Case 1	Current	0.33	0.35	0.69	1.85	0.025	0.033
	Efficient	(0.19, 0.30)	(0.28, 0.53)	0.52	1.74	0.004	0.027
Case 2	Current	0.37	0.24	0.59	1.76	0.018	0.028
	Efficient	(0.24, 0.29)	(0.28, 0.94)	0.47	1.74	0.003	0.027
Case 3	Current	0.49	0.49	1.94	1.94	0.039	0.039
	Efficient	(0.33, 0.55)	(0.33, 0.55)	1.74	1.74	0.027	0.027
Case 4	Current	0.33	0.36	0.69	1.90	0.025	0.035
	Efficient	(0.17, 0.30)	(0.28, 0.54)	0.52	1.77	0.004	0.027

Table 5
Efficient policies for eight treatment rooms with working routine NF–NF.

Case	Policies (δ_1, δ_2)
1	$(6, 6)^*$, (6, 5), (6, 4), $(5, 6)^*$, $(4, 6)^*$, (3, 6), (2, 6)
2	$(6, 6)^*$, $(6, 5)^*$, $(6, 4)^*$, (6, 3), (6, 2), $(5, 6)^*$, (4, 6), $(3, 6)^*$, (2, 6)
3	$(6, 6)^*$, $(6, 5)^*$, (6, 4), (5, 6), (4, 6)
4	$(6, 6)^*$, (6, 5), (6, 4), (5, 6), $(4, 6)^*$, $(3, 6)^*$, (2, 6)

* RPC is also *efficient* according to simulation results.

Moreover, the conditional expected sojourn time of a type 2 patient is higher when the type 1 doctor occupies more rooms at the instant the patient arrives.

The national norm on the sojourn time, including time for triage, is 3 h. From Fig. 4 we may conclude that the expected sojourn time of type 1 patients will exceed the norm if they arrive when there are already 10 patients of their type present and the ED doctors would both visit new patients first and use at most 3 rooms in parallel. In order to provide insight about the effect of using the optimal room assignment policy, we simulated the system for all room assignment policies and the same working routine (NF–NF). The results for room assignment policy (4, 2) are displayed in Fig. 5, which is a typical result for a non-optimal policy. For the (4, 2) policy, the norm will be exceeded for type 1 patients arriving in a system with 12 patients of their type present. For both policies, type 2 patients will never have an expected sojourn time exceeding the norm, but as the conditional expected sojourn time depends more on the number of type 2 patients in the system, it may exceed the norm for other values of (n_2, r_2, i_2).

5.2. Additional insights

We investigated several additional parameter settings in systems with a different number of rooms, in order to determine whether the insights found for the JBH case study also hold in general. In this subsection we present the most insightful results, based on a system with eight treatment rooms for four cases (see Table 3). In cases 1 and 2 the expected consultation time of type 1 patients is equal, but in case 2 the expected number of visits by the doctor is lower. In case 3 both types 1 and 2 patients have exactly the same characteristics. Case 4 is equal to case 1 except for ν_2.

From the analytical results it appears that also for the system with a different number of treatment rooms, the Pareto-front for the blocking probability and sojourn time consists of one coordinate. The results for all cases of Table 3 are displayed in Table 4, in which 'Current' implies $(\delta_1, \delta_2) = (8, 8)$ and the extreme values of the Pareto-front of the expected waiting time are given instead of all individual coordinates. The *efficient* working routine is NF–NF in all cases, and the *efficient* room assignment policies are displayed in Table 5.

From Table 4 it appears that for both patient types it is better if one of them requires less phases that each take longer, compare cases 1–3. However, when the rooms are divided according to the *efficient* assignment policies there is only an improvement for type 1 patients. When type 2 patients have to wait longer for the test results, this only increases their expected waiting and sojourn times and does not affect type 1 patients. From Table 5 it appears that all doctors should get at least two treatment rooms in parallel.

For all cases the performance of the system significantly improves when the *efficient* room assignment policies are used. In cases 1, 2 and 4 the expected sojourn time of type 1 patients decreases more than 20% and the blocking probability more than 80%. For case 3 the decrease is 10% for the expected sojourn time and 31% for the blocking probability.

We validated the results again by means of the simulation model with time-varying arrival rates. Most of the Pareto efficient RPCs are also found in the simulation results; 94% of the conclusions whether a RPC is optimal for a performance measure is the same in the results of both models. As expected, all performance measures are lower in the simulation results. The improvements by introducing an *efficient* RPC are in cases 1, 2 and 4 over 60% for the blocking probability and over 8% for the expected sojourn time of type 1 patients. For case 3 the decrease is over 10% for the blocking probability and 3% for the expected sojourn time. In addition to the *efficient* policies in Table 5, the simulation results indicate that policy (NF, 8, NF, 3) and all policies with $\delta_1 = 7$ and $\delta_2 \geq 2$ for routine NF–EF are *efficient*. We postulate that also for other parameter settings the performance measures can be improved significantly by implementing an RPC, and the efficient RPCs found by means of the analytical model are also found to be efficient in a system with time-varying arrival rates.

6. Discussion

The assignment of treatment rooms among doctors at EDs is often unbalanced, which possibly affects the blocking probability and waiting and sojourn times of patients. From the results of this study, we conclude that introducing threshold policies that indicate the maximum number of rooms that one doctor can use in parallel, can significantly improve the performance of the ED for all relevant performance measures. When two doctors interact on sharing treatment rooms, the working routine to visit new patients first is found to result in the best performance.

$$\mathbf{Q}\left(\mathbf{n};\mathbf{n}'\right)=\begin{cases}\lambda_1 & \mathbf{n}'=(n_1+1,r_1,1;n_2,r_2,1)\\ \lambda_2 & \mathbf{n}'=(n_1,r_1,1;n_2+1,r_2,1)\\ p_1\mu_1\mathbb{1}_{\{n_1=r_1\}} & \mathbf{n}'=(n_1-1,r_1-1,1;n_2,r_2,1)\\ p_1\mu_1\mathbb{1}_{\{n_1>r_1\}} & \mathbf{n}'=(n_1-1,r_1,1;n_2,r_2,1)\\ p_2\mu_2\mathbb{1}_{\{n_2=r_2\}} & \mathbf{n}'=(n_1,r_1,1;n_2-1,r_2-1,1)\\ p_2\mu_2\mathbb{1}_{\{n_2>r_2\}} & \mathbf{n}'=(n_1,r_1,1;n_2-1,r_2,1)\\ (1-p_1)\mu_1\mathbb{1}_{\{n_1>r_1,r_1<\delta_1,r_1+r_2<K\}} & \mathbf{n}'=(n_1,r_1+1,1;n_2,r_2,1)\\ (1-p_2)\mu_2\mathbb{1}_{\{n_2>r_2,r_2<\delta_2,r_1+r_2<K\}} & \mathbf{n}'=(n_1,r_1,1;n_2,r_2+1,1)\\ (1-p_1)\mu_1\mathbb{1}_{\{n_1=r_1\|r_1=\delta_1\|r_1+r_2=K\}}+(1-p_2)\mu_2\mathbb{1}_{\{r_2=\delta_2\|n_2=r_2\|r_1+r_2=K\}} & \mathbf{n}'=(n_1,r_1,1;n_2,r_2,1)\end{cases}$$

Box I.

We analytically expressed the expected sojourn time of a patient conditioned on the state of the system this patient encounters upon arrival. This conditional waiting and sojourn times information may be provided to the patients to inform them about their expected length of stay at the ED, which adds to patient satisfaction.

One limitation of the analytical model is that it suffers from an 'exploding state space'; it could be extended to incorporate more patient types, but for determining the conditional sojourn time this will make the state space too large and/or computation time very long.

In the analytical model we assume stationary arrival rates, which is typically not the case at an ED. Our simulation study indicates that this assumption does not affect the decisions whether a policy is optimal or not. Other limiting assumptions made in this model are: (1) no emergency arrivals, (2) doctors visiting their patients in random order, (3) when a doctor occupies two or more treatment rooms in parallel, the test results will be ready before the next phase completion of one of his patients, (4) the probability that a patient can go home after a phase completion is equal for all phases, and (5) the duration of one phase of service is Exponentially distributed. These assumptions are included for analytical tractability. Dropping any or all of these assumptions would be an interesting topic of further research.

Acknowledgments

The authors would like to thank the involved staff of the JBH for their cooperation and contributions to this research. Additionally, the authors are grateful for the valuable comments of two anonymous referees, which improved the paper significantly.

Appendix A. The generator

We will illustrate part of the generator matrix of the system described in Section 3 for the working routine in which both doctors visit a new patient first at phase completion. For all other working routines and generator entries, the generator matrix can be obtained in similar ways. Define additional parameters:

- K = total number of rooms,
- M_j = maximum number of patients in system, $j = 1, 2$,
- δ_j = the maximum number of rooms occupied, $j = 1, 2$,
- ν = the parameter of the exponential time a resident has to wait on test results.

There are six possible events when the system is in state $(n_1, r_1, 1; n_2, r_2, 1)$: arrival, departure and phase completion for both patient types. In this paper, $\mathbb{1}_{\{x\}}$ equals 1 if event x is true and 0 otherwise; $\mathbb{1}_{\{x,y\}}$ equals 1 if both x AND y are true; and $\mathbb{1}_{\{x\|y\}}$ equals 1 if x OR y is true. Let $\mathbf{n} = (n_1, r_1, 1; n_2, r_2, 1)$ be a state with $1 < n_1 < M_1$ and $1 < n_2 < M_2$, then: $\mathbf{Q}\left(\mathbf{n};\mathbf{n}'\right)$ is given in Box I. For a state with $n_1 = r_1 = n_2 = r_2 = 1$ we have:

$$\mathbf{Q}\left(\mathbf{n};\mathbf{n}'\right)=\begin{cases}\lambda_1\mathbb{1}_{\{i_1=0,r_1<\delta_1,r_1+r_2<K\}} & \mathbf{n}'=(n_1+1,r_1+1,1;n_2,r_2,i_2)\\ \lambda_1\mathbb{1}_{\{i_1=1\|r_1=\delta_1\|r_1+r_2=K\}} & \mathbf{n}'=(n_1+1,r_1,i_1;n_2,r_2,i_2)\\ \lambda_2\mathbb{1}_{\{i_2=0,r_2<\delta_2,r_1+r_2<K\}} & \mathbf{n}'=(n_1,r_1,i_1;n_2+1,r_2+1,1)\\ \lambda_2\mathbb{1}_{\{i_2=1\|r_2=\delta_2\|r_1+r_2=K\}} & \mathbf{n}'=(n_1,r_1,i_1;n_2+1,r_2,i_2)\\ p_1\mu_1\mathbb{1}_{\{i_1=1,\{i_2=1\|r_2=\delta_2\}\}} & \mathbf{n}'=(n_1-1,r_1-1,0;n_2,r_2,i_2)\\ p_1\mu_1\mathbb{1}_{\{i_1=1,i_2=0,r_2<\delta_2\}} & \mathbf{n}'=(n_1-1,r_1,0;n_2,r_2+1,1)\\ p_2\mu_2\mathbb{1}_{\{i_2=1,\{i_1=1\|r_1=\delta_1\}\}} & \mathbf{n}'=(n_1,r_1,i_1;n_2-1,r_2-1,0)\\ p_2\mu_2\mathbb{1}_{\{i_2=1,i_1=0,r_1<\delta_1\}} & \mathbf{n}'=(n_1,r_1+1,1;n_2-1,r_2,0)\\ (1-p_1)\mu_1\mathbb{1}_{\{i_1=1\}} & \mathbf{n}'=(n_1,r_1,0;n_2,r_2,i_2)\\ (1-p_2)\mu_2\mathbb{1}_{\{i_2=1\}} & \mathbf{n}'=(n_1,r_1,i_1;n_2,r_2,0)\\ \nu_1\mathbb{1}_{\{i_1=0\}} & \mathbf{n}'=(n_1,r_1,1;n_2,r_2,i_2)\\ \nu_2\mathbb{1}_{\{i_2=0\}} & \mathbf{n}'=(n_1,r_1,i_1;n_2,r_2,1).\end{cases}$$

Here, the first four lines represent arrivals; lines 5–8 represent departures; lines 9 and 10 represent phase completions without departures; and the last two lines represent the residents returning to their patients after waiting for the test results.

The generator-entries for other states are obtained following the same reasoning. The diagonal entries of $\mathbf{Q}$ are such that all row sums equal zero.

Appendix B. Conditional performance measures

We next summarize the steps in obtaining the conditional expected sojourn time of a patient. More details on this method can for example be found in [12,14]. Recall that the expected sojourn time of an admitted patient, conditioned on the state of the system at the instant the patient has arrived, is expressed by

$$\mathbf{S}_{\mathbf{n},j}=\int_0^\infty \sigma_{\mathbf{n},j}(x)f_j(x)\mathrm{d}x,$$
$$\text{with } \boldsymbol{\sigma}_{\mathbf{n},j}(x)=\boldsymbol{\alpha}_{\mathbf{n},j}+\sum_{\mathbf{n}'}\boldsymbol{\tau}_{\mathbf{n}',j}(x)\boldsymbol{\psi}_j(\mathbf{n},\mathbf{n}').$$

The following subsections respectively describe the analysis of the conditional expected waiting time $\boldsymbol{\alpha}_{\mathbf{n}}$, the conditional expected treatment time $\boldsymbol{\tau}_{\mathbf{n}}(x)$, and the transition probabilities $\boldsymbol{\psi}(\mathbf{n},\mathbf{n}')$.

B.1. The conditional expected waiting time

The conditional expected waiting time of a tagged type j patient that has just arrived, and upon arrival encountered a certain state $\mathbf{n}$, can be determined by considering an alternative system equal to the original system, but with the arrival rate of type j patients equal to zero. The time it takes for the alternative system to reach a state in which the queue for type j patients is empty equals the waiting time of a type j patient in the original system, since this implies the tagged patient will start his service.

For completeness, we summarize the results mentioned in [12]. Let $\tilde{\mathbf{Q}}$ denote the generator of the process obtained by letting $\lambda_j = 0$, and removing all states with an empty type j queue. The states with no type j patients queued represent the absorbing set. The time until absorption can then readily be determined by defining $\tilde{\boldsymbol{\alpha}}_{\mathbf{n}} := -1/diag(\tilde{\mathbf{Q}})$, with $diag(\tilde{\mathbf{Q}})$ the diagonal of $\tilde{\mathbf{Q}}$. Let $\tilde{\mathbf{P}}$ be the one-step transition probability matrix of the embedded jump chain of the process described by generator $\tilde{\mathbf{Q}}$. Then, $\boldsymbol{\alpha}_{\mathbf{n}} = \left(I - \tilde{\mathbf{P}}\right)^{-1} \tilde{\boldsymbol{\alpha}}_{\mathbf{n}}$, with I an identity-matrix of appropriate size. The inverse of $\left(I - \tilde{\mathbf{P}}\right)$ exists since $\tilde{\mathbf{P}}$ is by definition a substochastic matrix.

B.2. The conditional expected treatment time

The expected treatment time of a tagged patient with consultation time x entering a treatment room in state $\mathbf{n}$, $\boldsymbol{\tau}_{\mathbf{n}}(x)$, can be described by a system of differential equations, in which we assume that the tagged patient cannot leave the system for some small time Δ. We augment the state space by $z \in \{0, 1\}$, where $z = 1$ indicates the doctor is with the tagged patient.

We now formulate the differential equations for the conditional expected treatment time by conditioning on a small time Δ during which the tagged patient cannot complete its treatment. Note that the doctor still occasionally visits the tagged patient as if it were a regular patient, but after a visit the tagged patient will not leave the system. For the conditional expected treatment of a type 1 patient that is currently seen by the doctor in state $\tilde{\mathbf{n}} = (n_1, r_1, i_1; n_2, r_2, 1; 1)$ with $n_1, n_2 > 1$, $r_1 + r_2 < K$ (K the total number of treatment rooms), we get

$$\begin{aligned}\tau_{\tilde{\mathbf{n}}}(x) &= \Delta + \lambda_1 \Delta \tau_{\tilde{\mathbf{n}}+\mathbf{e}_1}(x - O(\Delta)) \\ &\quad + (1-p_1)\mu_2 \Delta \tau_{\tilde{\mathbf{n}}+\mathbf{e}_2-\mathbf{e}_7}(x - O(\Delta)) + \lambda_2 \Delta \tau_{\tilde{\mathbf{n}}+\mathbf{e}_4}(x - O(\Delta)) \\ &\quad + p_2 \mu_2 \Delta \tau_{\tilde{\mathbf{n}}-\mathbf{e}_4}(x - O(\Delta)) + (1-p_2)\mu_2 \Delta \tau_{\tilde{\mathbf{n}}+\mathbf{e}_5}(x - O(\Delta)) \\ &\quad + (1 - \lambda_1 - (1-p_1)\mu_1 - \lambda_2 - \mu_2)\Delta \tau_{\tilde{\mathbf{n}}}(x - \Delta) + o(\Delta).\end{aligned}$$

Here $\mathbf{e}_k$ denotes a vector with value 1 at position k and zeros elsewhere. The second and fourth line on the right-hand side respectively represent that in time Δ an arrival of type 1 and 2 has occurred, to a state with one additional type 1 patient ($\tilde{\mathbf{n}} + \mathbf{e}_1$) and one additional type 2 patient ($\tilde{\mathbf{n}} + \mathbf{e}_4$). Line three refers to a completion of the current phase of the tagged patient, where the doctor gets an extra room and the service of the tagged patient is temporarily paused ($\tilde{\mathbf{n}} + \mathbf{e}_2 - \mathbf{e}_7$). Line five represents a departure of type 2 ($\tilde{\mathbf{n}} - \mathbf{e}_4$), and line six a phase completion of a type 2 patient where the type 2 doctor treats a new patient in an additional treatment room ($\tilde{\mathbf{n}} + \mathbf{e}_5$). The last two lines represent, respectively, that nothing happens and that two events happen in time Δ. The equations for states with $z = 0$ and boundary states can be obtained in similar ways. We then obtain the derivative of $\tau_{\tilde{\mathbf{n}}}(x)$ by rearranging terms and letting $\Delta \to 0$.

This system of differential equations may equivalently be written in matrix notation. To this end, we introduce generator $\mathbf{Q}^*$, which is identical to $\mathbf{Q}$ excluding all states with $n_j = 0$ and the departure rate modified such that the permanent patient never leaves the system. All diagonal elements of $\mathbf{Q}^*$ are such that each row sums to zero. Since there are states in which the service rate of the tagged patient equals 0, we reorder the states such that the state space can be split into states with $z = 1$ (subscript $+$) and with $z = 0$ (subscript 0). We write $\mathbf{Q}^*$ as follows:

$$\mathbf{Q}^* = \begin{bmatrix} \mathbf{Q}^*_+ & \mathbf{Q}^*_{+0} \\ \mathbf{Q}^*_{0+} & \mathbf{Q}^*_0 \end{bmatrix}.$$

The system of differential equations can then be written as

$$\frac{\partial}{\partial x}\boldsymbol{\tau}_+(x) = \mathbf{1} + \mathbf{Q}^*_{+0}\boldsymbol{\tau}_{\mathbf{0}}(x) + \mathbf{Q}^*_+\boldsymbol{\tau}_+(x), \tag{B.1}$$

with $\mathbf{1}$ a vector with all entries equal to one. The initial condition of this system is $\boldsymbol{\tau}_+(0) = 0$, which indicates that the conditional expected treatment time of a patient with consultation time zero equals zero almost surely. Here $\boldsymbol{\tau}_{\mathbf{0}}$, $\boldsymbol{\tau}_+$ are vectors containing $\boldsymbol{\tau}_{\tilde{\mathbf{n}}}$ for states with $z = 0$ and $z = 1$ respectively, and $\boldsymbol{\tau}(x) = (\boldsymbol{\tau}_+(x), \boldsymbol{\tau}_{\mathbf{0}}(x))$. This system is similar to a system with one permanent patient.

Let $\boldsymbol{\pi}^*_+$ be a stationary distribution obtained by solving $\boldsymbol{\pi}^*_+\mathbf{Q}^*_+ = \mathbf{0}$. Then, the conditional expected treatment time of a patient with consultation time x is given by:

$$\boldsymbol{\tau}_+(x) = x\mathbf{1} + \left[I - \exp\left(x\left\{\mathbf{Q}^*_+ - \mathbf{Q}^*_{+0}(\mathbf{Q}^*_0)^{-1}\mathbf{Q}^*_{0+}\right\}\right)\right]\boldsymbol{\gamma}, \tag{B.2}$$

$$\boldsymbol{\tau}_{\mathbf{0}}(x) = -(\mathbf{Q}^*_0)^{-1}\left(\mathbf{1} + \mathbf{Q}^*_{0+}\boldsymbol{\tau}_+(x)\right),$$

with $\boldsymbol{\gamma}$ the unique solution of

$$-\left\{\mathbf{Q}^*_+ - \mathbf{Q}^*_{+0}(\mathbf{Q}^*_0)^{-1}\mathbf{Q}^*_{0+}\right\}\gamma = \mathbf{Q}^*_{+0}(\mathbf{Q}^*_0)^{-1}\mathbf{1},$$

$$\boldsymbol{\pi}^*_+\boldsymbol{\gamma} = 0.$$

Following the derivation in [12], we may show that (B.1) has a unique solution, and it is readily checked that (B.2) is the solution.

B.3. Transition probabilities

The conditional expected access and treatment time have to be linked together by the probability that a patient arriving in state $\mathbf{n}$ starts his service in state $\mathbf{n}'$. By analogy with the derivation above, we obtain the probability matrix $\boldsymbol{\psi}$ by tagging a patient and augmenting the state space with the location l of the tagged patient, which equals zero if the patient is no longer queued. In this system, all states where the tagged patient is no longer queued form an absorbing set. For completeness, we summarize these results from [12]. Denote with $\mathbf{P}^\bullet$ the one-step probability matrix of the embedded jump chain of the augmented Markov chain. Then, for all states $\mathbf{n}$ in the absorbing set $\mathbf{P}^\bullet(0, \mathbf{n}; 0, \mathbf{n}) = 1$. For all other states, $\mathbf{P}^\bullet(l, \mathbf{n}, l', \mathbf{n}')$ is non-zero only if $\mathbf{Q}(\mathbf{n}, \mathbf{n}') > 0$ and there is a transition from l to l'. $\mathbf{P}^\bullet$ can be written as

$$\mathbf{P}^\bullet \equiv \begin{pmatrix} I & \mathbf{O} \\ \mathbf{P}^\bullet_{\mathbf{0}} & \mathbf{P}^\bullet_+ \end{pmatrix},$$

with $\mathbf{O}$ the null-matrix, $\mathbf{P}^\bullet_{\mathbf{0}}$ a stochastic submatrix corresponding to transitions from the recurrent states into the absorbing set, and $\mathbf{P}^\bullet_+$ transitions among the recurrent states. We obtain the probability matrix $\boldsymbol{\psi}$ using $\boldsymbol{\psi}(\mathbf{n}, \mathbf{n}') = \boldsymbol{\psi}^\bullet(l, \mathbf{n}; 0, \mathbf{n}')$, and $\boldsymbol{\psi}^\bullet = \left(I - \mathbf{P}^\bullet_+\right)^{-1}\mathbf{P}^\bullet_{\mathbf{0}}$, where the inverse always exists since $\mathbf{P}^\bullet_{\mathbf{0}}$ is substochastic, see [12].

References

[1] N. Hoot, D. Aronsky, Systematic review of emergency department crowding: Causes, effects, and solutions, Ann. Emerg. Med. 52 (2) (2008) 126–136.

[2] S. Paul, M. Reddy, C. DeFlitch, A systematic review of simulation studies investigating emergency department overcrowding, Simulation 86 (8–9) (2010) 559–571.

[3] J. Wiler, R. Griffey, T. Olsen, Review of modeling approaches for emergency department patient flow and crowding research, Acad. Emerg. Med. 18 (12) (2011) 1371–1379.

[4] D. Koks, M. Zonderland, C. Heringhaus, Development of an observational instrument to determine variations in the patient care process and patient flow among emergency physicians and internists at the emergency department, Int. J. Emerg. Med. 6 (1) (2013) 1.

[5] E.D. Boudreaux, E.L. O'Hea, Patient satisfaction in the emergency department: a review of the literature and implications for practice, J. Emerg. Med. 26 (1) (2004) 13–26.

[6] P.J.H. Hulshof, N. Kortbeek, R.J. Boucherie, E.W. Hans, P.J.M. Bakker, Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS, Health Syst. 1 (2) (2012) 129–175.

[7] O. El-Rifai, T. Garaix, V. Augusto, X. Xie, A stochastic optimization model for shift scheduling in emergency departments, Health Care Manage. Sci. (2014) 1–14.

[8] N. Izady, D. Worthington, Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments, European J. Oper. Res. 219 (3) (2012) 531–540.
[9] G. Yom-Tov, A. Mandelbaum, Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing, Manuf. Serv. Oper. Manage. 16 (2) (2014) 283–299.
[10] S. Bhulai, G. Koole, A queueing model for call blending in call centers, IEEE Trans. Automat. Control 48 (8) (2003) 1434–1438.
[11] S. Borst, R. Núñez Queija, M. van Uitert, User-level performance of elastic traffic in a differentiated-services environment, Perform. Eval. 49 (1–4) (2002) 507–519.
[12] R. Litjens, R. Boucherie, Performance analysis of fair channel sharing policies in an integrated cellular voice/data network, Telecommun. Syst. 19 (2) (2002) 147–186.
[13] R. Litjens, H. van den Berg, R. Boucherie, Throughputs in processor sharing models for integrated stream and elastic traffic, Perform. Eval. 65 (2) (2008) 152–180.
[14] R. Núñez Queija, Processor-sharing models for integrated-services networks (Ph.D. thesis), Eindhoven University of Technology, 2000.
[15] MATLAB, version 7.11.0 (R2010b), The MathWorks Inc., Natick, Massachusetts, 2010.
[16] A. Law, W. Kelton, Simulation Modeling and Analysis, third ed., McGraw Hill, 2000.