European Journal of Operational Research 226 (2013) 602–614

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

A Markovian queueing model for ambulance offload delays

Eman Almehdawe *, Beth Jewkes, Qi-Ming He

Department of Management Sciences, University of Waterloo, 200 University Avenue West, Waterloo, Canada N2L 3G1

ARTICLE INFO

Article history:
Received 14 November 2011
Accepted 19 November 2012
Available online 5 December 2012

Keywords:
Queueing theory
Matrix-analytic method
Ambulance offload delay
Priority queues

ABSTRACT

Ambulance offload delays are a growing concern for health care providers in many countries. Offload delays occur when ambulance paramedics arriving at a hospital Emergency Department (ED) cannot transfer patient care to staff in the ED immediately. This is typically caused by overcrowding in the ED. Using queueing theory, we model the interface between a regional Emergency Medical Services (EMS) provider and multiple EDs that serve both ambulance and walk-in patients. We introduce Markov chain models for the system and solve for the steady state probability distributions of queue lengths and waiting times using matrix-analytic methods. We develop several algorithms for computing performance measures for the system, particularly the offload delays for ambulance patients. Using these algorithms, we analyze several three-hospital systems and assess the impact of system resources on offload delays. In addition, simulation is used to validate model assumptions.

1. Introduction

Ambulance offload time is the time taken to transfer a patient from an ambulance stretcher into the Emergency Department (ED) of a hospital. If an ED cannot accept care for an incoming ambulance patient, a common course of action is to let paramedics continue to provide patient care in the ambulance or on a stretcher in the ED until an ED bed becomes available. This delay in transfer of care is referred to as "offload delay". Patients experiencing offload delays prevent the ambulances and their crews from returning to service. According to a report by the Ontario Ministry of Health and Long Term Care [5] (Canada), the principal cause of ambulance offload delays is the congestion in downstream stages of patient care. i.e., the lack of capacity to treat hospital inpatients. Such a capacity shortage has a cascading impact – it contributes to ED overcrowding, to ambulance offload delays, and ultimately to a reduction in the EMS service level to the community.

Ambulance offload delays are a pressing health care concern in many countries and, in particular, an issue of growing concern to many communities in Canada. For example, the provincial government of Ontario invested $96 million in its comprehensive action plan to reduce the length of time paramedics wait to offload patients at hospital EDs in 2006. Despite such efforts, it was reported that offload delays still cost Toronto EMS approximately 180 ambulance hours per day in December 2007 [17]. In the Region of Waterloo (ROW), Ontario, a fleet of 18 ambulances and three hospitals serve a population of approximately 500,000 who live in three municipalities and four townships. According to the ROW EMS 2008 Master Plan [20], the region lost a maximum of 13.25 ambulance-days in a month in 2005, and 12.36 ambulance-days in a month in 2006. In December of 2007, a maximum of 22 offload delay incidents were reported in a single day.

Since offload delays increase both health care costs [21] and risks to patients [28], how to reduce ambulance offload delays has become an important issue to health care providers, and has attracted the attention of researchers and practitioners. Most research on offload delays is carried out by medical doctors who try to shed light on the importance of the problem and its implications. For instance, Ting [28] investigates the causes of ambulance offload delay and the impact of delayed ED care for patients. Taylor et al. [27] conduct an observational study to determine the difference between documented ambulance arrival times and the actual arrival times of patients from the ambulance into the emergency department. Silvestri et al. [25] carry out an observational study to examine the effect of ED bed availability on offload delays. Silvestri et al. [26] conduct an observational study to evaluate offload delay intervals and the association between out-of-hospital patient triage categorization and admission. The study concludes that delayed ambulances reduce EMS availability. Eckstein and Chan [6] investigate the effect of ED crowding on paramedic ambulance availability. Their empirical study suggests a direct link between ED crowding and the ability of EMS to provide a timely response to emergency calls.

The aforementioned observational studies indicate that there is a strong tie between offload delays and ED service capacity, in the form of hospital beds, for patients. Thus, to understand and to reduce offload delays, it is necessary to investigate the relationship analytically. A natural tool for such a study is queueing theory, since ambulances and patients form queues in the EMS-ED system.

* Corresponding author. Tel.: +1 519 888 4567.
E-mail address: ealmehda@uwaterloo.ca (E. Almehdawe).

0377-2217/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2012.11.030

In this paper, we introduce a queueing network that explicitly models the arrival, transition, and service processes of patients in an EMS-ED system. We use the queueing model to quantify offload delays as well as the impact of service congestion on ambulance waiting times in the EDs.

Queueing theory has been used extensively in the study of manufacturing, telecommunications, and service systems. The use of queueing theory in health care management has been growing in the past two decades (see the surveys by Formundam and Herrmann [9] and Green [11]). For example, Kao and Tung [14] study the problem of reallocating beds to services in order to minimize the expected overflows for a large public health care delivery system. They use a $M/G/\infty$ queueing model to approximate patient population dynamics. Creemers et al. [3] develop a queueing model to assign server time slots for different classes of patients. Gorunescu et al. [10] develop a loss queueing model to optimize the allocation and use of hospital beds. While the above models use classical queueing methods for analysis, we develop a Markov chain model to analyze the interaction between an EMS provider and multiple EDs in a region. On the other hand, most of the research on EMS operations focuses on the location of emergency units (e.g. Chaiken and Larson [1], Erkut et al. [7], and Erkut et al. [8]), or on the relocation and dispatching decisions (e.g. Schmid [23]).

In most ED settings, patients with life threatening injuries are given priority over patients with less severe conditions [9]. Siddhartan et al. [24] compare a First-Come-First-Serve (FCFS) admission discipline to a two class priority discipline for admitting patients into an ED. They study the waiting times and queue lengths for both classes of patients. Worthington [29] uses a three priority level system to analyze patient transfer from an outpatient physician to an inpatient physician. In our model, we assume that patients that arrive by ambulance have higher acuity levels than walk-in patients, and thus give the ambulance arrivals higher service priority. Recently, Mandelbaum et al. [18] develop a queueing model for the interface between an emergency department and internal wards of a hospital. Their inverted-V model structure is similar to our queueing model, except that Mandelbaum et al. [18] model uses the priority class for inpatient admission purposes.

In this study, we are primarily interested in modeling the flows of patients through a single EMS system into one of several emergency departments. We are concerned only with intermediate and acute care patients – those that consume ED beds – and we do not capture the lowest acuity patients that we assume receive care in a separate "minor treatment" area of the ED. We consider two types of patients: those that arrive to an emergency department by ambulance whom we refer to as ambulance patients, and those who arrive directly to an emergency department by other means whom we refer to as walk-in patients. Walk-in patients are assumed to have a lower acuity level than that of ambulance patients, and thus are given lower priority than ambulance patients.

To capture these characteristics, we introduce a queueing network with multiple servers and two priority classes of customers. Specifically, we assume that: (1) patients arrive to the EMS and EDs according to independent Poisson processes; (2) patient service times follow an exponential distribution; (3) ambulance patients have preemptive priority over walk-in patients; (4) the time taken by the ambulance to transport and transfer the patient into the ED is negligible compared to the time the patient spends in the ED. Although assumptions (2) and (4) appear to limit our model, we later demonstrate through simulation that they do not have a significant impact on our conclusions or on the applicability of the model.

In our model for the EMS-ED system, we introduce two Markov chains for the queueing processes of ambulance patients and walk-in patients. Offload delays are captured by the waiting times of ambulance patients. By using matrix-analytic methods, we develop several algorithms for computing system performance measures. Our goal is to develop a tool that can help decision makers evaluate the impact of resource allocation decisions at each hospital ED on offload delays and on system wide hospital congestion.

The primary contributions of this paper are twofold. First, continuous time Markov chains are introduced for analyzing queue lengths, waiting times, and sojourn times of ambulance and walk-in patients in all EDs. Efficient algorithms are developed for computing related performance measures such as the mean queue length and mean waiting times. Our second contribution is to apply the theoretical model to examine the impact of reallocating resources on system performance metrics.

The rest of the paper is organized as follows. In Section 2, we introduce the queueing model of interest. We analyze the model with ambulance patients only in Section 3. Then we investigate a model with both ambulance patients and walk-in patients in Section 4. For both models, we introduce a continuous time Markov chain and then use matrix-analytic methods for analysis. In Section 5, we numerically study several case studies with three emergency departments. Finally, Section 6 contains the results of a simulation study used to validate two of our modeling assumptions.

2. The stochastic model

We consider a queueing network with one EMS provider that serves K hospitals, each with a multiple-bed ED. The EMS has N ambulances. Fig. 1 illustrates a network consisting of three hospitals. In general, the flow of patients can be described as follows: high acuity patients call for an ambulance at a stationary Poisson rate. When a call arrives and there is an ambulance available, the patient is transported to one of the K EDs to receive service. These are referred to as ambulance patients. Alternatively, a patient may arrive to an ED for service by him/herself. We shall call these walk-in patients. A patient that arrives to an ED is either admitted immediately to a bed or joins a queue of patients waiting for service. When a bed becomes available, it is assigned to a waiting ambulance patient first, if any; otherwise, it is assigned to a waiting walk-in patient. We assume service for walk-in patients is preempted by an arriving ambulance patient if there are no beds available for the ambulance patient. All patients leave the ED immediately once their service is completed.

2.1. Arrival of patients

We assume that ambulance patients arrive to the system according to a Poisson process with rate λ_0. Walk-in patients arrive to the k^{th} ED according to a Poisson process with rate λ_k, for $k = 1,2,\ldots,K$. All Poisson processes are independent of each other. The Poisson assumption is supported by empirical studies (e.g., Channouf et al. [2] and the references therein). Although arrival processes in practice, depend on the time of the day, day of the week, and other factors, the use of a (stationary) Poisson process to approximate a non-stationary Poisson process has been justified in the literature (e.g., Lewis [16] and Kao and Tung [14], among others).

2.2. Ambulance routing

When a patient calls for an ambulance, if an ambulance is available, the patient is picked up and transported to the k^{th} ED with probability p_k. We call $\{p_k,\ k = 1,2,\ldots,K\}$ the routing probabilities. By the law of total probability, we have $p_1 + p_2 + \ldots + p_K = 1$. If all N ambulances are occupied when a call occurs, we assume that the patient is lost. In practice, this is a rare occurrence, and the call will actually be served by a neighboring EMS provider.

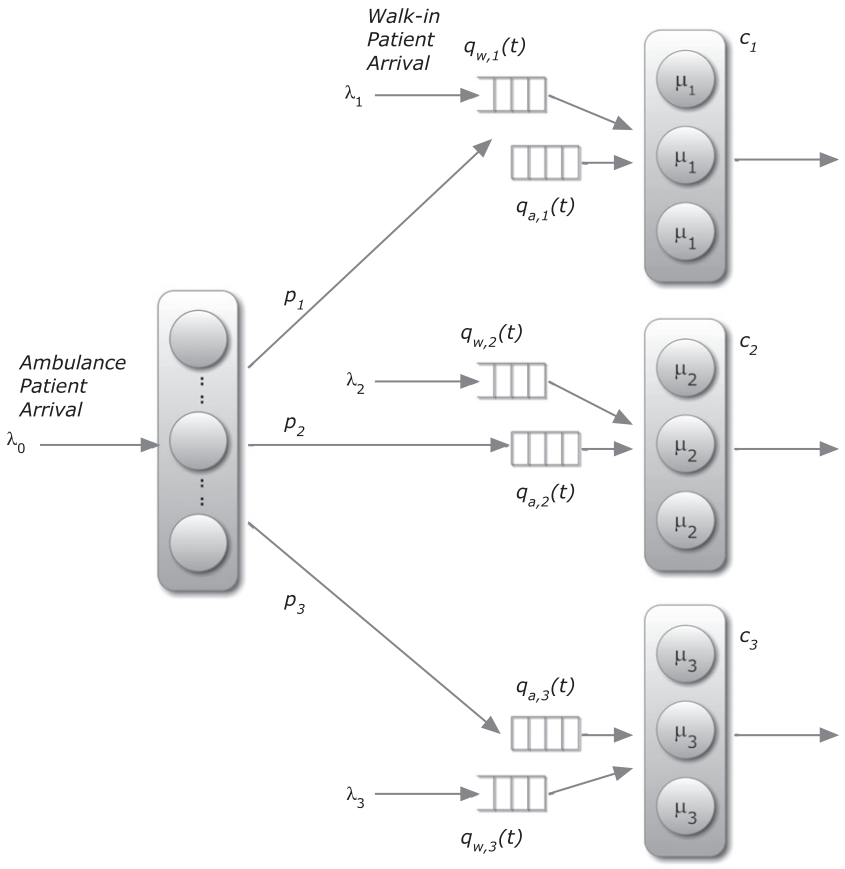

Fig. 1. EMS-ED queueing network diagram for $K = 3$.

Ambulance transit times include the time to reach a patient, to load the patient into an ambulance, and then to transfer the patient into an ED. In practice, the transit time is not zero, but it is short in comparison to the time that a patient spends in an ED. More importantly, offload delays, which are the focus of this work, mainly depend on the congestion within the emergency departments. We have therefore assumed that the transit time is zero. This simplification permits us to obtain many insights without overly complicating our model. In Section 6, through simulation, we demonstrate numerically that adding the EMS transit time into the model has little impact on offload delays and other performance measures of interest. In practice, the hospital to which the patient is taken may depend on their type of medical problem, or on which hospital is the closest in proximity. In our model, we do not take such details into consideration. Instead, we assume that the routing probabilities reflect the long term fraction of all patients sent to individual EDs.

2.3. Service capacity and service time at an ED

Each ED serves both ambulance and walk-in patients. We assume that the k^{th} ED has a service capacity of c_k units (or beds). That is: the k^{th} ED can serve c_k patients simultaneously. At the k^{th} ED, the service time of a patient, regardless of its type, is assumed to have an exponential distribution with parameter μ_k. We can view each server as a bed or the combination of resources (e.g. a bed, nurses, doctors, etc.) needed to serve a patient. Each unit of capacity operates independently of others. In Section 6, we examine the impact of the exponential assumption on system performance.

2.4. Service priority at an ED

We assume that ambulance patients have preemptive priority over walk-in patients. That is: if a bed becomes available, it will be assigned to an ambulance patient first. If an ambulance patient arrives at an ED and finds that all servers (beds) are occupied, then if there is one or more walk-in patient in service, the patient or one of the patients will be preempted by the arriving ambulance patient. When a bed becomes available in the future, and there are no ambulance patients waiting, the walk-in patient will return to service. Since the service times are assumed to be exponential, waiting times for walk-in patients are not affected by repeated/resumed services. We also assume that, within each priority class, patients are served on a first-come-first-served basis.

In practice, patients that arrive via ambulance typically have higher acuity levels than walk-in patients. Fig. 2, constructed with data from a local hospital in the Region of Waterloo, Ontario, Canada, shows that this assumption is reasonable. In Fig. 2, CTAS 1 (Canadian Triage Acuity Scale) represents patients with the most severe conditions who require immediate attention. CTAS 2, 3, and 4 patients have successively lower acuity medical problems. For this reason, we assume that ambulance patients have preemptive priority over walk-in patients. Preempting the service of a walk-in patient can be interpreted as preempting their care, as is the case when a severely ill patient arrives to the ED.

We summarize the model parameters as follows:

- N: total number of ambulances available in the system;
- K: number of regional hospitals (or EDs);
- λ_0: ambulance patient arrival rate to the EMS system;

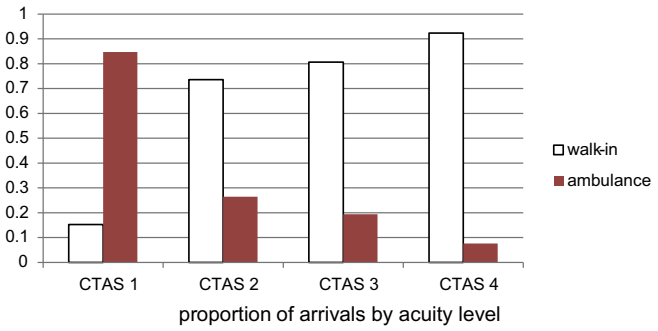

Fig. 2. Arrivals to an ED by acuity level and mode of arrival.

- p_k: probability that an ambulance patient is sent to the k^{th} ED, for $k = 1,2,\ldots,K$;
- μ_k: service rate per server in the k^{th} ED, for $k = 1,2,\ldots,K$;
- λ_k: arrival rate of walk-in patients at the k^{th} ED, for $k = 1,2,\ldots,K$;
- c_k: number of servers in the k^{th} ED, which corresponds to the service capacity at the k^{th} ED, for $k = 1,2,\ldots,K$.

In order to analyze the queueing network, we introduce two sets of state variables to describe the system state:

(1) $q_{a,k}(t)$: total number of ambulance patients in service or waiting in the k^{th} ED, at time t, for $k = 1,2,\ldots, K$;
(2) $q_{w,k}(t)$: total number of walk-in patients in service or waiting in the k^{th} ED, at time t, for $k = 1,2,\ldots, K$.

The fact that the service discipline at each hospital ED is assumed to be preemptive, where walk-in patients have lower priority and ambulance arrivals are assigned higher priority, allows us to analyze the queues of ambulance patients separately without the need to include walk-in patients. Thus, we first analyze the queues of ambulance patients only in Section 3. Then we analyze the queues for both types of patients in Section 4.

3. Ambulance patients

In this section, we analyze the stochastic model with only ambulance patients. The analysis consists of four parts. Section 3.1: a recursive method is introduced for constructing the infinitesimal generator for a continuous time Markov chain. Section 3.2: Matrix-analytic methods are used to develop an algorithm for computing the stationary distribution of the number of patients in the system. Section 3.3: a number of performance measures are derived. Section 3.4: a Markov chain is constructed for the waiting times of ambulance patients and the mean waiting times are obtained.

3.1. The Markov chain

Since the arrival process of ambulance patients to the EMS is Poisson and the service times are exponential, it is easy to see that the stochastic process $\{(q_{a,K}(t), q_{a,K-1}(t), \ldots, q_{a,1}(t)), t \geqslant 0\}$ is a continuous time Markov chain. The queue lengths, $q_{a,K}(t)$, $q_{a,K-1}(t)$, …, and $q_{a,1}(t)$, are finite, since $q_{a,K}(t) + q_{a,K-1}(t) + \ldots + q_{a,1}(t) \leqslant N + c_1 + c_2 + \ldots + c_K$. Thus, the state space Ω of $\{(q_{a,K}(t), q_{a,K-1}(t), \ldots, q_{a,1}(t)), t \geqslant 0\}$ is finite. In order to construct the infinitesimal generator of the Markov chain, we divide the states into subgroups according to the values of $q_{a,K}(t)$: $\Omega = \Omega_0 \cup \Omega_1 \cup \ldots \cup \Omega_{N+c_K}$, where, for $i_K = 0,1,\ldots,N + c_K$,

$$\begin{aligned}\Omega_{i_K} = \{(i_K, i_{K-1}, \ldots, i_2, i_1) : {} & 0 \leqslant i_j \leqslant c_j + \min\{N, c_K + N - i_K\}, 1 \leqslant j \leqslant K-1, \\ & 0 \leqslant i_{j_1} + i_{j_2} \leqslant c_{j_1} + c_{j_2} + \min\{N, c_K + N - i_K\}, 1 \leqslant j_1, j_2 \leqslant K-1, \\ & \ldots \\ & i_1 + i_2 + \ldots + i_{K-1} \leqslant c_1 + c_2 + \ldots + c_{K-1} + \min\{N, c_K + N - i_K\}\}.\end{aligned} \tag{1}$$

We shall call $q_{a,K}(t)$ the level variable and $(q_{a,K-1}(t), \ldots, q_{a,1}(t))$ the (vector) phase variable. The set of states in level i_K is Ω_{i_K}.

We observe that each of the state variables $q_{a,K}(t)$, $q_{a,K-1}(t)$, …, and $q_{a,1}(t)$ changes its value by at most one whenever an arrival or a service completion occurs. Then, $\{(q_{a,K}(t), q_{a,K-1}(t), \ldots, q_{a,1}(t)), t \geqslant 0\}$ is a level dependent quasi-birth-and-death (QBD) process with a finite number of levels. See Neuts [19] and Latouche and Ramaswami [15] for more details on QBD processes.

Due to the birth-and-death nature of the level variable $q_{a,K}(t)$, the infinitesimal generator for the Markov chain $\{(q_{a,K}(t), q_{a,K-1}(t), \ldots, q_{a,1}(t)), t \geqslant 0\}$ has the following structure:

$$Q_N^{(K)} = \begin{pmatrix} A_{(0,0)}^{(K)} & A_{(0,1)}^{(K)} & & & & \\ A_{(1,0)}^{(K)} & A_{(1,1)}^{(K)} & A_{(1,2)}^{(K)} & & & \\ & \ddots & \ddots & & \ddots & \\ & & A_{(N+c_K-1,N+c_K-2)}^{(K)} & A_{(N+c_K-1,N+c_K-1)}^{(K)} & A_{(N+c_K-1,N+c_K)}^{(K)} \\ & & & A_{(N+c_K,N+c_K-1)}^{(K)} & A_{(N+c_K,N+c_K)}^{(K)} \end{pmatrix}. \tag{2}$$

Intuitively, the matrices $A_{(i,i+1)}^{(K)}$, $A_{(i,i-1)}^{(K)}$, and $A_{(i,i)}^{(K)}$ give the transition rates by which the number of patients at the K^{th} ED, i.e., $q_{a,K}(t)$, increases by one, decreases by one, or does not change, respectively. Finding the matrix blocks in $Q_N^{(K)}$ is not straightforward. The main difficulty comes from the fact that the number of states in different levels varies. We observe that the number of states in each level is determined by the number of available ambulances to other EDs. Based on this observation, a recursive method is introduced for constructing the matrix blocks in the infinitesimal generator $Q_N^{(K)}$. The computational steps are summarized in Algorithm 4 in Appendix A.

3.2. Matrix-geometric solution

We denote by $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{N+c_K})$ the stationary distribution of the Markov chain $\{(q_{a,K}(t), q_{a,K-1}(t), \ldots, q_{a,1}(t)), t \geqslant 0\}$, where the elements of the row vector $\boldsymbol{\pi}_{i_K}$ are the steady state probabilities of the states in Ω_{i_K}, for $0 \leqslant i_K \leqslant c_K + N$. Since the Markov chain is irreducible, $\boldsymbol{\pi}$ exists and is the unique non-negative solution of the linear system:

$$\boldsymbol{\pi} Q_N^{(K)} = 0 \text{ and } \boldsymbol{\pi} \boldsymbol{e} = 1, \tag{3}$$

where $\boldsymbol{e}$ is a column vector of ones. Since the infinitesimal generator $Q_N^{(K)}$ has a block tridiagonal structure, a matrix-geometric solution can be obtained. First, for the levels $N + c_K$ and $N + c_K - 1$, we obtain

$$\boldsymbol{\pi}_{N+c_K} = \boldsymbol{\pi}_{N+c_K-1} R_{a,N+c_K}, \tag{4}$$

where

$$R_{a,N+c_K} = -A_{(N+c_K-1,N+c_K)}^{(K)} \left(A_{(N+c_K,N+c_K)}^{(K)}\right)^{-1}. \tag{5}$$

For levels $i = 1,2,\ldots,N + c_K - 1$, we have:

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_{i-1} R_{a,i} \tag{6}$$

where

$$R_{a,i} = -A_{(i-1,i)}^{(K)} \left(A_{(i,i)}^{(K)} + R_{a,i+1} A_{(i+1,i)}^{(K)}\right)^{-1}. \tag{7}$$

In order to find $\boldsymbol{\pi}$, we need to find $\boldsymbol{\pi}_0$ for level zero. The boundary balance equations and the normalization condition lead to the following linear system for finding $\boldsymbol{\pi}_0$:

$$\begin{aligned}&\boldsymbol{\pi}_0\left(A^{(K)}_{(0,0)}+R_{a,1}A^{(K)}_{(1,0)}\right)=0;\\&\boldsymbol{\pi}_0(\boldsymbol{e}+R_{a,1}\boldsymbol{e}+R_{a,1}R_{a,2}\boldsymbol{e}+\ldots+R_{a,1}\ldots R_{a,N+c_K}\boldsymbol{e})=1.\end{aligned}\tag{8}$$

We summarize the solution steps in Algorithm 1.

Algorithm 1. Stationary distribution of $\{(q_{a,K}(t),\ q_{a,K-1}(t),\ldots,\ q_{a,1}(t)),\ t\geqslant 0\}$

1. Use Algorithm 4 in Appendix A to generate matrix blocks in $Q^{(K)}_N$.
2. Find $R_{a,N+c_K}$ using Eq. (5).
3. Find $R_{a,i}$ recursively using Eq. (7), for $1\leqslant i\leqslant N+c_K-1$.
4. Find the vector π_0 using the boundary and normalization conditions in (8).
5. Find π_i using Eq. (6).

3.3. Performance measures

A number of performance measures can be derived directly from $\boldsymbol{\pi}$. We shall focus on the performance measures for the K^{th} ED. Performance measures for other EDs can be obtained from $\boldsymbol{\pi}$ as well, but the formulas are more involved.

1. In steady state, the distribution of the number of ambulance patients $q_{a,K}$ in the K^{th} ED is given by
$$P\{q_{a,K}=i\}=\pi^{(K)}(i)=\boldsymbol{\pi}_i\boldsymbol{e},\ \text{for } i=0,1,\ldots,N+c_K.\tag{9}$$
2. The mean number of ambulance patients in the K^{th} ED is given by
$$E[q_{a,K}]=\sum_{i=0}^{N+c_K}i\pi^{(K)}(i).\tag{10}$$
3. We define random variable $O^{(K)}$ as the number of ambulances in offload delay at the K^{th} ED. Since there are ambulances in offload delay at the K^{th} ED if and only if $q_{a,K}>c_K$, we have $O^{(K)}=\max\{0,q_{a,K}-c_K\}$. The probability distribution for the number of ambulances in offload delay can be calculated as follows:
$$P\{O^{(K)}=m\}=\begin{cases}\sum_{i=0}^{c_K}\pi^{(K)}(i), & \text{for } m=0;\\ \pi^{(K)}(m+c_K), & \text{for } m=1,2,\ldots,N.\end{cases}\tag{11}$$
The mean number of ambulances in offload delay in the K^{th} ED, $E[O^{(K)}]$, can be obtained accordingly.
4. For state $(i_K,\ldots,i_1)$, we denote by $\pi_{i_K,\ldots,i_1}$ its steady state probability, which is an element in the vector $\boldsymbol{\pi}$. The probability distribution of the total number of ambulances in offload delay, denoted by O, is given by
$$P\{O=m\}=\sum_{(i_K,\ldots,i_1)\in\Omega:\ \sum_{k=1}^{K}\max\{0,i_k-c_k\}=m}\pi_{i_K,\ldots,i_1},\quad \text{for } 0\leqslant m\leqslant N;\tag{12}$$
The mean total number of ambulances in offload delay, $E[O]$, can be obtained accordingly.
5. We refer to the probability that all ambulances are in offload delay as the loss probability, denoted as P_L. Then the loss probability is given by
$$P_L=P\{O=N\}=\sum_{(i_K,\ldots,i_1)\in\Omega:\ \sum_{k=1}^{K}\max\{0,i_k-c_k\}=N}\pi_{i_K,\ldots,i_1}.\tag{13}$$

3.4. Waiting times of ambulance patients (offload delays)

The waiting time $w_{a,K}$ of an ambulance patient arriving to the K^{th} ED depends on the number of ambulance patients waiting at the K^{th} ED. Denote by $\eta_i(K)$ the probability that i ambulance patients are in the K^{th} ED when an ambulance patient arrives in the K^{th} ED. Since an arriving patient can reach the K^{th} ED if and only if there is an ambulance available at the time of arrival, we have, for $0\leqslant i\leqslant c_K+N-1$,

$$\eta_i(K)=\frac{1}{1-P_L}\sum_{(i,i_{K-1},\ldots,i_1)\in\Omega:\ \max\{0,i-c_K\}+\sum_{k=1}^{K-1}\max\{0,i_k-c_k\}<N}\pi_{i,i_{K-1},\ldots,i_1}.\tag{14}$$

Let $\boldsymbol{\alpha}(K)=(\eta_{c_K}(K),\ldots,\eta_{c_K+N-1}(K))$. Note that $\eta_i(K)$ is the probability that an arriving ambulance patient to the K^{th} ED has to wait for the service completion of $i-c_K+1$ patients before getting a bed, for $i\geqslant c_K$. In the K^{th} ED, there are c_K beds for all patients, each with an exponential service time with parameter μ_K. If all beds are occupied, the time until the next service completion is exponentially distributed with parameter $c_K\mu_K$. Thus, if all c_K servers are busy, the total time to serve i patients has an Erlang distribution of order i and rate $c_K\mu_K$. Consequently, when an ambulance patient arrives to hospital K, the waiting time $w_{a,K}$ has a generalized Erlang distribution with a phase-type representation $(\boldsymbol{\alpha}(K),c_K\mu_K J_N)$, where

$$J_N=\begin{pmatrix}-1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1\end{pmatrix}_{N\times N}.\tag{15}$$

The distribution function of the waiting time $w_{a,K}$ is given by

$$P\{w_{a,K}<t\}=1-\boldsymbol{\alpha}(K)\exp\{-c_K\mu_K J_N t\}\boldsymbol{e}.\tag{16}$$

By routine calculations, we obtain

$$E[w_{a,K}]=\sum_{i=1}^{N}\frac{i\eta_{c_K-1+i}(K)}{c_K\mu_K}.\tag{17}$$

The mean waiting time $E[w_{a,K}]$ and the mean queue length $E[q_{a,K}]$ satisfy Little's law: $E[q_{a,K}]=\lambda_0(1-P_L)p_K(E[w_{a,K}]+1/\mu_K)$, where $\lambda_0(1-P_L)p_K$ is the arrival rate to the K^{th} ED. We use Little's law to verify the accuracy of computations.

Denote by w_a the waiting time of an arbitrary ambulance patient who enters the system (i.e., the patient is not lost). Since arriving ambulance patients are sent to individual hospitals with probabilities $\{p_1,\ldots,p_K\}$, the mean waiting time of an arbitrary ambulance patient who actually enters a hospital is given by $E[w_a]=\sum_{k=1}^{K}p_kE[w_{a,k}]$. Since the service time in the k^{th} ED has an exponential distribution with parameter μ_k, the mean sojourn time of an ambulance patient at the k^{th} ED is given by $E[w_{a,k}]+1/\mu_k$. The mean sojourn time of an arbitrary ambulance patient who enters the system can be calculated by $\sum_{k=1}^{K}p_k(E[w_{a,k}]+1/\mu_k)=E[w_a]+\sum_{k=1}^{K}p_k/\mu_k$.

4. Walk-in patients

To account for the walk-in patients who arrive to the hospital EDs with lower acuity ailments, we utilize the Markov chain defined in Section 3 to develop a new Markov chain that includes both ambulance and walk-in patients. Due to the facts that the arrival processes of walk-in patients to individual hospitals are independent and the service priority is preemptive, without loss of generality, we can focus on the walk-in patient queue in one ED.

4.1. A modified Markov chain

We add $q_{w,K}(t)$ to the Markov chain considered in Section 3 to obtain a continuous time Markov chain $\{(q_{w,K}(t), q_{a,K}(t), q_{a,K-1}(t), \ldots, q_{a,1}(t)), t \geqslant 0\}$, which has an infinite state space. Since the level variable $q_{w,K}(t)$ changes its value by at most one at each transition, the process $\{(q_{w,K}(t), (q_{a,K}(t), q_{a,K-1}(t), \ldots, q_{a,1}(t))), t \geqslant 0\}$ is a QBD process with an infinite number of levels. Every level, which consists of all states with fixed $q_{w,K}(t)$, has the same number of states as that in Ω (defined in Section 3.1). Since the service discipline is preemptive, walk-in patients have no impact on the service of ambulance patients. Thus, the infinitesimal generator $Q_{w,K}$ has the following structure:

$$Q_{w,K} = I \otimes \left(Q_N^{(K)} - \lambda_K I\right) + \begin{pmatrix} 0 & \lambda_K I & & & & & \\ M_{K,1} & -M_{K,1} & \lambda_K I & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & M_{K,c_K} & -M_{K,c_K} & \lambda_K I & & \\ & & & M_{K,c_K} & -M_{K,c_K} & \lambda_K I & \\ & & & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{18}$$

where $I \otimes \left(Q_N^{(K)} - \lambda_K I\right)$ is the Kronecker product of I (which is infinite in size) and $Q_N^{(K)} - \lambda_K I$, $Q_N^{(K)}$ is defined in Eq. (2), and $M_{K,n}$ is a diagonal matrix that includes service rates for walk-in patients conditioning on the number of ambulance patients in the K^{th} ED: for $n = 1, 2, \ldots, c_K$,

$$M_{K,n} = \begin{array}{c} 0 \\ 1 \\ \vdots \\ c_K - 1 \\ c_K \\ \vdots \\ c_K + N \end{array} \begin{pmatrix} \min\{n, c_K\}\mu_K I & & & & & \\ & \min\{n, c_K - 1\}\mu_K I & & & & \\ & & \ddots & & & \\ & & & \mu_K I & & \\ & & & & 0 & \\ & & & & & \ddots \\ & & & & & & 0 \end{pmatrix}, \tag{19}$$

Note that, an interpretation of n in $M_{k,n}$ is $n = \min\{c_K, q_{w,K}(t)\}$, i.e., the number of walk-in patients that could be in service. The diagonal elements of $M_{K,n}$ indicate the number of walk-in patients in service, which depends on the number of available beds and the number of walk-in patients in the K^{th} ED, and is given by $\max\{0, \min\{n, c_K - q_{a,K}(t)\}\}$. It is easy to see that the Markov chain $\{(q_{w,K}(t), q_{a,K}(t), q_{a,K-1}(t), \ldots, q_{a,1}(t)), t \geqslant 0\}$ is level dependent up to level c_K. Beyond level c_K, the Markov chain has a level independent structure. This allows us to find its stationary probability distribution using matrix-analytic methods.

4.2. Mean queue length for walk-in patients

Let $\boldsymbol{\phi} = (\boldsymbol{\phi_0}, \boldsymbol{\phi_1}, \ldots)$ be the stationary probability distribution of $\{(q_{w,K}(t), (q_{a,K}(t), q_{a,K-1}(t), \ldots, q_{a,1}(t))), t \geqslant 0\}$. The stationary distribution exists if and only if the Markov chain is ergodic. Since the Markov chain of interest is irreducible and has a QBD structure, by Neuts [19], the Markov chain is ergodic if and only if $\lambda_K \boldsymbol{\pi e} < \boldsymbol{\pi} M_{K,c_K} \boldsymbol{e}$, which can be simplified to

$$\lambda_K + p_K \lambda_0 (1 - P_L) < c_K \mu_K. \tag{20}$$

Intuitively, the left hand side of Eq. (20) is the total arrival rate to the K^{th} ED and the right hand side is the potential service capacity at the K^{th} ED. Eq. (20) ensures that there is enough capacity to serve all patients arriving to the K^{th} ED. In the rest of this paper, we assume that Eq. (20) holds. The stationary probability distribution $\boldsymbol{\phi}$ can thus be obtained by solving the linear system

$$\boldsymbol{\phi} Q_{w,K} = 0 \text{ and } \boldsymbol{\phi e} = 1. \tag{21}$$

By Neuts [19], the stationary distribution has a matrix geometric form:

$$\boldsymbol{\phi}_n = \boldsymbol{\phi}_{c_K} R_w^{n - c_K}, \text{ for } n \geqslant c_K \tag{22}$$

where the rate matrix R_w is the minimal nonnegative solution to the nonlinear equation:

$$\lambda_K I + R_w \left(Q_N^{(K)} - \lambda_K I - M_{K,c_K}\right) + R_w^2 M_{K,c_K} = 0. \tag{23}$$

The above equation can be solved using the logarithmic reduction algorithm of [15]. For the level dependent part of the Markov chain (i.e., levels $0, 1, \ldots, c_K$), the probabilities can be obtained by solving a finite level QBD process. Details for computing $\boldsymbol{\phi}$ are given in Algorithm 2.

Algorithm 2. Computation of stationary distribution for $\{(q_{w,K}(t), ((q_{a,K}(t), q_{a,K-1}(t), \ldots, q_{a,1}(t))), t \geqslant 0\}$

1. Check stability of the Markov chain using the condition (20). If the system is stable, continue with step 2; Otherwise the stationary probability distribution does not exist.
2. Find R_w by solving Eq. (23).
3. Set $R_{w,c_K} = R_w$.
4. Find $R_{w,n}$ for $1 \leqslant n < c_K$ recursively starting from $n = c_K - 1$ using the equation: $R_{w,n} = -\lambda_K \left(Q_N^{(K)} - \lambda_K I - M_{K,n} + R_{w,n+1} M_{K,n+1}\right)^{-1}$
5. Find the vector ϕ_0 using the boundary and normalizing conditions: $\boldsymbol{\phi}_0 \left(Q_N^{(K)} - \lambda_K I + R_{w,1} M_{K,1}\right) = 0$, $\boldsymbol{\phi}_0 (I + R_{w,1} + R_{w,1}R_{w,2} + \ldots + R_{w,1}R_{w,2} \ldots R_{w,c_{K-1}} + R_{w,1}R_{w,2} \ldots R_{w,c_K} (I - R_w)^{-1}) \boldsymbol{e} = 1$.
6. For $1 \leqslant n \leqslant c_K$, find ϕ_n starting from $n = 1$ up to $n = c_K$ using equation: $\phi_n = \phi_{n-1} R_{w,n}$.
7. For $n > c_K$, find ϕ_n using Eq. (22).

By routine calculations, the mean queue length of walk-in patients in the K^{th} ED can be obtained as

$$E[q_{w,K}] = \sum_{n=0}^{c_K - 1} n \boldsymbol{\phi}_n \boldsymbol{e} + \boldsymbol{\phi}_{c_K} (R_w (I - R_w)^{-2} + c_K (I - R_w)^{-1}) \boldsymbol{e}. \tag{24}$$

4.3. Sojourn times for walk-in patients

We now construct a continuous time Markov chain for analyzing the sojourn time of a walk-in patient. Since a walk-in patient may get a bed and then lose it a number of times prior to leaving the hospital, we focus on the sojourn time, $w_{w,K}$, the total time that a walk-in patient spends in the K^{th} ED.

To find the distribution of the sojourn time, we construct an absorbing Markov chain for the sojourn time of a tagged walk-in patient. To do so, we only need to consider those walk-in patients who arrived before the tagged walk-in patient. That is: there is no arrival of walk-in patients in the absorbing Markov chain for the sojourn time. The Markov chain is terminated when the tagged walk-in patient completes its service. If the tagged walk-in patient occupies a bed, the service is completed at the rate μ_K. The tagged walk-in patient may be pushed out of a bed a number of times by ambulance patients before the completion of service. Again, we recall that the service to ambulance patients is not affected by that of walk-in patients. We define, for $0 \leqslant n \leqslant c_K - 1$,

$$T_{n,w} = \begin{pmatrix} Q_N^{(K)} - M_{K,1} & & & \\ M_{K,1} & Q_N^{(K)} - M_{K,2} & & \\ & \ddots & \ddots & \\ & & M_{K,n} & Q_N^{(K)} - M_{K,n+1} \end{pmatrix}, \tag{25}$$

and, for $n \geqslant c_K$,

$$T_{n,w} = \begin{array}{c} 0 \\ 1 \\ \vdots \\ c_K \\ \vdots \\ n \end{array} \begin{pmatrix} Q_N^{(K)} - M_{K,1} & & & & & & \\ M_{K,1} & Q_N^{(K)} - M_{K,2} & & & & & \\ & \ddots & \ddots & & & & \\ & & M_{K,c_K} & Q_N^{(K)} - M_{K,c_K} & & & \\ & & & \ddots & \ddots & & \\ & & & & & M_{K,c_K} & Q_N^{(K)} - M_{K,c_K} \end{pmatrix}. \tag{26}$$

Given that there are n walk-in patients already in the K^{th} ED when a tagged walk-in patient arrives, the tagged patient's sojourn time has a phase-type distribution with matrix representation $((0,\ldots,0,\boldsymbol{\phi}_n/(\boldsymbol{\phi}_n\mathbf{e}),T_{n,w})$. Note that, if the level within $T_{n,w}$ is $c_K - 1$ or less, the tagged patient may be in service, depending on the number of ambulance patients in the K^{th} ED, and may complete its service earlier than other patients in service. Then we obtain the conditional probability distribution of the sojourn time as:

$$P\{w_{w,K} \leqslant t|n\} = 1 - (0,\ldots,0,\boldsymbol{\phi}_n/(\boldsymbol{\phi}_n\mathbf{e}))\exp\{T_{n,w}t\}\mathbf{e}. \tag{27}$$

The distribution of the sojourn time of an arbitrary walk-in patient can be obtained as

$$P\{w_{w,K} \leqslant t\} = 1 - \sum_{n=0}^{\infty}(0,\ldots,0,\boldsymbol{\phi}_n)\exp\{T_{n,w}t\}\mathbf{e}. \tag{28}$$

By using truncation, the above formula can be used to compute the distribution of the sojourn time. Furthermore, the following explicit formula can be obtained for the mean sojourn time, where the computation can be done in a finite number of steps, as long as the matrix R_w can be obtained. Define

$$D_K = -\left(Q_N^{(K)} - M_{K,c_K}\right)^{-1} M_{K,c_K}, \text{ and } A_K = -\left(Q_N^{(K)} - M_{K,c_K}\right)^{-1}. \tag{29}$$

For $0 \leqslant n \leqslant c_K - 1$, define

$$\begin{aligned} B_n = &-\left(Q_N^{(K)} - M_{K,n+1}\right)^{-1} + \left(Q_N^{(K)} - M_{K,n+1}\right)^{-1} M_{K,n}\left(Q_N^{(K)} - M_{K,n}\right)^{-1} \\ &+ \ldots + (-1)^{(n+1)}\left(Q_N^{(K)} - M_{K,n+1}\right)^{-1} M_{K,n}\left(Q_N^{(K)} - M_{K,n}\right)^{-1} \\ &\ldots M_{K,1}\left(Q_N^{(K)} - M_{K,1}\right)^{-1}. \end{aligned} \tag{30}$$

By routine calculations, the conditional mean sojourn time can be found as, for $0 \leqslant n \leqslant c_K - 1$,

$$E[w_{w,K}|n] = -(0,\ldots 0,\boldsymbol{\phi_n}/(\boldsymbol{\phi_n}\mathbf{e}))T_{n,w}^{-1}\mathbf{e} = \frac{\boldsymbol{\phi}_n}{\boldsymbol{\phi}_n\mathbf{e}}B_n\mathbf{e}, \tag{31}$$

and, for $n \geqslant c_K$,

$$E[w_{w,K}|n] = \frac{\boldsymbol{\phi}_n}{\boldsymbol{\phi}_n\mathbf{e}}\left(A_K + D_KA_K + D_K^2A_K + \ldots + D_K^{n-c_K}A_K + D_K^{n-c_K+1}B_{c_K-1}\right)\mathbf{e}. \tag{32}$$

Note that D_K is an irreducible stochastic matrix. Then there exists a unique stochastic vector $\boldsymbol{\theta}_{D_K}$ satisfying $\boldsymbol{\theta}_{D_K}D_K = \boldsymbol{\theta}_{D_K}$ and $\boldsymbol{\theta}_{D_K}\mathbf{e} = 1$. It can be shown that $I - D_K + \mathbf{e}\boldsymbol{\theta}_{D_K}$ is invertible. By routine calculations, Eq. (32) can be reduced to

$$\begin{aligned} E[w_{w,K}|n] = \frac{\boldsymbol{\phi}_n}{\boldsymbol{\phi}_n\mathbf{e}}\Big(&\left(I - D_K^{n-c_K+1}\right)(I - D_K + \mathbf{e}\boldsymbol{\theta}_{D_K})^{-1}A_K \\ &+(n - c_K + 1)\mathbf{e}\boldsymbol{\theta}_{D_K}A_K + D_K^{n-c_K+1}B_{c_K-1}\Big)\mathbf{e}. \end{aligned} \tag{33}$$

For an arbitrary walk-in patient at the K^{th} ED, we obtain:

$$\begin{aligned} E[w_{w,K}] = &\sum_{n=0}^{c_K-1}\boldsymbol{\phi_n}\mathbf{e}E[w_{w,K}|n] + \boldsymbol{\phi_{c_K}}(I - R_w)^{-1}(I - D_K \\ &+ \mathbf{e}\boldsymbol{\theta}_{D_K})^{-1}A_K\mathbf{e} + \boldsymbol{\phi}_{c_K}(I - R_w)^{-2}\mathbf{e}\boldsymbol{\theta}_{D_K}A_K\mathbf{e} \\ &+ \boldsymbol{\phi}_{c_K}\left(\sum_{n=0}^{\infty}R_w^nD_K^n\right)D_K\left(B_{c_K-1} - (I - D_K + \mathbf{e}\boldsymbol{\theta}_{D_K})^{-1}A_K\right)\mathbf{e}. \end{aligned} \tag{34}$$

The infinite summation in Eq. (34) can be transformed into the following form by using the direct-sum $f(\cdot)$:

$$f\left(\sum_{n=0}^{\infty}R_w^nD_K^n\right) = \sum_{n=0}^{\infty}f(I)(R_w' \otimes D_k)^n = f(I)(I - R_w' \otimes D_K)^{-1}. \tag{35}$$

We note that (1) the direct-sum $f(X)$ of matrix X is a row vector and is obtained by concatenating the rows of X starting from the first row and; (2) $R_w' \otimes D_K$ is the Kronecker product of matrices R_w' and D_K. Consequently, computing $E[w_{w,K}]$ involves only finite summations and can be done without truncation. The procedure to compute $E[w_{w,K}]$ is summarized in Algorithm 3.

Algorithm 3. Computation of $E[w_{w,K}]$

1. Find R_w by solving Eq. (23).
2. Compute D_K and A_K by Eq. (29).
3. Use Eq. (35) to find $f(\sum_{n=0}^{\infty}R_w^nD_K^n)$ and $\sum_{n=0}^{\infty}R_w^nD_K^n$.
4. Use Eq. (34) to find $E[w_{w,K}]$.

Similar to the mean queue length and mean waiting time for ambulance patients, Little's law applies to the mean queue length $E[q_{w,k}]$ and mean sojourn time $E[w_{w,k}]$, i.e., $E[q_{w,k}] = \lambda_k E[w_{w,k}]$. Thus, computing one gives the other. Little's law can be used for an accuracy check if both are computed separately. Since all computations in this section, as well as in Section 3, involve large matrices, it is important to compute both $E[q_{w,k}]$ and $E[w_{w,k}]$ and use Little's law to check the accuracy of the computations.

Remark 1. We note that the waiting time of a tagged walk-in patient (i.e., the time from the arrival of the patient until the first time that the patient gets a bed) can be studied similarly. Absorbing Markov chains can be constructed in the same way, except that only states without a bed available to the tagged patients are kept. The details are omitted here. Kao and Narayanan [13] consider a multiprocessor single node queue and two types of jobs with one having preemptive priority over the other. To find the waiting time distribution for the low priority jobs, they find the distributions of two random variables: the time spent waiting in the queue until reaching a server, and the time elapsed between the epoch when the job reaches the server for the first time and the epoch it departs the system. Our approach described above is direct and more efficient.

5. Case studies

In this section, we use the methods developed in Sections 3 and 4 to analyze three cases that have been developed to reflect a scaled down version of a real EMS-ED system from southwestern Ontario, Canada.

5.1. Parameter selection

As noted above, parameters for the three case studies are guided by scaling down a real EMS-ED system to reflect the capacity of its single EMS provider and the acute and intermediate patient care areas of the three regional hospitals served by the EMS. The case studies are developed with the following features:

(1) Case study 1 represents a small network (i.e., small numbers of ambulances and beds) that experiences infrequent offload delays.
(2) Case study 2 represents a somewhat larger EMS-ED network with greater arrival rates, in which significant offload delays are experienced. For this case study, we also investigate the effect of ambulance routing probabilities on total offload delays experienced by the EMS.

(3) Case study 3 represents parameter values closest to the real EMS-ED system this work is based on. For this case study, we investigate the effect of service rates on offload delays.

More specifically, individual system parameters for the three case studies are selected as follows.

5.1.1. Number of ED servers

The number of ED servers was chosen to reflect a scaled down version of the actual number of beds in the real EMS-ED system. The reason we used a smaller number of servers than the number of actual beds is that service to patients is constrained by resources such as nurses and doctors, and also the fact that ED beds are routinely occupied by patients that have been admitted to the hospital and they are waiting for an inpatient bed.

5.1.2. Patient arrival rates

Real ED utilization rates and known proportions of walk-in patients vs ambulance arrivals were used to select the arrival rates for ambulance and walk-in patients. We then varied the ambulance patient arrival rates to generate different EMS workloads.

5.1.3. Routing probabilities

Actual data on routing probabilities were used to select the values of $\{p_1, p_2, p_3\}$ for the three EDs. We note that one of the EDs receives up to 45% of the ambulance arrivals and had a disproportionate overall arrival rate of patients, as compared to its overall capacity.

5.1.4. Service rates at EDs

The service rate for each ED, μ_k, was selected based on real Length of Stay (LOS) data. The LOS is approximately 6 hour, which is equivalent to $\mu_k = 1/6$. For case study 3, we varied ED service rates to observe its impact on ED performance measures.

To compare ED performance in each case study, we define two types of server utilization for the k^{th} ED, for $1 \leqslant k \leqslant K$:

- ED utilization for ambulance patients $\rho_{a,k}$: since the service of ambulance patients is not affected by walk-in patients, we can define the server utilization for ambulance patients. Define $\rho_{a,k} = \min\{1, \lambda_0 p_k(1 - P_L)/(c_k\mu_k)\}$, where $\lambda_0 p_k(1 - P_L)$ is the arrival rate of ambulance patients to the k^{th} ED, and $c_k\mu_k$ is the total service capacity at the k^{th} ED.
- ED total utilization ρ_k: considering both types of patients, server utilization can be defined as $\rho_k = \min\{1, (\lambda_0 p_k(1 - P_L) + \lambda_k)/(c_k\mu_k)\}$.

5.2. Case study 1

The system parameters used in this case are recorded in Table 1. The results are reported in Table 2.

Results in Table 2 show the dramatic difference between waiting times for ambulance and walk-in patients. For ambulance patients, the mean waiting times (offload delays) are almost zero. For walk-in patients, the mean sojourn times are more than 11 hour in all three EDs. Ambulance patients consume slightly less than 30% of the ED capacity, but since they get priority over walk-in patients, they have much shorter waiting times. The overall ED utilization is close to 90%, which, together with the priority service discipline, causes much longer waiting times for the lower priority walk-in patients. The results show clearly the effect of the priority service discipline on the waiting times of all patients and the offload delays of ambulances.

This case study shows that the priority based admitting policy has a great impact on patient waiting times. Assigning a higher priority to ambulance patients ensures short waiting times and minimal offload delays at the cost of long waiting times for walk-in patients.

Table 1
System parameters for case study 1.

Parameter set	Value
N	6
(λ_0) patients/hour	1.5
$(\lambda_1, \lambda_2, \lambda_3)$ patient/hour	(1.7, 1.4, 0.8)
(μ_1,μ_2,μ_3) patient/hour	(1/6, 1/6, 1/6)
(c_1,c_2,c_3)	(15, 12, 8)
(p_1,p_2,p_3)	(0.45, 0.29, 0.26)
$(\rho_{a,1},\rho_{a,2},\rho_{a,3})$	(27%, 22%, 29%)
(ρ_1, ρ_2, ρ_3)	(95%, 91.75%, 89.25%)

Table 2
Performance measures for case study 1.

Measures	Matrix analytic results		
	$k = 1$	$k = 2$	$k = 3$
$E[q_{a,k}]$	4.05	2.61	2.34
$E[O^{(k)}]$	8.7×10^{-6}	5.4×10^{-6}	1.3×10^{-3}
$E[w_{a,k}]$	1.29×10^{-6}	1.25×10^{-6}	3.2×10^{-3}
$E[q_{w,k}]$	24.10	16.06	10.44
$E[w_{w,k}]$	14.17	11.47	13.06
P_L	1.35×10^{-6}		

5.3. Case study 2

In this case study, a slightly smaller ED capacity is used, and we study the impact of varying the routing probabilities $\{p_1, \ldots, p_K\}$ on system performance. We consider two scenarios. The first scenario reflects the unbalanced routing probabilities present in the real system studied. The imbalance is a result of heuristic routing policies used by the emergency control center staff, as well as the need to send patients with certain illnesses to a specific hospital because of the services it provides. We have not captured this in our model except through an imbalance in the routing probabilities to each ED. The second scenario corresponds to a system in which the routing probabilities are proportional to ED capacity. Specifically, we set $p_k = c_k\mu_k/(c_1\mu_1 + c_2\mu_2 + c_3\mu_3)$ for $k = 1, 2, 3$. The same patient arrival rates are used for both scenarios, as shown in Table 3. They were chosen to accentuate the impact of the imbalanced routing probabilities on offload delays.

The results, recorded in Table 4 for both scenarios, show how balancing the ED utilization for ambulance patients $\{\rho_{a,1},\rho_{a,2},\rho_{a,3}\}$, has balanced the number of ambulances in offload delays at the EDs. More interestingly, the expected total number of ambulances in offload delay (i.e., $\sum_{k=1}^{3} E[O^{(k)}]$) is decreased from 3.42 (=1.68 + 0.16 + 1.58) in the current scenario to 2.92 (=0.83 + 0.93 + 1.16) ambulances in the balanced scenario, which corresponds to a 14% decrease in the number of ambulances in offload delays. The total expected offload delay (i.e., $\sum_{k=1}^{3} p_k E[w_{a,k}]$) is decreased from 0.54 hour to 0.45 hours in the balanced scenario. This corresponds to a 9.9% decrease in the total hours of offload delays experienced in the region. The loss probability P_L is decreased from 6.93% in the current scenario to 4.98% in the balanced scenario. These loss probabilities are higher than what is experienced in the real system; our interest was in demonstrating the impact of routing decisions on offload delays.

From the EMS perspective, decision makers are interested in finding routing probabilities for which the total number of ambulances in offload delay is minimized. Fig. 3 presents the distributions of ambulances in offload delay under both the initial (unbalanced) and balanced routing scenarios. Under the unbalanced scenario,

Table 3
System parameters for case study 2.

Parameter set	Value
N	9
(λ_0) Patient/hour	7
(λ_1, λ_2, λ_3) Patient/hour	(0.3, 0.6, 0.23)
(μ_1, μ_2, μ_3) Patient/hour	(1/6, 1/6, 1/6)
(c_1, c_2, c_3)	(20, 17, 12)

Table 5
System parameters for case study 3.

Parameter set	Value
N	16
(λ_0) Patient/hour	7
(λ_1,λ_2,λ_3) Patient/hour	(0.75, 0.9, 0.5)
(c_1, c_2, c_3)	(24, 21, 16)
(p_1, p_2, p_3)	(0.45, 0.29, 0.26)

the probability of zero ambulances in offload delay is 29%, while under the balanced scenario this probability is 35%. This represents a significant increase in the availability of ambulances, and will result in better coverage and lower operating costs for the EMS provider.

While the benefit to ambulance patients is clear, the impact of balancing the routing probabilities had a negative effect on the waiting times for walk-in patients in the second ED. As shown in Table 4, the total utilization of the second ED is 100% for the balanced scenario. Then the queue of walk-in patients can be very long. Consequently, the routing mechanism has to be adjusted for implementation in practice. Nevertheless, the results indicate a possible direction for reducing offload delays of ambulance patients, without increasing service capacity.

5.4. Case study 3

In this case study, we increase the number of ambulances to 16. We set the number of servers at each ED to be roughly 60% of the number of beds available within the real system being studied to reflect a realistic throughput rate for patients when they have a mean LOS of 6 hours. To study the impact of changing patient LOS, we vary the mean service time from (1/6, 1/6, 1/6) to (1/5, 1/5, 1/5). Increasing the service rate or increasing the number of servers have similar effects on the performance measures because both variations correspond to increasing the service capacity (i.e., $c_k\mu_k$) at the EDs. The system input parameters for this case study are reported in Table 5.

The results recorded in Table 6 indicate that for the current situation, EMS provides enough ambulances and the three hospitals provide ample capacity to serve ambulance patients. The waiting times (offload delays) for ambulance patients are short, but the queue lengths and waiting times of walk-in patients are significant. This corresponds fairly well to the real system we studied.

We also record the results when the service rate of each of the three EDs is increased from 1/6 to 1/5 in Table 6. As expected, total offload delays, walk-in patient sojourn times and expected queue lengths decrease as the service capacity increases. Compared to that of ambulance patients, the sojourn time for walk-in patients decreases more drastically. Further, we observe that the benefit of adding capacity is greater for EDs with higher utilization. As shown in Table 6, the improvement in the first ED performance is the highest and the change in the second ED is the lowest. This is also expected given the relationship between waiting times and system utilization.

This case study shows how our model can be used to assess the effect of adding more capacity to the system. It also shows where to add resources in order to improve the system performance the most.

Remark 2. Using formula (45) in Appendix A, the sizes of the matrix blocks (e.g., $Q_N^{(K)}, M_{K,C_K}, R_w$) are 5276 for case 1, 14835 for case 2, and 39174 for case 3. It is clear that the space complexity for computing the matrix-geometric solution increases quickly as

Table 4
Performance measures for case study 2.

Performance measure	Current			Balanced		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
p_k	45%	29%	26%	40.82%	34.69%	24.49%
$\rho_{a,k}$	87.95%	66.68%	84.69%	81.45%	81.44%	81.45%
ρ_k	96.95%	87.86%	96.19%	90.45%	100%	92.95%
$E[q_{a,k}]$	19.27	11.50	11.74	17.12	14.78	10.93
$E[O^{(k)}]$	1.68	0.16	1.58	0.83	0.93	1.16
$E[w_{a,k}]$	0.60	0.09	0.93	0.32	0.43	0.71
$E[q_{w,k}]$	18.12	7.46	15.34	5.33	–	7.75
$E[w_{w,k}]$	60.40	12.43	66.70	17.77	–	33.70
P_L		6.93%			4.98%	

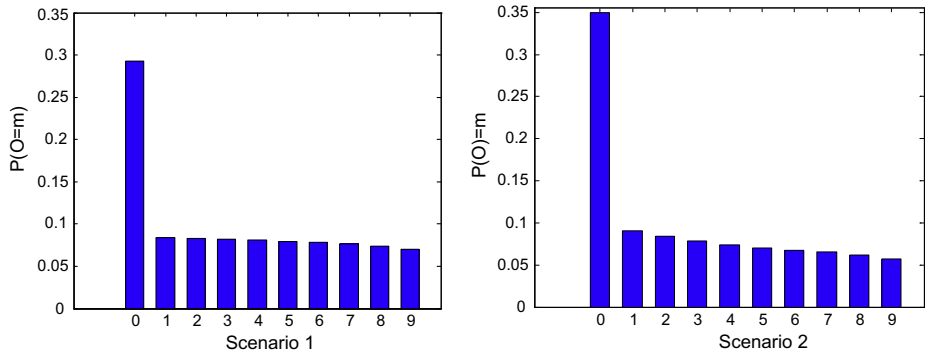

Fig. 3. The distribution for the total number of ambulances in offload delay.

Table 6
Performance measures for case study 3.

Performance measure	Current			Increased capacity		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
μ_k	1/6	1/6	1/6	1/5	1/5	1/5
$\rho_{a,k}$	78.75%	58.00%	68.25%	65.63%	48.33%	56.88%
ρ_k	97.50%	83.71%	87.00%	81.25%	69.76%	72.50%
$E[q_{a,k}]$	19.52	12.19	11.14	15.82	10.15	9.14
$E[O^{(k)}]$	0.64	0.02	0.23	0.07	0.00	0.04
$E[w_{a,k}]$	0.20	0.01	0.13	0.02	9.32×10^{-5}	0.04
$E[q_{w,k}]$	20.85	7.10	5.98	4.74	4.69	2.90
$E[w_{w,k}]$	27.80	7.89	11.95	6.32	5.21	5.79
P_L	9.01×10^{-4}			1.6×10^{-5}		

the numbers of ambulances and ED beds grow. Further exploring the tridiagonal structure in those matrix blocks may make the algorithms more efficient.

6. Model validation

The queueing network model developed in this paper is based on two main assumptions: transit times of ambulances are negligible, and service times at EDs are exponentially distributed. These assumptions have been made to gain insights into system performance without making the models highly complex. In this section we created a simulation model to show that relaxing these assumptions does not have a large impact on the nature of our results. We show that transit times have a negligible effect on offload delays as they occur at an upstream stage of the network. We are also able to show that the performance measure results for ambulance patients are less sensitive to the service time distribution than for walk-in patients. In other words, offload delays are not significantly affected by the service time distribution assumed.

We use the three case studies in Section 5 as the base models for model validation. We add transit time to the queueing network or change the service time distribution from exponential to more general distributions. The validation models are then analyzed through simulation. Performance measures are collected for the original models and for the validation models. Then we compare the results. The assumptions are validated if the performance measures collected for the original and validation models are close to each other.

6.1. Transit time assumption

First, we consider an extended model in which the transit time of ambulance patients is nonzero. Real transit time data was found to be well captured by a beta distribution with parameters ($\alpha = 2.75, \beta = 22.9$) and a mean of 0.73 (Stat-Fit was used to conduct the statistical fitting). In the queueing literature, the exponential distribution is often used to model ambulance transit times or service time (e.g., [22] and [12]). We also used an exponential distribution with parameter $\mu = 1/0.73$ as a second alternative for the transit time distribution.

We define the utilization of ambulances in the EMS, u_A, as the long-term percentage of time a random ambulance is being used. For the zero transit time case, an ambulance is busy only when it is experiencing offload delays. Mathematically, $u_A = E[O]/N$. For

Table 7
Effects of nonzero transit time (*Note*: the 95% confidence interval half widths for simulation in parentheses).

System performance measure	Case study 1								
	Zero transit time			Beta (2.75, 22.9)			Exponential (0.73 hour)		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
$E[q_{a,k}]$	4.05	2.61	2.34	4.04(0.01)	2.61(0.01)	2.33(0.01)	4.04(0.01)	2.61(0.01)	2.33(0.01)
$E[O^{(k)}]$	8.7×10^{-6}	5.4×10^{-6}	1.3×10^{-3}	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
$E[w_{a,k}]$	1.29×10^{-6}	1.25×10^{-6}	3.2×10^{-6}	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
$E[q_{w,k}]$	24.10	16.06	10.44	24.99(0.56)	16.01(0.24)	10.44(0.19)	25.23(0.61)	15.96(0.19)	10.51(0.24)
$E[w_{w,k}]$	14.17	11.47	13.06	14.70(0.33)	11.42(0.18)	13.05(0.24)	14.84(0.36)	11.40(0.14)	13.14(0.29)
P_L	1.35×10^{-6}			8.88×10^{-4} (0.43×10^{-4})			8.10×10^{-4} (0.41×10^{-4})		
u_A	0.02%			18.61% (0.03×10^{-2})			18.27% (0.03)		
Case study 2									
$E[q_{a,k}]$	19.27	11.50	11.74	17.10(0.03)	10.72(0.03)	10.20(0.03)	17.17(0.02)	10.74(0.03)	(10.25(0.02)
$E[O^{(k)}]$	1.68	0.16	1.58	0.59(0.01)	0.07(0.01)	0.66(0.01)	0.62(0.01)	0.07(0.00)	0.68(0.01)
$E[w_{a,k}]$	0.60	0.09	0.93	0.23(0.01)	0.05(0.01)	0.43(0.01)	0.22(0.01)	0.04(0.01)	0.42(0.02)
$E[q_{w,k}]$	18.12	7.46	15.34	5.25(0.07)	5.56(0.04)	5.08(0.12)	5.43(0.06)	5.63(0.05)	5.26(0.11)
$E[w_{w,k}]$	60.40	12.43	66.70	17.53(0.25)	9.26(0.07)	22.11(0.51)	18.08(0.23)	9.38(0.08)	22.87(0.47)
P_L	6.93%			12.58% (0.03×10^{-2})			12.40% (0.04×10^{-2})		
u_A	38.00%			65.22% (0.05×10^{-2})			64.83% (0.05×10^{-2})		
Case study 3									
$E[q_{a,k}]$	19.52	12.19	11.14	19.33(0.05)	12.14(0.05)	11.07(0.04)	19.33(0.06)	12.14(0.02)	11.07(0.04)
$E[O^{(k)}]$	0.64	0.02	0.23	0.52(0.01)	0.02(0.00)	0.20(0.01)	0.52(0.01)	0.02(0.01)	0.20(0.01)
$E[w_{a,k}]$	0.20	0.01	0.13	0.17(0.01)	0.01(0.01)	0.11(0.02)	0.17(0.01)	0.01(0.01)	0.11(0.02)
$E[q_{w,k}]$	20.85	7.10	5.98	26.98(0.25)	6.83(0.04)	5.89(0.08)	27.29(1.39)	6.82(0.05)	5.86(0.08)
$E[w_{w,k}]$	27.80	7.89	11.95	36.50(0.40)	7.71(0.03)	12.08(0.06)	36.93(0.47)	7.66(0.02)	12.11(0.06)
P_L	9.01×10^{-4}			4.8×10^{-3} (0.00)			4.7×10^{-3} (0.00)		
u_A	5.56%			36.42% (0.08)			36.42% (0.02)		

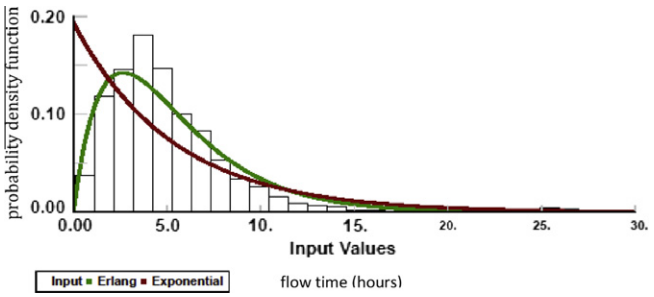

Fig. 4. The fitted distribution for patient flow time.

the nonzero transit time case, an ambulance is busy if it is either transferring a patient or experiencing offload delays. The EMS utilization rates for the two non-zero transit time cases were collected from the simulation output. The results are presented in Table 7 for all three case studies. We have the following observations:

- The results in Table 7 support the assumption that zero transit time has negligible effect on the offload delays experienced by ambulances for case studies 1 and 3, where the ambulance utilization, u_A, is small or moderate (i.e., 18% and 36%). This is consistent with [4], who noted that the departure process of an M/G/1 queue is Poisson, and thus if N is not small, the arrival processes to the EDs are also Poisson, and is not affected by the transit time.
- When the ambulances are highly utilized as in case study 2 (i.e., 65%), the probability of losing patients increases significantly when the transit time becomes nonzero. The offload delays do not change significantly, but the waiting times of walk-in patients are changed dramatically. In fact, due to losing about 12% of the ambulance patients, walk-in patients get served more quickly (i.e., $E[w_{w,k}]$ is smaller).
- Both the beta and exponential distributions give similar results for the system performance measures.
- For case study 1 (low offload delays case) the simulation does not capture well the small possibility of offload delays at the three EDs. The analytic method finds that, for example, expected offload delays in the third ED are 3.2×10^{-3} hours, which corresponds to 13.82 ambulance hours per month. The simulation gives zero for the expected offload delays after 500 replications and 3.3 machine hours (on a ThinkPad W500 16 GB RAM computer). This demonstrates a limitation of the simulation approach, which is the difficulty in capturing rare events.

6.2. Service time assumption

The second assumption we validate is the exponential service time for serving patients at the EDs. The data we have from one of the regional hospitals is for the flow time of patients, so it includes patients' delays and service time. To approximate the service time distribution, we fitted flow time data using the Stat-Fit package. The resulting distribution is Erlang of order 2 and is shown in Fig. 4. We assume that the service time has a similar distribution to the flow time but with a different mean. Then the candidate for the service time distribution is the Erlang distribution.

Since the Erlang distribution does not have the memoryless property, the preemptive repeat and the preemptive resume service discipline give different results. We assume a preemptive resume service discipline for walk-in patients in this section, which is closer to the practice in the EDs. In Table 8, analytical and simulation results are reported for all three case studies of Section 5, where the service time is Erlang with the same mean as the exponential distribution. We have the following observations:

- The results in Table 8 support the assumption that the exponential service time has negligible effects on the the offload delays experienced by ambulances for the three cases studies considered.
- Due to the smaller coefficient of variation for the Erlang distribution, expected queue lengths and consequently, expected waiting times for both ambulance and walk-in patients are slightly lower under the Erlang service time distribution (for case studies 1 and 3 only). Thus, our assumption of exponentially distributed service time leads to an upper bound on the system performance measures.
- Another observation we have with respect to case study 2 is the significant increase in walk-in patients' expected sojourn time and queue lengths at all EDs when the service time distribution

Table 8
Service time distribution effect (95% confidence interval half widths in parentheses).

Performance measure	Exponential			Erlang $M = 2$		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
Case study 1						
$E[q_{a,k}]$	4.05	2.61	2.34	4.05(0.01)	2.61(0.01)	2.34(0.01)
$E[O^{(k)}]$	8.7×10^{-6}	5.4×10^{-6}	1.3×10^{-3}	0.00(0.00)	0.00(0.00)	0.00(0.00)
$E[w_{a,k}]$	1.29×10^{-6}	1.25×10^{-6}	3.2×10^{-6}	0.00(0.01)	0.00(0.01)	0.00(0.01)
$E[q_{w,k}]$	24.10	16.06	10.44	22.44(0.43)	14.78(0.24)	9.63(0.17)
$E[w_{w,k}]$	14.17	11.47	13.06	13.20(0.24)	10.56(0.14)	12.03(0.21)
P_L		1.35×10^{-6}			$1.00 \times 10^{-6}(1.17 \times 10^{-6})$	
Case study 2						
$E[q_{a,\,k}]$	19.27	11.50	11.74	19.5(0.03)	11.63(0.03)	11.83(0.02)
$E[O^{(k)}]$	1.68	0.16	1.58	1.69(0.01)	0.15(0.01)	1.54(0.01)
$E[w_{a,\,k}]$	0.60	0.09	0.93	0.57(0.01)	0.08(0.01)	0.90(0.01)
$E[q_{w,\,k}]$	18.12	7.46	15.34	27.76(1.65)	7.54(0.11)	21.15(0.93)
$E[w_{w,\,k}]$	60.40	12.43	66.70	86.39(2.34)	12.57(0.20)	92.00(3.98)
P_L		6.93×10^{-2}			$5.77 \times 10^{-2}(5.0 \times 10^{-4})$	
Case study 3						
$E[q_{a,\,k}]$	19.52	12.19	11.14	19.43(0.03)	12.20(0.02)	11.09(0.03)
$E[O^{(k)}]$	0.64	0.02	0.23	0.54(0.01)	0.02(0.00)	0.18(0.01)
$E[w_{a,\,k}]$	0.20	0.01	0.13	0.17(0.01)	0.01(0.00)	0.10(0.01)
$E[q_{w,\,k}]$	20.85	7.10	5.98	32.63(1.97)	6.94(0.05)	5.58(0.07)
$E[w_{w,\,k}]$	27.80	7.89	11.95	44.29(2.57)	7.74(0.04)	11.52(0.11)
P_L		9.01×10^{-4}			$4.5 \times 10^{-4}(3.1 \times 10^{-7})$	

is Erlang. This is because under the Erlang distribution service time, which has a smaller coefficient of variation, more high priority ambulance patients are accepted (P_L decreased). As a result, the low priority walk-in patients queue lengths and waiting times increase significantly.

In summary, if the loss probability is small, performance measures for both types of patients are not affected significantly by adding the transit time or by changing the service time distribution. In reality, ambulances usually operate at around $u_A = 35\%$ utilization [20] (including transit time), which is similar to case study 3. For such a case, the loss probability is small. This indicates that the queueing network introduced in this paper is robust as long as the system of interest is working under normal operating conditions. In other words, the analysis in this section indicates that the assumptions made in Section 2 are appropriate as long as the ambulance utilization is not too high, which is the actual condition under which the EMS operates.

Acknowledgements

The authors thank two anonymous referees for their valuable suggestions and comments that lead to significant improvement in the exposition of the paper.

Appendix A

To construct $Q_N^{(K)}$, we first construct $Q_n^{(1)}$, for $n = 0, 1, \ldots, N$. Then, recursively, we construct $Q_n^{(k)}$, for $k = 2,3, \ldots, K$. We stop the recursion when $Q_N^{(K)}$ is obtained.

In the following construction, the variable k, $1 \leqslant k \leqslant K$, implies that hospitals 1, 2, …, and k are involved, and the variable n, $0 \leqslant n \leqslant N$, represents the number of available ambulances. For $k = 1$, we have, for $n = 0$,

$$Q_0^{(1)} = \begin{matrix} 0 \\ 1 \\ \vdots \\ c_1 \end{matrix} \begin{pmatrix} 0 & & & \\ \mu_1 & -\mu_1 & & \\ & \ddots & \ddots & \\ & & c_1\mu_1 & -c_1\mu_1 \end{pmatrix}; \tag{36}$$

and, for $n \geqslant 1$,

$$Q_n^{(1)} = \begin{matrix} 0 \\ 1 \\ \vdots \\ c_1 \\ \vdots \\ c_1+n-1 \\ c_1+n \end{matrix} \begin{pmatrix} -p_1\lambda_0 & p_1\lambda_0 & & & & & \\ \mu_1 & -\mu_1-p_1\lambda_0 & p_1\lambda_0 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & c_1\mu_1 & -c_1\mu_1-p_1\lambda_0 & p_1\lambda_0 & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & c_1\mu_1 & -c_1\mu_1-p_1\lambda_0 & p_1\lambda_0 \\ & & & & & c_1\mu_1 & c_1\mu_1 \end{pmatrix}. \tag{37}$$

Note that, if $n = 0$, there is no ambulance available. Thus, there can be no arrival of patients in $Q_0^{(1)}$. If $n \geqslant 1$, the total arrival rate of patients is λ_0 and the arrival rate to the first ED is $p_1\lambda_0$. The service rate is determined by $\min\{c_1, q_1(t)\}\mu_1$.

We also define the following matrices:

$$U_0^{(1)} = (0)_{(c_1+1)\times(c_1+1)};$$
$$U_n^{(1)} = \begin{pmatrix} I_{(c_1+n)\times(c_1+n)} & 0 \\ 0 & 0 \end{pmatrix}_{(c_1+n+1)\times(c_1+n+1)}, \text{ for } n \geqslant 1. \tag{38}$$

$$V_n^{(1)} = \begin{pmatrix} I_{(c_1+n)\times(c_1+n)} \\ 0 \end{pmatrix}_{(c_1+n+1)\times(c_1+n)}, \text{ for } n \geqslant 1. \tag{39}$$

To indicate the size of a matrix, we have used subscripts. For example, $(0)_{(c_1+1)\times(c_1+1)}$ is a square matrix of zeros of size $c_1 + 1$.

We define

$$U_n^{(k)} = \begin{matrix} 0 \\ \vdots \\ c_k \\ c_k+1 \\ \vdots \\ c_k+n \end{matrix} \begin{pmatrix} U_n^{(k-1)} & & & & & \\ & \ddots & & & & \\ & & U_n^{(k-1)} & & & \\ & & & U_{n-1}^{(k-1)} & & \\ & & & & \ddots & \\ & & & & & U_0^{(k-1)} \end{pmatrix}, \text{ for } n \geqslant 0. \tag{40}$$

$$V_n^{(k)} = \begin{matrix} 0 \\ \vdots \\ c_k \\ c_k+1 \\ \vdots \\ c_k+n-1 \\ c_k+n \end{matrix} \begin{pmatrix} V_n^{(k-1)} & & & & & \\ & \ddots & & & & \\ & & V_n^{(k-1)} & & & \\ & & & V_{n-1}^{(k-1)} & & \\ & & & & \ddots & \\ & & & & & V_1^{(k-1)} \\ & & & & & V_0^{(k-1)} \end{pmatrix}, \text{ for } n \geqslant 1. \tag{41}$$

For $2 \leqslant k \leqslant K$, we have, for $n \geqslant 0$ and $0 \leqslant i \leqslant n + c_k$,

$$A_{n(i,i)}^{(k)} = Q_{n-\max(0,i-c_k)}^{(k-1)} - \min(i, c_k)\mu_k I - p_k\lambda_0 U_{n-\max(0,i-c_k)}^{(k-1)}. \tag{42}$$

If $i_k = i$, the number of ambulances available to hospitals 1, 2, .., and $k-1$ is $\max\{0, i - c_k\}$. Thus, the transitions of $(q_{k-1}(t), \ldots, q_1(t))$ are described by $Q_{n-\max\{0,i-c_k\}}^{(k-1)}$. The transitions of $q_k(t)$ are determined by $\min\{i,c_k\}\mu_k I$ for decreasing its value by one, and by $p_k\lambda_0 U_{n-\max\{0,i-c_k\}}^{(k-1)}$ for increasing its value by one.

$$A_{n(i,i+1)}^{(k)} = \begin{cases} p_k\lambda_0 U_n^{(k-1)}, & \text{for } 0 \leqslant i \leqslant c_k - 1; \\ p_k\lambda_0 V_{n-(i-c_k)}^{(k-1)}, & \text{for } c_k \leqslant i \leqslant n + c_k - 1. \end{cases} \tag{43}$$

Note that, for levels i and $i+1$, if $i \geqslant c_k$, they have different number of states. The reason is that if $i \geqslant c_k$, for level $i+1$, there is one less ambulance available for hospitals 1, 2, …, and $k-1$.

$$A_{n(i,i-1)}^{(k)} = \begin{cases} \min(i, c_k)\mu_k I, & \text{for } 1 \leqslant i \leqslant c_k; \\ \min(i, c_k)\mu_k \left(V_{n+1-(i-c_k)}^{(k-1)}\right)', & \text{for } c_k + 1 \leqslant i \leqslant n + c_k. \end{cases} \tag{44}$$

where $\left(V_{n+1-(i-c_k)}^{(k-1)}\right)'$ is the transpose of $\left(V_{n+1-(i-c_k)}^{(k-1)}\right)$.

Then $Q_n^{(k)}$ is constructed from $A_{n(i,i)}^{(k)}, A_{n(i,i+1)}^{(k)}$, and $A_{n(i,i-1)}^{(k)}$ by letting $A_{(i,i)}^{(k)} = A_{n(i,i)}^{(k)}, A_{(i,i+1)}^{(k)} = A_{n(i,i+1)}^{(k)}$, and $A_{(i,i-1)}^{(k)} = A_{n(i,i-1)}^{(k)}$ in Eq. (2),

Algorithm 4. Computing matrix blocks in $Q_N^{(K)}$

1. Based on Eqs. (37)–(39), compute matrices $\left\{Q_n^{(1)}, \text{ for } 0 \leqslant n \leqslant N\right\}, \left\{U_n^{(1)}, \text{for } 0 \leqslant n \leqslant N\right\}$, and $\left\{V_n^{(1)}, \text{for } 1 \leqslant n \leqslant N\right\}$. Set $k = 2$.
2. If $k \leqslant K$, go to step 3; Otherwise, Stop.
3. Based on Eqs. (42)–(44), compute $\left\{A_{n(i,i)}^{(k)}, \text{for } 0 \leqslant n \leqslant N \text{ and } 0 \leqslant i \leqslant n + c_k\right\}, \left\{A_{n(i,i+1)}^{(k)}, \text{ for } 0 \leqslant n \leqslant N \text{ and } 0 \leqslant i \leqslant n + c_k - 1\right\}, \left\{A_{n(i,i-1)}^{(k)}, \text{ for } 0 \leqslant n \leqslant N \text{ and } 1 \leqslant i \leqslant n + c_k\right\}$. Then compute $\left\{Q_n^{(k)}, \text{ for } 0 \leqslant n \leqslant N\right\}, \left\{U_n^{(k)}, \text{ for } 0 \leqslant n \leqslant N\right\}$, and $\left\{V_n^{(k)}, \text{ for } 1 \leqslant n \leqslant N\right\}$. Set $k = :k + 1$, Go to step 2.

Denote by $\xi_{K,N}$ the number of states of the Markov chain $\{(q_{a,K}(t), q_{a,K-1}(t), \ldots, q_{a,1}(t)), t \geqslant 0\}$. Let $\delta(.)$ be the indicator function. Then $\xi_{K,N}$, which is also the size of the matrix $Q_N^{(K)}$, can be obtained as

$$\xi_{K,N} = \sum_{(N_0,N_1,\ldots,N_K):\ \sum_{j=0}^{K} N_j = N, N_j \geqslant 0, 0 \leqslant j \leqslant K} \Pi_{j=1}^{K} (c_j + 1)^{1-\delta(N_j>0)}. \quad (45)$$

References

[1] J. Chaiken, R. Larson, Methods for allocating urban emergency units: a survey, Management Science 19 (1972) 110–130.

[2] N. Channouf, P. L'Ecuyer, A. Ingolfsson, A.N. Avramidis, The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta, Health Care Management Science 10/1 (2007) 25–45.

[3] S. Creemers, J. Beliën, M. Lambrecht, The optimal allocation of server time slots over different classes of patients, European Journal of Operational Research 219 (2012) 508–521.

[4] D.J. Daley, Queueing output processes, Advances in Applied Probability 8/2 (1976) 395–415.

[5] H.E. Department and A.E.W. Group, Improving access to emergency services: a system commitment, Ministry of Health And Long Term Care, Tech. Rep., 2005.

[6] M. Eckstein, L.S. Chan, The effect of emergency department crowding on paramedic ambulance availability, Annals Of Emergency Medicine 43/1 (2004) 100–105.

[7] E. Erkut, A. Ingolfsson, G. Erdogan, Ambulance location for maximum survival, Naval Research Logistics 55 (2008) 42–58.

[8] E. Erkut, A. Ingolfsson, T. Sim, G. Erdogan, Computational comparison of five maximal covering models for locating ambulances, Geographical Analysis 41 (2009) 43–65.

[9] S. Fomundam, J. Herrmann, A survey of queueing theory applications in healthcare, The Institute for Systems Research, Tech. Rep. 2007-24, 2007.

[10] F. Gorunescu, S.I. McClean, P. Millard, A queueing model for bed-occupancy management and planning of hospitals, The Journal of the Operational Research Society 53/1 (2002) 19–24.

[11] L. Green, in: R.W. Hall (Ed.), Queueing Analysis in Healthcare, Patient Flow: Reducing Delay in Healthcare Delivery, Springer, New York, 2006. ch. 10..

[12] S.I. Harewood, Emergency ambulance deployment in Barbados: a multi-objective approach, The Journal of the Operational Research Society 53/2 (2002) 185–192.

[13] E. Kao, K. Narayanan, Modeling a multiprocessor system with preemptive priorities, Management Science 37/2 (1991) 185–197.

[14] E. Kao, G. Tung, Bed allocation in a public health care delivery system, Management Science 27/5 (1981) 507–520.

[15] G. Latouche, V. Ramaswami, An Introduction to Matrix Analytic Methods in Stochastic Modeling. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999.

[16] P. Lewis, Recent Results in the Statistical Analysis of Univariate Point Processes in Stochastic Point Processes, Wiley, New York, 1972.

[17] P. Macintyre, Hospital offload delay status update, Toronto EMS, Tech. Rep., January 2009.

[18] Mandelbaum, P. Momčilovič, Y. Tseytlin, On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers, Management Science 58 (2012) 1273–1291.

[19] F.M. Neuts, Matrix Geometric Solutions in Stochastic Methods: An Algorithmic Approach, Dover Publications, Mineola, NY, USA, 1995.

[20] R. of Waterloo Public Health, Emergency medical services master plan, Tech. Rep., December 2007.

[21] J. Prno, Offload delaysn++ more than just an EMS issue, <http://www.cchse.org/assets/hamiltonandarea/ER_Wait_Times_Panel_Dr_John_Prno.pdf>, April 2010.

[22] M. Restrepo, S. Henderson, H. Topaloglu, Erlang loss models for the static deployment of ambulances, Health Care Management Science 12 (2009) 6779.

[23] V. Schmid, Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming, European Journal of Operational Research 219 (2012) 611–621.

[24] K. Siddhartan, W. Jones, J. Johnston, A priority queueing model to reduce waiting times in emergency care, International Journal of Health Care Quality Assurance 9/5 (1996) 10–16.

[25] S. Silvestri, G. Ralls, L. Papa, M. Barnes, Impact of emergency department bed capacity on emergency medical services unit off-load time, Academic Emergency Medicine 13/5 (2006) 70–71.

[26] S. Silvestri, G. Ralls, K. Shah, G. Parrish, Evaluation of patients in delayed emergency medical services unit off-load status, Academic Emergency Medicine 13/5 (2006) 70.

[27] C. Taylor, D. Williamson, A. Sanghvi, When is a door not a door? the difference between documented and actual arrival times in the emergency department, British Medical Journal 23/6 (2006) 442–443.

[28] J.Y. Ting, The potential adverse patient effects of ambulance ramping, a relatively new problem at the interface between pre hospital and ED care, Journal of Emergency, Trauma, and Shock 1/2 (2008) 129.

[29] D. Worthington, Hospital waiting list management models, The Journal of the Operational Research Society 42 (1991) 833–843.