Omega 65 (2016) 148–158

Contents lists available at ScienceDirect

Omega

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

Analysis and optimization of an ambulance offload delay and allocation problem ☆

Eman Almehdawe [a,*], Beth Jewkes [b], Qi-Ming He [b]

[a] Faculty of Business Administration, University of Regina, 3737 Wascana Parkway, Regina, Canada S4S 0A2
[b] Department of Management Sciences, University of Waterloo, 200 University Avenue West, Waterloo, Canada N2L 3G1

ARTICLE INFO

Article history:
Received 12 June 2014
Accepted 11 January 2016
Available online 25 January 2016

Keywords:
EMS
Offload delays
Queueing networks
Optimization

ABSTRACT

Ambulance offload delays have recently become one of the most significant operational challenges for Emergency Medical Services (EMS) providers. Offload delays occur when an ambulance arriving at a hospital Emergency Department (ED) is blocked until a bed becomes available for the patient. To formally investigate the effect of patient routing decisions on EMS offload delays, we introduce a stylized queueing network model with blocking. Following a decomposition approach, we develop an approximation scheme to find explicit solutions that can be used to find proper patient allocation policies to multiple hospitals in a region. We introduce a Markov chain representation for a single ED network and solve for its exact steady state distribution. A comprehensive numerical study is carried out to validate the approximation approaches and to gain insight into ambulance offload delays. By keeping the total offload delays at minimal levels, we observe that it is better to load larger EDs more heavily than smaller ones due to resource pooling.

1. Introduction

Emergency Medical Services (EMS) are responsible for transferring patients to Emergency Departments (ED) within a target response time. Sometimes, upon arrival at a highly congested ED, an ambulance is forced to wait to offload a patient until a bed becomes available. This waiting time is referred to as *offload delay* in North America, or *access block* in Australia. In some countries, such as the United States, an ED can declare *diversion* status if it is overcrowded [1]. For EMS management, diversion means that patients should be routed to other less crowded EDs. Diversion, or reallocating patients to another regional hospital, can be key in alleviating overall offload delays experienced, but entails higher costs to healthcare systems. In addition, offload delays increase actual EMS response times and waste scarce resources. Thus, offload delays are a pressing concern for EMS management.

Ambulance offload delays have attracted the attention of researchers and practitioners in the past decade. Studies on ambulance offload delays can be categorized as either observational or analytical. Observational studies focus on identifying the relationships between offload delays and hospital congestion. A survey of such studies can be found in Ting [2] and Taylor et al. [3]. More details can be found in Silvestri et al. [4] and Silvestri et al. [5]. Analytical studies on offload delays utilize queueing theory. Specifically, such studies are related to queueing networks with blocking and priority. Queueing models are widely used in service systems analysis to improve customer service. We refer the reader to Formundam and Herrmann [6] and Green [7] for a comprehensive review on the use of queueing theory in healthcare systems. We also refer to Almehdawe et al. [8] for a summary of some queueing works related to hospital bed use and allocation. Other related references are Kao and Tung [9], Gorunescu et al. [10], Davies and Davies [11], Masselink et al. [12], Côté and Stein [13], Knight et al. [14] and Gorunescu et al. [15].

The problem of ambulance allocation to regional hospitals was studied recently by Leo et al. [16]. In their Mixed Integer Programming model, they consider allocating ambulance and walk-in patients simultaneously to regional hospitals by minimizing their travel and waiting times. Then they recommend reorganization of the EMS network based on those results. Compared to the model developed in this paper, we assume that only patients arriving by an ambulance can be allocated by the EMS dispatcher, while walk-in patients select by themselves the ED to which they will go.

The queueing network investigated in this paper is introduced for the analysis and design of EMS and is similar in structure to that in Almehdawe et al. [8], although the objectives, model assumptions, and methodologies are different. While the objective of Almehdawe et al. [8] is to conduct a performance analysis of EMS, the objective of this paper is to develop an optimization

☆This manuscript was processed by Associate Editor Ghate.

* Corresponding author. Tel.: +1 306 585 4728.
E-mail address: Eman.Almehdawe@uregina.ca (E. Almehdawe).

http://dx.doi.org/10.1016/j.omega.2016.01.006
0305-0483/

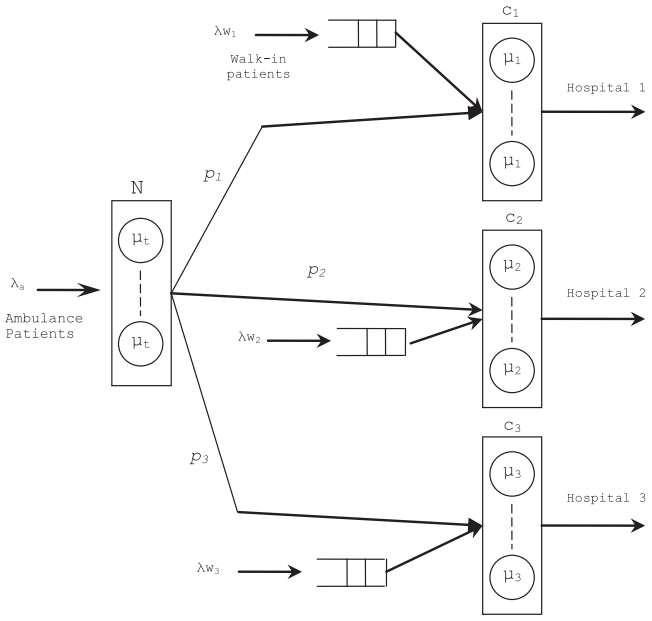

Fig. 1. An EMS-ED queueing network for a region of 3 hospitals.

method for the design of EMS. In Almehdawe et al. [8], the preemption priority is assumed for ambulance patients. In this paper, the non-preemption priority is assumed. In Almehdawe et al. [8], the ambulance transit time is considered to be negligible. In this paper, the ambulance transit time is assumed to have an exponential distribution. As a result, the queueing network becomes far more complicated and the method developed in Almehdawe et al. [8] to derive the steady state probability distributions does not work well due to the curse of dimensionality. In addition, the method used in Almehdawe et al. [8] is not effective in optimization. Thus, a decomposition approach is utilized in this paper, and the matrix-analytic methods and some classical queueing results are applied, at both the hospital level and the region level, to analyze two levels of the queueing network individually.

In the queueing network developed in this paper (see Fig. 1), two types of patients are served: ambulance patients and walk-in patients. We assume that ambulance patients have higher service priority over walk-in patients (see [8] for details on the validity of this assumption). Our approach is to decompose the queueing network into subsystems, each containing one ED. First, we introduce an approximation model for each individual system so that explicit (but approximate) results can be obtained for performance measures such as the mean waiting time of ambulance patients (offload delays). The results are used to select system parameters to minimize the mean offload delays in a given region. Second, we study individual subsystems analytically. Under certain assumptions, we use results from individual subsystems to produce various performance measures of the queueing network. The exact results for individual subsystems are used to check the quality of the parameters selected by using the approximate models. In summary, in this paper, we develop a method to find how to allocate ambulance patients to different hospitals in a given region. We make the following contributions:

- We model the complex problem of offload delays in terms of hospital congestion and EMS system congestion.
- We develop an approximation scheme for individual EDs performance measures and validate the approximation via simulation.
- We construct and solve an optimization problem to find the optimal allocation of ambulance patients to each ED in a region.
- We find explicitly the waiting time distribution for a multi-server queueing system with non-preemptive priorities and blocking.

The rest of the paper is organized as follows. In Section 2, we introduce the EMS system of interest and describe the steps for model approximation and the optimization of the allocation of ambulance patients. In Section 3, we introduce an $M[2]/M/c$ non-preemptive priority queue and an optimization problem for the allocation of ambulance patients. In Section 4, we define a One-ED network and apply matrix-analytic methods to validate the approximation scheme of Section 3. A numerical analysis is carried out in Section 5, where issues such as model validation and optimal allocation of ambulance patients are addressed. Section 6 concludes the paper. Some technical details are collected in Appendices 1–3.

2. The EMS system and solution approaches

We consider an EMS system with N ambulances that serve K hospitals, each with an ED. When the dispatching center receives an emergency call requiring an ambulance, an ambulance is dispatched to the call scene, if one is available. Upon arrival, the paramedic team apply the basic life saving procedure. If the patient requires transport to a hospital, the paramedics load him into the ambulance. Then they transfer the patient to one of the K regional hospitals. The time to reach the patient, load him into the ambulance and then transfer him into the ED is referred to as the patient *transit time*. We refer to such patients as *ambulance patients*. Patients may alternatively go to one of the K EDs by themselves. We refer to such patients as *walk-in patients*. In each ED, both ambulance patients and walk-in patients are served. Ambulance patients have higher service priority. That is: when an ED bed becomes available, it will be assigned to a waiting ambulance patient. If there is no waiting ambulance patient, the bed becomes available to walk-in patients. The service of both types of patients cannot be interrupted. Thus, if an ambulance patient sees all ED beds are occupied upon arrival (i.e., no resource to serve the patient), the patient and its ambulance have to wait, and this waiting time is referred to as *ambulance offload delay*. The ambulance becomes available when the ambulance patient is admitted to the ED. Within each priority group of patients at one ED, patients are served on a first-come-first-served basis. A patient leaves the system immediately after his service is done.

Focusing on the movement of ambulances and patients, the EMS system can be modeled as a queueing network as shown in Fig. 1, which is defined explicitly as follows:

Patient arrival processes: Ambulance patients call the EMS according to a Poisson process with parameter λ_a. Walk-in patients arrive at the k-th ED according to a Poisson process with parameter $\lambda_{w,k}$, for $k=1,2,\ldots,K$. All the Poisson processes are independent. The assumption of Poisson arrivals is supported by empirical studies (see Channouf et al. [17] and the references therein). Justification of the arrival processes can also be found in Almehdawe et al. [8].

Routing probabilities: Upon arrival, an ambulance patient will be transferred to the k-th ED with routing probability p_k, for $k=1,2,\ldots,K$, if an ambulance is available at that moment; otherwise, the patient is lost. Thus, we must have $p_1+p_2+\cdots+p_K=1$. One of the main issues addressed in this paper is how to choose the routing probabilities to minimize the overall offload delays.

Patient transit time: The transit time from dispatching the ambulance to the call scene until it arrives to a hospital is exponentially distributed with parameter $\mu_{T,k}$. Mateo Restrepo and Topaloglu [18] and the references therein use this assumption and

argue that the time spent by an ambulance at the scene typically dominates the travel time.

ED service capacity: At the k-th ED, there are c_k beds available to serve patients. A bed is considered to be the combination of a real hospital bed, nurses, doctors, and others. The service time of one bed (to be referred to as a server) at the k-th ED is exponentially distributed with parameter μ_k.

Patient service priority: In each ED, ambulance patients have non-preemptive service priority over walk-in patients. Thus, when a bed becomes available, it is assigned to a waiting ambulance patient, if there is one. The available bed is assigned a waiting walk-in patient only if there is no waiting ambulance patient. Once a walk-in patient enters a bed, his service will not be interrupted. See [8] for more details on this assumption.

To analyze the queueing network described, a quasi-birth-and-death (QBD) process can be introduced and matrix-geometric solutions can be obtained for computing performance measures (see Neuts [19]). Since the total number of ambulances is finite, the operations in all K EDs are not independent. Since the service discipline is non-preemption, the service of ambulance patients is affected by the service of walk-in patients. Thus, it is a tedious process to introduce proper Markov chains for the EMS system and, even if that can be done, the Markov chain is too big to be analyzed efficiently.

In this paper, we develop two approximation approaches to study the system:

- First, we introduce an $M[2]/M/c$ non-preemptive priority queue to approximate individual EDs, which shall be defined and analyzed explicitly in Section 3. The $M[2]/M/c$ queueing model gives explicit expressions for an estimate of the mean ambulance offload delays. The explicit results of the approximation models can be used to optimize the selection of routing probabilities to reduce the mean ambulance offload delays.
- Second, we consider the special EMS system with $K=1$ in Section 4. We call this case the One-ED network. By using the classical matrix-analytic methods, exact results can be obtained for various performance measures. The results for the One-ED network are considered approximations to that of the individual EDs in the EMS system with $K>1$.

The approximation approaches work well if the K EDs operate independently, which is guaranteed if the loss probability of ambulance patients is zero. Thus, throughout this paper, we make the following assumption:

Assumption: In the EMS system, the loss probability of ambulance patients is close to zero.

The quality of the approximation approaches, which will be validated through simulation in Section 5, depends largely on this assumption. In practice, an EMS normally operates at low utilization levels (30% or less) [20]. That is: the probability of having all the ambulances busy is small. Thus, we assume that at least one ambulance is available all of the time. Therefore, the assumption is justified from application's point of view.

3. An $M[2]/M/c$ non-preemptive priority queue and ambulance routing

If the loss probability of ambulance patients is close to zero, the arrival rate of ambulance patients to an ED is given approximately by $p\lambda_a$, where p is the routing probability to the ED. Ignoring the transit stage, the arrival process of ambulance patients to the ED can be approximated by a Poisson process with parameter $p\lambda_a$. This indicates that the ED part of a One-ED network can be approximated by an $M[2]/M/c$ non-preemptive priority queue, which is defined as follows:

- Ambulance patients arrive according to a Poisson process with parameter $p\lambda_a$.
- Walk-in patients arrive according to a Poisson process with parameter λ_w, which is independent of the ambulance patients' arrival process.
- The service time of a patient, regardless of its type (ambulance patient or walk-in patient), has an exponential distribution with parameter μ.
- There are c identical servers.
- Ambulance patients have non-preemptive service priority over walk-in patients. That is: if a server becomes available, it will first be assigned to an ambulance patient, if there is one. However, the service of a walk-in patient will not be interrupted by the arrivals of ambulance patients.

The $M[2]/M/c$ non-preemptive priority queue has been studied extensively and explicit formulas have been obtained for its mean waiting times and mean queue lengths. In this paper, we are interested in the waiting times of the two types of customers. Denote by $\widehat{W}_a$ the waiting time in the queue of an arbitrary ambulance patient (i.e., a high priority customer) and $\widehat{W}_w$ the waiting time in the queue of an arbitrary walk-in patient (i.e., a low priority customer). Let $\sigma=p\lambda_a/(c\mu)$ and $\rho=(p\lambda_a+\lambda_w)/(c\mu)=\sigma+\lambda_w/(c\mu)$. If $\rho<1$, it is well-known that (see Gross et al. [21]):

$$E[\widehat{W}_a]=\frac{1}{1-\sigma}\left(c!(1-\rho)c\mu\sum_{n=0}^{c-1}\frac{(c\rho)^{n-c}}{n!}+c\mu\right)^{-1};$$

$$E[\widehat{W}_w]=\frac{1}{(1-\sigma)(1-\rho)}\left(c!(1-\rho)c\mu\sum_{n=0}^{c-1}\frac{(c\rho)^{n-c}}{n!}+c\mu\right)^{-1}. \quad (1)$$

The mean queue lengths for both types of patients, denoted as $E[\widehat{q}_a]$ and $E[\widehat{q}_w]$, can be obtained by using Little's Law:

$$E[\widehat{q}_a]=p\lambda_a E[\widehat{W}_a] \quad \text{and} \quad E[\widehat{q}_w]=\lambda_w E[\widehat{W}_w]. \quad (2)$$

The mean waiting time of ambulance patients $E[\widehat{W}_a]$ is used to approximate the mean ambulance offload delay in the ED. The function $E[\widehat{W}_a]$ is also used in an optimization problem for the allocation of ambulance patients, where the following property plays an important role.

Lemma 1. *The function $E[\widehat{W}_a]$ is convex in σ for $0\le\sigma<1$.*

Proof. First note that the expression for $E[\widehat{W}_a]$ in Eq. (1) can be rewritten, in terms of the well-known Erlang delay formula $B(c,\rho)$, as

$$E[\widehat{W}_a]=\frac{1}{c\mu(1-\sigma)}B(c,\rho), \quad (3)$$

where

$$B(c,\rho)=\left(1+c!(1-\rho)\sum_{n=0}^{c-1}\frac{(c\rho)^{n-c}}{n!}\right)^{-1}. \quad (4)$$

By [22] and [23], $B(c,\rho)$ is convex in ρ. Since $\rho=\sigma+\lambda_w/(c\mu)$, it is easy to see that the Erlang delay formula is strictly increasing in ρ (then in σ). The first order derivative of $B(c,\rho)$ is positive. The second order derivative of [3] is given by:

$$\frac{d^2}{d\sigma^2}\left(\frac{B(c,\rho)}{(1-\sigma)}\right)=\frac{2B(c,\rho)}{(1-\sigma)^3}+2\frac{\frac{d}{d\sigma}B(c,\rho)}{(1-\sigma)^2}+\frac{\frac{d^2}{d\sigma^2}B(c,\rho)}{(1-\sigma)} \quad (5)$$

which is nonnegative. Consequently, $E[\widehat{W}_a]$ is convex in σ. This completes the proof of the lemma.□

Allocation of ambulance patients: We now use the explicit results above to optimize the allocation of ambulance patients by finding the best routing probabilities, with respect to offload delays. To obtain the offload delays of ambulances, all K EDs have to be considered simultaneously. Recall that the total arrival rate of ambulance patients to the K EDs is λ_a. Then the ambulance patients' arrival rate for the k-th ED is $p_k\lambda_a$. For the k-th ED, recall that (i) the number of servers is c_k; (ii) the service rate is μ_k; and (iii) the arrival rate of walk-in patients is $\lambda_{w,k}$. In Eq. (1), we add subscript k to variables $\widehat{W}_a$, μ, ρ, c, and p. Then an expression for the mean waiting time of ambulance patients (or the mean offload delay) can be obtained from Eq. (1). Consequently, the average offload delays of ambulances can be obtained as the weighted average: $\sum_{k=1}^{K} p_k E[\widehat{W}_{a,k}]$.

To minimize the average offload delay, we need to ensure that the mean offload delays in individual EDs are finite. To achieve that, we must ensure that

(a) the system has enough capacity to serve all ambulance patients who are not lost; and
(b) individual EDs have enough capacity to serve all ambulance patients arrived to them.

Condition (a) is equivalent to assuming that $\sigma_k<1$, where $\sigma_k=p_k\lambda_a/(c_k\mu_k)$, for $k=1,2,\ldots,K$. Let $p_k^{(\max)}=c_k\mu_k/\lambda_a$. Then condition (a) holds if $p_k<p_k^{\max}$. To ensure condition (b), we must have

$$\lambda_a<\sum_{k=1}^{K} c_k\mu_k. \tag{6}$$

If condition (6) holds, then the set of *feasible routing probabilities* $\{(p_1,\ldots,p_K): p_k<p_k^{(\max)}, k=1,\ldots,K, \sum_{k=1}^{K} p_k=1\}$ is nonempty.

Now, we are ready to propose the following optimization problem to find the routing probabilities $\{p_1,\ldots,p_K\}$ that minimize the average ambulance offload delays:

$$\min_{(p_1,\ldots,p_K)} \sum_{k=1}^{K} p_k E[\widehat{W}_{a,k}]=\sum_{k=1}^{K}\frac{p_k}{1-\sigma_k}\left(c_k!(1-\rho_k)c_k\mu_k\sum_{n=0}^{c_k-1}\frac{(c_k\rho_k)^{n-c_k}}{n!}+c_k\mu_k\right)^{-1}$$

$$s.t.\quad \sum_{k=1}^{K} p_k=1;$$

$$0\le p_k\le p_k^{(\max)},\quad \text{for}\quad k=1,2,\ldots,K. \tag{7}$$

where $\rho_k=\sigma_k+\lambda_{w,k}/(c_k\mu_k)$, for $k=1,2,\ldots,K$. In optimization problem (7), without loss of generality, the condition $p_k<p_k^{\max}$ is related to $p_k\le p_k^{\max}$. Under condition (6), an optimal solution to (7) exists.

By Lemma 1 and $\rho_k=\sigma_k+\lambda_{w,k}/(c_k\mu_k)$, for $k=1,\ldots,K$, the objective function in (7), which is separable, is convex in $\{\sigma_1,\ldots,\sigma_K\}$ and, consequently, convex in $\{p_1,\ldots,p_K\}$. The constraints in (7) are linear. Thus, the optimization problem is a convex program, which can be solved effectively by existing methods. For the numerical examples presented in Section 5, we solve the above optimization problem using the *fmincon* solver in Matlab where the *interior-point* algorithm is used.

4. The One-ED system and matrix-geometric solutions

The One-ED system is defined in exactly the same way as the EMS system with $K=1$ defined in Section 2 (See Fig. 2). For notational convenience, in this section, we remove subscript k from system parameters $\mu_{T,k}$, μ_k, c_k, p_k, $\lambda_{w,k}$, and other system variables. We still use parameter λ_a for the total arrival rate of ambulance patients, and $p\lambda_a$ for the arrival rate of ambulance patients to the One-ED system. To analyze the One-ED system, we define three state variables:

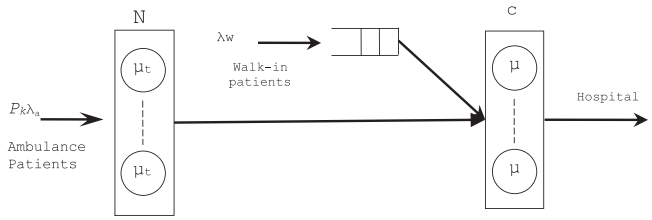

Fig. 2. One ED network.

$q_T(t)$: The number of ambulance patients in transit at time t.
$q_s(t)$: The number of both ambulance and walk-in patients in service and ambulance patients waiting for service at time t.
$q_w(t)$: The number of walk-in patients waiting for service at time t, if $q_s(t)\ge c$; otherwise (i.e., $q_s(t)<c$), $q_w(t)=-1$. We call $q_w(t)$ the queue length of waiting walk-in patients with the understanding that, if $q_w(t)=-1$, the queue length is zero.

It is easy to see that $\{(q_w(t),q_s(t),q_T(t)),t\ge 0\}$ is a continuous time Markov chain (CTMC). We call $q_w(t)$ the *level variable* and $\{q_s(t),q_T(t)\}$ the *phase variable*. Let Ω be the state space of the Markov chain $\{(q_w(t),q_s(t),q_T(t)),t\ge 0\}$. Let Ω_n be set of states in level n, i.e., the subset of Ω such that $q_w(t)=n$, for $n=-1,0,1,2,\ldots$.

$$\Omega_{-1}=\{(-1,i,j):\ 0\le i\le c-1,\ 0\le j\le N\};$$
$$\Omega_n=\{(n,i,j):\ c\le i\le c+N,\ 0\le j\le N, i-c+j\le N\},\quad \text{for } n=0,1,2,\ldots. \tag{8}$$

Then we have $\Omega=\Omega_{-1}\cup\Omega_0\cup\Omega_1\cup\ldots$. Note that levels $n=0,1,2,\ldots$ have the same number of states. The infinitesimal generator of $\{(q_w(t),q_s(t),q_T(t)),t\ge 0\}$ can be given as

$$Q=\begin{matrix}-1\\0\\1\\\vdots\end{matrix}\begin{pmatrix}A_{-1,-1} & A_{-1,0} & & & \\ A_{0,-1} & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots\end{pmatrix}, \tag{9}$$

where the transition blocks $\{A_{-1,-1},A_{-1,0},A_{0,-1},A_0,A_1,A_2\}$ are given in Appendix A. We note that $A_{-1,-1}$ and A_1 describe transitions within each level (i.e., transitions between states in Ω_n), $A_{0,-1}$ and A_2 for transitions from level n to level $n-1$, and $A_{-1,0}$ and A_0 for transitions from level n to level $n+1$. The transition blocks satisfy equalities: $A_{-1,-1}\mathbf{e}+A_{-1,0}\mathbf{e}=0, A_{0,-1}\mathbf{e}+(A_1+A_0)\mathbf{e}=0$, and $(A_2+A_1+A_0)\mathbf{e}=0$, where $\mathbf{e}$ is a column vector of ones.

First, we find the stationary distribution of $\{(q_w(t),q_s(t),q_T(t)),t\ge 0\}$. Let $\boldsymbol{\pi}=(\boldsymbol{\pi}_{-1},\boldsymbol{\pi}_0,\boldsymbol{\pi}_1,\ldots)$ be the stationary probability distribution of $\{(q_w(t),q_s(t),q_T(t)),t\ge 0\}$, where $\boldsymbol{\pi}_n$ includes all the limiting probabilities of the states in level Ω_n. Let $A=A_0+A_1+A_2$. Matrix A is an irreducible infinitesimal generator. Let $\boldsymbol{\theta}$ satisfy $\boldsymbol{\theta}A=0$ and $\boldsymbol{\theta}\mathbf{e}=1$. 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

$$\boldsymbol{\theta}A_0\mathbf{e}<\boldsymbol{\theta}A_2\mathbf{e}, \tag{10}$$

which is guaranteed under $\lambda_w+p\lambda_a<c\mu$. If the ergodicity condition is satisfied, then $\boldsymbol{\pi}$ exists and it is the unique non-negative solution for the linear system: $\boldsymbol{\pi}Q=0$ and $\boldsymbol{\pi}\mathbf{e}=1$. Since the infinitesimal generator Q has a level independent QBD structure, a matrix-geometric solution can be obtained for its stationary distribution:

$$\boldsymbol{\pi}_n=\boldsymbol{\pi}_0R^n,\quad \text{for } n\ge 0 \tag{11}$$

where R is called the rate matrix and is the minimal nonnegative solution to nonlinear matrix equation:

$$A_0 + RA_1 + R^2A_2 = 0, \tag{12}$$

and boundary probabilities $(\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0)$ can be obtained by solving the linear system

$$(\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0)\begin{pmatrix} A_{-1,-1} & A_{-1,0} \\ A_{0,-1} & A_1 + RA_2 \end{pmatrix} = 0;$$
$$\boldsymbol{\pi}_{-1}\mathbf{e} + \boldsymbol{\pi}_0(I-R)^{-1}\mathbf{e} = 1, \tag{13}$$

where I is the unit matrix.

Next, we use the matrix geometric solution $\{\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0, R\}$ to find performance measures at the ED level. Performance measures to be used in Section 5 are presented in this section. Some additional performance measures are presented in Appendix B.

1. *ED utilization*: The maximum arrival rate to the ED node is $\lambda_w + p\lambda_a$ and the available capacity for those arrivals is $c\mu$. As a result, the ED utilization, ρ, can be defined as follows:

$$\rho = \frac{\lambda_w + p\lambda_a}{c\mu} \tag{14}$$

2. *Loss probability of ambulance patients*: Let p_{loss} be the probability that no ambulance is available when an ambulance patient arrives (i.e., the probability that the ambulance patient is lost). An ambulance patient is lost if every ambulance is either in transit or waiting in the ED. Then we have

$$p_{loss} = \sum_{i=0}^{c-1} (\boldsymbol{\pi}_{-1})_{(i,N)} + \sum_{i=c}^{N+c} (\boldsymbol{\pi}_0(I-R)^{-1})_{(i,N+c-i)}. \tag{15}$$

3. *Distribution of the number of walk-in patients waiting in the queue (related to $q_w(t)$)*: Let q_w be the number of walk-in patients waiting in the queue in steady state. It is easy to see that the distribution of q_w is given by $\{\boldsymbol{\pi}_{-1}\mathbf{e} + \boldsymbol{\pi}_0\mathbf{e}, \boldsymbol{\pi}_n\mathbf{e}, n = 1, 2, \ldots\}$. The mean number of walk-in patients in the queue can be calculated by

$$E[q_w] = \sum_{n=0}^{\infty} n\boldsymbol{\pi}_n\mathbf{e} = \sum_{n=1}^{\infty} n\boldsymbol{\pi}_0R^n\mathbf{e} = \boldsymbol{\pi}_0\left(\sum_{n=1}^{\infty} nR^n\right)\mathbf{e} = \boldsymbol{\pi}_0R(I-R)^{-2}\mathbf{e}. \tag{16}$$

4. *Distribution for the number of ambulances in offload delay at the ED*: Recall that $q_s(t)$ is defined as the number of both ambulance patients and walk-in patients in service and blocked. If $q_s(t)$ is greater than c, then there are $q_s(t)-c$ ambulances blocked outside the ED. Then $q_a = \max\{0, q_s(t)-c\}$ is the number of ambulances in offload delay, defined for the steady state, and its probability distribution $\boldsymbol{\xi} = (\xi(0), \xi(1), \ldots, \xi(N))$ can be found as:

$$\xi(m) = \begin{cases} \sum_{j=0}^{j=c} \eta(j), & \text{for } m = 0 \\ \eta(m+c), & \text{for } m = 1, \ldots, N \end{cases} \tag{17}$$

where $\eta(j)$ is defined in Appendix B. The mean queue number of ambulances in offload delay can be found by $E[q_a] = \sum_{m=0}^{N} m\xi(m)$.

5. *Offload delay distribution*: Offload delays (or waiting time) w_a of an ambulance patient arriving at an ED depends on the number of ambulance patients at the ED and the number of walk-in patients already in service which are captured in the state variable $q_s(t)$. Note that $\omega(j)$, which is defined in Appendix B, the probability that j patients are waiting in the ED right after an ambulance patient arrives to the ED, for $j = 1, 2, \ldots, N$. Since there are c beds for all patients in the ED, each with an exponential service time with parameter μ, if all beds are occupied, the time to serve one patient has an exponential distribution with parameter $c\mu$. Thus, the total time to serve j patients has an Erlang distribution of order j. Consequently, when an ambulance patient arrives to the hospital ED, the waiting time has a generalized Erlang distribution with a *PH*-representation $(\boldsymbol{\alpha}, T_a)$ of order N, where $\boldsymbol{\alpha} = (\omega(1), \omega(2), \ldots, \omega(N))$, and

$$T_a = c\mu\begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}_{N\times N}. \tag{18}$$

The distribution function of the waiting time W_a is given by $P\{W_a < t\} = 1 - \boldsymbol{\alpha}\exp\{T_at\}\mathbf{e}$. The expected offload delays can be found using the formula:

$$E[W_a] = \frac{1}{c\mu}\sum_{j=1}^{N} j\omega(j) \tag{19}$$

Note here that Little's Law holds for the queueing model, i.e., $E[q_a] = p\lambda_a(1-p_{loss})E[W_a]$, which can also be used to check the accuracy of the results obtained.

6. *Waiting time distribution for the walk-in patients*: The waiting time W_w of a walk-in patient arriving at an ED depends on the number of ambulance patients in service and waiting outside the ED. It also depends on the number of walk-in patients already in service that is captured in the state variable $q_s(t)$. Since walk-in patients have a lower priority than patients arriving by ambulance, a walk-in patient cannot get admission unless there is a bed available for him, i.e., $q_s(t) < c$. According to Ozawa [24], the waiting time W_w of an arbitrary walk-in patient has a phase-type distribution. By taking a different approach from that in Ozawa [24], we find an explicit *PH*-representation $(\boldsymbol{\alpha}_w, T_w)$ for W_w, where

$$\boldsymbol{\alpha}_w = \phi(I)(\boldsymbol{\pi}_0(I-R)^{-1}) \otimes \mathbf{e}';$$
$$T_w = \Lambda^{-1}(I \otimes (A_1 + A_0) + R' \otimes A_2)\Lambda, \tag{20}$$

where $\phi(I)$ is the direct-sum of I (i.e., a row vector obtained by stringing out all rows in I starting from the first row), $\mathbf{e}'$ and R' are the transposes of $\mathbf{e}$ and R, respectively, and Λ is a matrix with elements of the vector $(\boldsymbol{\pi}_0(I-R)^{-1}) \otimes \mathbf{e}'$ on its diagonals and all other elements are zero. A proof of (20) is given in Appendix C.

The mean waiting time is given by $E[W_w] = -\boldsymbol{\alpha}_w(T_w)^{-1}\mathbf{e}$. According to Little's Law, we must have $E[q_w] = \lambda_wE[W_w]$, which is useful for checking accuracy of computation. The sojourn time of a walk-in patient is the sum of the its waiting time and its service time. Since both the waiting time and the service time are phase-type random variables and are independent, the sojourn time of a walk-in patient has a phase-type distribution as well. Details are omitted.

5. Numerical analysis of the EMS system

In Section 3, the ambulance offload delay is denoted as $\widehat{W}_a$ for the $M[2]/M/c$ queue. In Section 4, the ambulance offload delay is denoted as W_a for the One-ED system. In this section, the ambulance offload delay from a simulation model is denoted as $\tilde{W}_a$. The same convention is applied to other quantities of interest $\{p_{loss}, q_a, q_w, W_w, W_{w,k}, W_{a,k}\}$. Due to Little's Law between the mean queue length and mean waiting time, we shall present results for the waiting times only. In Section 5.1, we validate the approximations developed in Sections 3 and 4 through simulation. For that purpose, we develop two examples; Example 5.1 represents a

Table 1
Approximations and simulation results for the first ED.

λ_a	$\tilde{p}_{loss}$	$\hat{\sigma}_1$	$\hat{\rho}_1$	$E[W_{a,1}]$	$E[\widehat{W}_{a,1}]$	$E[\tilde{W}_{a,1}]$	$E[W_{w,1}]$	$E[\widehat{W}_{w,1}]$	$E[\tilde{W}_{w,1}]$
0.5	0.0001	0.0670	0.6920	0.1461	0.1444	0.1477	0.4790	0.4688	0.4808
1.0	0.0018	0.1340	0.7590	0.2125	0.2109	0.2158	0.9069	0.8752	0.9153
1.5	0.0093	0.2011	0.8261	0.2931	0.2983	0.3008	1.7616	1.7148	1.7915
2.0	0.0345	0.2681	0.8931	0.3855	0.4119	0.3954	3.7189	3.8523	3.7267

Table 2
Percent difference results for the first ED.

λ_a	$\Delta E[\widehat{W}_{a,1}]$ (%)	$\Delta E[\tilde{W}_{a,1}]$ (%)	$\Delta E[\widehat{W}_{w,1}]$ (%)	$\Delta E[\tilde{W}_{w,1}]$ (%)
0.5	−1.16	1.10	−2.13	0.38
1.0	−0.75	1.55	−3.50	0.93
1.5	1.77	2.63	−2.66	1.70
2.0	6.85	2.57	3.59	0.21

network that consists of three hospitals. We focus on one hospital and vary the ambulance patients' arrival rate to observe the validity of the approximation with the increased utilization on the network. Example 5.2 represents a network of five hospitals. In Section 5.2, we use the approximations to conduct a case study for a regional EMS system that serves three hospitals. In Section 5.3, we extend the case study by performing some sensitivity analysis.

5.1. Model validation through simulation

Example 5.1. Assume that $K=3$. We focus on the first ED (i.e., $k=1$) and look at the mean waiting times of both ambulance patients and walk-in patients obtained by the three methods: The $M[2]/M/c$ queue, the One-ED model, and simulation. Then we only need parameters for the first ED: $N=5$, $p_1=1/3, \mu_{T,1}=2$, $\lambda_w=1.5, c_1=6$, and $\mu_1=0.4$. For $\lambda_a=0.5, 1, 1.5, 2$, we compute the loss probability, mean waiting times and other quantities and present them in Table 1. In Table 2, we calculate the percent difference between approximation and simulation results versus exact results. For example, $\Delta E[\tilde{W}_{a,1}]=\frac{E[\tilde{W}_{a,1}]-E[W_{a,1}]}{E[W_{a,1}]}*100\%$.

Tables 1 and 2 demonstrate that

- When the ED utilization is low to medium, as represented by λ_a between 0 and 1.5 patients per hour, then the approximation results are within 5% of the exact results for both ambulance and walk-in patients.
- When the ED utilization is high, as represented by λ_a from 1.5 to 2.0, the approximations are different from the exact results for ambulance patients only. The main reason for this deviation is due to the loss probability. However, walk-in patients results are still within 5% of the exact results.

Thus, the two approximations can provide good estimates of system performance measures if the loss probability of ambulance patients is small. In addition, Table 1 shows clearly that:

- The two approximation approaches provide consistent estimates of performance measures. Intuitively, it should be true as long as the loss probability of ambulance patients, p_{loss}, for the One-ED model is not significant.
- When σ becomes large, the loss probability $\tilde{p}_{loss}$ becomes significant. That leads to a significant difference between simulation results and the two approximations. Based on our numerical experiments, if the loss of ambulance patients is considered properly in the $M[2]/M/c$ queue, the approximation can be close to simulation results. For the One-ED approximation, if the loss of ambulance patients is transformed properly into the loss of ambulances in the One-ED model, then the approximation can be close to the simulation results. How to properly adjust the number of available ambulances and how to adjust the ambulance patient arrival rate are two interesting issues for future research.

Table 3
The 95% confidence intervals for the mean waiting times for Example 5.1.

λ_a	$E[\tilde{W}_{a,1}]$ (lower, upper)	$E[\tilde{W}_{w,1}]$ (lower, upper)
0.5	0.1477 (0.1467, 0.1487)	0.4808 (0.4758, 0.4858)
1.0	0.2158 (0.2145, 0.2170)	0.9153 (0.9040, 0.9265)
1.5	0.3008 (0.2990, 0.3026)	1.7915 (1.7666, 1.8163)
2.0	0.3954 (0.3929, 0.3979)	3.7267 (3.6516, 3.8017)

Table 4
System parameters for Example 5.2.

k	p_k	$\mu_{T,k}$	$\lambda_{w,k}$	c_k	μ_k
1	0.1	1.5	1	5	0.5
2	0.2	1.5	1.5	4	1
3	0.2	2	2.5	4	1
4	0.25	2	2	3	2
5	0.25	2	1.5	5	1

The simulation results with the 95% confidence intervals of the mean waiting times (with 50 replications) are presented in Table 3.

Example 5.2. Assume that $K=5$, $N=10$ and $\lambda_a=6.5$. Other system parameters are given in Table 4. By simulation, we have $\tilde{p}_{loss}=0.0139$. Other performance measures are presented in Table 5.

Table 5 indicates that the approximation works well for a system with $\hat{\sigma}_k$ as large as 0.3. The traffic intensity $\hat{\rho}_k$ can be as large as 0.95. The 95% confidence intervals of the simulated mean waiting times (with 50 replications) are shown in Table 6.

5.2. Optimization of the allocation of ambulance patients

In this subsection, we apply the framework developed in Sections 3 and 4 to analyze a case study that reflects an EMS system that serves a region in southwestern Ontario, Canada. The EMS system considered serves three local hospitals (i.e., $K=3$). The input parameters considered for this case study are obtained or estimated by using available data.

- The number of ambulances available in the region is $N=17$.
- The arrival rate of ambulance patients is $\lambda_a=3$ per hour.
- The arrival rates of walk-in patients $(\lambda_{w,1}, \lambda_{w,2}, \lambda_{w,3})$ are given in Table 7.

- The ambulance transit rates are $\mu_{T,1}=\mu_{T,2}=\mu_{T,3}=2$ per hour. That is: the average transit time of ambulance patients to EDs is half an hour.
- The number of servers at individual EDs (c_1,c_2,c_3) are given in Table 7. The number of servers at each ED reflects the number of beds at the ED.
- The routing probabilities (p_1,p_2,p_3) are given in Table 7. The routing probability for each ED is calculated as the percentage of patients who were transferred to that ED with respect to the total number of patients who traveled by an ambulance for the same period of time.
- The service rates at EDs (μ_1,μ_2,μ_3) are given in Table 7. We use the reciprocal of the Length Of Stay (LOS) of patients in each hospital ED to find the average service rate.

We first calculate the estimated ED utilization and average waiting times for both types of patients based on the current routing probabilities. The results are shown in Table 7. Then, we use our optimization problem in (7) to find the optimal routing probabilities. We also compute the estimated ED utilization and average waiting times for patients in individual EDs based on the optimal routing decision and record the results in Table 8.

Based on the current routing probabilities (45%, 29%, 26%) the total expected offload delays (i.e., $\sum_{k=1}^{3} p_k E[\widehat{W}_{a,k}]$) are 0.1105 h. While under the optimal routing probabilities (36.27%, 30.78%, 32.95%) the average expected offload delays are 0.0973 h. This corresponds to a 11.9% decrease in the total offload delays experienced by ambulances in the region. Although the main purpose of the optimization problem is to mitigate the overall offload delays experienced by EMS, we notice that the average expected waiting time for walk-in patients (i.e., $\sum_{k=1}^{3}\lambda_{w,k}E[\widehat{W}_{w,k}]/(\sum_{n=1}^{3}\lambda_{w,n})$) has decreased significantly by 42.5% when the optimal routing decision is used.

Although the objective of this paper is to find proper allocation of ambulance patients in the long run, two simulation studies are carried out to compare the optimal policy with two other possible policies: ED capacity contribution policy and a state-dependent policy.

1. *ED capacity contribution policy*: In this policy, patients are routed to EDs based on the ED capacity contribution to the regional capacity. The routing probability is derived by setting $p_k=\frac{c_k}{\sum_{k=1}^{K}c_k}$. When this policy is applied for the EMS system defined in this case study, we get $(p_1,p_2,p_3)=(0.3723,0.3191,0.3085)$. As a result, the mean waiting times of ambulance patients are estimated by simulation as $(E[\tilde{W}_{a,1}],E[\tilde{W}_{a,2}],E[\tilde{W}_{a,3}])=(0.0997,0.1095,0.0842)$ and the mean waiting times of walk-in patients are $(E[\tilde{W}_{w,1}],E[\tilde{W}_{w,2}],E[\tilde{W}_{w,3}])=(1.0454,1.1013,0.5637)$. The average waiting time of an arbitrary patient is 0.0981 for an ambulance patient, and 0.9146 for a walk-in patient. Comparing those results to the optimal results recorded in Table 8, we notice that the optimal routing policy slightly outperforms the simple policy suggested here in terms of ambulance patients waiting and walk-in patients waiting times. We expect the superiority of our solution to increase as the gap between the walk-in patients arrival rate to respective EDs capacity increases. Consider for example a similar setting to the case study above but with different walk-in patients arrival rates, $\lambda_{w,k}=(3.4,3.7,3.8)$. Based on our model, the optimal allocation policy is $p_k^*=$ (0.5282, 0.2639, 0.2080). This results

Table 5
Approximation and simulation results for individual EDs.

k	$\widehat{\sigma}_k$	$\widehat{\rho}_k$	$E[W_{a,k}]$	$E[\widehat{W}_{a,k}]$	$E[\tilde{W}_{a,k}]$	$E[W_{w,k}]$	$E[\widehat{W}_{w,k}]$	$E[\tilde{W}_{w,k}]$
1	0.2600	0.6600	0.1713	0.1713	0.1628	0.5038	0.5038	0.4776
2	0.3250	0.7000	0.1587	0.1588	0.1510	0.5290	0.5292	0.4968
3	0.3250	0.9500	0.3301	0.3302	0.3292	6.6002	6.6031	5.8709
4	0.2708	0.6042	0.0823	0.0823	0.0795	0.2079	0.2079	0.2001
5	0.3250	0.6250	0.0794	0.0795	0.0746	0.2119	0.2119	0.1979

Table 6
The 95% confidence intervals for the mean waiting times for Example 5.2.

k	$E[\tilde{W}_{a,1}]$ (lower, upper)	$E[\tilde{W}_{w,1}]$ (lower, upper)
1	0.1628 (0.1608, 0.1648)	0.4776 (0.4691, 0.4862)
2	0.1510 (0.1497, 0.1523)	0.4968 (0.4904, 0.5032)
3	0.3292 (0.3177, 0.3207)	5.8709 (5.6550, 6.0869)
4	0.0795 (0.0791, 0.0799)	0.2001 (0.1985, 0.2017)
5	0.0746 (0.0740, 0.0752)	0.1979 (0.1958, 0.2001)

Table 7
Case study input parameters and performance measures results.

k	$\lambda_{w,k}$ (patient/hr)	c_k	μ_k	p_k (%)	$\widehat{\rho}_k$	$\widehat{\sigma}_k$	$E[\widehat{W}_{a,k}]$	$E[\widehat{W}_{w,k}]$
1	4.2	35	1/6	45.00	0.9514	0.2314	0.1542	3.1738
2	3.5	30	1/6	29.00	0.8740	0.1740	0.0904	0.7178
3	3.2	29	1/6	26.00	0.8234	0.1614	0.0572	0.3242

Table 8
Case study optimization and approximation results.

k	p_k^* (%)	$\widehat{\rho}_k$	$\widehat{\sigma}_k$	$E[\widehat{W}_{a,k}]$	$E[\widehat{W}_{w,k}]$	$E[\tilde{W}_{a,k}]$(lower, upper)	$E[\tilde{W}_{w,k}]$(lower, upper)
1	36.27	0.9065	0.1865	0.0985	1.0540	0.0986 (0.0975, 0.0997)	1.0468(1.0209, 1.0727)
2	30.78	0.8847	0.1847	0.1010	0.8761	0.1016(0.1006, 0.1025)	0.8845(0.8609, 0.9081)
3	32.95	0.8666	0.2045	0.0925	0.6936	0.0920 (0.0908, 0.0932)	0.6899(0.6695, 0.7103)

Table 9
Effect of ambulance patients' arrival rate on optimal decisions.

λ_a	p_1^* (%)	p_2^* (%)	p_3^* (%)	ρ_1 (%)	ρ_2 (%)	ρ_3 (%)	σ_1 (%)	σ_2 (%)	σ_3 (%)	$E[\widehat{W}_a]$	$E[\widehat{W}_w]$
1.0	35.40	28.29	36.31	78.07	75.66	73.72	6.07	5.66	7.51	0.0198	0.0842
1.5	35.79	29.49	34.72	81.20	78.85	76.98	9.20	8.85	10.77	0.0312	0.1524
2.0	36.01	30.12	33.87	84.35	82.05	80.22	12.35	12.05	14.01	0.0471	0.272
2.5	36.16	30.51	33.33	87.50	85.26	83.45	15.50	15.26	17.24	0.0688	0.4863
3.0	36.27	30.78	32.95	90.65	88.47	86.66	18.65	18.47	20.45	0.0973	0.8911
3.5	36.36	30.98	32.66	93.82	91.69	89.86	21.82	21.69	23.65	0.1342	1.7522
4.0	36.44	31.14	31.13	96.98	94.91	93.05	24.98	24.91	26.84	0.1807	4.2408

in $(E[\tilde{W}_{a,1}], E[\tilde{W}_{a,2}], E[\tilde{W}_{a,3}]) = (0.0657, 0.1072, 0.1190)$. The average waiting time of an arbitrary patient is 0.0878 for an ambulance patient. While based on the above ED capacity contribution policy, which does not change when we change the walk-in patients arrival rates, the estimated ambulance patients waiting times are $(E[\tilde{W}_{a,1}], E[\tilde{W}_{a,2}], E[\tilde{W}_{a,3}]) = (0.0176, 0.1429, 0.2172)$. This corresponds to an average waiting time of an arbitrary patient equal to 0.1192. In both of the above scenarios, the optimal allocation policy outperforms the ED capacity contribution policy. In the first scenario, the percentage of improvement was 0.82%. In the second scenario the improvement was 26.34% on the total offload delays.

2. *A state-dependent policy*: The state-dependent policy is developed for the routing of ambulance patients: when an ambulance patient arrives and there is an ambulance available, the patient will be sent to the ED with the minimum number of ambulances in the ED or in transit to the ED; if there is a tie, then the ambulance patient is sent to the ED with the biggest number of servers (i.e., the ED which has the highest service capacity). When the state-dependent policy is applied to the EMS system defined in the case study, the mean waiting times of ambulance patients are estimated by simulation as $(E[\tilde{W}_{a,1}], E[\tilde{W}_{a,2}], E[\tilde{W}_{a,3}]) = (0.1533, 0.0735, 0.0242)$ and the mean waiting times of walk-in patients are $(E[\tilde{W}_{w,1}], E[\tilde{W}_{w,2}], E[\tilde{W}_{w,3}]) = (2.9455, 0.5130, 0.1145)$. Compared to the results presented in Table 8, the mean waiting time at ED 1 is higher for the system with the state-dependent policy. The percentages of ambulance patients sent to EDs are $(p_1, p_2, p_3) = (0.4715, 0.3283, 0.2001)$. Then the average waiting time of an arbitrary ambulance patient can be obtained from (p_1, p_2, p_3) and $(E[\tilde{W}_{a,1}], E[\tilde{W}_{a,2}], E[\tilde{W}_{a,3}])$ as 0.1019, which is greater then the estimated average waiting time of an arbitrary customer under the optimal routing probabilities (p_1^*, p_2^*, p_3^*), which is 0.0963 (see Table 8).

Table 10
Simulation results for the case study.

λ_a	$E[\tilde{W}_a]$(lower, upper)	$E[\tilde{W}_w]$(upper, lower)
1.0	0.0197 (0.0190, 0.0204)	0.0835 (0.0793, 0.0876)
1.5	0.0315 (0.0307, 0.0327)	0.1535 (0.1468, 0.1602)
2.0	0.0464 (0.0451, 0.0476)	0.2647 (0.2511, 0.2782)
2.5	0.0694 (0.0676, 0.0712)	0.4915 (0.4646, 0.5184)
3.0	0.0962 (0.0944, 0.0979)	0.8641 (0.8147, 0.9135)
3.5	0.1352 (0.1326, 0.1378)	1.7626 (1.6634, 1.8618)
4.0	0.1802 (0.1790, 0.1813)	4.0501 (3.9070, 4.1931)

Another observation we have from this case study is related to resource pooling. We observe that larger EDs should be loaded more heavily than smaller EDs. For example, if we consider ED1 which has the largest capacity, we notice that it should be loaded more heavily (highest utilization of 90.65%) because it has the highest capacity in terms of the number of beds. Next, we perform some sensitivity analysis on the results of the case study.

5.3. Sensitivity analysis

In this subsection, we increase the ambulance patients' arrival rate gradually and observe its effect on the optimal routing decisions (p_1^*, p_2^*, p_3^*) and total expected waiting time for both ambulance patients and walk-in patients. We use the case study of Section 5.2 as the base scenario. The results are reported in Table 9, while the simulation results are shown in Table 10. From these results we draw the following observations:

- As the high priority load increases on the EMS-ED network, as represented by increased λ_a, the expected delays of patients increase in all regional hospitals. However, walk-in patients experience severe consequences of this increased load as we can see from Fig. 3.
- As the ambulance patients arrival rate increases, we notice that it is optimal to send a higher percentage of those arrivals to the larger EDs (ED 1 and ED 2) and less to the smaller ED (ED 3) (see Fig. 4). As λ_a increases, p_2 and p_1 increase, while p_3 decreases.

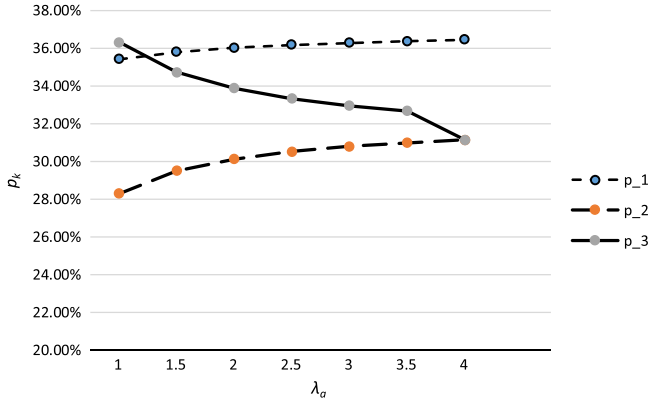

Fig. 4. Effect of ambulance patients' arrival rate on allocation probabilities (p_1, p_2, p_3).

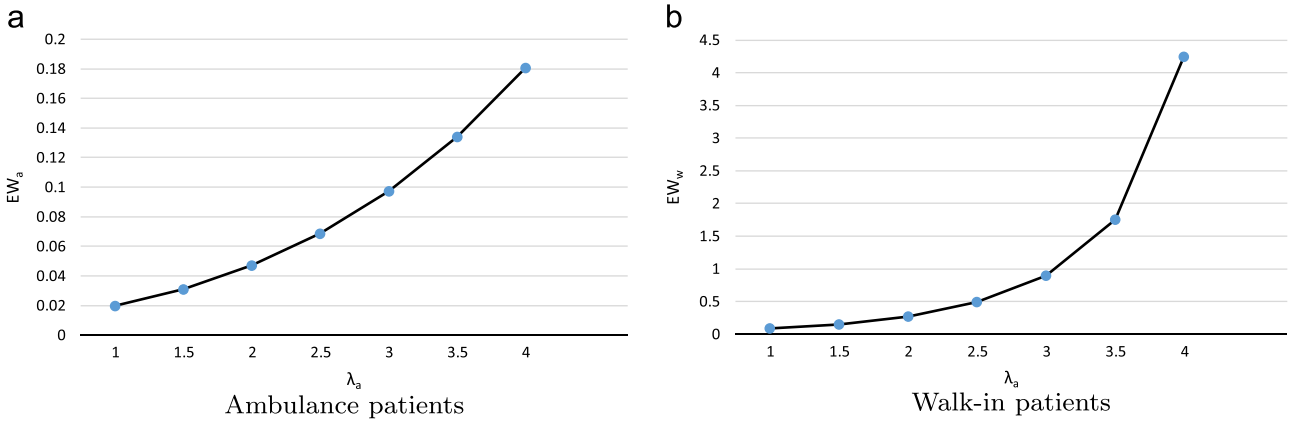

Fig. 3. Effect of ambulance patients' arrival rate on total expected waiting times.

- As the load increases on the EMS-ED system, the deviation between the approximation results and the simulation results increases. While this is true for both arrival streams, we notice that the model results for the ambulance patients are closer than the walk-in patients results. This is because σ_k for $k=1,2,3$ is low (less than 30% for all EDs).

6. Conclusions

In summary, the optimization model in this paper is robust under normal operating conditions as supported by the numerical analysis. Our model results can be used to guide EMS dispatchers on how to allocate patients to hospital EDs by using those allocation probabilities as targets that they aim to achieve in the long run when they make their dispatching decisions. However, it does not limit their capability to deciding where to send the next patient. In that sense, we see this model as a decision support tool. It can also be used to evaluate the consequences between various policy alternatives, e.g. comparing the effect of adding or decreasing capacity in a hospital ED on the total offload delays experienced by EMS. Or, the effect of a given allocation policy on the total offload delays experienced in a region. It can also be used to determine the capacity requirements needed to achieve certain goals, e.g. how many beds are needed to decrease the offload delays by 20% in the next year.

Another finding of this work is related to the effect of resource pooling and work consolidation on offload delays. Larger EDs can mitigate the effect of high utilization on both walk-in patients and ambulance patients. However, the higher priority ambulance patients benefit more from those larger EDs.

The decomposed model and the solution methodology developed in this paper perform well under normal operating conditions (low EMS utilization). When we compare the optimization model results with a state-dependent routing policy, the model developed in this paper gave better results. In future research, we are interested in developing an iterative scheme that can be used to adjust the decomposed queueing model parameters when the loss probability is high. The results can be useful in other application areas such as telecommunication systems.

Appendix A

In this appendix, we present the blocks of the infinitesimal generator Q defined in Section 4. First, we define three sets of matrices used in the construction of the transition blocks in Q. For $i=0,1,\ldots,N+c$, let

$$a_i = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N-(i-c)^+-1 \\ N-(i-c)^+ \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & \ldots & N-(i-c)^+ \end{array} \\ \begin{pmatrix} * & p\lambda_a & & & \\ & * & p\lambda_a & & \\ & & * & p\lambda_a & \\ & & & \ddots & \ddots \\ & & & & * & p\lambda_a \\ & & & & & * \end{pmatrix}_{(N-(i-c)^+)\times(N-(i-c)^+)} \end{array}, \tag{21}$$

where $*$ is calculated such that the rows of the matrix Q sum to zero: $(a_i)_{(j,j)} = -\lambda_w - \min(i,c)\mu - p\lambda_a - j\mu_0$, for $j=0,1,\ldots,N-(i-c)^+-1$, and $(a_i)_{(N-(i-c)^+,N-(i-c)^+)} = -\lambda_w - \min(i,c)\mu - (N-(i-c)^+)\mu_0)$. Let, for $i=0,1,\ldots,c-1$,

$$b_i = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & \ldots & & N \end{array} \\ \begin{pmatrix} \lambda_w & & & & & \\ \mu_T & \lambda_w & & & & \\ & 2\mu_T & \lambda_w & & & \\ & & \ddots & \ddots & & \\ & & & & N\mu_T & \lambda_w \end{pmatrix}_{N\times N} \end{array}, \tag{22}$$

and, for $i=c,c+1,\ldots,N+c-1$,

$$b_i = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N+c-i-1 \\ N+c-i \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & \cdots & & N-(i-c)-1 \end{array} \\ \begin{pmatrix} 0 & & & & & \\ \mu_T & 0 & & & & \\ & 2\mu_T & 0 & & & \\ & & \ddots & \ddots & & \\ & & & (N-(i-c)-1)\mu_T & 0 & \\ & & & & (N-(i-c))\mu_T \end{pmatrix}_{(N+c-i)\times(N+c-i)} \end{array}, \tag{23}$$

The third set of matrices are:

$$d_i = i\mu I_{N+1-(i-c)^+}, \quad \text{for } 1 \le i \le c$$

$$d_i = c\mu(I_{N+1-(i-c)}\ 0), \quad \text{for } c+1 \le i \le N+c. \tag{24}$$

The blocks $A_{-1,-1}, A_{-1,0}$, and $A_{0,-1}$ for boundary transitions are given as:

$$A_{-1,-1} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ c-2 \\ c-1 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \ldots & c-2 & c-1 \end{array} \\ \begin{pmatrix} a_0 & b_0 & & & & \\ d_1 & a_1 & b_1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & d_{c-2} & a_{c-2} & b_{c-2} \\ & & & & d_{c-1} & a_{c-1} \end{pmatrix} \end{array};$$

$$A_{0,-1} = \begin{array}{c} \\ c \\ c+1 \\ \vdots \\ N+c \end{array} \begin{array}{c} \begin{array}{cccc} 0 & \ldots & c-2 & c-1 \end{array} \\ \begin{pmatrix} 0 & \ldots & 0 & d_c \\ 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots\vdots & \\ 0 & \ldots & 0 & 0 \end{pmatrix} \end{array};$$

$$A_{-1,0} = \begin{array}{c} \\ 0 \\ \vdots \\ c-2 \\ c-1 \end{array} \begin{array}{c} \begin{array}{cccc} c & c+1 & \ldots & N+c \end{array} \\ \begin{pmatrix} 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 \\ b_{c-1} & 0 & \ldots & 0 \end{pmatrix} \end{array}, \tag{25}$$

The rate at which the number of walk-in patients increase is defined in the matrix A_0. We note that the queue size increases by one only when the beds at the destination ED are full; otherwise it does not change. Then A_0 is

$$A_0 = \begin{array}{c} \\ c \\ c+1 \\ \vdots \\ N+c \end{array} \begin{array}{c} \begin{array}{ccccc} c & & c+1 & \ldots & N+c \end{array} \\ \begin{pmatrix} \lambda_w I_{N+1} & & & & \\ & \lambda_w I_N & & & \\ & & \ddots & & \\ & & & & \lambda_w I_1 \end{pmatrix} \end{array}, \tag{26}$$

where I_n is the unit matrix of order n. Matrix A_1 includes transitions that do not affect the walk-in patients queue length; it includes service completions of ambulances, service completions of ambulance patients, and ambulance patients arrival to the EMS

service. The details of A_1 are given as follows:

$$A_1 = \begin{array}{c} \\ c \\ c+1 \\ \vdots \\ N+c-1 \\ N+c \end{array} \begin{array}{c} \begin{array}{ccccc} c & c+1 & \cdots & N+c-1 & N+c \end{array} \\ \begin{pmatrix} a_c & b_c & & & \\ d_{c+1} & a_{c+1} & b_{c+1} & & \\ & \ddots & \ddots & \ddots & \\ & & d_{c+N-1} & a_{c+N-1} & b_{c+N-1} \\ & & & d_{c+N} & a_{c+N} \end{pmatrix} \end{array} \tag{27}$$

The matrix A_2 represents the rate at which the walk-in patients queue decreases by one. Because those patients possess lower priority than patients arriving by an ambulance, a walk-in patient cannot be admitted unless there are no patients of the higher priority waiting for a bed, or simply when $q_2(t) > c$. The details of A_2 are given as follows:

$$A_2 = \begin{array}{c} \\ c \\ c+1 \\ \vdots \\ N+c \end{array} \begin{array}{c} \begin{array}{cccc} c & c+1 & \cdots & N+c \end{array} \\ \begin{pmatrix} c\mu I_{N+1} & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} \end{array}. \tag{28}$$

Appendix B

In this appendix, additional performance measures for the One-ED model considered in Section 4 are presented.

1. *Distribution of the sum of the number of patients in service and the number of waiting ambulance patients* $q_s(t)$: The distribution can be obtained from the two vectors $\{\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0(I-R)^{-1}\}$. Let $\{\eta(i), 0 \le i \le c+N\}$ be the distribution of $q_s(t)$. Then

$$\eta(i) = \begin{cases} \sum_{j=0}^{N} (\boldsymbol{\pi}_{-1})_{(i,j)}, & \text{for } i = 0, 1, \ldots, c-1; \\ \sum_{j=0}^{N} (\boldsymbol{\pi}_{-1})_{(i,j)} + \sum_{j=0: i-c+j \le N}^{N} (\boldsymbol{\pi}_0(I-R)^{-1})_{(i,j)}, & \text{for } i = c, c+1, \ldots, c+N. \end{cases} \tag{29}$$

Note that $(\boldsymbol{\pi}_{-1})_{(i,j)}$ is the component of vector $\boldsymbol{\pi}_{-1}$ that corresponds to state $(-1, i, j)$, and $(\boldsymbol{\pi}_0(I-R)^{-1})_{(i,j)}$ is the component of vector $\boldsymbol{\pi}_0(I-R)^{-1}$ that corresponds to states $\{(n, i, j), n = 0, 1, \ldots\} \in \Omega$.

2. *Distribution of the number of ambulances in transit* $q_T(t)$: Let $\boldsymbol{\gamma} = (\gamma(0), \gamma(1), \ldots, \gamma(N))$ be the distribution of the number of ambulances in transit in steady state. Then we have, for $j = 0, 1, 2, \ldots, N$,

$$\gamma(j) = \sum_{i=0}^{c-1} (\boldsymbol{\pi}_{-1})_{(i,j)} + \sum_{i=c}^{N+c-j} (\boldsymbol{\pi}_0(I-R)^{-1})_{(i,j)}. \tag{30}$$

3. *Distribution of the number of waiting ambulances (or ambulance patients) in the ED right after the arrival of an ambulance patient to the ED*: Let $\boldsymbol{\omega} = (\omega(0), \omega(1), \ldots, \omega(N))$ be the distribution of interest. First, we note that, right after an ambulance in transit arrives to the ED, the number of waiting ambulances seen by the ambulance is different from $\boldsymbol{\xi}$. Also note that the arrival rate of ambulances to the ED is $j\mu_T$, given that $q_T(t) = j$. By renewal theory, we obtain

$$\omega(i) = \begin{cases} \dfrac{1}{\omega_{all}} \sum_{i=0}^{c-1} \sum_{j=1}^{N} (\boldsymbol{\pi}_{-1})_{(i,j)} j\mu_T, & \text{for } i = 0; \\ \dfrac{1}{\omega_{all}} \sum_{j=1}^{N+1-i} (\boldsymbol{\pi}_0(I-R)^{-1})_{(i+c-1,j)} j\mu_T, & \text{for } i = 1, 2, \ldots, N, \end{cases} \tag{31}$$

where

$$\omega_{all} = \sum_{i=0}^{c-1} \sum_{j=0}^{N} (\boldsymbol{\pi}_{-1})_{(i,j)} j\mu_T + \sum_{i=1}^{N} \sum_{j=0}^{N+1-i} (\boldsymbol{\pi}_0(I-R)^{-1})_{(i+c-1,j)} j\mu_T. \tag{32}$$

Appendix C

We show (20) by finding the Laplace–Stieltjes transform (LST) of W_w defined in Section 4. Note that walk-in patients arrive according to a Poisson process. By the well-known PASTA (Poisson arrival see time average), the distribution of the system at the arrival epoch of an arbitrary walk-in patient is the same as that of an arbitrary time, which is $\{\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots\}$. By conditioning on the system state at arrival, the LST of W_w can be derived as follows, for $s > 0$,

$$\begin{aligned} E[e^{-sW_w}] &= \boldsymbol{\pi}_{-1}\mathbf{e} + \sum_{n=0}^{\infty} \boldsymbol{\pi}_n \left((sI - A_1 - A_0)^{-1} A_2\right)^n (sI - A_1 - A_0)^{-1} A_{0,-1} \\ \mathbf{e} &= \boldsymbol{\pi}_{-1}\mathbf{e} + \sum_{n=0}^{\infty} \boldsymbol{\pi}_0 R^n \left((sI - A_1 - A_0)^{-1} A_2\right)^n (sI - A_1 - A_0)^{-1} A_{0,-1}\mathbf{e} \\ &= \boldsymbol{\pi}_{-1}\mathbf{e} + \phi\left(\sum_{n=0}^{\infty} R^n \left((sI - A_1 - A_0)^{-1} A_2\right)^n\right)(\boldsymbol{\pi}_0' \\ &\otimes (sI - A_1 - A_0)^{-1} A_{0,-1}\mathbf{e}) = \boldsymbol{\pi}_{-1}\mathbf{e} + \phi(I)(sI - I \otimes (A_1 + A_0) - R' \\ &\otimes A_2)^{-1}(\boldsymbol{\pi}_0' \otimes A_{0,-1}\mathbf{e}) = \boldsymbol{\pi}_{-1}\mathbf{e} + \phi(I)\Lambda\left(sI - \Lambda^{-1}(I \otimes (A_1 + A_0) - R' \right. \\ &\left. \otimes A_2)\Lambda\right)^{-1}\Lambda^{-1}(\boldsymbol{\pi}_0' \otimes A_{0,-1}\mathbf{e}) = \boldsymbol{\pi}_{-1}\mathbf{e} + \boldsymbol{\alpha}_w(sI - T_w)^{-1}\Lambda^{-1}(\boldsymbol{\pi}_0' \\ &\otimes A_{0,-1}\mathbf{e}). \end{aligned} \tag{33}$$

Then W_w has a phase-type distribution with *PH*-representation $(\boldsymbol{\alpha}_w, T_w)$ if the followings properties can be verified:

- T_w is a *PH*-generator;
- $\boldsymbol{\pi}_{-1}\mathbf{e} + \boldsymbol{\alpha}_w\mathbf{e} = 1$;
- $T_w\mathbf{e} + \Lambda^{-1}(\boldsymbol{\pi}_0' \otimes A_{0,-1}\mathbf{e}) = 0$.

The first property can be verified easily. The second is shown as:

$$\begin{aligned} \boldsymbol{\pi}_{-1}\mathbf{e} + \phi(I)\Lambda\mathbf{e} &= \boldsymbol{\pi}_{-1}\mathbf{e} + \phi(I)((\boldsymbol{\pi}_0(I-R)^{-1})' \otimes \mathbf{e}) \\ &= \boldsymbol{\pi}_{-1}\mathbf{e} + \boldsymbol{\pi}_0(I-R)^{-1}\mathbf{e} = 1. \end{aligned} \tag{34}$$

The third property can be proved as:

$$\begin{aligned} T_w\mathbf{e} + \Lambda^{-1}(\boldsymbol{\pi}_0' \otimes A_{0,-1}\mathbf{e}) \\ = \Lambda^{-1}\Big((\boldsymbol{\pi}_0(I-R)^{-1})' \otimes (A_1 + A_0)\mathbf{e} + (\boldsymbol{\pi}_0(I-R)^{-1}R)' \\ \otimes A_2\mathbf{e} + \boldsymbol{\pi}_0' \otimes A_{0,-1}\mathbf{e}\Big) = \Lambda^{-1}(\boldsymbol{\pi}_0' \otimes (-A_2\mathbf{e} + A_{0,-1}\mathbf{e})) = 0. \end{aligned} \tag{35}$$

The last equality is due to $A_2\mathbf{e} = -(A_1 + A_0)\mathbf{e} = A_{0,-1}\mathbf{e}$.

References

[1] Deo S, Gurvich I. Centralized vs. decentralized ambulance diversion: a network perspective. Management Science 2011;57(7):1300–1319.

[2] Ting JY. 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 2008;1(2):129.

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

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

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

[6] Fomundam S, Herrmann J. A survey of queueing theory applications in healthcare. Technical Report 2007–24, The Institute for Systems Research; 2007.
[7] Green L. In: Hall RW, editor. Queueing analysis in healthcare, patient flow: reducing delay in healthcare delivery. New York: Springer; 2006 [chapter 10].
[8] Almehdawe E, Jewkes B, He Q-M. A Markovian queueing model of ambulance offload delays. European Journal of Operational Research 2013;226:602–614.
[9] Kao E, Tung G. Bed allocation in a public health care delivery system. Management Science 1981;27(5):507–520.
[10] Gorunescu F, McClean S, Millard P. Using a queueing model to help plan bed allocation in a department of geriatric medicine. Health Care Management Science 2002;5:307–312.
[11] Davies R, Davies H. Modelling patient flows and resource provision in health systems. Omega 1994;22(2):123–131.
[12] Masselink IH, van der Mijden TL, Litvak N, Vanberkel PT. Preparation of chemotherapy drugs: planning policy for reduced waiting times. Omega 2012;40(2):181–187.
[13] Côté MJ, Stein WE. An Erlang-based stochastic model for patient flow. Omega 2000;28(3):347–359.
[14] Knight VA, Harper PR, Smith L. Ambulance allocation for maximal survival with heterogeneous outcome measures. Omega 2012;40(6):918–926.
[15] Gorunescu F, McClean S, Millard P. A queueing model for bed-occupancy management and planning of hospitals. The Journal of the Operational Research Society 2002;53(1):19–24.
[16] Leo G, Lodi A, Tubertini P, Di Martino M. Emergency department management in Lazio, Italy. Omega 2016;58:128–138.
[17] Channouf N, LEcuyer P, Ingolfsson A, Avramidis AN. The application of forecasting techniques to modeling emergency medical system calls in calgary, alberta. Health Care Management Science 2007;10(1):25–45.
[18] Mateo Restrepo SH, Topaloglu H. Erlang loss models for the static deployment of ambulances. Health Care Management Science 2009;12:67–79.
[19] Neuts FM. Matrix geometric solutions in stochastic methods: an algorithmic approach. Mineola, NY, USA: Dover Publications; 1981.
[20] Region of Waterloo Public Health, Emergency medical services master plan. Technical Report; December 2007.
[21] Gross D, Shortle J, Thompson J, Harris C. Fundamentals of queueing theory. New Jersey, USA: John Wiley and Sons; 2008.
[22] Lee HL, Cohen MA. A note on the convexity of performance measures of $m/m/c$ queueing systems. Journal of Applied Probability 1983;20:920–923.
[23] Grassmann W. The convexity of the mean queue size of the $m/m/c$ queue with respect to the traffic intensity. Journal of Applied Probability 1983;20(4):916–919.
[24] Ozawa T. Sojourn time distributions in the queue defined by a general QBD process. Queueing Systems 2006;53(4):203–211.