Operations Research for Health Care 11 (2016) 13–19

Contents lists available at ScienceDirect

Operations Research for Health Care

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

Offload zone patient selection criteria to reduce ambulance offload delay

Corine M. Laan [a], Peter T. Vanberkel [b,*], Richard J. Boucherie [a], Alix J.E. Carter [c]

[a] Stochastic Operations Research, University of Twente, The Netherlands
[b] Industrial Engineering, Dalhousie University, Canada
[c] Emergency Health Services; Department of Emergency Medicine; Division of Emergency Medical Services, Dalhousie University, Canada

ARTICLE INFO

Article history:
Received 14 January 2016
Accepted 5 September 2016
Available online 13 September 2016

Keywords:
Offload delay
Continuous time Markov chain
Emergency Medical Services
Offload zone

ABSTRACT

Emergency department overcrowding is a widespread problem and often leads to ambulance offload delay. If no bed is available when a patient arrives, the patient has to wait with the ambulance crew. A recent Canadian innovation is the offload zone—an area where multiple patients can wait with a single paramedic–nurse team allowing, the ambulance crew to return to service immediately. Although a reduction in offload delay was anticipated, it was observed that the offload zone is often at capacity. In this study we investigate why this is the case and use a continuous time Markov chain to evaluate how interventions can prevent congestion in the offload zone. Specifically we demonstrate conditions where the offload zone worsens offload delay and conditions where the offload zone can essentially eliminate offload delay.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

The design of Emergency Medical Services (EMS) using operations research methods has a rich history beginning in the 1960s [1–4]. The largest body of work focuses on dispatching strategies (i.e. selecting which ambulance to send to which call) and determining ambulance base locations. The objective is typically to improve the tradeoff between responsiveness and costs. The interface with, and transfer of patients to, Emergency Department (ED) has seen less attention. However, when EDs are congested (as is increasingly becoming the norm [5]) the time to transfer a patient from EMS to the ED can be significant [6] and can negatively affect response time. Formally, this delay period is referred to as Offload Delay (OD)—a delay between an ambulance's arrival at the ED and the transfer of patient care, resulting in a prolonged hospital stay for the ambulance [7,8].

In Nova Scotia, Canada, OD is worsening: the 90th percentile for the time an ambulance stays at the hospital has increased from 24 min in 2002 to 109 min in 2007 [9]. By 2010 two of Nova Scotias most affected urban EDs, The Queen Elizabeth II Health Sciences Centre and Dartmouth General Hospital reported OD times of 114 min and 142 min 90% of the time respectively [10]. Similar OD has been experienced in Ontario and is reported by [11–13].

Delaying the admission of a patient to an ED can result in poor pain control, delayed time to antibiotics, increased morbidity and potentially increased mortality [14,15]. While an ambulance crew is delayed at a hospital, they are unavailable for emergency response in the community which diminishes service [16]. Preliminary evaluation work in Alberta found considerable improvement in EMS efficiency and cost-effectiveness was possible if OD is reduced [17].

A common mitigation strategy to reduce OD is "diversion" [18]. When an ED declares diversion status ambulances are rerouted away from that ED and instead to a less crowded ED elsewhere [11]. Due to extended travel times and patient safety concerns this practice has become less common [19]. A second strategy, which is the focus of this paper, is the implementation of a monitored holding area for patients who arrive by ambulance which frees the ambulance to return to service. In Nova Scotia this area is called an Offload Zone (OZ) but similar concepts by different names can be found in Ontario [20,21].

The Queen Elizabeth II Health Sciences Centre and Dartmouth General Hospital, in collaboration with EMS, implemented OZs in 2012 [10]. In the OZ there are two dedicated staff members, one nurse and one paramedic who receive patients and monitor them until they can be admitted to the ED. Once the transfer of care has been made by the ambulance crew to the OZ staff the ambulance

* Corresponding author.
E-mail address: peter.vanberkel@dal.ca (P.T. Vanberkel).

http://dx.doi.org/10.1016/j.orhc.2016.09.001
2211-6923/© 2016 Elsevier Ltd. All rights reserved.

crew can return to service and be available for another emergency response. The OZ can serve multiple patients and eliminates the need for an ambulance crew to wait with each patient.

Two years after opening the two OZs we completed a Health Care Failure Mode and Effect Analysis (HFMEA) study to identify risks to patient safety and process efficiency [19]. In this study a detailed process map of the OZ functions and its relationship with EMS and the ED was developed through consensus by paramedics and nurses. From this map, staff identify potential hazards and prioritized them based on the likelihood of occurrence and the potential severity. The primary goal of HFMEA is to be proactive in identifying risks and hazards [22]. The following conclusion drawn from the HFMEA study motivates the research described herein:

> One unexpected finding of the process map was that the real life functioning of the OZ deviated significantly from the original protocol. The original intent of the OZ, was to monitor up to six ambulance patients at once in order to reduce the need for one paramedic crew to remain for each patient, therefore allowing the paramedics to return to the community. The steps in the original OZ protocol did not include providing patient care (beginning investigations, etc.) in the OZ; however process mapping has shown that the OZ evolved to an area of extensive patient care. Major steps [such as,] Patient assessment in OZ and Patient care in OZ, consist of diagnostics, procedures, treatments and even MD assessments. The highest hazard score for an effect on process efficiency was related to medical care in the OZ: 'Patient not placed in ED from OZ because patient already receiving care in OZ '. It is thought that this is due to a lack of incentive to move the patient to the ED from the OZ because the patient is already receiving diagnostics/physician assessments and would not directly benefit from moving to the ED. In this model the OZ simply becomes an extension of the ED. [This] has the potential to create a backlog of arriving ambulance patients and could lead to a significant increase in OD, subsequently reducing the quality and timeliness of care for patients in the community awaiting an ambulance" [19].

In this paper we investigate how this lack of incentive to move patients to the ED from the OZ impacts OD. Specifically, we compare a scenario *without* an OZ to scenarios *with* an OZ while varying the degree of 'incentive' to admit OZ patients. To analyse these scenarios we use a Continuous Time Markov Chain (CTMC) to model the OZ.

CTMCs have been used to analyse many service industries with applications in call centres [23] and health care systems. Almehdawe et al. [24] use a CTMC to model offload delay across a network of hospitals. Specifically, they compute a variety of performance measures subject to different resource levels. They analyse the CTMC with matrix-geometric solutions using a probability matrix with a block structure. Dobson et al. [25] also applied a CTMC to a health care flow problem. The authors model a medical teaching facility and the complex patient interactions that occur to facilitate student, resident and attending patient exams. They address the question of how to prioritize work and batch patients to improve throughput. A general multi-class multi-server priority queueing system with customer priority upgrades is examined using a CTMC by He et al. [26]. The general model has various applications with the emergency health care application emphasized. CTMC as a modelling approach to health care problems is demonstrated in [27]. In addition to application of CTMC, formal presentations of their properties are presented by [28,29]. Our paper used a CTMC to model an operational decision made within the ED that impacts the performance of EMS. We solve the CTMC numerically with the method of iterative bounds [30] implemented in MatLab r2013b.

The paper is organized as follows: In Section 2 we introduce the patient flow process in greater detail and formulate the CTMC model. In Section 3 we present numerical results and provide general conclusions in Section 4.

2. Model

2.1. Patient flow

Patients arrive at the ED by either one of two methods. Most arrive by their own means (e.g. by car or by walking) and are referred to as "walk-in" patients. The remaining patients arrive by ambulance and are referred to as ambulance arrivals. Both patient types are triaged according to the Canadian Triage Acuity Score (CTAS); a scoring based on a 1–5 rating with 1 being Resuscitative and 5 being Non-urgent [31]. Resuscitative patients are taken directly to a trauma room for treatment regardless of their means of arrival. Walk-in patients with CTAS 2–5 register and then proceed to the waiting room. Ambulance arriving patients with CTAS 4 or 5 are registered and then proceed to the waiting room also. Ambulance arriving patients with CTAS 2 or 3 are registered but are not placed in the waiting room. These patients wait either in the ambulance with the paramedic crew or in the OZ. This is a general description of the patient flow process and may change in some circumstances. For example, the pathways governed by CTAS can be overruled based on a patient's condition or caregiver judgement. The patient flow process is summarized in Fig. 1.

Patients wait for admission until an ED bed becomes available and they are selected. In general, the lowest CTAS is admitted first. However, when there are ties in CTAS (as is common), the process for breaking ties has been found to be different before and after the implementation of an OZ [19]. Prior to the implementation of the OZ, tie breaking priority was given to patients waiting with an ambulance to allow the ambulance to return to service. After the implementation of the OZ, this pressure to free the ambulance crew dissipated and the tie breaking priority changed. This leads to the primary research question to be addressed by the model: How does patient selection affect the performance of the OZ?

Using the CTMC described in Section 2.2, we compute the number of ambulances waiting in a variety of scenarios. As a baseline scenario, we compute the number of ambulances waiting prior to the implementation of the OZ with patient selection based on CTAS and tie breaking priority given to patients waiting with an ambulance. The next scenario includes the OZ with tie breaking priority *always* given to patients in the OZ (extreme 1). Then a scenario with tie breaking priority *always* given to walk-in patients (extreme 2) is considered. Finally, we consider a range of scenarios between these two extremes where tie breaking priority is given to patients waiting in the OZ with priority p_{OZ}, $0 \leq p_{OZ} \leq 1$. Patients waiting in the waiting room are given tie breaking priority with probability $(1 - p_{OZ})$.

2.2. Model definition

We model the ED with a finite CTMC. The state of our system is completely described by the queue length per patient type and the number of ED beds in use by each patient type. Therefore, we define the following parameters: $N_{i,a}$ (where $i = 1, \ldots, 5$ indicates the CTAS which does not change the longer patients wait) is the number of patients who arrived by ambulance that are waiting, $N_{i,w}$ is the number of walk-in patients waiting, and $N_{i,b}$ is the number of ED beds in use by patients of type i. The state space is given by:

$$S = [N_{1,a}, N_{1,w}, N_{1,b}, \ldots, N_{5,a}, N_{5,w}N_{5,b}].$$

The number of ED beds available is given by c, the service rate per patient type is μ_i, the arrival rate per patient type is $\lambda_{i,w}$ and $\lambda_{i,a}$ respectively for walk-ins and ambulance arrivals. The total arrival rate is given by $\lambda = \sum_i \lambda_{i,a} + \lambda_{i,w}$. The arrival process is assumed to be Poisson which has been shown by [32] to be well suited for modelling non-scheduled arrivals in health care systems.

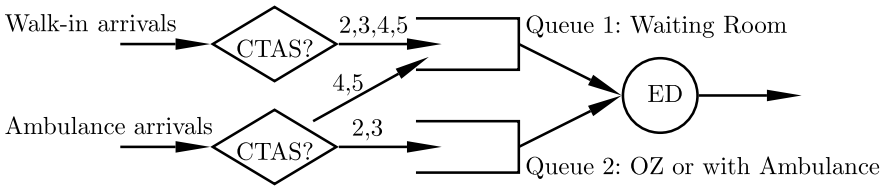

Fig. 1. Summary of patient flow.

The service rate is assumed to be exponentially distributed, which is appropriate when the coefficient of variation is close to 1 and is a good approximation when the actual variance is less than the parameter of the exponential distribution [33]. Others have likewise found exponential distributions well suited to model ED service times [11,34].

Exponential distributions are "ubiquitous in stochastic modelling …because of their ability to model random lengths of time reasonably well" [27]. These advantages and their properties within a health care setting are expanded upon in [35]. With respect to our model, the exponential service time allows the state space to consist of the number of patients waiting and excludes how long they have been waiting. This smaller state space makes the model tractable and the analysis of scenarios very fast. This is particularly important in our study as the tie breaking priority ranges from 0 to 100% (i.e. $0 \leq p_{OZ} \leq 1$) leading to considerable computational effort for each scenario.

In each state, there are events that cause a transition to another state. The time of a transition from state s to state s' ($s \neq s'$) is exponentially distributed with the following transition rates:

- If there are no patients waiting, the number of used beds decreases with rate $\sum_i N_{i,b}\mu_i$;
- If the number of used beds is strictly smaller than c, the number of used beds increases with rate λ;
- If there are patients waiting, the number of patients in the highest priority queue decreases with rate $\sum_i N_{i,b}\mu_i$;
- If there are patients waiting, the number of patients in queue i, a [and queue i, w] increases with rate $\lambda_{i,a}$ and [$\lambda_{i,w}$].

A generator matrix can be constructed for this Markov process [29]. However, because this generator matrix is extensive for ten patient types and computationally challenging, we construct the generator matrix for the case where there are only three patient types: type (1) ambulance arrivals with a high acuity (CTAS 2 or 3), type (2) walk-in arrivals with high acuity (CTAS 2 or 3) and type (3) patients with a low acuity (CTAS 4 or 5). Note that type 3 includes patients who arrive by ambulance or walk-in and that CTAS 1 patients are not modelled since they are uncommon [36] and do not wait in the OZ. This aggregation of patient types is chosen to reflect the different service times observed from CTAS 2/3 versus CTAS 4/5 patients [37]. It follows that patient type 3 has the lowest priority for admission, type 1 has priority for admission with probability p_{OZ} and type 2 has priority for admission with probability $1 - p_{OZ}$.

The ambulance and walk-in patients with a high acuity (types 1 and 2, respectively) have a service rate of μ_H and the patients with a low acuity (type 3) have a service rate of μ_L. Let λ_1, λ_2 and λ_3 respectively be the mean arrival rate for types 1, 2 and 3. Let N_b^H be the number of patients in a bed with high acuity (types 1 and 2) and $N_b - N_b^H$ be the number of patients in a bed with low acuity (type 3). Let N_1 be the number of type 1 patients waiting, N_2 be the number of type 2 patients waiting and N_3 be the number of type 3 patients waiting.

The condensed state space is $S = (N_b, N_b^H, N_1, N_2, N_3)$. Each state $s \in S$ corresponds uniquely with a number of beds in use and the queue length for each patient type. The generator matrix Q is constructed with the elements $q_{s,s'}$; the diagonal elements of Q are constructed as follows:

$$q_{s,s} = -\sum_{s' \neq s} q_{s,s'}.$$

The generator matrix is constructed with the following transition rates which are categorized to improve clarity. The first four rates represent transitions occurring when there are no patients waiting and not all ED beds occupied (see Box I).

To find the equilibrium distribution π of the CTMC, we use balance equations [29]. The probability that the system is in state s, is given by π_s, and when a system is in state s, the rate out of that state is given by $\sum_{s' \neq s} q_{s,s'}$. The flow out of each state is given by $\pi_s \sum_{s' \neq s} q_{s,s'}$. This has to be equal to the flow into the state: $\sum_{s' \neq s} \pi_s' q_{s',s}$. In stationarity, the balance equations hold for all states s:

$$\pi_s \sum_{s' \neq s} q_{s,s'} = \sum_{s' \neq s} \pi_{s'} q_{s',s} \quad \forall s \in S.$$

Some mathematical rewriting yields $\pi Q = 0$. To find the probability distribution, we solve $\pi Q = 0$ together with the normalization equation $\sum_{s \in S} \pi_s = 1$. Note that from all states we may return to the empty state so that we have a unique solution.

The size n of the (square) generator matrix Q increases quickly as the number of patients allowed in the queue increases:

$$n = c + \Pi_{t=1}^{3} M_t. \tag{1}$$

To overcome the size of this matrix we use a numerical method [38]. We solve this numerically with an iterative method, while making a sparse matrix Q. The iterative method we implemented is the method of iterative bounds [30]. This method is implemented in MatLab r2013b.

From π the number of waiting ambulances and patients of each type is determined. Patients of type 2 and 3 wait in the ED's waiting room. Patients of type 1 wait in the OZ *or* with an ambulance when the OZ is full or when there is no OZ. In the baseline scenario each waiting type 1 patient corresponds to a waiting ambulance as there is no OZ. In the remaining scenarios, each waiting type 1 patient corresponds to a patient waiting in the OZ, when the OZ is at or below capacity, and an ambulance otherwise. The distribution for the number of waiting ambulances (X) is,

$$\mathbb{P}(X = z) = \begin{cases} \sum_{s \in S'(z)} \pi_s & \text{when } z > 0 \\ 1 - \sum_{k=1}^{\infty} \mathbb{P}(X = k) & \text{when } z = 0 \end{cases}$$

where $S'(z)$ is the set of states where $N_1 = z + b_{OZ}$ and b_{OZ} is the capacity of the OZ measured in beds. Note that the number of patients in the OZ is the $\min(N_1, b_{OZ})$ and the number of ambulances waiting is the $\max(N_1 - b_{OZ}, 0)$. It follows that the expected number of ambulances waiting is,

$$\mathbb{E}[X] = \sum_{z=0}^{\infty} z\mathbb{P}(X = z).$$

In a similar manner the number of patients waiting in the waiting room and average waiting times (using Little's Law [39]) can be computed.

$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b+1,N_b^H,N_1,N_2,N_3)} = \lambda_3 \qquad (N_b^H \le N_b < c; N_1, N_2, N_3 = 0),$$
$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b+1,N_b^H+1,N_1,N_2,N_3)} = \lambda_1 + \lambda_2 \qquad (N_b^H \le N_b < c; N_1, N_2, N_3 = 0),$$
$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b-1,N_b^H,N_1,N_2,N_3)} = (N_b - N_b^H)\mu_L \qquad (N_b^H \le N_b \le c; N_1, N_2, N_3 = 0),$$
$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b-1,N_b^H-1,N_1,N_2,N_3)} = N_b^H \mu_H \qquad (N_b^H \le N_b \le c; N_b^H > 0; N_1, N_2, N_3 = 0).$$

The next three rates represent transitions occurring when a patient arrives and finds all ED beds occupied:

$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b,N_b^H,N_1+1,N_2,N_3)} = \lambda_1 \qquad (N_b^H \le N_b = c; N_1 \le M_1),$$
$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b,N_b^H,N_1,N_2+1,N_3)} = \lambda_2 \qquad (N_b^H \le N_b = c; N_2 \le M_2),$$
$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b,N_b^H,N_1,N_2,N_3+1)} = \lambda_3 \qquad (N_b^H \le N_b = c; N_3 \le M_3).$$

Note that M_1, M_2, and M_3 are respectively the maximum queue size for the three patient types and are required for numerical purposes. The remaining rates correspond with patients being admitted to an ED bed and, by extension, being removed from their queue. The formulation ensures the priority scheme used in the model is adhered to. The following rates ensure that type 3 patients are only admitted when there are no type 1 and 2 (i.e. higher priority) patients waiting.

$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b,N_b^H,N_1,N_2,N_3-1)} = (N_b - N_b^H)\mu_L \qquad (N_b^H \le N_b = c; N_1, N_2 = 0, N_3 > 0),$$
$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b,N_b^H-1,N_1,N_2,N_3-1)} = N_b^H \mu_H \qquad (N_b^H \le N_b = c; N_b^H > 0; N_1, N_2 = 0, N_3 > 0).$$

The following rates ensure that type 2 patients are admitted when there are no type 1 patients waiting.

$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b,N_b^H,N_1,N_2-1,N_3)} = N_b^H \mu_H \qquad (N_b^H \le N_b = c; N_b^H > 0; N_1 = 0, N_2 > 0),$$
$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b,N_b^H+1,N_1,N_2-1,N_3)} = (N_b - N_b^H)\mu_L \qquad (N_b^H \le N_b = c; N_1 = 0, N_2 > 0).$$

The following rates ensure that type 1 patients are admitted when there are no type 2 patients waiting.

$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b,N_b^H,N_1-1,N_2,N_3)} = N_b^H \mu_H \qquad (N_b^H \le N_b = c; N_b^H > 0; N_1 > 0, N_2 = 0),$$
$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b,N_b^H+1,N_1-1,N_2,N_3)} = (N_b - N_b^H)\mu_L \qquad (N_b^H \le N_b = c; N_1 > 0, N_2 = 0).$$

The following rates ensure that type 1 patients are admitted with probability p_{OZ} when type 2 patients are also waiting.

$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b,N_b^H,N_1-1,N_2,N_3)} = p_{OZ} N_b^H \mu_H \qquad (N_b^H \le N_b = c; N_b^H > 0; N_1, N_2 > 0),$$
$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b,N_b^H+1,N_1-1,N_2,N_3)} = p_{OZ}(N_b - N_b^H)\mu_L \qquad (N_b^H \le N_b = c; N_1, N_2 > 0).$$

The following rates ensure that type 2 patients are admitted with probability $1 - p_{OZ}$ when type 1 patients are also waiting.

$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b,N_b^H,N_1,N_2-1,N_3)} = (1 - p_{OZ})N_b^H \mu_H \qquad (N_b^H \le N_b = c; N_b^H > 0; N_1, N_2 > 0),$$
$$q_{(N_b,N_b^H,N_1,N_2,N_3),(N_b,N_b^H+1,N_1,N_2-1,N_3)} = (1 - p_{OZ})(N_b - N_b^H)\mu_L \qquad (N_b^H \le N_b = c; N_1, N_2 > 0).$$

Box I.

3. Results

Model data was derived from three different publicly available data sources and summarized in Table 1. High acuity patients that arrive by ambulance (type 1) make up $0.6 * 0.21 * 100 = 12.6\%$ of the total patients. High acuity walk-in patients (type 2) make up $0.6 * (1 - 0.21) * 100 = 47.4\%$ of the total patients and low acuity patients (type 3) make up the remaining $0.4 * 100 = 40.0\%$ of patients. Formally,

$$\lambda = \lambda_1 + \lambda_2 + \lambda_3$$
$$\lambda_1 = p_H * f_H * \lambda \tag{2}$$
$$\lambda_2 = p_H * (1 - f_H) * \lambda \tag{3}$$
$$\lambda_3 = (1 - p_H) * \lambda. \tag{4}$$

The load for the system is,

$$\rho = \frac{\lambda}{\mu c} \tag{5}$$

Table 1
Model data.

	CTAS 2/3	CTAS 4/5
Patient type distribution [36]	$p_H = 0.6$	$p_L = 0.4$
Fraction arriving by ambulance [40]	$f_H = 0.21$	$f_L = 0.38$
Service rate (patients/h) [37]	$\mu_H = 0.45$	$\mu_L = 2$

and the mean service time $(1/\mu)$ is,

$$\frac{1}{\mu} = \frac{p_H * f_H}{\mu_H} + \frac{p_H * (1 - f_H)}{\mu_H} + \frac{(1 - p_H)}{\mu_L}. \tag{6}$$

For a given ρ we can compute λ from (5) and λ_1, λ_2 and λ_3 from (2), (3) and (4).

3.1. Small hospital

Initially we consider a small hospital with $c = 5$ beds in the ED and an OZ with capacity to hold $b_{OZ} = 2$ patients. The CTMC implicitly assumes that the underlying system is stationary. Across a 24-hour period, EDs are not stationary and have busy and slow periods. However, the OZ is only operational during the

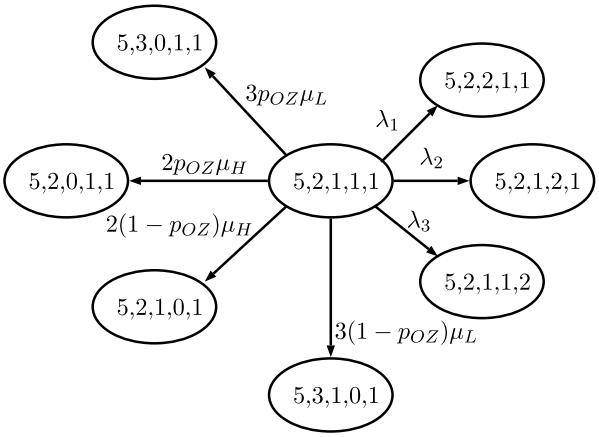

Fig. 2. Truncated state transition diagram.

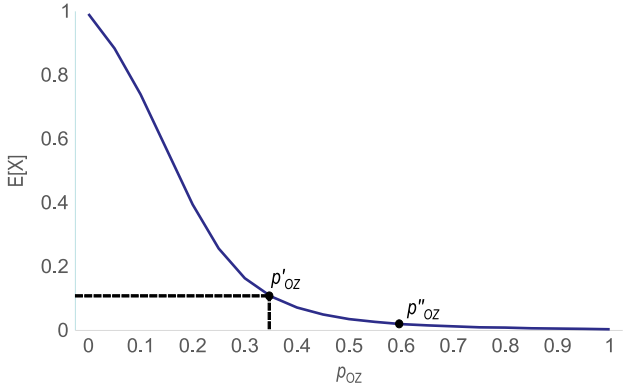

Fig. 3. The expected number of waiting ambulances as a function of p_{OZ}.

busy period which is assumed to be stationary. To reflect this busy period when the ED is congested we set $\rho = 0.98$ so that $\lambda_1 = 0.4$, $\lambda_2 = 1.53$ and $\lambda_3 = 1.25$. Finally, $M_1 = 10$, $M_2 = 10$ and $M_3 = 25$ are chosen to balance computational time and model accuracy. Larger values of M allow for a larger number of states and increases model accuracy but at the expense of computation time. This tradeoff is investigated in Section 3.2.

Continuing with this example, Fig. 2 provides a truncated state transition diagram illustrating the state with 5 ED beds occupied (2 by high acuity patients) and 1 patient waiting in each queue, i.e. state 5, 2, 1, 1, 1. For this state we illustrate all outgoing adjacent states and their transition rates.

In the baseline scenario with no OZ the expected number of waiting ambulances is $\mathbb{E}[X] = 0.153$. The distribution for the number of ambulances waiting is $\mathbb{P}(X \geq 1) = 0.13$, $\mathbb{P}(X \geq 2) = 0.02$ and $\mathbb{P}(X \geq 3) = 0.004$. Extreme 1, representing a scenario with an OZ and no change in how patients are selected ($p_{OZ} = 1$), results in $\mathbb{E}[X] = 0.004$ and $\mathbb{P}(X \geq 1) = 0.0047$ and for all practical purposes eliminates OD and waiting ambulances. Extreme 2, representing a scenario with an OZ and tie breaking priority going to waiting room patients ($p_{OZ} = 0$), results in $\mathbb{E}[X] = 0.992$ and $\mathbb{P}(X \geq 1) = 0.27$ and shows how the OZ has worsened OD considerably. Fig. 3 plots $\mathbb{E}[X]$ as a function of p_{OZ}.

From Fig. 3 it is clear that the relationship between $\mathbb{E}[X]$ and p_{OZ} is nonlinear and that there are diminishing returns. Points p'_{OZ} and p''_{OZ} of Fig. 3 require further discussion. $p'_{OZ} = 0.35$ is the p_{OZ} value which gives the same $\mathbb{E}[X]$ value as in the baseline scenario. It indicates that tie breaking priority must be given to OZ patients at least 35% of the time in order to maintain the same performance as in the baseline scenario. If priority is given less than 35% of the time then OD will actually become worse after the implementation of the OZ.

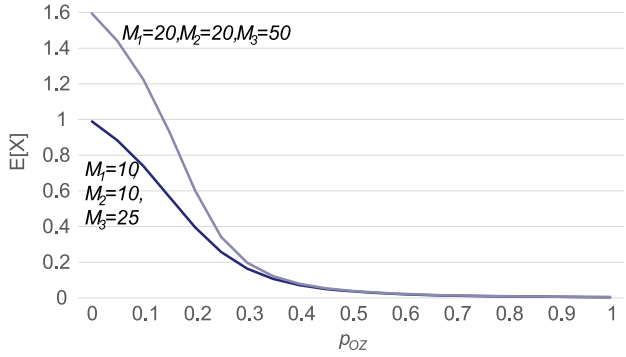

Fig. 4. The expected number of waiting ambulances as a function of p_{OZ} and M_1, M_2, and M_3.

Table 2
p'_{OZ} and p''_{OZ} results for larger values of M_1, M_2 and M_3.

Run time (s)	ED beds	OZ beds	M_1, M_2, M_3	p'_{OZ}	p''_{OZ}
128	5	2	10, 10, 25	0.35	0.6
6663	5	2	20, 20, 50	0.35	0.6

The point, $p''_{OZ} = 0.6$ indicates when the decrease in the number of ambulances waiting is negligible for larger values of p_{OZ}. In other words, giving tie breaking priority to OZ patients more than 60% of the time will not decrease the number of ambulances waiting by a meaningful amount. In this case, p''_{OZ} was chosen subjectively by inspection but it could be easily formalized by defining the level by which management feel decreases in $\mathbb{E}[X]$ are negligible.

3.2. Sensitivity

In this section the sensitivity of p'_{OZ} and p''_{OZ} to the model parameters are investigated. First larger values of M are considered, then a larger ED, then a larger OZ and finally a less busy ED.

Table 2 compares p'_{OZ} and p''_{OZ} values when the M_1, M_2, and M_3 are doubled from 10, 10, 25 to 20, 20, 50 respectively. This change increases the state space considerably (see (1)) and consequently increased the model run time from 128 s to 6663 s. The p'_{OZ} and p''_{OZ} values however do not change. A plot of the expected ambulances waiting as a function of p_{OZ} (Fig. 4) helps explain this.

In Fig. 4 the number of ambulance waiting is larger when $M_1 = 20$, $M_2 = 20$, and $M_3 = 50$ but only significantly larger when $p_{OZ} < 0.3$. When $p_{OZ} > 0.3$ the difference appears negligible. To explain this, consider that a finite M_1 value means that the number of type 1 patients waiting in the model cannot exceed M_1 (similarly for the other patients types and M_2 and M_3) and when that occurs those patients are essentially dropped from the model. From Fig. 4 this only appears concerning when $p_{OZ} < 0.3$ which is what you would expect since smaller p_{OZ} means more type 1 patients waiting. For larger p_{OZ} values the difference between the plotted lines is negligible illustrating that $M_1 = 10$, $M_2 = 10$, and $M_3 = 25$ are sufficiently large for computing p'_{OZ} and p''_{OZ} in this case.

Table 3 summarizes the results for a larger ED with 30 beds. For this analysis, ρ has been maintained at 0.98 and λ has been adjusted accordingly. The p'_{OZ} and p''_{OZ} are different from those found for the smaller ED (Fig. 3), which is to be expected. However, similarly to the results reported in Table 2, they are insensitive to larger M_1, M_2, and M_3 values. Finally, this table also demonstrates how large problems ($c = 30$, $M_1 = 20$, $M_2 = 20$, and $M_3 = 50$), which is typical for a regional hospital (and consistent in size with the QEII ED) can be accommodated with the model.

Table 4 analyzes the same large ED considered in Table 3 but varies the capacity of the OZ. Specifically, we evaluate p'_{OZ} and

Table 3
p'_{OZ} and p''_{OZ} results for a larger ED.

Run time (s)	ED beds	OZ beds	M_1, M_2, M_3	p'_{OZ}	p''_{OZ}
3,335	30	2	10, 10, 25	0.3	0.55
173,197	30	2	20, 20, 50	0.3	0.55

Table 4
p'_{OZ} and p''_{OZ} results for larger OZs.

ED beds	OZ beds	M_1, M_2, M_3	p'_{OZ}	p''_{OZ}
30	1	10, 10, 25	0.4	0.7
30	2	10, 10, 25	0.3	0.55
30	3	10, 10, 25	0.35	0.45
30	4	10, 10, 25	0.2	0.4
30	5	10, 10, 25	0.15	0.35

Table 5
p'_{OZ} and p''_{OZ} results for smaller ρ values.

ED beds	OZ beds	M_1, M_2, M_3	ρ	p'_{OZ}	p''_{OZ}
5	2	10, 10, 25	0.98	0.4	0.7
5	2	10, 10, 25	0.95	0.3	0.55
5	2	10, 10, 25	0.9	0.35	0.45
5	2	10, 10, 25	0.8	0.2	0.4

p''_{OZ} for b_{OZ} values of 1, 2, 3, 4, and 5. Note that in all cases, a single paramedic–nurse team provides sufficient coverage. As the capacity of the OZ increases both p'_{OZ} and p''_{OZ} decrease implying that less tie breaking priority needs to be given to OZ patients. This is expected since a OZ with a large capacity relieves a lot of waiting ambulances regardless of the tie breaking priority.

Finally in Table 5 we report results for a variety of ρ values. Smaller ρ decreases both p'_{OZ} and p''_{OZ} implying that less tie breaking priority needs to be given to OZ patients. Again this is to be expected since hospitals which are less busy have fewer ambulances waiting and hence can comparably operate at smaller p'_{OZ} and p''_{OZ} values.

4. Conclusion

Physicians and staff at our partner organizations have observed the OZ in operation for a few years and anecdotal evidence suggested that it does not always decrease OD. Our previous study [19] found that ED staff had little incentive to admit patients who were waiting in the OZ and instead admitted patients from the waiting room. This led to the OZ often being at capacity and unable to relieve OD. In this paper we modelled the "incentive to admit OZ patients" using a Bernoulli distribution with parameter p_{OZ}—the probability that a patient from the OZ is admitted when a patient of equal acuteness is waiting in the waiting room. How this parameter impacts the number of waiting ambulances was investigated.

Consistent with staff observations, our model found that implementing an OZ will result in longer OD if admission priority is disproportionately given to walk-in patients. Specifically, when p_{OZ} is not greater than a certain threshold (in our case $p'_{OZ} = 0.35$) OD will become worse. This threshold is sensitive to the capacity of the OZ (b_{OZ}) and the clinical load (ρ). A larger b_{OZ} allows a smaller p'_{OZ} and a smaller ρ also allows a smaller p'_{OZ}. For EMS, this means the ED's incentive to admit OZ patients impacts OD less when there is a large OZ and when the ED is less busy.

Our model also found that there are diminishing returns and that p_{OZ} need not exceed a certain threshold (in our case $p''_{OZ} = 0.6$) as beyond this threshold the decrease in OD is insignificant. This second threshold is practically relevant for other reasons. Most notably, it implies that the OZ adds a certain flexibility for the ED. OZ patients do not *always* need to be given tie breaking priority and walk-in patients can be chosen instead (e.g. due to excessive queues) without negatively impacting OD. This flexibility becomes greater (i.e. p''_{OZ} becomes smaller) as the capacity of the OZ is increased and as the clinical load decreases.

Our study has a few limitations that should be considered particularly before applying our model in other settings. The data used on the model is publicly available from different sources. It represents what one would generally expect from an ED but does not reflect a specific ED. The methodology can easily be applied to specific hospitals with specific data. In completing such case studies, care should be taken to ensure the model's underlying assumptions are correct for the hospital under study. In particular, the service times should be suitably modelled with independent exponential distributions and the time period under study should exhibit stationary behaviour, e.g. time-invariant arrivals. Such assumptions are typical in application of queueing theory to health care as discussed in detail in [41]. Time-dependent arrivals are incorporated into a CTMC model for staff dimensioning in [42].

For the hospitals, this research demonstrated and quantified an anecdotally reported phenomenon, namely that patient selection practices by the ED can negatively impact the expected benefits of the OZ. In ongoing reviews of the OZ design and potential, this research forms one of the considerations, along with concerns related to patient safety [19], scopes of practice/human resources and cost effectiveness. Further research for the hospital includes incorporating site specific data and the development of an interface for the model to automate the evaluation.

List of Acronyms

EMS	Emergency Medical Services
ED	Emergency Department
OD	Offload Delay
OZ	Offload Zone
HFMEA	Health Care Failure Mode and Effect Analysis
CTMC	Continuous Time Markov Chain

References

[1] J.B. Goldberg, Operations research models for the deployment of emergency services vehicles, EMS Manag. J. 1 (1) (2004) 20–39.
[2] L. Brotcorne, G. Laporte, F. Semeta, Ambulance location and relocation models, European J. Oper. Res. 147 (3) (2003) 451–463.
[3] L. Aboueljinane, E. Sahin, Z. Jemai, A review on simulation models applied to emergency medical service operations, Comput. Ind. Eng. 66 (4) (2013) 734–750.
[4] M.K. Delgado, L.J. Meng, M.P. Mercer, J.M. Pines, D.K. Owens, G.S. Zaric, Reducing ambulance diversion at hospital and regional levels: systemic review of insights from simulation models, West. J. Emerg. Med. 14 (2013) 489–498.
[5] G. Dickinson, Emergency department overcrowding, Can. Med. Assoc. J. 140 (3) (1989) 270–271.
[6] D.R. Cooney, S. Wojcik, N. Seth, C. Vasisko, K. Stimson, Evaluation of ambulance offload delay at a university hospital emergency department, Int. J. Emerg. Med. 6 (15) (2013) 1–4.
[7] D.W. Spaite, T.D. Valenzuela, H.W. Meislin, E.A. Criss, P. Hinsberg, Prospective validation of a new model for evaluating emergency medical services systems by in-field observation of specific time intervals in prehospital care, Ann. Emerg. Med. 22 (4) (1993) 638–645.
[8] D.C. Cone, S.J. Davidson, Q. Nquyen, A time-motion study of the emergency medical services turnaround interval, Ann. Emerg. Med. 31 (2) (1998) 241–246.
[9] S. Karim, A. Carter, J. Ferguson, et al., The evolution of offload delay over a six year period in a provincial EMS system (abstract), Prehospital Emerg. Care 13 (1) (2009).
[10] Capital District Health Authority: Quarterly performance report emergency departments and system flow, 2011. Available at: http://www.cdha.nshealth.ca/system/files/sites/343/documents/quarterly-performance-emergency-february-2011.pdf (accessed June 2015).
[11] E. Almehdawe, Queueing network models of ambulance offload delays (Ph.D. thesis), University of Waterloo, 2012.
[12] P. Macintyre, Hospital offload delay status update. Tech. Rep., Toronto EMS, 2009.
[13] Region of Waterloo Public Health: Emergency medical services master plan. Tech. rep., Region of Waterloo Public Health, 2009.
[14] R.W. Derlet, J.R. Richards, R.L. Kravitz, Frequent overcrowding in US emergency departments, Acad. Emerg. Med. 8 (2) (2001) 151–155.

[15] R.W. Derlet, J.R. Richards, Emergency department overcrowding in Florida, New York, and Texas, Southern Med. J. 95 (8) (2002) 846–849.
[16] D.R. Cooney, M.G. Millin, A. Carter, B.J. Lawner, J.V. Nable, H.J. Wallus, Ambulance diversion and emergency department offload delay: Resource document for the national association of EMS physicians position statement, Prehospital Emerg. Care 15 (4) (2011) 555–561.
[17] A.D. McRae, D. Wang, I.E.B. , et al., Upstream relief: Benefits on EMS offload delay of a provincial ED overcapacity protocol aimed at reducing ED boarding (abstract), Can. J. Emerg. Med. 14 (S1) (2012).
[18] S. Deo, I. Gurvich, Centralized vs. decentralized ambulance diversion: A network perspective, Manage. Sci. 57 (7) (2011) 1300–1319.
[19] A.J. Carter, J.B. Gould, P. Vanberkel, J.L. Jensen, J. Cook, S. Carrigan, M.R. Wheatley, A.H. Travers, Offload zones to mitigate emergency medical services (EMS) offload delay in the emergency department: a process map and hazard analysis, Can. J. Emerg. Med. 17 (6) (2015) 1–9.
[20] K. Newell, A. Hemlin, G. Furlong, Offload delay - returning paramedic unit hours to the street: the Ottawa approach, Can. Paramed. 36 (2013) 20–22.
[21] L. Hendry, $$$ for ambulance offloading, 2012. Available at: http://www.intelligencer.ca/2012/08/30/-for-ambulance-offloading (accessed June 2015).
[22] J. DeRosier, E. Stalhandske, J.P. Bagian, T. Nudell, Using health care failure mode and effect analysis^{TM}: the VA national center for patient safety's prospective risk analysis system, Joint Comm. J. Qual. Patient Saf. 28 (5) (2002) 248–267.
[23] A. Deslauriers, P. L'Ecuyer, J. Pichitlamken, A. Ingolfsson, A.N. Avramidis, Markov chain models of a telephone call centre with call blending, Comput. Oper. Res. 34 (6) (2006) 1616–1645.
[24] E. Almehdawe, B. Jewkes, Q.M. He, A Markovian queueing model for ambulance offload delays, European J. Oper. Res. 226 (3) (2013) 602–614.
[25] G. Dobson, H.H. Lee, A. Sainathan, V. Tilson, A queueing model to evaluate the impact of patient batching on throughput and flow time in a medical teaching facility, Manuf. Serv. Oper. Manag. 14 (4) (2012) 584–599.
[26] Q.M. He, J. Xie, X. Zhao, Priority queue with customer upgrades, Nav. Res. Logist. 59 (5) (2012) 362–375.
[27] M. Fackrell, Modelling healthcare systems with phase-type distributions, Health Care Manage. Sci. 12 (1) (2009) 11–26.
[28] J. Norris, Markov Chains, Cambridge University Press, Cambridge, 1997.
[29] S.M. Ross, Introduction to Probability Models, Academic Press, 2007.
[30] J.V.D. Wal, P.J. Schweitzer, Iterative bounds on the equilibrium distribution of a finite Markov chain, Probab. Engrg. Inform. Sci. 1 (01) (1987) 117–131.
[31] R. Beveridge, B. Clarke, L. Janes, N. Savage, J. Thompson, G. Dodd, Canadian emergency department triage and acuity scale: implementation guidelines, Can. J. Emerg. Med. 1 (3 suppl) (1999) S2–28.
[32] L. Green, Patient Flow: Reducing Delay in Healthcare Delivery, Springer, 2006, pp. 281–307. (chapter 10).
[33] M.E. Zonderland, R.J. Boucherie, Handbook of Healthcare System Scheduling, in: International Series in Operations Research and Management Science, Springer Verlag, New York, 2012.
[34] L.V. Green, J. Soares, J.F. Giglio, R.A. Green, Using queueing theory to increase the effectiveness of emergency department provider staffing, Acad. Emerg. Med. 13 (1) (2006) 61–68.
[35] L. Green, Handbook of Healthcare Delivery Systems, Taylor & Francis, London, 2011, pp. 1–16. (chapter 3).
[36] Capital District Health Authority.: Capital district emergency services council quarterly report, 2012. Available at: http://www.cdha.nshealth.ca/system/files/sites/343/documents/quarterly-report-emergency-services-council-oct-2012.pdf (accessed June 2015).
[37] Canadian Institute for Health Information: Understanding emergency department wait times, 2005. Available at: https://secure.cihi.ca/free_products/Wait_times_e.pdf (accessed June 2015).
[38] D.A. Bini, Numerical solutions of large markov chains, Rend. Semin. Mat. Univ. Politec. Torino 64 (2) (2006) 121–142.
[39] W.L. Winston, Operations Research: Applications and Algorithms, third ed., International Thomson Publishing, Tampa, 1994.
[40] Ontario Hospital Association, the Ontario Medical Association and the Ontario Ministry of Health and Long-Term Care: Improving access to emergency care: Addressing system issues, 2006. Available at: http://www.health.gov.on.ca/en/common/ministry/publications/reports/improving_access/improving_access.pdf (accessed June 2015).
[41] J.L. Wiler, R.T. Griffey, T. Olsen, Review of modeling approaches for emergency department patient flow and crowding research, 18, (12), pp. 1371–1379, 2011.
[42] J.L. Vile, J.W. Gillard, P.R. Harper, V.A. Knight, A queueing theoretic approach to set staffing levels in time-dependent dual-class service systems, Decis. Sci. (2016) (in press).