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.

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/

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.