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

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

A Markovian queueing model for ambulance offload delays

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

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

ARTICLE INFO

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

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

ABSTRACT

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

1. Introduction

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2. The stochastic model

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

2.1. Arrival of patients

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

2.2. Ambulance routing

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