IIE Transactions on Healthcare Systems Engineering

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uhse20

Operations research/management contributions to emergency department patient flow optimization: Review and research prospects

Soroush Saghafian\textsuperscript{a}, Garrett Austin\textsuperscript{a} & Stephen J. Traub\textsuperscript{b}

\textsuperscript{a} Industrial Engineering, School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ, 85281, USA
\textsuperscript{b} Department of Emergency Medicine, Mayo Clinic, Phoenix, AZ, 85054, USA

Published online: 09 Jun 2015.

To cite this article: Soroush Saghafian, Garrett Austin & Stephen J. Traub (2015) Operations research/management contributions to emergency department patient flow optimization: Review and research prospects, IIE Transactions on Healthcare Systems Engineering, 5:2, 101-123

To link to this article: http://dx.doi.org/10.1080/19488300.2015.1017676

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions
Operations research/management contributions to emergency department patient flow optimization: Review and research prospects

SOROUSH SAGHAIFAN1,∗, GARRETT AUSTIN1 and STEPHEN J. TRAUB2

1Industrial Engineering, School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ 85281, USA
E-mail: soroush.saghafian@asu.edu
2Department of Emergency Medicine, Mayo Clinic, Phoenix, AZ 85054, USA

Received August 2014 and accepted February 2015

In recent years, Operations Research/Management (OR/OM) has had a significant impact on improving the performance of hospital Emergency Departments (EDs). This includes improving a wide range of processes involving patient flow from the initial call to the ED through disposition, discharge home, or admission to the hospital. We review approximately 350 related papers to (i) demonstrate the influence of OR/OM in EDs, and (ii) assist both researchers and practitioners with the OR/OM techniques already available to optimize ED patient flow. In addition, we elaborate on some practical challenges yet to be addressed. By shedding light on some less studied aspects that can have significant impacts on ED operations, we also discuss important possibilities for future OR/OM researchers.

Keywords: Operations research/management, emergency department, patient flow, operational efficiency, patient safety

1. Introduction

Growth in healthcare expenditures as a percentage of the United States Gross Domestic Product (GDP) has been explosive, outpacing even past estimates of exponential increases. In 2006, Hall et al. (2006) anticipated an increase in healthcare costs to 15.9% of the GDP by 2010. In fact, the World Health Organization reported that expenditures in 2010 were at 17.6% of GDP (WHO, 2010). This highlights an unquestioned need to improve the efficiency of healthcare delivery methods.

In the United States, Emergency Departments (EDs) are the gate to hospitals through which 50% of non-obstetrical admissions occur (Pitts et al., 2008). Considering that admitted patients create about one-third of the U.S. healthcare bill each year (Abelson, 2013), improving ED operations may have a significant impact on U.S. healthcare expenditures. Indeed, while the direct aggregate spending on emergency care in the United States is estimated to be 5% to 10% of national health expenditures (Lee et al., 2013), considering the fact that ED is the first point of contact for nearly half of all hospital admissions (Schuur and Venkatesh, 2012, and Pitts et al., 2010), improving ED operations and related decisions can have broader impacts. Operations Research/Management (OR/OM) techniques seem able to contribute to such improvements.

To better understand the need for improving ED operations, we note that a considerable percentage of patients report experiencing a delay in the ED, with more than half of those citing long waiting times as a cause (Kennedy et al., 2004). Long waiting times are partially caused by a mismatch between “supply” and “capacity”: the annual number of ED visits increased from 90.3 million to 119.2 million visits between 1996 and 2006, while the number of hospital EDs has decreased from 4019 to 3833 (Pitts et al., 2008). This increasing strain has placed EDs in a state of overcapacity approximately 50% of the time (Geer and Smith, 2004). In a 2009 report to the U.S. Senate, the U.S. Government Accountability Office (GAO) emphasized that crowding continued to occur in EDs, and some patients waited longer than recommended time frames (GAO, 2009). Zilm et al. (2010, p. 296) note that “The ‘ripple’ effect of high inpatient occupancies (particularly by day of week), and delays in discharges, has extended lengths of stays in the ED, frequently resulting in ‘grid lock’ with few ED treatment stations available to maintain patient flow.”
These issues affect not only the timeliness of serving patients, but the ability to serve altogether: Burt and McCaig (2005) found that 44.9% of all U.S. EDs experienced a period of diversion over the course of each year with as many as 1886 ambulances diverted each day. These are only a few indicators of a tremendous need for focus on improving ED operations. Such improvements are essential for increasing profit (Falvo et al., 2007; McConnell et al., 2006), improving patient satisfaction (Thompson et al., 1996; Boudreaux and O’Hea, 2004), and—most importantly—improving patient safety (Mayer, 1979; Trzeciak and Rivers, 2003).

On a larger scale, the problem is more significant than simply managing increasing volumes of patients; some overcrowding issues can be attributed to potentially modifiable use of the ED. A high proportion of patients incorrectly seek out the ED as their first source of care (Burnett and Grover, 1996). Inappropriate ambulance use also puts increasing strain on ED resources (Richards and Ferrall, 1999). Baer et al. (2001) reports that a notable percentage of ED patients are recently discharged ones: “frequent flyers,” considered as patients with five or more visits per year, constitute 14% of total ED visits (Huang et al., 2003). Boarding patients (patients who cannot be moved to inpatient units due to lack of inpatient bed availability) represent up to 18% of all U.S. EDs experienced a period of diversion over the course of each year with as many as 1886 ambulances diverted each day. These are only a few indicators of a tremendous need for focus on improving ED operations. Such improvements are essential for increasing profit (Falvo et al., 2007; McConnell et al., 2006), improving patient satisfaction (Thompson et al., 1996; Boudreaux and O’Hea, 2004), and—most importantly—improving patient safety (Mayer, 1979; Trzeciak and Rivers, 2003).

Historically, a common method of dealing with the inability to serve patients is to “close the doors” (e.g., through ambulance diversion) and focus on patients already in the system. Beyond the negative impact of diverting patients on overall care, however, there are financial considerations: given that 84% of a hospital’s costs are fixed (Roberts et al., 1999), there is little financial incentive to be on diversion. Moreover, some hospitals do not use mechanisms such as ambulance diversion, and in some states diversion is not legal. As a practical matter, barring an internal disaster no ED can truly close completely, as the Emergency Medical Treatment and Labor Act (EMTALA) passed in 1986 requires EDs to serve all patients who present to the facility, regardless of their insurance or financial status. Thus, EDs must focus on improving their patient flow process as a main mechanism for combating the aforementioned issues.

Welch et al. (2006) and Welch et al. (2011) list various metrics by which ED operations can be measured. Among the widely used metrics are LOS (length of stay), LWBS (% of patients who leave without being seen), door to diagnostic evaluation by a qualified medical professional (arrival time to provider contact time, also known as “door-to-doc” time) and ambulance diversion (amount of time ambulances are diverted away from the ED). Olsheker and Rathlev (2006, p. 354) provide a valuable, cautionary perspective regarding LOS and diversion, though the message holds true with all metrics: “[No] measure is universally applicable as a marker of overcrowding and should be used with caution when comparing performance between institutions. Diversion is not an option in some EMS systems and throughput time is ED specific and dependent on the complexity of the case mix. In spite of this, the measures have specific value in tracking individual institutional performance over time.”

It is worth noting that ED operations metrics are also intertwined with hospital quality measures. For instance, the Centers for Medicare & Medicaid Services (CMS) offer an Electronic Health Record (EHR) incentive program for hospitals maintaining quality measures, and the proposed clinical quality measures for the 2014 CMS EHR Incentive Program have a strong focus on boarding time and ED LOS (CMS, 2013). We believe it is important to note, however, that there are challenges to OR/OM in the ED when it comes to assessing tradeoffs between operational improvement and quality. While ED operations have well-defined metrics, ED quality—particularly as it pertains to counting “defects”—does not. Most of the quality events brought to the attention of ED managers are rare, significant adverse events identified through physician self-report, or notification from other services. This ad-hoc approach, while useful for identifying major issues, is insufficiently sensitive to capture small changes in quality that may result from day-to-day changes in ED operations.

Nevertheless, OR/OM techniques have significantly helped various parts of hospitals (and especially the ED) to improve their performance gauging metrics (see, e.g., Ozcan, 2009; Hopp and Lovejoy, 2013; and Green, 2012). However, to the best of our knowledge, there is no review that comprehensively describes and unifies the OR/OM contributions in EDs. To help researchers and practitioners involved in improving ED operations, we first provide such a review. We then provide some important research prospects by shedding light on some fruitful research directions for future research. We also discuss some practical challenges yet to be addressed. These require a focus on the complex details of OR/OM techniques used to model ED patient flow. With an understanding of the depth of OR/OM tools used in the ED, some of the managerial challenges in the ED can be addressed. However, we argue that although common techniques such as Mathematical Programming, Queueing Theory, Simulation Analysis, Markov models and Game Theory have already addressed many challenges, we must also look beyond traditional OR/OM methods; there are still various unanswered questions in the ED. This lends itself to our perspective regarding possibilities for future researchers on valuable but less studied aspects that can have significant impacts on the practice of ED operations.

In closing this section, we note that while the use of OR/OM to solve fundamental social problems evolved mainly from World War II (Green and Kolesar, 2004), it took more than two decades for OR/OM to be used in ED operations. An early related example of using OR/OM tools is the work of Savas (1969), where response time and round-trip time were the performance metrics studied in a computer simulation of ambulance quantity and location. In Goldman et al. (1968), a computer simulation is used...
to reallocate hospital beds and develop a modified usage policy. Time and motion studies were also among pioneers to identify issues with ED patient flow (Heckerling, 1985; Saunders, 1987). More modern techniques can be found in studies such as Green (2006), Green et al. (2006), Armony et al. (2011), Saghafian et al. (2012, 2014), Huang (2013), Huang et al. (2013), and many others that we will review. In reviewing such studies, we hope to provide a resource for both researchers and practitioners to familiarize them with past related, valuable contributions, and to invoke to plan future studies that can help EDs reach new levels of both operational efficiency and patient safety.

The remainder of the paper is organized as follows. Section 2 discusses three components of ED patient flow: flow into, within, and out of the ED. Section 3 describes various OR/OM tools for modeling and optimizing ED patient flow. Section 4 concludes and presents some important research prospects by discussing some less studied aspects of ED patient flow with high potential impact on ED operations.

2. Three components of ED patient flow: Flow into, within, and out of ED

Patient flow can be viewed from two perspectives: operational and clinical (see, e.g., Côté, 2000; Côté and Stein, 2000; and Marshall et al., 2005). The former perspective refers to the physical movement of patients through a set of locations, and the latter refers to the progression of their health status. In this paper, we use the term patient flow to mean the physical movement of patients.

In what follows, we focus on patient flow into, in, and out of the ED, and discuss OR/OM contributions to each of these categories separately.

2.1. Patient flow into the ED

We start by considering contributions that allow altering the arrival process to the ED. Early work in this area includes studies that challenged traditional ambulance dispatch practices (e.g., Carter et al., 1972). Since then, innovations on altering arrival to the ED have been ample. We categorize the related studies in this vein into subsections below.

2.1.1. Ambulance deployment and location

Ambulance response time is a key metric used to evaluateprehospital emergency medical services (Peleg and Pliskin, 2004). This, coupled with the understanding that shaving minutes off of response time has a great potential to save lives (Mayer, 1979; BBC, 2002), has led to a focus on optimizing this first step in the patient flow process. Policies regarding ambulance deployment and location have become commonplace. The EMS Act of 1973 mandates that 95% of requests be served within 10 minutes in rural areas, and 30 minutes in urban areas (Ball and Lin, 1993). Such stipulations on service times are not exclusive to the United States (Gendreau et al., 2001; Galvão et al., 2005). A survey of more than 3000 calls in Ireland showed that only 81% of calls had a response within 15 minutes (Breen et al., 2000), and England and Australia struggle with response benchmarks as well (Kelly et al., 2002; Stoykova et al., 2004). Lownthian et al. (2011) report a global increase in the number of ambulance arrivals to EDs, and a 7.0–12.5% annual increase in ambulance response times, further exacerbating the problem.

Ambulance deployment and location represents some of the earliest OR/OM work not only in emergency response services, but in healthcare. Savas (1969) and Šwoveland et al. (1973) present some of the first research on the subject of ambulance location, while Fitzsimmons (1973) tackles early work on ambulance deployment. Brandeau and Chiu (1989) summarize the original contributions to the field of ambulance location. Lee (2011) strategically organizes EMS decision-making into three categories: ambulance location, ambulance relocation and ambulance dispatching. We acknowledge contributions in these realms below, since all of them affect patient flow into the ED.

Brotcorne et al. (2003) and Goldberg (2004) provide a detailed review of literature on ambulance location with additional focus on relocation and dispatching. The review of Goldberg (2004) provides an in-depth summary of the Hypercube queueing model, introduced by Larson (1974), where a queueing network is characterized by a class of Markov models. Predetermined values of decision variables can be altered to reach desired performance levels for ambulance location. The robustness of the Hypercube model is evident with immediate extensions to make it less computationally rigorous (Larson, 1975), additions to address “ties” between potentially dispatched ambulances (Burwell et al., 1993), through present day application where it has been used to analyze ambulance decentralization (Takeda et al., 2007). Ingólfsson et al. (2008) analyze ambulance location optimization with a more in-depth definition of response time, incorporating often-neglected delays such as the duration of the phone call, time spent dispatching ambulances, time to contact paramedics and more. Peleg and Pliskin (2004) use a geographic decision support tool for ambulance location, improving operations as evidenced by an additional 60% of calls meeting the established 8-minute threshold. A similar effort is seen in Peters and Hall (1999), Singer and Donoso (2008) model ambulance deployment with Queueing Theory, allowing for an in-depth analysis of the trade-offs of fleet size and cycle time. Repede and Bernardo (1994) utilize a time variation coverage location approach for a 13% increase in coverage and 36% decrease in response time, without the addition of resources.

Rajagopalan et al. (2008) generate search algorithms that allow for exponentially quicker decision making on ambulance relocation. Gendreau et al. (2006) focus on relocation as a dynamic model, which relocates ambulances as they are
being dispatched. Gendreau et al. (2001) also use a dynamic model through parallel computing to perform relocation. They suggest that the downside, the potential need to relocate the fleet at every event, is outweighed by the benefit of maximized coverage.

Andersson and Värbrand (2007) provide dynamic ambulance dispatch and relocation analysis for the most complex ambulance control situation in Sweden. This is based on a quantification of “preparedness,” an effort that is continued with great detail by Lee (2011). Deo and Gurvich (2011) utilize a queueing network model to test five different decision making strategies, two of which employ the knowledge of a centralized “social planner” to aid in real-time ambulance routing decisions.

2.1.2. Ambulance diversion

Ambulance diversion (AD) is a technique utilized to reduce ED arrival rates by diverting incoming ambulances to other nearby hospitals (Deo and Gurvich, 2011). First reported at a New York City hospital in the 1990s, it was initially established as a flow management technique, rarely used to relieve the strain on overburdened EDs (Handel, 2011; Asplin, 2003). However, over the next two decades, U.S. hospitals would see nearly 30% more patients per year with a 12% decline in the number of EDs. Not coincidentally, diversion increased: this increase in volume resulted, for instance, in an average per ED increase in diversion hours from 57 to 190 per month in Los Angeles County (Sun et al., 2006). The effect: 91% of ED directors in the United States reported overcrowding as a recurring day-to-day issue, making AD a prominent fixture in hospitals (Olshaker and Rathlev, 2006).

Despite the negative public perception surrounding the technique, AD is designed to improve patient safety and network performance. However, recent anecdotal and empirical evidence suggests that EDs across the nation are not seeing any significant improvements in their wait times while on diversion (Deo and Gurvich, 2011; Mihal and Moilanen, 2005; Kowalczyk, 2008). This discrepancy between ideal AD outcomes and actual observations has increased to the point where some EDs elect to forego diversion process altogether (Kowalczyk, 2008).

Nonetheless, large volumes of research are underway, trying to determine how to best implement AD. The American College of Emergency Physicians (ACEP) has developed guidelines by which hospitals declare, and exist in, the state of diversion in order to minimize its negative effects (Brennan et al., 2000). We focus in the remainder of this section on research geared at understanding and reducing ambulance diversion to better accommodate patient arrivals.

An extensive diversion program implemented in Patel et al. (2006) attempts to adhere to AD guidelines to reduce diversion as a whole. The intervention resulted in a reduction in diversion hours to 7143, a 74% decrease (Patel et al., 2006). Vilke et al. (2004) studies two EDs as they commit to eliminating AD entirely. Adding staff to the ED coupled with a strong focus on AD during a test period reduced diversion hours from 19.7 to 1.4 for one hospital and 27.7 to 0 for another. Pham et al. (2006) identifies at least four other successful studies that primarily enforce policies to reduce diversion.

The addition of resources is common when tackling diversion. California saw a 45% statewide decrease in AD from 2003 to 2007 (Borders et al., 2009). The best practices that lead to successful diversion reduction were consistent with other findings, particularly the use of a “bed czar,” where an individual is responsible for evaluating strategies focused on reducing diversion hours (Geer and Smith, 2004). Ingalls Hospital, a large urban Chicago hospital, saw its diversion hours decrease 79% in one year, largely through the introduction of an admission and discharge room (Geer and Smith, 2004).

Although the majority of evidence is positive with respect to adding resources to decrease diversion (Warden et al., 2003; McConnell et al., 2005; Burt and McCAig, 2006), some negative results on diversion (after the addition of beds (Han et al., 2007) and staffing (Schull et al., 2003) suggests that ED expansion alone is insufficient to decrease diversion, unless such an expansion is done along with removing other hospital bottlenecks. Interestingly, a significant number of diversion efforts focus on expanding ICU resources, despite evidence that the ED is the bottleneck in more than 80% of AD cases (Allon et al., 2013). In addition to expansion-based approaches, some studies analyze other diversion policies, with a focus on minimizing patient waiting (Ramirez-Nafarrate et al., 2012) and using a network perspective whereby AD policies are centralized (Ramirez-Nafarrate et al., 2011).

2.1.3. Ambulance alternatives

We turn briefly to alternate roles that an ambulance can play in patient flow into the ED. Namely, we identify work with great detail by Lee (2011). Deo and Gurvich (2011) utilize a queueing network model to test five different decision making strategies, two of which employ the knowledge of a centralized “social planner” to aid in real-time ambulance routing decisions.

Snooks et al. (2004), however, express skepticism towards the use of ambulance alternatives into the ED, citing a lack of established literature in non-conveyance and the ability of ambulance crews to triage with accuracy. Indeed, a scarcity of research on patients not being transported to the ED is unsurprising. EMS-initiated refusal of transport...
is present in only 7.0% of the largest cities (Knapp et al., 2009), down from 17.0% a decade earlier (Jaslow et al., 1998). We identify a common theme of a lack of reliability on EMS triage in the literature (Schaefer et al., 2000; Pointer et al., 2001; Hauswald, 2002; Levine et al., 2006). In closing this section, we also mention that, given the high percentage of fixed cost in a hospital’s total expenses, these ambulance alternatives would purely be to relieve congestion without regard to revenue (Roberts et al., 1999). As a result, more research establishing ambulance alternatives is required before such a practice can be considered as an operational alternative.

2.2. Patient flow within the ED

In this section, we address the literature that aims at optimizing patient flow inside the ED. Miró et al. (2003) note that patient flow within the ED has a large effect on overcrowding, in conjunction with internal factors and external pressures. Purnell (1991) provides an early review of triage and fast track systems in the ED, touching on many important topics like patient classification and the skills of triage personnel. Wiler et al. (2011) review modeling approaches from a technical perspective, focusing on research on patient flow and crowding. Wiler et al. (2010) provide a less technical review of operations improvements in the front end of the ED – registration, triage, and fast track. Oredsøn et al. (2011) provides a review of triage-related flow improvements, primarily recognizing fast tracks, streaming and triage as interventions.

2.2.1. General triage interventions

The first concepts of triage began in World War I (Keen, 1917), with widespread research and publication on triage in circulation for the greater part of the past two decades. The concept of evaluating and prioritizing patients in the ED is not a new one (Meislin et al., 1988; Wright et al., 1992). Iserson and Moskop (2007) note three conditions that must be satisfied to constitute triage, which we simplify:

1. At least a modest scarcity of health care resources exists.
2. A health care worker assesses each patient’s medical needs, usually based on a brief examination.
3. The triage officer uses an established system or plan, usually based on an algorithm or a set of criteria, to determine specific treatment or treatment priority for each patient (pp. 275–276).

Work on triage has been remarkably extensive, making a comprehensive review well outside the scope of this work. Triage in most U.S. EDs is typically managed through Emergency Severity Index (ESI); it is, however, not universal. For instance, the Australasian Triage Scale (ATS), the Manchester Triage Scale (MTS), and the Canadian Triage Acuity Scale (CTAS) are used in other countries (Beveridge, 1998; Richardson, 1998; Saghaian et al., 2014). Despite the differences, the relative successes of these triage scales has led to a move away from the once popular three-level scale in the United States. Today, the five-level ESI system – proposed by Wuerz et al. (2000), which combines urgency with an estimate of resource requirements – has become the most common algorithm (Fernandes et al., 2005; Chonde et al., 2013). A five-level scale is better than a three-level scale according to Wang (2004), where a queueing system that models high risk patients concludes that patients should be split into as many classes as possible. FitzGerald et al. (2010) revisit past efforts and evaluates the future direction of triage, concluding that the five-level scale is now firmly established.

Reviews studying the effects of triage have been as extensive as the implementation of triage has been global. Outside of the aforementioned countries, we see triage employed at 97% of EDs in Switzerland in 2010 (Farrokhnia and Goransson, 2011), an innovative triage in South Africa (Gottschalk et al., 2006), and modifications to established systems in Portugal (Martins et al., 2009) and the Netherlands (van der Wulp et al., 2008). Göransson et al. (2005) provides a review of the use of triage in a fair majority of Swedish EDs. Fernandes et al. (1999) conducts a study that finds ED triage reliable when employed under the proper circumstances, while providing several sources that suggest mixed results with triage. Harding et al. (2011) provides a non-technical review to study the overall effect of triage on patient flow. Results from different studies were also mixed, suggesting a triage system tailored to the patient mix may be necessary.

Although traditional triage uses a nurse to evaluate patients at triage, research has found the investment of a physician at triage to have a benefit to a combination of common performance metrics such as LOS, LWBS, and diversion levels (Partovi et al., 2001; Han et al., 2010; Russ et al., 2010). From an OR/OM perspective, the main trade-off is between (i) using the physician (an expensive resource) at triage who might be better used to treat patients in the rooms, (ii) gaining more accurate information upfront, and (iii) issuing discharge or appropriate tests early on. Traub et al. (2014a) performed a mechanistic analysis of the effects of a physician in triage finding (in a single facility study) that the overall reduction in LOS was a function of rapid discharge of low-acuity patients much more so than of placing orders for patients who were ultimately seen in the main ED by another physician. Triage has been found to help reduce LWBS rates and patient LOS, even in the midst of increased patient census (Chan et al., 2005; Sanchez et al., 2006; Ruohonen et al., 2006). One concern with this approach is that the benefit of decreased waiting times might be at the expense of quality, as significant inconsistencies in patient classification may be a byproduct of triage (Wuerz et al., 1998). Batt and Terwiesch (2012) study the phenomena of state-dependent service times as seen in human-paced service systems, transportation and telecommunications; they find that the use of triage reduces
service time in periods of crowding in the ED, although some thought must be given to the financial cost of triage.

A large number of Emergency Medicine studies note the high percentage of non-urgent patients who utilize the ED (see, e.g., Lowe and Bindman, 1997; Koziol-McLain et al., 2000; and Carret et al., 2009), and efforts have been undertaken to address this. Indeed, Derlet et al. (1995) assessed and referred more than 30,000 patients out of the ED, finding that nonemergency patients can be triaged out of the ED to relieve stress. Other work has been done validating the relocation of patients to this end (Washington et al., 2000; Derlet and Richards, 2008). At least one study cautions against this, however, as the lower acuity patients on the five-level scale who are being deferred actually make up some portion of admissions (Vertesi, 2004). Young et al. (1996) note that 5.5% of patients classified as non-urgent at triage were later admitted to the hospital.

2.2.2. Complexity-augmented triage

With global use of triage in the ED, research into the use of variables beyond a traditional triage scale is common. From a queueing perspective, if patients can be distinguished based on measures related to their service times, then prioritization algorithms such as shortest processing time first (that are used in many industries, including manufacturing) can improve performance metrics. Observing this, the complexity-augmented triage proposed by Saghaian et al. (2014) notes that an additional complexity evaluation at triage would only take a matter of seconds, but its benefit could be significant. Through simulation analysis calibrated with hospital data and various queueing models, Saghaian et al. (2014) show that complexity-augmented triage does indeed benefit ED performance, both for patient safety (measured by risk of adverse events) and operational efficiency (measured by LOS). Saghaian et al. (2014) also investigate several patient flow designs that can be utilized after the complexity-augmented triage is implemented. Ieraci et al. (2008) also assert that patient complexity should be factored into triage and streaming. Although only a single case is presented, their conclusions are backed with a marked improvement (a 58% reduction) in waiting time. Sprivulis (2004) takes research into complexity a step further, designating clear complexity groups for patients based on the number of procedures or investigations required, with an additional partitioning based on patient age. Evidence regarding the accuracy of classifying patients by complexity is positive, welcome news given mixed history regarding traditional acuity-based triage (Vance and Sprivulis, 2005).

2.2.3. Patient streaming

The innovation and implementation of patient streaming was pioneered in an Australian hospital, Flinders Medical Center. King et al. (2006) and Ben-Tovim et al. (2008) discuss restructuring patient flow in the ED of this hospital based on whether a patient will be discharged or admitted. This involved processing non-urgent patients in a First-In-First-Out (FIFO) manner and using traditional prioritization methods on admitted patients. We see an identical streaming methodology in Kinsman et al. (2008), resulting in continual reduction in LOS times for both patient types. Inspired by the work of King et al. (2006), Kelly et al. (2007) segregates patient flow. It is recognized that the admitted group of patients may have different barriers to overcome than the discharged group. Though it was accompanied with a resource reallocation effort, improvements were seen in some key metrics, namely ambulance turn away and waiting time (Kelly et al., 2007). In at least one case, patient streaming is seen to have significant benefit only to discharged patient metrics, although inpatient care is not adversely affected (O’Brien et al., 2006). As reported with patient complexity, staff are relatively accurate at predicting patient disposition (Kosowsky et al., 2001; Holdgate et al., 2007).

OR/OM contributions in patient streaming are similar to interventions in call centers, and more generally to resource pooling in applications where resources with noticeably different process times can be optimally partitioned to benefit a particular customer class (e.g., Whitt, 1999; Hu and Benjaafar, 2009). It is noted in Hu and Benjaafar (2009) that prioritization can be performed as an alternative to partitioning, though this would require reliable patient evaluation, which the literature suggests is difficult to achieve (Wuerz et al., 1998; Fernandes et al., 1999). Peck and Kim (2009) use a simulation with a fast track with nurses evaluating acuity and disposition, showing that patient waiting could be reduced by upwards of 50% with the use of streaming. Saghaian et al. (2012) provide detailed queueing-based analysis in the realm of disposition-based ED patient streaming where patients are sent to separate tracks based on a prediction of their disposition (admit or discharge) made by triage nurses. Comparing the two systems of physical patient streaming and traditional patient pooling in Saghaian et al. (2012) shows a strong advantage for patient pooling due to a low resource utilization in physical streaming flow designs known as the “anti-pooling effect.” Observing this, the study of Saghaian et al. (2012) develops a virtual streaming flow design, which does not require restructuring the ED, but one which vastly outperforms pooling flow designs. This suggests caution towards past successes on patient streaming in which resources are physically separated and introduces virtual streaming as a new paradigm that can effectively achieve the advantages of both streaming and resource pooling.

2.2.4. ED fast track

Fast track in the ED is a dedicated stream of resources to process lower acuity patients more quickly. Welch (2009) notes that a fast track dedicated for minor injuries has been a mainstay in EDs since the 1980s. Given that about 80% of ED visits are non-urgent, the use of an ED fast track lane is a great aid in serving lower acuity patients and reducing
overcrowding (Williams, 2006). Fast track implementation has found success in numerous ED environments, such as in an urban pediatric ED (Simon et al., 1996; Hampers et al., 1999), or a teaching hospital (Meislin et al., 1988).

Roche and Cochran (2007) apply a queueing methodology to eight EDs, implementing fast tracks with different patient acuity, volume, and anticipated LOS in an effort to minimize patients LWBS. Cochran and Roche (2009) split patients into two tracks by acuity level to reduce patient LWBS and increase ED effectiveness. O’Brien et al. (2006) experienced a 20.3% reduction in patient wait time with the implementation of a fast track. A simulation by García et al. (1999) investigated redistributing resources towards a fast track, reporting the potential to reduce LOS by 25% for lower acuity patients without negatively impacting other patients. Fernandes et al. (1997) also reports successful results in lowering LOS and LWBS for lower acuity patients. The implementation of a fast track in Considine et al. (2008) saw the percentage of patients discharged within 2 hours and 4 hours increase to 53% and 92%, respectively, up from 44% and 84%. Cooke et al. (2002) reduced the probability of a patient waiting more than an hour by 32%, down to 59.2%.

Given the overlap between triage and fast track efforts, we see research reporting improvements in performance metrics through the implementation of both, despite an increase in patient census (Sanchez et al., 2006; Kwa and Blake, 2008). Fast track benefit to LOS is ultimately aligned with a time-tested understanding in OR/OM of the benefits of processing time prioritization (Lawler and Moore, 1969; Davis and Patterson, 1975). In the rare case where performance metrics were not met (Nash et al., 2007), customer satisfaction improvements legitimized the implementation of the fast track. In fact, when surveyed, there were marked improvements in categories from LOS to overall patient satisfaction (Rodi et al., 2006). However, we warn that creating a fast track may also result in the “anti-pooling” effect discussed earlier for patient streaming mechanisms. Hence, careful analyses must be performed before creating a fast track to make sure resources are assigned to the fast track in an appropriate way.

2.2.5. Bed planning

As Saghafian et al. (2012, p. 1080) notes: “The most direct way to alleviate crowding and improve responsiveness [in the ED] is by adding resources. However, because this is also the most expensive approach, it is generally not the preferred option.” This sheds light on an important connection between OR/OM techniques and ED operations, as OR/OM techniques are widely used for resource allocation in various industries. Resource allocation is indeed a well-known problem in OR/OM with a long history (Karchere and Hoeber, 1953; Cooper, 1963; Arrow and Hurwicz, 1977). Richardson (2003) recognizes that the discussion regarding resources is not surrounding their addition, but rather the proper allocation. In the discussion of bed planning, most research has occurred outside of the ED, namely in the ICU. We refuse to omit these papers, however, because changes outside of the ED can have significant impact within the ED. For example, an increase of 20 beds (43%) in one ICU resulted in a 66% reduction in ambulance diversion hours and 9.7% reduction in ED LOS for admitted patients (McConnell et al., 2005). In fact, the lack of ICU beds seems to be a common bottleneck for patient flow (GAO, 2003; Burt and McCaig, 2006; Pham et al., 2006). However, a cross-sectional study of California hospitals found that the ICU was the bottleneck in only 34 of 181 hospitals (Allon et al., 2013). Thus, any conclusions relating to expanding bed capacity must be justified. Relative size of the ED and ICU must be considered prior to expanding bed capacity. A failed example of this is provided by Han et al. (2007), where an effort to expand the ED to improve ambulance diversion had no tangible positive impact.

First defining the issues with current bed planning practices and the necessity of proper bed management, we start with the established target occupancy level that has been in place for more than 25 years. Eighty-five percent occupancy is the standard target capacity for beds, being the minimum level to increase the number of hospital beds (Green, 2006). Green and Nguyen (2001) address the issue with using occupancy levels for bed planning, pointing out a number of issues. Occupancy levels are based on the number of certified beds, which may differ from the number of staffed beds. Occupancy is typically measured as the midnight census, generally the lowest level of the day. Lastly, occupancy levels are yearly averages and do not incorporate weekly or seasonal variations. Green (2003) and de Bruin et al. (2007) both critique traditional occupancy levels, with Green (2003) also detailing the severity of bed delay and the need to plan properly. To further stress the importance of bed planning, we highlight that the lack of staffed critical care beds is the number one reason for ambulance diversion (American Hospital Association, 2007). The issue of staffed beds illustrates the need to properly allocate staffing and beds together.

Having established the necessity of accounting for external factors in bed planning, as well as the woes of current practice, we now summarize work done to combat the issue. Queueing Theory (Kao and Tung, 1981; Worthington, 1987; Huang, 1995) and simulation (Goldman et al., 1968; Dumas, 1984) are established interventions in the realm of bed planning, and we will discuss them in more details in Section 3. Given the incorporation of variability present in these methods, we see them as the preferred approach of determining bed capacity and optimizing bed allocation. Green (2003) succeeds in this regard by determining bed capacity for an obstetrics and an ICU unit, where insufficient capacity exists 40% and 90% of the time, respectively. Modeling bed demand with a Poisson distribution for a pediatrics unit in Milne and Whitty (1995) provided more accurate results than formerly averaged data. Results from a simulation in Bagust et al. (1999) recommend to invest.
in a spare capacity of beds to minimize risk of bed shortage. Simulation is also used in Pines et al. (2011), where dynamic inpatient bed management reduces ED boarding times, providing significant financial benefit. Recent work has been done to identify bottlenecks and optimize bed allocation (de Bruin et al., 2007; Elbeyli and Krishnan, 2000). For example, Elbeyli and Krishnan (2000) study the allocation of inpatient beds to different departments to maximize the effect on patient times. A simulation model in Harper and Shahani (2002) allow the user to perform a “what if” analysis, balancing hospital bed capacity with refused admissions. Cochran and Bharti (2006) create a queueing network and perform a discrete event simulation (DES) to optimize hospital bed allocation, resulting in 10% increase in throughput. Cochran and Roche (2008) generate a decision making tool using Queueing Theory to optimize inpatient bed planning.

Lovejoy and Desmond (2011) propose a simple solution to the issue of bed congestion by using a method more practical for physicians, requiring less hard-to-gather data and ultimately delivering speedy results: Little’s Law. The analysis in Lovejoy and Desmond (2011) justifies the purchase of less costly dedicated beds for an observation unit to free up inpatient beds, thus relieving congestion upstream to the ED. A simple ratio method (Plati et al., 1996) frequently used to calculate nursing or bed requirements based on patient census fails to incorporate variability of arrivals and service times in decision making. These methods are effective relative to their ease-of-use, but some caution should be taken, given that they do not incorporate variability. This is addressed in Nguyen et al. (2005), where a new method outperforms the past ratio method by minimizing the mean and standard deviation of occupancy-based parameters.

From a less technical operations perspective, active management of hospital beds, a task or dedicated full time job typically performed by professionals and RNs, has had a positive impact on patient flow, as well as benefiting safety and satisfaction (Howell et al., 2008; Borders et al., 2009). These decisions performed by “planners” or “czars” typically involve reallocating beds between the ED and internal wards based on demand. For example, multiple ICU and ED assessments performed by physicians acting as bed managers in Howell et al. (2010) saw a 28% reduction in LOS for ED patients admitted into ICU or CCU. The use of a dedicated bed planner is frequently coupled with other improvements as well. A dedicated RN bed planner, as well as the implementation of a bed management database and streamlined ED-ICU communication lowered one hospital’s diversion hours by 63% (Hemphill and Nole, 2005). Deeper analysis into the use of a bed manager shows great potential, and the investment of resources into training staff for such a task may be a wise decision (Proudlove et al., 2003, 2007).

2.2.6. Staffing and scheduling

Personnel are typically the deciding factor in moving a patient through the ED effectively and efficiently. As personnel account for two thirds of a hospital’s entire expenses (Warner, 2006), the importance of appropriate staffing and scheduling cannot be overstated. The overall wait to see a physician in the ED increased to 30 minutes in 2004, up from 22 minutes in 1997 (Green, 2008). It is common to see patient LWBS rates above 6% due to physician unavailability (Ding et al., 2006). Also, patient discharge is often delayed because staff are tied up with more urgent patients (Kelly et al., 2007), indicating that staffing and scheduling have a widespread effect on all areas of the ED. Green et al. (2007, p. 34) address the common issue with staffing in healthcare: “Hospital managers, while aware of the variability over the day, have not used queueing models, but instead allocate staff based on general perceptions and intuition.” The lack of OR/OM driven decision making pointed out by Green et al. (2007) is also identified by Carter and Lapierre (2001) after interviewing physicians to determine their staffing methods. Similar to Beaulieu et al. (2000), their solution is to formulate a mathematical program that incorporates all the rules that should govern a physician’s schedule. Below, we identify additional literature that focuses on OR/OM related techniques for ED staffing and scheduling to improve patient flow.

Considering physicians, nurses and exam rooms as variables in a simulation, Duguay and Chetouane (2007) test numerous variable settings to improve key performance indicators. Adding a doctor and nurse during regular business hours was found to have the best impact on patient waiting. Two heuristic algorithms are used to staff physicians, nurses and technicians in Sinreich et al. (2012). The two algorithms developed efficient work schedules which reduced patient waiting between 20% and 64% and LOS between 7% and 29%. A linear optimization model is used in Sinreich and Jabali (2007) to find a resource’s contribution to ED operations, allowing for a reduction in LOS while also reducing staffing levels. Green et al. (2006) and Green et al. (2001) identify the variation of patient arrival through the day and use a Lag SIPP approach to create a weekday and weekend staffing model. The result in pursuing a demand-based staffing model was more than a 20% decrease in LWBS, despite an increase in patient volume. Yankovic and Green (2011) build a queueing model with nurses as servers to minimize delay probability through staffing, with the added benefit of a tool to identify overcrowding bottlenecks. Fullam (2002) presents an ED staffing success, where nurse staffing levels are determined by acuity data. In Patel and Vinson (2005), ED staff members are organized into teams consisting of one physician, two nurses and one technician in one suburban ED. The result of such a team assignment system was notable decreases in patient wait time and LWBS (Patel and Vinson, 2005). Traub et al. (2014b) study the rotational assignment of patients to physicians, finding a decrease in both LOS and LWBS. Increased satisfaction for patient waiting in DeBehnke and Decker (2002) further validates the use of patient care teams.
Finally, we note that several studies provide methods to forecast surges in ED volume, which can also be used to improve staffing and scheduling methods. For instance, Chase et al. (2012) consider the ratio of new patients requiring treatment over total physician capacity (a metric termed the care utilization ratio (CUR)), and finds it to be a robust and promising predictor.

### 2.3. Patient flow out of the ED

In this section, we focus on the final leg of ED patient flow: flow out of the ED. With a holistic systems approach, lean thinking has put focus on patient flow out of the ED so as to avoid congestion (Ben-Tovim et al., 2008). Khare et al. (2009) uses a simulation model and finds that reducing admitted patients boarding time in the ED was the biggest influence of reduced congestion. Schneider et al. (2001) confirms these findings, where rapid removal of inpatients from the ED was the greatest relief of overcrowding in their study. With this, we recognize that improving patient flow out of the ED is every bit as essential to ED operations as is patient arrival or flow within the ED.

#### 2.3.1. Patient discharge

Optimizing patient discharge is as crucial to improving patient flow as any other aspect of ED operations. It is shown that a lower admission-to-discharge ratio in the ED is crucial for a low LOS (Vermeulen et al., 2009). Fatovich and Hirsch (2003) identify improving the discharge process as one of five major steps for addressing ED overcrowding. Despite a scarcity of data, Black and Pearson (2002) recognize that delayed discharge is a serious issue. The use of a discharge lounge can free up a needed inpatient bed while patients ready for discharge have their prescriptions filled, wait for transportation, receive care education or schedule follow-up appointments (Williams, 2006). Geer and Smith (2004) represent a success story in this regard; the implementation of a discharge room was part of several process improvements that resulted in a 79% reduction in diversion hours. This is also backed by Moskop et al. (2009b), who even endorse taking discharge further by having a “reverse triage” system for early discharge of hospital inpatients. This innovative solution is originally proposed for inpatients by Kelen et al. (2009b), where a five-category scale is used to identify patients for early discharge. Although “triate at discharge” has been proposed as a crisis measure (Hick et al., 2009; Kelen et al., 2009a), Kelen et al. (2009b) note that such a system could be used in daily ED operations.

Powell et al. (2012) illustrate the need to take a system wide approach in the hospital, particularly in discharge. Peck et al. (2012) achieve a creative success in this regard where a generalized linear regression model is one of a few ways to accurately predict inpatient admissions based on information gathered at ED triage. Several models of improvements in inpatient discharge time were shown to have positive impact towards ED boarding. Kravet et al. (2007) take a similar approach, whereby discharging inpatients earlier ultimately reduced crowding. Unfortunately, there has been limited research on improving patient flow through discharge. In the medical field, Samuels-Kalow et al. (2012) review literature relating to patient quality at discharge, particularly how discharge information is received.

#### 2.3.2. Patient routing

Inspired by the success and simplicity of Erlang models in call centers, Armony et al. (2011) attempt to model routing from the ED to the inpatient ward with a series of time dependent processes. It is shown in Armony et al. (2011) that only 4.9% of patients were admitted to their internal ward within 30 minutes of being assigned. Lack of resource availability and poor routing decisions were identified as root causes, which may be optimized with queueing-based analysis.

Mandelbaum et al. (2012) introduce a RMI (randomized most-idle) policy to route from the ED to internal wards, achieving the same idle levels between server pools as LISF (longest-idle-server-first), without requiring idle times or pool capacity information. Given that patient flow out of the ED is greatly affected by staff unavailability (Kelly et al., 2007), abandoning an LISF policy will not force staff to spend additional time collecting patient data. Ultimately, however, research on routing has lacked an emphasis on the ED. Armony and Mandelbaum (2011) provide a case of routing with homogeneous impatient customers and heterogeneous servers for large service systems, which can be applied to the ED.

#### 2.3.3. Bed-block

“Bed-Block,” or the patient boarding phenomenon, relates to ED patients admitted to the hospital who are unable to be transferred out of the ED due to unavailability of inpatient beds. In the medical literature, the discussion regarding ED flow problems does not begin without mentioning boarding, especially since boarded patients block ED beds and prevent from seeing new patients in a smooth and timely manner. Derlet and Richards (2008, p. 24) express the significance of this issue, stating “boarding of inpatients in the ED is unquestionably the leading cause of crowding.” Decreasing boarding times has been found to be a major factor in reducing LOS (Khare et al., 2009; Moskop et al., 2009a). In fact, the number of patients who are boarded is such a significant problem that it is perhaps the most common trigger for ambulance diversion (Epstein and Tian, 2006; Ramirez et al., 2009; Allon et al., 2013). Pines et al. (2011) note that financial decisions further exacerbate the problem, enabling high boarding levels: hospital revenue is higher for non-ED admissions than for ED admissions, leading to non-ED admitted patients having a higher priority. ED success is largely gauged by the ability to manage boarding: the Institute of Health Improvement, et al.
challenges: Mathematical Programming, Queueing Theory, Simulation, Markov models (and Markov Decision Processes, MDPs), and Game Theory.

3.1. Mathematical programming and optimization

We start by reviewing related work using Location Theory techniques such as the Maximal Covering Location Problem (MCLP) and the Maximum Availability Location Problem (MALP). For the purposes of this paper, we only highlight work that immediately affects patient flow into the ED – primarily ambulance decisions. The Maximum Covering Location Problem (MCLP), introduced by Church and ReVelle (1974) and having undergone many early extensions (see, e.g., Daskin, 1982, 1983; Batta et al., 1989), seeks to maximize coverage within a certain distance by establishing fixed location points. Hogan and Revelle (1986) build on the MCLP by recognizing the need for secondary, or backup coverage. Murray et al. (2010) provide a recent enhancement of classic location problems. Genetic algorithms have been endorsed for their ability to arrive at near optimal solutions in a much more realistic time frame than integer programs or heuristics for MEXCLP (Maximum Expected Covering Location Problem) (Aytug and Saydam, 2002). MALP is structured as an integer program extending the MEXCLP (ReVelle and Hogan, 1989). It seeks to maximize the population that will be reached within a given time. As we will next briefly describe, these techniques have been instrumental in research related to transferring patients to EDs.

Multiple tabu search (TS) heuristics are frequently used to determine ambulance location (Adenso-Díaz and Rodriguez, 1997; Gendreau et al., 1997, 1999, 2001). One of these is a reactive TS that iterates through an algorithm to improve ambulance coverage, stopping after additional iterations would mean negligible improvement (Rajagopalan et al., 2011). Erdogan et al. (2010) use a TS not only for ambulance location, but also for crew scheduling, outperforming past literature on processing time and performance. The TS extends into the ED as well, where it is used by Gendreau et al. (2007) for physician scheduling. A comparison of genetic algorithms, simulated annealing and TS concluded that all three are robust, each with unique merits, though favor is expressed towards TS (Youssef et al., 2001; Arosteegui Jr. et al., 2006). TS is utilized to address the patient flow within the ED as well, where it helps set favorable physician schedules (Carter and Lapierre, 2001). For instance, Gendreau et al. (2007) compares mathematical programming, column generation, TS and constraint programming approaches for physician scheduling in the ED.

Mathematical programming tools are also used for ED staffing in other forms. For a recent example, we refer interested readers to Wang (2013), where a model based on separated continuous linear programming (SCLP) is used to provide high-level staffing guidelines. Karnon et al. (2009)
summarize several case studies about types of mathematical models used to improve health care. Beaulieu et al. (2000) take a mathematical programming approach to ED physician scheduling, where a large number of scheduling constraints are incorporated into a program to staff 20 physicians, ultimately outperforming the “expert scheduler.” Similar to Queueing Theory efforts, forecasting is often used along with mathematical programming to develop staffing and scheduling models that also incorporate variability of hour-by-hour patient arrivals. Mathematical programming tools have also been used to develop schedules for medical residents (see, e.g., Cohn et al., 2009) and cyclic schedules for ED physicians (see, e.g., Ferrand et al., 2011).

3.2. Queueing theory

Queueing Theory has had a prominent role in research related to patient flow optimization. Wang et al. (2013, p. 341) sum up the simplicity and efficiency of Queueing Theory: “Although analytical methods contain less details than simulation, and are based on simplified models, it could provide quick results and an opportunity to investigate system properties more efficiently under appropriate assumptions.” Fomundam and Herrmann (2007) and Lakshmi and Iyer (2013) review Queueing Theory applications in healthcare, ranging from a single department to the regional healthcare level. Wiler et al. (2011) reviews modeling applications concerning patient flow and crowding in the ED, with a dedicated section on Queueing Theory. Green (2006) details some basic queueing models with healthcare application, such as M/M/s, M/G/1, and G/G/s.

We begin with some interesting extensions of the M/M/s queue used in general (not necessarily ED) patient flow applications, where a queue with s servers follows a Poisson arrival distribution and exponential service distribution. Two interacting queueing networks are set up in Yankovic and Green (2011) with no blocking or balking. A closed queueing system is used in de Véricourt and Jennings (2011) for staffing purposes, where there is a finite population of n patients within the M/M/s/n model. Singer and Donoso (2008) use an M/G/s approximation to identify key performance indicators in ambulance services. Broyles and Cochran (2007) use an M/M/1/K to measure the financial impact of patient reneging. Patient balking with an M/M/1/K is also seen in Cochran and Broyles (2010). They point out that the accuracy coupled with the minimal data required for a queueing model makes it the preferred method for modeling over regression. An M/M/s/K addresses patient flow in Roche and Cochran (2007), where s is calculated by the arrival rate, length of service and desired bed utilization level. The use of M/M/s in bed planning throughout the hospital is detailed in Green and Nguyen (2001) and Green (2003), and within the ICU in Ridge et al. (1998) and Kim et al. (1999) (we point out later how bed management in the ICU affects ED operations). Extensions of M/M/s are also seen in Green et al. (2001), Green et al. (2006), and Green et al. (2007), where arrivals are time dependent and are based on a Lag SIPP (Stationary Independent Period by Period) approach. Au et al. (2009) use a six-hour moving average to represent time-dependency in an M/M/s model for predicting overflow. Au-Yeung et al. (2007) develop a queueing model with an approximate generating function analysis, designed to accommodate a larger state space than traditional models. This is modeled with a network of M/M/s queues where patients are identified by arrival and acuity. Au-Yeung et al. (2006) turn a network of M/M/s queues into a simulation model, although there is an expressed concern about the validity of some of the assumptions.

Infinite-server queues are also used to optimize the patient flow. For instance, an M/G/∞ model is used to analyze hospital bed allocation and minimize overflows in Kao and Tung (1981). Ultimately, the appropriateness and suitability of the type of the queueing model used depends on (i) the goal of the study, and (ii) the underlying assumptions made. Generally speaking, we observe four main deficiencies in the typical queueing theory models used for optimizing ED patient flow: (i) they ignore blocking issues in the ED, (ii) they assume stationary arrival and service processes while these processes are indeed non-stationary in reality, (iii) they ignore abandonment issues in the ED, and (iv) they assume state-independent service processes. Furthermore, as we will discuss in Section 4, human behavioral elements of service delivery are also widely ignored in current queueing models. In what follows, we overview papers that try to address these deficiencies.

Some papers have recognized that the patient queue needs to be modeled as a finite queue with blocking to be robust (Cochran and Bharti, 2006; Osorio and Bierlaire, 2009). Koizumi et al. (2005) incorporate blocking into a queueing model and extend upon the work of a traditional single server model by modeling with a M/M/s multi-server network. Brethauer et al. (2011) point out flaws in past blocking models, however, criticizing the use of an infinite queue capacity. A heuristic is used to predict blocking probabilities from which optimal capacity can be derived.

As discussed earlier, an important part of flow out of the ED is the bed-block phenomenon, which refers to situations in which ED patients who need to be hospitalized cannot be transferred to their inpatient units due to lack of bed availability. It prevents EDs from serving new patients in a timely manner, and results in longer Length of Stay (LOS) as well as a percentage of patients who Left Without Being Seen (LWBS). Some OR/OM studies such as Saghaian et al. (2012) consider the effect of the bed-block phenomenon on various patient flow optimization techniques including “virtual streaming.” Shi et al. (2013) provides a detailed study of issues related to altering discharge times in inpatient units, which directly affects the ED bed-block durations.

While the majority of queueing models assume time-stationarity, there is also some work that considers time dependency. Armony et al. (2011) study the ED as a
queueing model representing a section of a bigger queueing network, i.e., the whole hospital. Three Markovian queueing models are analyzed to fit ED occupancy along with a simulation framework. Armony et al. (2011) discuss that the time dependent (i.e., non-stationary) model, $M_t/M_t/\infty$, is only accurate when the ED is occupied with less than 15 patients. The state dependent model, $M_t/M_t/\infty$, in Armony et al. (2011) is found to be very accurate for a queueing model. These results are intuitive as the state dependent model incorporates important factors beyond time, such as ambulance diversion. Armony et al. (2011) then suggest modeling ED occupancy as a black-box birth and death process with state dependent arrival and service distributions. An $M_t/M_t/\infty$ is also used in de Bruin et al. (2007) to model flow of emergency cardiac patients. Time dependency is also captured in the modeling of clinical wards, where it is recognized that understanding variability outside of the ED is essential for capacity planning (Bekker and de Bruin, 2009).

Wiler et al. (2013) incorporate patient abandonment (LWBS) by using a $M/GI/r/s + GI$ model introduced by Whitt (2005), where patient arrival follows a Poisson distribution, service times are a general distribution i.i.d, with $r$ bed servers, $s$ waiting area capacity and abandonment times are a general distribution i.i.d. Batt and Terwiesch (2013) provide innovative work in patient waiting, noting that factors beyond waiting time affect abandonment. Among these are the observed number of patients waiting, the flow of patients in and out, and the inferred severity of waiting patients—all visual information acquired by a patient in the waiting room.

Cochran and Roche (2009) address many complexities that are often ignored in high-level queueing models, taking into account patient acuity, variation of arrival, and different resource consumption across several EDs. This is also seen in Roche and Cochran (2007) to test an “extreme” fast track in the ED. Queueing Theory has also been applied as an extension to the Maximum Availability Location Problem (MALP), where it is used to relax the assumption that server availability is independent (Marianov and ReVelle, 1996; Ghani, 2012). Huang (2013) and Huang et al. (2013) split patients into two queueing networks, new patients and WIP patients, to optimize physician decisions of which patients to service. Gallivan et al. (2002) simplify the patient flow process through a heavy-traffic deterministic system, assuming that the number of patients per day, the probability of “success,” and patient LOS is always the same. For some other studies regarding applications of queueing theory in the ED, we refer interested readers to Alavi-Moghaddam et al. (2012) and the references therein.

### 3.3. Simulation analysis

Simulation has provided strong decision making tools for ED operations even before global accessibility to computers and the development of widely available software. We see the use of an animated simulation that factors in random arrivals in Saunders et al. (1989), including individual service times and patient acuity, factors not incorporated into early ED queueing models. The ability to model processes in a great level of detail makes simulation a potential tool for virtually every aspect of the ED that impacts patient flow. There exists great depth in DES research in staffing and scheduling, fast track implementation, ambulance diversion, streaming, sequencing, performance tracking, and overall process improvement. It is the “what-if” analysis—the ability to test a high number of scenarios in a minimal amount of time—that have made simulation a widespread tool in the ED. Kolker (2008, p. 391) endorses the use of simulation above other methods: “Process model simulation approach seems to be much more flexible and versatile. It is free from assumptions of the particular type of the arrival process (Poisson or not), as well as the service time (exponential or not). The system structure (flow map) could be of any complexity, and custom action logic can be built in to mimic practically any features of the real system behavior.” While simulation is widely used in various ED patient flow studies, we note that as Günal and Pidd (2010) discuss, there is a lack of generality in simulation studies: the objective, scope, level of details, and calibrations performed vary considerably among such studies, making simulation more of a “case-by-case” approach rather than a generically available tool.

With an emphasis on patient flow, we identify a number of reviews that survey the use of simulation in the ED (Jun et al., 1999; Fone et al., 2004; White, 2005; Brailsford, 2007; Günal and Pidd, 2010). Paul et al. (2010) review over 90 papers focused on the investigation of ED crowding with simulation. Singreich and Marmor (2005) provide a walkthrough of how simulation can be used in the ED. How to define research objectives, gather and classify data, and validate the simulation are explained in a general manner to be applied to any ED. High-fidelity simulation analysis of ED patient flow calibrated with hospital data can also be found in studies such as Saghaﬁan et al. (2012), Saghaﬁan et al. (2014), and the references therein.

Simulation is a strong tool for ambulance services as it allows for robust models in the absence of parameters (Goldberg, 2004). In one case, three ambulance diversion policies are compared with a simulation, with one policy being based off of a MDP (Ramirez-Nafarrate et al., 2012). A number of alternatives are tested in order to reduce LOS in McGuire (1994) and patient waiting time in Komashie and Mousavi (2005). LOS is a common focal point of simulation, with the implementation of triage (Ruohononen et al., 2006), proper expansion of resources and a fast track (Samaha et al., 2003) or a combination of changes (Wang et al., 2012) providing an operational benefit. Macdonald et al. (2005) use simulation to test alternatives for how changes within the ED can improve LWBS rates and other performance indicators. The experimental power of simulation was able to test 21 alternatives on nine performance criteria, finding that different alternatives are optimal depending on the desired improvement. Hoot et al.
Markov models and Markov decision processes (MDPs) are common in ambulatory research to support data-driven decisions for ambulance location and deployment (Maxwell et al., 2010). For instance, the Hypercube model, used for decisions in the ambulance system, is based on a multidimensional Markov chain with multiple queues (Brandeau and Larson, 1986). In Ramirez-Nafarrate et al. (2012), an ambulance diversion policy is based on a MDP. The use of MDPs is simplified in Ramakrishnan et al. (2005), where continuous time Markov chains and discrete time Markov chains are used to model the ED and internal wards, respectively. MDPs are also extensively used in Saghafian et al. (2012) and Saghafian et al. (2014) (and some references therein) to gain insights into effective ED patient flow designs, triage and prioritization. Similar objectives are followed in Zayas-Caban et al. (2013), where MDP and sample path analyses are used to determine how ED patients should be dynamically prioritized. Overall, research involving Markov models present diverse application, with the ability to predict inpatient LOS (Kapadia et al., 2000), make admission decisions (Nunes et al., 2009), represent bed queues (Au et al., 2009), represent renegeting (Cochran and Broyles, 2010) and model patient flows (Davies and Davies, 1994; Wang et al., 2013).

3.5. Game theory

Game Theory, a study of rational decision making that has been widely used in the field of economics, has been mostly obscure in ED operations. In rare, scattered cases, Game Theory offers a supplemental tool to a number of decision based applications. We identify those limited contributions below.

The review provided by Brandeau and Chiu (1989) recognizes that Game Theory models have been integrated with location models as an additional application. Hagtvedt et al. (2009) and Deo and Gurvich (2011) model a hospital’s decision to go on ambulance diversion using Game Theory. Game Theory has also been proposed as an aid to measuring transport reliability (Bell, 2000). Mandelbaum et al. (2012) provide a unique perspective, suggesting that Game Theory could be used to model intricate decisions made by hospital staff such as decreasing or increasing service rates and/or the quality of care.

3.6. Summary and level of use of OR/OM tools

Table 1 summarizes the use of the OR/OM tools (discussed in the previous sections) in addressing the three components of ED flow. Areas indicated with “S” or “N/A” in this table also indicate opportunities for further research.
In recent years, OR/OM tools have been widely used to optimize ED patient flow, and this paper provides a comprehensive survey of such contributions. In order to demonstrate the potential impact of OR/OM tools in the ED, we classified operational improvements into three categories: flow into the ED, flow within the ED, and flow out of the ED. The range of papers identified that model patient flow speak to the breadth of work in OR/OM. We identified significant problems facing the ED such as ambulance diversion, triage, sequencing, streaming, resource planning, scheduling, staffing, discharge, routing and bed-block, and identified successes in combating these issues.

While OR/OM studies in the ED are ample, we note that there still seems to be a lack of implementation (see also the related discussion in Brailsford et al., 2009). This is mainly due to the low level of close collaboration between ED managers, hospital stakeholders, and OR/OM researchers. We believe involving the ED managers and other hospital stakeholders early on in the process of developing appropriate flow models can have a significant impact on the implementation of the results. Our view is also aligned with the recent President’s Council of Advisors on Science and Technology (PCAST) report (PCAST, 2014) that identifies various ways to improve health care delivery and lower the costs. Specifically, aligned with the several suggestions provided by PCAST, we believe that sharing lessons learned from successful OR/OM implementations, increasing the access of OR/OM researchers to appropriate data sets, better use of information technology and health analytics by system modelers, and training health professionals in using and even developing OR/OM approaches for their practice are important changes that can result in higher levels of implementation in the near future.

Our review also suggests that the term “operations research” may have different meanings depending on whether it is used in the Emergency Medicine literature or in the OR/OM literature. In the Emergency Medicine literature, operations research may refer to simple before-and-after studies or limited logistic regressions that show an improvement or decline in operational variables (such as length of stay) after an intervention, whereas in the OR/OM literature the term is usually employed to describe robustly controlled interventions suggested by rigorous mathematical models. In the Emergency Medicine literature, there

---

**Table 1.** The use of OR/OM tools in addressing challenges in ED patient flow optimization (“A”: Ample; “S”: Scarce; “N/A”: not available)

<table>
<thead>
<tr>
<th>Tool</th>
<th>Flow Into Flow</th>
<th>Flow Within Flow</th>
<th>Flow Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematical Programming</td>
<td>A</td>
<td>A</td>
<td>S</td>
</tr>
<tr>
<td>Queueing Theory</td>
<td>A</td>
<td>A</td>
<td>S</td>
</tr>
<tr>
<td>Simulation</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>Markov Models &amp; MDPs</td>
<td>A</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>Game Theory</td>
<td>S</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
is seldom a discussion of mathematical models or engineering principles; in the OR/OM literature, there is a premium placed on them. In the day-to-day management of EDs – which is done almost exclusively by nurses and physicians without the input of engineers – anecdotal evidence suggests that the mathematical models developed in the OR/OM literature are often downplayed in favor of perceived experience or an administrative gestalt of “how things work.” This disconnection, in our opinion, represents an extraordinary opportunity to improve ED operations through joint collaborations between OR/OM researchers and ED managers or other hospital stakeholders.

Successful OR/OM work in the ED is most prominent in the sections of ambulance management, fast track, patient streaming, bed planning, staffing and scheduling. Outside of these areas, depth of work is somewhat scarce. Acuity-based triage, a wide source of publication in the non-technical arena, has not seen abundant focus in OR/OM (with only a few exceptions such as Konrad et al., 2013). We look for more prominent use upcoming in the area of triage, especially with the recent successful innovations reported in both OR/OM and Emergency Medicine literatures in areas such as complexity-based triage (Saghafian et al., 2014), disposition-based triage and patient streaming (King et al., 2006; Kelly et al., 2007; Kinsman et al., 2008; Saghafian et al., 2012), and telemedicine-based triage (Traub et al., 2013).

Towards the back end of the ED service process, new processes involving reverse triage, while unestablished, show promise for future work. We believe focusing on the final part of ED flow, flow out of the ED, is an important research direction for future studies. In particular, there has not been enough focus on issues of outflow such as “bed block/access block” (Proudlove et al., 2003; Khare et al., 2009; Saghafian et al., 2012; Shi et al., 2013), and we believe that this should change in the near future, particularly as flow out of the ED is often the bottleneck, causing poor overall ED flow. With a shockingly low percentage of OR/OM work focusing on effective ways of moving patients out of the ED (see, e.g., Table 1), a paradigm shift should be imminent.

Process improvement methodologies with origins in automotive and electronics industry seek to further immerse themselves in healthcare. The use of Lean in healthcare, which was virtually nonexistent a decade ago, is beginning to become prominent (Jimmerson et al., 2005; King et al., 2006; Ben-Tovim et al., 2008; Decker and Stead, 2008; Dickson et al., 2009; Eller, 2009; Ng et al., 2010; Holden, 2011; Piggott et al., 2011). This has recently been coupled with Six Sigma to simultaneously eliminate waste and improve quality (Langabeer et al., 2009; Berwald et al., 2010). Both of these methodologies also stress the importance of a system-wide optimization approach that would greatly benefit ED operations. Hence, combining them with OR/OM techniques can be another fruitful path for research.

Strong correlation between ED and inpatient LOS suggests that continuing work in improving ED patient flow can have further impact on downstream operations. We also see in Kelen et al. (2001) and McConnell et al. (2005) how the addition of resources or focus on process improvement outside of the ED still has a benefit on traditional ED metrics. On a broader scale, decisions made by neighboring hospitals may also have a major impact on patient flow within the ED (Deo and Gurvich, 2011). The effect of providing better access to primary care on improving ED metrics is another research direction that deserves more study in the future. In general, having a comprehensive knowledge of the system beyond a narrow scope of the ED will improve operations across the board, and we expect to see more contributions from OR/OM researchers in this vein in the future. Additionally, successes in patient flow improvement processes need to be validated across multiple institutions. An overwhelming majority of process optimization research presents results on a single hospital. Results spanning multiple institutions such as Borders et al. (2009) show promise but lack geographic diversity. There is a definite need for research on multiple institutions in disparate geographic locations to gain broader insights on effective ED interventions.

We also hope to see further advancements in OR/OM tools that can better represent ED patient flow. One fruitful direction is to incorporate behavioral aspects in care delivery in OR/OM models, which are important for an accurate representation of patient flow but are currently largely overlooked. Future research may also continue to develop more advanced queueing models that better represent such complex care delivery and patient flow.

In the near term, we believe that information technology will help to move OR/OM forward in the ED. If “Big Data” in the ED refers to, inter alia, precise time stamps to record significant events (such as patient admission/discharge/transfer, patient movement through the ED, and order entry and completion), then Big Data will provide the raw materials needed to feed the theoretical constructs of OR/OM in order to develop appropriate solutions to extraordinarily complex problems. It is, therefore, essential to develop data-driven OR/OM techniques that can take advantage of this new opportunity.

In conclusion, we expect to see more contributions in the future from (i) close collaborations between OR/OM researchers and hospital stakeholders, (ii) innovative work for moving patients out of the ED, (iii) combining the process improvement methodologies with origins in automotive and electronics industry (e.g., Lean, Six Sigma, etc.) with OR/OM techniques, (iv) comprehensive views of the system (e.g., the effect of hospital inpatient units, neighboring hospitals, access to primary care, etc.), and (v) improved OR/OM patient flow models that benefit from data-driven and/or behavioral-driven approaches. If carefully developed and implemented, these can have significant impacts on ED operations.
Acknowledgments

The authors are grateful for invaluable comments and discussions provided by Wallace Hopp. The authors are also thankful for the helpful comments provided by the DE, anonymous AE, and referees.

Funding

This work was partially supported by Mayo Clinic through grant XSS0133.

References


