Improving the Performance of Emergency Departments: A Survey from an Operations Management Perspective

Agapitos DIAKOGIANNIS

Soutenu le 21 mars 2014

Encadrants :  
Oualid JOUNI  
Karim GHANES  
Zied JEMAI

Jurys :  
Vincent MOUSSEAU  
Wassila OUERDANE  
Oualid JOUNI  
Karim GHANES
Abstract

Emergency Department congestion is a crucial phenomenon that provokes serious problems in the performance of healthcare systems. Researchers have tried to bring into light the causes of these problems and aim to find solutions to this important issue. This survey summarizes the articles that propose ways to improve the key performance indicators related to the emergency departments with the use of operations research tools. The key performance indicators offer the ability to evaluate the system in a way that responds to the basic questions posed by the most significant stakeholders, such as healthcare policy makers, patients and employees. Operations research tools are divided into two categories: simulation and analytical methods (queuing and game theory, linear programming, statistical analysis, etc.), which differ mainly in their applicability in certain systems and in their ability to capture the data of a system. The articles studied either propose a method that ameliorates the performance of the ED (e.g. decreasing the waiting time of patients by a certain percentage or time period) or introduce a formula that can assist policy makers to optimize the quality of service of EDs in general.

*Keywords: Emergency Department, Key Performance Indicators, Operations Management*
# Contents

1. Abbreviations .................................................................................................................. 3
2. Introduction .......................................................................................................................... 4
3. Background-Emergency Department Modeling ................................................................. 5
4. Key Performance Indicators Analysis
   4.1 Length of Stay .................................................................................................................. 7
   4.2 Time to First Treatment ................................................................................................. 10
   4.3 Left Without Being Seen ............................................................................................. 13
   4.4 Ambulance Diversion .................................................................................................... 15
   4.5 Fairness .......................................................................................................................... 20
   4.6 Combination of Key Performance Indicators ............................................................... 22
5. Discussion ............................................................................................................................ 30
6. Conclusions & Future Work ............................................................................................... 32
7. References .......................................................................................................................... 35
1. Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD:</td>
<td>Ambulance Diversion</td>
</tr>
<tr>
<td>ED:</td>
<td>Emergency Department</td>
</tr>
<tr>
<td>IU:</td>
<td>Internal Unit or Internal Ward</td>
</tr>
<tr>
<td>KPI:</td>
<td>Key Performance Indicator</td>
</tr>
<tr>
<td>LOS:</td>
<td>Length of Stay</td>
</tr>
<tr>
<td>LWBS:</td>
<td>Left Without Being Seen</td>
</tr>
<tr>
<td>OR:</td>
<td>Operations Research</td>
</tr>
<tr>
<td>OM:</td>
<td>Operations Management</td>
</tr>
<tr>
<td>QoS:</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>TTFT:</td>
<td>Time to First Treatment or Time to First See or Waiting Time</td>
</tr>
</tbody>
</table>
2. Introduction

Studies focusing on the operation of healthcare systems are becoming more vigorous in the last years. The importance of a high service quality, as well as the tendency of a reduction of resources due to their high cost, motivate scientists to seek for alternatives that can give a solution to this significant problem.

One of the medical sectors that have been seriously damaged are the Emergency Departments (EDs). The inability of the ED resources to cope with the rapid changes observed in demand leads to high levels of congestion in the ED waiting rooms. Scarcity is observed both in resources able to change in the short term, such as nurses and physicians, as well as resources that are rather fixed, such as bed capacity. The above problem is caused because of the random arrivals of patients in the EDs as well as the random treatment duration; in an ideal world with perfect information about patients’ arrivals and treatment times, it would be feasible to allocate the exact amount of resources required for the above service.

This randomness renders operations management (OM) very useful for the analysis of EDs, as it is a scientific domain that is able to propose measures that ameliorate the management of random phenomena. This literature review focuses on the analysis of the Key Performance Indicators (KPIs) of EDs with the use of operations research (OR) tools, such as analytical methods and simulation. The applicability of the solutions proposed by simulations has motivated many researchers to use this tool for improving performance in the ED, with a useful guide for constructing such models being introduced by Sinreich and Marmor (2005). On the other hand, analytical methods, such as queueing theory, game theory, statistical analysis and mathematical programming models, have been less explored because of their complexity and their inability to capture all the elements of a certain system (i.e. ED). Nevertheless, the contribution of an analytical method in the research domain is much more important than the contribution of a simulation, because the first is able to express a general model that can be applied in any given system that has the basic characteristics of an ED.

The KPIs stand as a descent metric for the performance of an ED, as they count the most significant elements that characterize this service. Furthermore, these KPIs are a tool that can be easily understood by all interested parties, such as researchers, physicians, ED managers, etc.,
due to their simplicity. In more detail, papers analyzed in this survey either introduce a significant formula that assists in the optimization of ED services or propose a method that has a positive impact on the operation of EDs, both based on an improvement of one or more KPIs. The selection of these indicators was done in a manner that approaches the ED holistically and is able to respond to the concerns of various stakeholders, such as patients or ED policy makers.

The scientific domains that deal with the EDs are numerous and include a very wide range of sciences, such as medicine, psychology, architecture, finance, etc. Each domain is able to contribute in different aspects required in the EDs; doctors decide the way of the treatment, psychologists assist in the stress reduction of patients in the waiting rooms, architects design the EDs in a way that facilitates the patient’s path and finally financial analysts help ED managers raise the revenues of the medical center. Healthcare systems in general have been treated by many researchers with the use of OR tools (Brailsford and Vissels (2011)). In this survey, an emphasis will be given to the ED operations.

The rest of the paper is organized as follows. Section 2 describes the way that EDs operate and their background. Section 3 presents papers that analyze the KPIs selected for this review in detail. Section 4 discusses the results derived from this survey. Section 5 summarizes this study and indicates propositions of possible future work.

### 3. Background - ED Modelling

In this section the reader is able to understand the background of EDs, as well as the patient path in this medical unit. Furthermore, each KPI that is analyzed in the body of this survey is defined and explained briefly.

The profile of the patient in the EDs is also random, even though some people, such as insured or people of certain age groups, are more frequent users than the other ones (LaCalle and Rabin (2010)). Depending on the country, a difference in the treatment procedure between insured and uninsured patients may be observed, as well as discrimination on the basis of race or nationality (Heron et al. (2006)). It should be mentioned that children are served in a separate pediatric department.
The patient’s path in an ED begins from the process of triage, where in most cases a nurse diagnoses the severity of the situation. The patient is assigned a severity code and proceeds to the waiting room; the most severe incidents usually receive immediate treatment. The policy followed by physicians is mostly First Come First Served (FCFS), so whenever a physician becomes idle, the treatment of the first patient in the queue is initiated. Based on the diagnosis of the physician, the patient must pass through several examinations; waiting in the corresponding queues is generated. In the inter-space between examinations, physicians might have to check the results in order to gain additional information on the situation of the patient and sometimes ask him to perform additional examinations based on the results. At this point it should be mentioned that the treatment of a patient is usually performed by a certain physician, because understanding the exact situation of a patient is a time consuming process. After the completion of all examinations, the patient is either discharged from the hospital or admitted in an internal ward.

The above path is the one that determines the most significant KPIs in the EDs. The best known metric for the measurement of the performance is Length of Stay (LOS), as it refers to the time period spent by the patient in the ED. The LOS is a descent indicator for crowding in the system. A KPI that is very useful for patients that require immediate medical treatment is the Time To First Treatment (TTFT) or waiting time, used to count the time interval between the arrival in the ED and the first treatment by a physician. Crowding in the waiting room is the main reason for the Left Without Being Seen (LWBS), a metric that counts the percentage of patients that depart from the ED after the process of triage, probably because they estimate that their waiting time will be too lengthy. The above congestion might urge the medical staff to declare that arrivals must be reduced in order to be able to serve the patients in the waiting room, fact that leads to Ambulance Diversion. The above measures the hours that ambulances were signaled to seek for an alternative because of overcrowding in the specific department. Lastly, the metric of Fairness in the EDs, both from the perspective of clients (fair service policy), as well as from the perspective of employees (fair working environment) is analyzed.
4. KPI Analysis

In what follows, the main analysis of each KPI is performed separately, as well as a subsection of analysis of combinatoric studies (using more than one KPI). Each subsection commences with a small introduction for each KPI and continues with the list of relative literature. The analysis of each paper includes both basic elements, such as brief explanation of the method used, as well as reference to the main result of this study, which is either an improvement on a KPI, expressed usually in percentages, or the introduction of a formula that proposes a way of that optimizes the performance of the ED based on the KPI analyzed.

Before commencing the analysis of studies that suggest improvements for the KPIs analyzed, a short referral to studies that have dealt with the factors influencing the KPIs is done. These studies are very important in order to reveal the problems found in reality for each of the KPIs and inform researchers that are focusing on the operations management domain. There are many research papers dealing with the above, but this is not the subject of this study. The reader is referred to Sorup et al. (2013) for more articles on this field.

4.1 Length of Stay (LOS)

The majority of patients attending an ED are of medium severity, fact that means that their situation is neither negligible nor urgent; they do not need to see a doctor immediately, but they have to receive treatment before being discharged. Therefore the most important factor for them is the total sojourn in the ED, counted by the LOS, as they want to complete this (obligatory, but not critical) procedure as soon as possible. The importance of this KPI has brought into light some important facts concerning data that is related to LOS. Sometimes policy makers set a maximum limit of LOS in order to ‘satisfy’ the social demand for rapid service, with the most known example being the 4 hour target in the UK, which states that 98% of patients must be discharged, transferred or admitted in an IU within 4 hours (Mayhew and Smith (2008)). Based on the above, Izadi and Worthington (2012) have used queueing and simulation models to determine the staff requirements for achieving the 4 hour target. However, setting completion time targets might cause some inconvenience in the treatment procedure and downgrade the quality of service (Orr (2008)). Additionally, it seems that a very small
percentage of severe incidents have a multiple impact on the mean value of LOS observed in hospitals in general (LaCalle and Rabin (2010)), making it important to add medians and several percentiles in the statistical studies focusing on this KPI (Ding et al. (2010)).

Huang et al. (2012) study the control of patient flow in EDs with the use of queueing theory. Their goal is to assist the decision taken by physicians in case of high traffic in the ED; the physician must choose between patients that will be treated for the first time (triage) or the patients that have already seen a doctor and return to him after the completion of an examination (in-process). The objective function that is minimized in their model includes waiting costs that are directly correlated to the total LOS of a patient. The results of the study lead to a formula that shows the point where the decision of the physician must change: there is a threshold that determines whether IP-patients or triage patients will be served when the physician becomes idle. One should also mention that the study takes into consideration elements of advanced triage, as the prediction of whether a patient will be admitted (A) in the hospital or discharged (D) after the ED. The case study of the paper stands as a practical implementation of the first (theoretical) part. The scientists investigate in the difference of the objective function used for the queueing model of the first part when triage information varies in three different scales: no information, partial information (where only the number of IP phases is known) and full information (besides previous element, the triage notes whether a patient is A or D). The results of the three levels of information show that partial information ameliorates the objective function by 18%, whereas full information gives 27% better optimal solution.

Song et al. (2013) focus on the diseconomies of queue pooling in ED performance. The authors test their hypothesis that a dedicated queueing system can improve the ED LOS based on a sample of 231,081 patients of the Kaiser Permanente South Sacramento Medical Center’s ED. Their model intervenes in the traditional pooling based triage, where nurses assign a severity index to patients who afterwards queue in the waiting room (pool) and wait to be served by one of the physicians. Alternatively, they propose a model where the triage determines a-priori the physician that is going to serve a specific patient (stream). In both cases, priority is given to the most urgent incidents. The results of their study show that their proposition reduces the LOS of patients by 10.01%, or more practically reduces the LOS of a mean severity patient served by a mean performing physician by 32 minutes. A sensitivity analysis is also performed in this paper,
as a supplementary effort to determine whether any defect in the methods affects the final result. Several fixed variables, such as quality of care or rate of patients’ admission to the internal wards, are investigated, with the analysis proving that all differences are statistically insignificant.

Wang (2013) works on a separated continuous linear programming (SCLP) approach in ED staffing. Even though the final goal is to minimize an objective function that is expressed in terms of financial costs, the model proposes an alternative way to examine the LOS. The author divides the ED into 3 stages: the time spent in the waiting room, the period waiting for an examination and the time spent to see the physician again after the examination. Knowing that the time spent on treatment cannot be optimized and that the examination procedures are of fixed duration, the sum in the objective function demonstrates the variable constitutes of LOS that can be minimized. The constraints of the model secure that the flow equilibrium is maintained and that treatment and examination procedures occur only when staff is disposable. In the second phase of the paper, an additional optimization problem is proposed that determines the optimal staff number required in order to minimize the above mentioned costs.

Rossetti et al. (2004) use simulation in order to determine the optimal attending physician staffing schedules. Their goal is to find the schedule that minimizes the total LOS and their study is conducted in the ED of Virginia Medical Center, which has close to 60,000 visits per year. Four different scenarios of staffing were used for the comparative analysis:

1) ... determined by the ED manager (based on his experience and intuition).
2) ... determined by the arrival rate (data collection process of 1,175 patients).
3) ... adding an additional shift.
4) ... changing the schedule only in weekdays based on the above arrival rate.

The results indicate that scenario 2 is the optimal, as adapting to historical data seems to be a reliable solution. The simulation modeled indicates that by adding a physician in the peak hours (10 a.m. to 6 p.m.) is a strategy that reduces the ALOS by 14.5 minutes per patient.

Simulation is the domain of OR that can propose high impact solutions in the short term and thus it has been widely used for optimizing LOS. Some of the studies that use simulation for the improvement of LOS are: McGuire (1997), Samaha et al. (2003), Khare et al. (2009), Wang et al. (2012).
4.2 Time to First Treatment (TTFT)

In case of severe incidents, EDs must be able to respond immediately because even small amount of time is crucial. Guttmann et al. (2011) conclude that mortality is associated to waiting times in the EDs. Therefore, one can easily understand that TTFT is one of the most significant metrics when dealing with the performance of EDs. It should be also mentioned that there are several treatment procedures taking place after the initial examination of the patient by a physician, but this KPI focuses only on the time required for the first treatment of a patient by a physician.

Zayas-Caban et al. (2013) investigate in the optimal control of an emergency room triage and treatment process. Their model divides the procedure after triage into two stages: patients with severe indices enter phase 1, which includes examination tests and treating, whereas patients with less severe indices enter phase 2, where they have to queue with other patients for a physician in the waiting room. The hospital in which the study is implemented, Lutheran Medical Center in New York, applies the Treatment-Triage-and-Release (TTR) model, which proposes a triage performed by a physician. The authors use analytical methods, such as the queueing and dynamic programming methods, and simulation in order to validate their results. The model equations assist in deciding policies that lead to the optimal solution; in different cases (e.g. changing the rate of arrival of patients) alternative policies must be applied.

Cooke et al. (2009) introduce a separate stream for minor injuries on accident and ED waiting times. The retrospective analysis is based on a 10 week trial, 5 weeks examined with regular triage system and 5 weeks with the application of the new stream. The experiment is conducted in an ED in the UK and the data was driven from a sample of 13,606 patients. The proposition gives a positive result in all the time segments that were examined, and thus improving the scenario of waiting more than 60 minutes (time described as a critical bound) by a total of 32% in the 4 last weeks of the study; else put, they decrease the percentage of patients that had a TTFT more than an hour. A similar method was followed be Cochran and Roche (2009), who used a Split Patient Flow (SPF) in order to improve performance in EDs. The authors divide the ED into several stages and form a continuous time Markov chain with the corresponding transition matrices in order to determine the percentage of total time spent into each stage. The
study also includes demand forecasting elements that help them specify the peak periods in the ED. The way of optimizing the KPI is by setting a TTFT target and then trying to analyze the ED and see what operations (e.g. bed capacity in each stage) are required in order to achieve this goal. The two previous methodologies have been a subject of study for several articles that focus on the reduction of waiting time by changing some internal elements of an ED, such as Miró et al. (2003), Lau and Leung (1997) and Subash et al. (2004).

<table>
<thead>
<tr>
<th>Waiting less than...</th>
<th>Number of Patients</th>
<th>% of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>After</td>
</tr>
<tr>
<td>30 minutes</td>
<td>2,517</td>
<td>2,979</td>
</tr>
<tr>
<td>60 minutes</td>
<td>4,610</td>
<td>5,164</td>
</tr>
</tbody>
</table>

Lin et al. (2013) estimate the waiting time of multi-priority emergency patients with downstream blocking. The scientists include the LOS of patients in the Internal Unit (IU) in order to study the TTFT in the ED, constructing two queueing models for the patient flow in each of them. The queueing theory is verified in each step by a Monte Carlo simulation and uses data from a local hospital. The aim is to study the effect of increasing the resources on the ED and the IU while satisfying the upper bound time limits for the treatment of patients based on their acuity. The first outcome of this study states that whenever the LOS in the IU increases, it is preferable to boost the resources in the IU rather than in the ED, fact that can be seen in the difference of steepness of the plane in figure 1. Boosting the resources is also preferable whenever the IU LOS is fixed and the arrival of patients in the ED is considered as variable. Therefore, the study concludes that the TTFT strongly depends on the boarding effect, and thus boosting resources in the IU can be a manner of meeting the goals set for the waiting time of patients.
The study also includes the analysis of whether a fast-track has a positive effect on the TTFT. Even though the total TTFT of patients is reduced, the increase of waiting time for patients of higher severity renders this method as undesirable. However, the above explains the impact of TTFT outliers on the waiting times observed in general, meaning that the average TTFT is mostly due to a small percentage of patients that wait for a long period of time for their treatment.

Alavi-Moghaddam et al. (2012) study whether the application of queueing methods can assist in decreasing waiting times in EDs. Their study collects data from the Imam Hosein Hospital based in Tehran that accommodates annually 50,000 ED patients. Using the above data, they construct a simulation model in order to perform a sensibility analysis of critical factors that could affect the quality of service in the ED. The scientists test 8 different scenarios that include the increase of two types of resources: infrastructure (laboratory and consultation capacity) and staff capacity. Even though scenarios testing the increase of infrastructure can have a positive effect on the performance of the ED, it is rather difficult to implement them because infrastructure is considered as a fixed variable in the short or medium term. An interesting scenario manages to decrease the waiting time of patients from 26 to 18 minutes by adding a nurse to take the electro-cardio gram in each of the 12-hours shift, result that was statistically significant for P<0.05. In general, simulation is a very frequent tool for improving TTFT. Several examples of
studies using simulation for reducing waiting time are: Laskowski et al. (2009), Connelly and Bair (2004), Duguay and Chetouane (2007), Ajamiet al. (2011) and Sinreich et al. (2012).

4.3 Left without being seen (LWBS)

In several visits in the EDs, patients decide that they are no longer willing to wait for their treatment, probably because they estimate that waiting time is too long for their standards. These standards differ between countries, regions, races, social strata, age levels, etc. and have been subject of investigation for many researchers. Nevertheless, it is very difficult to measure the exact time that each patient waits before leaving, because the ED realizes that he has left only when he is called by a physician; there is no indicator that states when a patient leaves from the waiting room and therefore the exact time of his sojourn remains unknown. The data concerning the time of waiting before leaving that is known is the range of the sojourn; the recorded time before abandonment for each patient stands as an upper bound for the actual time spent in the waiting room. This KPI counts the percentage of patients that decided to depart before the seeing a physician at all. There are other metrics that count the percentage of patients that decide to abandon the system while waiting for the results of an examination (treatment has already begun). However, the LWBS is selected for analysis because patients that have seen a doctor sometimes leave the system because the supplementary studies needed are performed for preventive purposes, which are of minor importance. The following table shows some of the studies focusing on data related to the LWBS.
Table 2. Some data on LWBS.

<table>
<thead>
<tr>
<th>% of LWBS</th>
<th>Sample</th>
<th>Country</th>
<th>Year</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.00</td>
<td>USA</td>
<td>1991</td>
<td>Bindman et al.</td>
</tr>
<tr>
<td>2</td>
<td>4.20</td>
<td>USA</td>
<td>1994</td>
<td>Stock et al.</td>
</tr>
<tr>
<td>3</td>
<td>1.40</td>
<td>Canada</td>
<td>1994</td>
<td>Fernandes et al.</td>
</tr>
<tr>
<td>4</td>
<td>0.10</td>
<td>Taiwan</td>
<td>2001</td>
<td>Liao et al.</td>
</tr>
<tr>
<td>5</td>
<td>0.83</td>
<td>USA</td>
<td>2003</td>
<td>Arendt et al.</td>
</tr>
<tr>
<td>6</td>
<td>3.37</td>
<td>USA</td>
<td>2004</td>
<td>McMullan and Vesser</td>
</tr>
<tr>
<td>7</td>
<td>7.20</td>
<td>UK</td>
<td>2005</td>
<td>Goodacre and Webster</td>
</tr>
<tr>
<td>8</td>
<td>3.57</td>
<td>Canada</td>
<td>2005</td>
<td>Monzon et al.</td>
</tr>
<tr>
<td>9</td>
<td>9.00</td>
<td>USA</td>
<td>2006</td>
<td>Vieth and Rhodes</td>
</tr>
<tr>
<td>10</td>
<td>1.10</td>
<td>USA</td>
<td>2006</td>
<td>Johnson et al.</td>
</tr>
<tr>
<td>11</td>
<td>4.50</td>
<td>Canada</td>
<td>2006</td>
<td>Rowe et al.</td>
</tr>
<tr>
<td>12</td>
<td>8.60</td>
<td>Australia</td>
<td>2007</td>
<td>Mohsin et al.</td>
</tr>
<tr>
<td>13</td>
<td>1.70</td>
<td>USA</td>
<td>2009</td>
<td>Pham et al.</td>
</tr>
<tr>
<td>14</td>
<td>4.43</td>
<td>USA</td>
<td>2011</td>
<td>Guttmann et al.</td>
</tr>
<tr>
<td>15</td>
<td>11.23</td>
<td>Australia</td>
<td>2012</td>
<td>Tropea et al.</td>
</tr>
<tr>
<td>16</td>
<td>5.70</td>
<td>Guyana</td>
<td>2013</td>
<td>Parekh et al.</td>
</tr>
<tr>
<td>17</td>
<td>4.18</td>
<td>Switzerland</td>
<td>2013</td>
<td>Grosgrin et al.</td>
</tr>
<tr>
<td>18</td>
<td>13.12</td>
<td>Pakistan</td>
<td>2013</td>
<td>Fayyaz et al.</td>
</tr>
</tbody>
</table>

The study of Baker et al. (1991) explains why LWBS is a very significant KPI for EDs. Among others, their study concludes that about half (46%) of the patients that LWBS were judged to require immediate medical attention and about a third (29%) of the above patients would require medical care within one or two days. Therefore, it is of high importance for EDs to manage to reduce the percentage of patients that LWBS.

The majority of studies focus on the statistical analysis of factors influencing the LWBS, which serve basically for informative purposes. However, several studies try to find solutions to the problems caused by contacting the patients that LWBS. Skaikh et al. (2012) investigate on how long patients that LWBS would be willing to wait for their treatment before leaving. The results show the percentages of patients that admitted they would wait for certain time limits, with the most interesting result being the fact that half the patients were willing to wait up to two hours. Arendt et al. (2003) study the factors that would prevent patients from leaving. Results showed that about 85% of LWBS would like to be updated more frequently on how long they should wait and 7 out of 10 stated that they would prefer immediate temporary treatments.
Green et al. (2006) studies the schedule of staffing in order to increase effectiveness in EDs. The authors analyze the performance of an urban hospital in New York that serves about 25,000 ED patients per year. After examining the performance of the ED during a period of 39 weeks, the scientists propose an alternative staff scheduling based on the results of the queueing model used to simulate the ED. The alternative proposition of their analysis is implemented in the same hospital for the same period of time (39 weeks) and the results were compared to the first ones. Despite the fact that the number of admissions increase between the two time periods by 1,078 patients, the number of patients that LWBS decreased by 258 units. One should also mention that the proposition of the authors also includes an increase in staff provider hours of 12 hours per week. However, the fact that admission of patients increase more than the working hours of the hospital staff (6.3% compared to 3.1%) validates the model of the authors that manages to decrease the KPI examined by 22.9%.

Figure 2. Before and after provider staffing (weekdays). (Green et al., 2006)

4.4 Ambulance Diversion (AD)

The high traffic phenomena in EDs cause high congestion levels that might urge medical staff to apply ambulance diversion policies. However, the above is not the only blocking caused in EDs:
whenever patients occupying a bed in EDs have to be transferred to an internal ward of the hospital and this ward is already fully occupied, patients remain idle in the ED and still utilize part of the resources (such as beds). This phenomenon is known as boarding and provokes a blocking situation in the ED, which might also end up to diversion policies as well. Nevertheless, a big increase in the rate of arrivals of patients in the ED without any boarding phenomena can be a sufficient reason for AD.

The importance of AD can be understood by examining an example of a patient of high severity that requires immediate treatment. The time period needed for his treatment is divided in two categories: the time spent in the ambulance for his transportation to the ED and the TTFT. If the hospital that the patient is closest to is applying AD at that moment, then the crucial time period of transportation will be increased; this increase depends on the position of the next best alternative. Even though it is not part of the operations that are done in the ED, the placement of ambulances in order to minimize the response time to calls of patients is an interesting domain that has been extensively studied with the use of OR tools. Some of the relative literature is the following: Budge et al. (2009), Brotcorne et al. (2003), Marianov et al. (1996) and Borrás and Pastor (2002).

Vilke et al. (2004) make an approach to decrease ED ambulance diversion hours. Their study focuses on two neighboring hospitals that serve in total about 84,000 patients per year and are located in San Diego, California. The fact that these two ED facilities are the only ones in a radius of 5 miles leads to the hypothesis that there can be a correlation between their ambulance diversion hours. In fact, the authors observe that whenever Hospital A goes on diversion, Hospital B diverts ambulances after a small period of time. Therefore, the study focuses on the performance of each hospital for 3 weeks, where in the 1st and 3rd weeks regular resources are used, whereas the 2nd week Hospital A works with supplementary resources in order to eliminate ambulance diversion phenomena. Even though Hospital B works with the same resources as in the regular weeks, the results show that both EDs manage to reduce the hours of ambulance diversion.

Table 3. Results of Vilke et al. (2004).

<table>
<thead>
<tr>
<th>Ambulance Diversion (Hours)</th>
<th>Hospital A</th>
<th>Hospital B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular</td>
<td>19.6</td>
<td>27.7</td>
</tr>
<tr>
<td>Study</td>
<td>1.4</td>
<td>0</td>
</tr>
</tbody>
</table>
Allon et al. (2009) study the impact of size and occupancy of hospital on the extent of AD. Their study includes both the use of analytical methods and simulation in order both to derive results that are applicable and to be able to validate the theoretical part of their investigation. The initial queueing approach is depicted in the following figure.

Figure 3. A two-station queueing model (Allon et al., 2013)

\[ \lambda_{a} (w) \]: rate of arrival of ambulance patients (walk-in patients)
\[ p_{a} (w) \]: rate of ambulance patients (walk-in patients) admitted to the hospital’s IP
\[ N_{1} \]: number of patients in the IP (ED)

After simplifying the queueing model, the authors come up with the following results:

1) The capacity of the IP has a negative correlation with the time period that the ED diverts ambulances.

2) The threshold of non-used beds at which the ED applies the AD policy is positively correlated to the fraction of time spent on diversion. (This threshold signifies that whenever the ED has \( X \) units of beds that are not utilized, it will divert ambulances).

The 2nd proposition is obviously contradictory, as increasing the threshold of unused beds would harm other performance indicators of the ED, such as the LOS and the TTFT, and this is why the study includes the measure of delay probability, as well. The use of simulation uses data for
arrival and discharge rates given by federal hospitals in California. The table showing the difference of the results between the analytical method and the simulation also includes the number of IP beds (N) and the intensity of the ED ($p_2$).

Table 4. Estimation of fraction of time on diversion: approximation vs simulation (Allon et al., 2013)

<table>
<thead>
<tr>
<th>N</th>
<th>$p_2$</th>
<th>Approximation</th>
<th>Simulation</th>
<th>Difference (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>0.98</td>
<td>0.103</td>
<td>0.107</td>
<td>3.44</td>
</tr>
<tr>
<td>40</td>
<td>0.86</td>
<td>0.052</td>
<td>0.055</td>
<td>4.61</td>
</tr>
<tr>
<td>45</td>
<td>0.76</td>
<td>0.024</td>
<td>0.0258</td>
<td>5.59</td>
</tr>
<tr>
<td>50</td>
<td>0.69</td>
<td>0.0104</td>
<td>0.0115</td>
<td>8.86</td>
</tr>
<tr>
<td>55</td>
<td>0.62</td>
<td>0.00423</td>
<td>0.0049</td>
<td>13.5</td>
</tr>
<tr>
<td>60</td>
<td>0.57</td>
<td>0.00163</td>
<td>0.0019</td>
<td>14.1</td>
</tr>
</tbody>
</table>

Finally the study contains a statistical analysis that counts diversion hours in hospitals in California based on different characteristics (e.g. whether a hospital is rural or urban, trauma or not, etc.).

Broyles and Cochran (2011) investigate in a queuing base statistical approximation of hospital ED boarding. The authors model the ED and IP wards as a two multi-server continuous time Markov chain in series, as shown in the following figure:

Figure 4. The ED and IP displayed as two queues in series (Broyles and Cochran, 2011)

- $\lambda_{ED}$: rate of arrival of patients in the ED (IP)
- $f_A \cdot \lambda_{ED}$: rate of admission from the ED to the IP
- $\lambda_{DA} + f_A \cdot \lambda_{ED}$: rate of discharge of patients from the IP
The application of the model is performed with the use of data (the above rates as well as capacity) given by a large hospital in Arizona. The results show that there is a strong correlation between ED boarding time and the IP LOS for patients. This means that even a small reduction in the IP LOS can provoke a great reduction in the ED boarding time.

Figure 5. % reduction of ED boarding time (Broyles and Cochran, 2011)

Studies that do not deal with internal operations in an ED but strongly affect the AD metric for each of them have also been developed. The two following articles describe the situation prevailing in the decision making policies of EDs and propose methods that can improve the final outcome.

Hagtvedt et al. (2009) focus on cooperative strategies as a manner of reducing AD. Their study is divided in three approaches, using a Markov chain (birth-death process), a simulation and a game theory respectively. The authors discuss that the first two parts of the study are unlike to be applicable in reality because of the ethical issues that arise (such as patient discrimination). Nevertheless, the game theory explains how cooperation between hospitals can be achieved. The prisoner’s dilemma for only two hospitals shows that each hospital will try to divert ambulances before it would have actually needed it. This change stems from the pressure that
each hospital feels from its rival one and therefore they are forced to solutions that are suboptimal. Subsequently, it is natural that a larger number of N hospitals will face even greater difficulties when trying to manage the AD issue. The authors suggest that the cost of diverting ambulances should be adequately big in order to promote cooperation between different EDs.

Deo and Gurvich (2011) compare centralized with decentralized AD from a network perspective. The authors use queueing and game theory in order to develop a model that explains the difference between the two methods concerning AD. It should also be mentioned that hospitals apply AD whenever at least K inpatient beds are occupied. In the decentralized method, each ED aims to maximize its own utility function, where in this case the optimization refers to the reduction of waiting time for each ED separately. The game theory shows that the optimal solution in this way is to set K=0 for each ED, meaning that they would signal that they are always on diversion. However, legislation has set the ADND (All on Diversion, Nobody on Diversion) guide, which underlines that whenever all hospitals are on diversion, then the initial itinerary of each ambulance will be maintained. In this way, the EDs try to optimize their utility function separately, preventing the beneficial pooling that AD could lead to. Contrariwise, the centralized method analyzed is the one that a social planner makes the decision of when each ED should go on diversion and thus optimizes a holistic utility function. This method operates with the benefit of pooling and the queueing and game theory models analyzed help the planner decide the optimal policy. Results of a numerical example where thresholds were equal in the two hospitals examined show that there is an improvement in the waiting times of both EDs.

4.5 Fairness

The quality of the environment in which an employee works and a client is served always affects the outcome of the service offered. In the case of EDs, the ability to secure fairness between employees (medical staff) might stand as an alternative way to improve the QoS in the EDs. Fairness has been extensively studied from the perspective of patients, with the analysis of possible triage methods, such as the Canadian Triage and Acuity Scale (CTAS), the Manchester Triage System, Emergency Severity Index (ESI), etc. In the above cases, fairness is based on the logic that the most severe cases must be seen first, because otherwise they might be in danger.
However, whenever there is no life in danger, it is preferred to serve patients of lower acuity first, because they require less time for their treatment. Nevertheless, patients might try to beat a queue by asking a friend who is working in the hospital (Friedman et al. (2007)) to prioritize them, and thus violate the fairness criteria.

Even though fairness is a priori a KPI that is not as concrete as the previous ones, it has been used in this survey because it is believed to be an alternative method that can propose crucial solutions in the future concerning ED performance.

Mandelbaum et al. (2012) study the fair routing of patients from EDs to Internal Wards (IW), based on data given by a large Israeli hospital that serves approximately 75,000 ED patients per year. The incentive for this study is provoked by data showing that one of the five wards of the hospital was experiencing a very high patient per bed ratio compared to the other four wards, a phenomenon that was not only experienced in this particular hospital, but in other hospitals as well. Therefore, the authors of this paper use models that are able to restore a more fair working environment for the medical staff. The objective of this study is to minimize the deviation of work rate between employees. Queueing theory methods are used in order to derive a model that shows how fairness is achieved for the medical staff. The deviation of work rate stems from the fact that some servers (employees) are more efficient than their colleagues. However, higher efficiency means that at the same time period, the more efficient servers are able to serve more clients compared to the less efficient servers, creating the unfair environment described above. Using the randomized-most-idle (RMI) routing policy, the authors propose a formula that possesses two important elements:

1) The more efficient staff continue to serve more patients than the less efficient ones (in order to maintain an efficient system).
2) The more efficient staff is “rewarded” with more breaks during their work time (in order to create a fairer environment between medical staff).

The logical question that rises from the above is whether efficiency of the ED is harmed. The initial answer, taking into consideration only mathematical models with fixed service rates would be positive. However, without applying the model proposed, the medical staff would have no motive to reduce the service time, as this would not affect their utility function, which is
the lowest possible work. Therefore, applying this model would motivate the medical staff to work more efficiently because in this case they would be “rewarded” with more breaks. Consequently, in the long run the quality of the service would be improved, as long as factors such as the LOS would be reduced.

### 4.6 Combination of KPIs

The above KPIs depict some important factors that EDs must deal with. However, several studies investigate the EDs from a more complex point of view; they try to integrate various metrics in the same study. The above effort is very important for the domain of EDs, as it ensures a more holistic approach in the field.

Burströmet al. (2012) study how staff allocation in triage can affect the quality of service in EDs, focusing on the KPIs of LOS and TTFT. The authors perform a statistical analysis using data from 3 big Swedish EDs, summing up to a total of about 150,000 of patients. The three different triage models used are:

- **Physician-led team triage**
  
The physician is the head of a number of smaller teams that consist of a pair of a junior doctor and a nurse. The operating subgroups act based on a standard protocol.

- **Nurse/Emergency physician triage**
  
A nurse performs the triage procedure and an Emergency physician deals with the patient treatment.

- **Nurse/Junior triage**
  
The most commonly used process, where a nurse performs triage and the examination is performed by a junior physician.
The results of the statistical analysis of the above study show that Physician-led team triage outperforms the other alternatives. More specifically, TTFT was improved by 56.5% and 49.5% respectively, whereas LOS was improved by 15.7% and 3.1% respectively. The above calculations are based on the mean values of the results that are illustrated in the following figure:

Figure 7. Time to Physician and Length of Stay in three emergency departments. (Burström et al., 2012)
Even though results show a significant improvement on the two KPIs, the authors discuss the fact that the above might occur not only because of the change in the triage method, but also because of other factors that influence the outcome. For example, sex and age are two elements that differ in the three EDs studied and this might be a reason for the improved results.

Hoot et al. (2009) make an effort on forecasting ED crowding. The authors use simulation as a forecasting tool in order to predict several KPIs in a tertiary-care ED that serves more than 50,000 patients per year. The sample in their 3-month study adds up to 13,248 patients. Their aim is to conclude whether this simulation could supply them with valuable forecasts. The results of their experiment show that predictions are not that accurate for all of the indicators investigated. More specifically:

**Figure 8. Calibration of the Simulation in the Forecasting Operational Data** (Hoot et al., 2009)

<table>
<thead>
<tr>
<th></th>
<th>2 h Ahead</th>
<th>4 h Ahead</th>
<th>6 h Ahead</th>
<th>8 h Ahead</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waiting count (# of patients)</td>
<td>0.0 ± 4.5</td>
<td>0.9 ± 5.8</td>
<td>1.6 ± 6.5</td>
<td>2.2 ± 7.0</td>
</tr>
<tr>
<td>Waiting time (hours)</td>
<td>-0.1 ± 0.6</td>
<td>0.1 ± 0.9</td>
<td>0.3 ± 1.1</td>
<td>0.5 ± 1.3</td>
</tr>
<tr>
<td>Occupancy level (% of beds)</td>
<td>2.4 ± 9.6</td>
<td>2.5 ± 11.2</td>
<td>2.9 ± 12.1</td>
<td>3.3 ± 12.9</td>
</tr>
<tr>
<td>Length of stay (hours)</td>
<td>-0.7 ± 1.0</td>
<td>-0.8 ± 1.3</td>
<td>-0.8 ± 1.5</td>
<td>-0.8 ± 1.6</td>
</tr>
<tr>
<td>Boarding count (# of patients)</td>
<td>0.3 ± 2.5</td>
<td>0.2 ± 3.1</td>
<td>0.1 ± 3.6</td>
<td>0.1 ± 3.9</td>
</tr>
<tr>
<td>Boarding time (hours)</td>
<td>-6.6 ± 2.7</td>
<td>-7.3 ± 3.1</td>
<td>-7.7 ± 3.3</td>
<td>-7.8 ± 3.5</td>
</tr>
</tbody>
</table>

*The forecasting residuals are summarized with the mean ± standard deviation.

As shown in the above figure, boarding time forecasts are not accurate at all, as the model seems to predict less hours of boarding, probably because of a systematic bias. Nevertheless, predictions for the remaining indicators seem to be more reliable, especially for a prediction made up to 4 hours ahead. The following figure shows how the model predicts AD phenomena:

**Figure 9. Time series plot of the 6-hour forecast of the ambulance diversion status** (Hoot et al., 2009)

The grey parts represent the amount of time that AD occurred. The simulation is able to predict this phenomenon accurately 6 hours ahead, based on the peaks shown whenever a grey area is
seen. The ability of simulation to take into consideration several elements enables the authors to conclude that congestion phenomena might stem from different factors, such as an increased arrival rate of patients or large numbers of boarding patients.

In general, forecasting the demand in an ED has been the subject of interest for several studies. Scientists are trying to derive accurate models in order to assist the actual performance of the EDs. The reader is referred to some relative articles: Jones et al. (2008), Wargon et al. (2009), Channouf et al. (2007) and Mistry et al. (2008). Despite forecasting demand of patients, it is also important to adjust the supply of medical staff, such as the number of nurses and physicians required, to the corresponding demand. The reader is referred to several articles that are addressing the issue of optimizing staff scheduling using either simulation or analytical methods: Green et al. (2013), Yankovic and Green (2011), Al-Najjar and Ali (2011), Xiao et al. (2010).

Kelen et al. (2001) conduct an experiment in a hospital that serves about 54,000 patients per year. The authors introduce a supplementary acute care unit (ACU) that served the more severe incidents of the ED. Despite being hosted in a part of the IU of the hospital, this ACU was entirely supplied with resources of the ED. The interaction between the original ED and the ACU was multiple, as patients could be sent to the new department for primary evaluation, for admission processing etc. The study, which lasted for 10 weeks (in total 10,871 patients), uses as control variables two sets of data: one given by the operation of the hospital at the same month a year before the experiment and one given by data of the hospital in the last two weeks before the study. The results concerning LWBS and AD are shown in the following table:

<table>
<thead>
<tr>
<th></th>
<th>LWBS (%)</th>
<th>AD (hours/100 patients*week)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 year before</td>
<td>9.4</td>
<td>-</td>
</tr>
<tr>
<td>2 weeks before</td>
<td>10.1</td>
<td>6.7</td>
</tr>
<tr>
<td><strong>10 week experiment</strong></td>
<td><strong>5.0</strong></td>
<td><strong>2.8</strong></td>
</tr>
</tbody>
</table>

Saghafian et al. (2012) propose patient streaming as a mechanism for improving responsiveness in EDs. Their study is based on a very simple fact: the KPI of interest depends on the genre of the patient, because different types of patients might require different types of treatment. More specifically, the study focuses on the patients with ESI 2 & 3 (which form approximately the 80% of the patient body in their study case) and introduce a supplementary triage element: a
prediction of whether a patient will be admitted in the hospital or not after the ED. Consequently, they study whether this streaming technique has a positive result on the KPI of interest for each client of the system. Patients that will be admitted in the hospital (A patients) require a small TTFT, as safety is the most important factor for severe cases and they thus need immediate medical treatment. On the other hand, patients that will be discharged (D patients) after the treatment in the ED require a small LOS (for the patients in the first case the LOS in the ED will be a small percentage of their total LOS in the hospital so this is not an important metric for them). Furthermore, setting an “LOS” goal for this group of patients secures that the TTFT will also be small.

The scientists use both analytical methods and simulation in order to derive their results. Three types of groups are formed after triage:

1) Simple pooling, where all patients form one group of people waiting for treatment.
2) Streaming, where patients are separated into two groups, depending on the prediction of whether they will be admitted in hospital or not after the completion of the treatment in the ED.
3) Virtual streaming, which is the same with Streaming, without the physical constraints of the separate paths (for example available resources such as physicians or beds of one path can be used in order to serve the other path in case of high demand).

Despite the fact that pooling is more efficient than Streaming, Virtual Streaming seems to offer some very interesting characteristics that could improve responsiveness in EDs. In fact, Virtual Streaming is able to balance the need for low TTFT for A patients and low LOS for D patients in a better way than pooling can.
The above study possesses a very interesting sensitivity analysis that explains in which cases Virtual Streaming should be preferred to Pooling. The results of this analysis are summarized in the following figure.

Figure 11. ED patient flow design strategy based on key environmental characteristics of the ED. (Saghafian et. al, 2012)
Generally, there is a wide range of studies focusing on interventions in the triage of patients. An operation that has been widely used is the introduction of a fast-track for patients of lower severity. The main background behind the fast-tracks is that patients that will either not occupy resources at all, such as beds, or will occupy some resources for a small period of time, such as physicians (due to the simplicity of their treatment), should be not held in the waiting room for too long because of the complexity of the severe patients. In other words, this is an operation that aims to render the environment fairer for the majority of patients, by trying to eliminate the effect of severe incidents (outliers) on the performance of an ED. Some examples of studies that investigate in this operation are: Devkaran et al. (2009), Considine et al. (2008) Darrab et al. (2006) and O’Brien et al. (2006). The fast-track application has an effect mainly on three KPIs: LOS, TTFT and LWBS, all of which are improved.

Saghafian et al. (2013) focus on a complexity-augmented triage as tool that can improve patient safety and increase operational efficiency. The authors use both analytical methods and simulation using data from University of Michigan ED in order to validate their model. The study deals with the LOS as well as the risk of adverse events (ROAE) of patients and is separated in two phases, phase one being the waiting procedure between triage and treatment and phase two being the whole procedure of treatment. In terms of the KPIs used in this review, risk of adverse events of patients is related to the TTFT, because in phase one, while patients are waiting for their treatment, the ROAE is a function of TTFT and in phase two, while patients are waiting for the results of their examinations, the ROAE is reduced by 60% compared to phase 1, meaning that is of minor importance. The difference between ordinary triage and the method of triage proposed by the authors is shown in the following figure.
The queueing methods analyzed in the paper determine policies that should be implemented by physicians concerning the selection of which patient should be treated next in both phases (patients categorized as urgent (U) and non-urgent (N)).

The simulation of the study states that by applying the alternative triage method, the results show that there is an improvement to the ROAE and LOS by 18.0% and 21.3% respectively. The study also includes an optimization of the use of complex-augmented triage, in which patients are considered to be more complex when they have to follow more examination during the treatment process.

Summarizing, complex-augmented triage should be preferred when:
- complex patients are considered the ones requiring at least two examinations
- there is a high bed and/or staff utilization
5. Discussion

This section summarizes the key points of articles analyzed in this survey. The first part points out the recurrent data collection issues and the remaining part focuses on issues stated in the articles concerning OR, as well as several general significant outcomes observed in the survey of EDs.

One of the most important factors in studies is the quality of results. In the articles examined, the problem of data collection was often addressed, as it is difficult to collect data in such a complex system. For example, it is feasible to count the number of LWBS, as it is the difference between patients that were triaged and patients that were examined. However, it is rather difficult to collect data stating when these patients decided to leave the ED and for what reason. Furthermore, most of the articles had to exclude some data from the study due to medical reasons (for example psychiatric patients were excluded due to the fact that their requirements were highly unpredictable). Therefore, the sample of the studies was automatically reduced. In addition, the impact of outliers on KPIs was extensively referred in several articles. Studies that face outliers require deeper statistical analysis, as the use of mean value on its own is not sufficient. Accordingly, elements such as standard deviation, median and percentiles are required in order to have a more holistic view on the subject.

Operations research seems to be a domain that can assist the performance of EDs in a substantial manner. The studies presented in this literature review prove that scientists’ propositions are able to ameliorate the pressing issue of congestion and low quality service in most EDs. However, the optimization of a service is always limited by its constraints and in this case resources such as physical persons (nurses, physicians, etc.) and bed capacity seem to be the bottleneck; given the resources, optimization of a system can help by giving the best possible solution, but even this solution might not be acceptable. Therefore, besides the tools offered by operations research, EDs need to be equipped with sufficient resources as well. The following table shows which were the tools that were used in the studies examined in this literature review. The column “other” stands for tools that were not used so frequently, such as mathematical programming and game theories.
Table 7. OR tools of articles examined in the review.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Simulation</th>
<th>Queueing</th>
<th>Statistics</th>
<th>Other</th>
<th>Case Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alavi-Moghaddam et al. (2012)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allon et al. (2013)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Broyles &amp; Cochran (2011)</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Burstrom et al. (2012)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Cochran &amp; Roche (2009)</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Deo &amp; Gurvich (2011)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Cooke et al. (2013)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green et al. (2006)</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Hagtvedt et al. (2009)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hoot et al. (2009)</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huang et al. (2012)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kelen et al. (2001)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Lin et al. (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Mandelbaum et al. (2012)</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rossetti et al. (2004)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Saghafian et al. (2012)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saghafian et al. (2013)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Song et al. (2013)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vilke et al. (2004)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Wang (2013)</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zayas &amp; Caban et al. (2013)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

The methods of articles can be divided into two big categories: the ones that use simulation and the ones that use analytical methods. The difference between the two categories is based on several factors. Analytical models are very difficult to construct and require a set of hypotheses in order to finally derive a useful formula, whereas simulation models can capture every single detail of the system without requiring any hypothesis. In contrast, the results of each study have different applications; the results of a simulation are useful only for the system examined, whereas results of an analytical method can be applied in any given system. Consequently, one understands that the contribution of analytical methods in the research area is more significant, despite being less implementable.

An important fact for the above studies is the difference between theory and practice. From the scientist’s point of view, the lack of experience in the environment of EDs might render him unable to fully understand the system he is trying to optimize. Therefore, theoretical propositions might not be able to be applied in reality. On the other hand, the above problem might stem from the fact that medical staff try to optimize their personal utility when they
deliver the service to patients, a phenomenon known as “gatekeeping”. Despite the fact that they are instructed to act in a certain way, some employees may act in a different way, either because they trust their intuition more than the instructions given or because they want to serve patients in a way that is more profitable for them (profitable not only in terms of earnings).

Concerning the KPIs used, it very useful to address the relationship between each of them. LOS consists of the time spent by a patient in the waiting room (TTFT) and the time spent in the treatment process, fact that means that TTFT is a constituent of LOS. However, optimizing one of them doesn’t necessarily mean that there will be a positive effect on the other one as well. For example, an operation that reduces the TTFT by 10 minute might simultaneously increase the treatment procedure by 20 minutes, meaning that even though TTFT is reduced, LOS is actually increased. Concerning the KPIs of the previous example, the way that Sinreich et al. (2012) have approached these KPIs can stand as a descent pattern. Furthermore, it is increase patient safety by decreasing the LWBS, but it is obvious that this will have a negative impact on other performance metrics, such as LOS and AD.

In general, one can conclude that the ED is not an isolated system and strongly depends on its environment. From geographical point of view, the existence or not of hospitals close to an ED can have an effect on its performance, mainly because of the effect on AD. Even in the same hospital, the performance of the IU is a significant variable for the ED, causing serious problems such as patient boarding. Furthermore, policy makers, either internal such as ED managers or external such as government legislators, have an immediate impact on the performance of the ED. In addition, the performance of the ED also depends on the cultural environment; the attitude and the expectations of people waiting for treatment is not the same for all nationalities.

6. Conclusions & Future Work

ED congestion is a crucial phenomenon that provokes serious problems in the QoS of healthcare systems. Scientists have studied extensively the way of improving KPIs dealing with ED using the tool of OR which is able to optimize systems facing randomness. This survey has summarized the
articles that propose ways to improve the performance of KPIs related to the EDs with the use of OR. In the first parts of this review, some introductory information concerning EDs and their function is listed, as well as their functional background. The selection of KPIs was based on the interest of stakeholders that make decisions concerning the EDs. Articles that were analyzed in the core of the survey include at least one of the following: they either propose a way of treating the ED that ends up with a real case improvement in the performance of the ED or they propose a theoretical formula which can be used for the optimization of the ED performance. In the section of discussion, the main recurrent key points of these articles are stated as well as a significant table that shows the tools of OR that were used in the studies. In what follows, possible future work opportunities are proposed.

One of the most important elements in the healthcare services domain is to thoroughly understand what clients demand. Different types of patients demand different service, which means that some KPIs are more important for them than others. A possible solution to this problem would be to divide patients into control groups and try to satisfy them in different manners, because they would be easier to manage in this way.

A method that has been used in several studies is to forecast whether a patient will be discharged or admitted in an IU after his treatment in the ED. These studies propose solutions that are based on these predictions, fact that renders it very valuable. One possible way to investigate in the quality of prediction is to construct models with the use of multi-criteria decision making. Sometimes such predictions might be insecure and could finally have a negative effect on the performance of EDs. Multi-criteria decision making models will be able to determine when such information has a positive effect on the performance of the ED.

Focusing on each KPI, one can see that there is just a small number of articles that are trying to improve LWBS with the use of tools of OR, even though there are many statistical studies that focus on how one can collect data on this KPI, fact that signifies that there is space for supplementary research on this domain in the future. Furthermore, fairness seems to be a domain that is not yet explored, but can be an alternative way of improving a system. The pioneering approach of Mandelbaum et al. (2012) could be the basis of extended research on the KPI of fairness. KPIs such as LOS, TTFT and AD have been given much attention by scientists working on the domain. However, further improvements can be explored, as well. For example, EDs use different criteria in order to determine when they should divert ambulances; some are
based on the boarding phenomenon whereas other divert whenever they see that the rate of
arrival passes a certain limit. Therefore, studies that determine the optimal policy, such as in the
AD case, could be performed in the future. In certain cases, it is important for studies that do
not deal with operations in the ED to capture some elements from the ED. An example for the
above could be the studies focusing on the optimal location of ambulances that try to minimize
the response time to patient calls, as they should also take into consideration the effect of AD in
the total time period required for a patient to receive treatment in an ED.

Finally, one should underline the importance of combinatory studies that are able to capture all
the above KPIs. Sometimes KPIs are correlated, and thus improving one of those can improve
the other one as well (e.g. TTFT and LOS). Nevertheless, other KPIs might have a negative
correlation (e.g. improving LWBS might increase the LOS due to higher congestion). Therefore it
is very useful to examine combinations of KPIs so as to be able to determine which KPI is more
important and apply the relative method.

Acknowledgements

I would like to thank my supervisors, Oualid Jouini and Karim Ghanes, for their useful advice and
for their unlimited patience with my recurrent questions. I would also like to express my deep
appreciation to professor Georgios Saharidis, who believed in me and inspired me with his
professionalism and his interest in students’ progress. Finally, I wish to acknowledge the help
offered by my parents, Thaleia and Giannis, and foremost I want to thank specially Constantina,
who has been offering to me invaluable support on a daily basis.
7. References


