Evaluation of Response Capacity to Patient Attention Demand in an Emergency Department

Eva Bruballa
Tomàs Cerdà Computer Science School
Universitat Autònoma de Barcelona
Barcelona, Spain
eva.bruballa@eug.es

Alvaro Wong, Dolores Rexachs, Emilio Luque
Computer Architecture and Operating Systems Department
Universitat Autònoma de Barcelona
Barcelona, Spain
alvaro@caos.uab.es,
dolores.rexachs@uab.es,
emilio.luque@uab.es

Francisco Epelde
Short Stay Unit
Parc Taulí Hospital Universitari
Institut d'Investigació i Innovació
Parc Taulí I3PT
Universitat Autònoma de Barcelona
Sabadell, Spain
fepelde@tauli.cat

***Abstract*— The progressive growth of aging, increased life expectancy and a greater number of chronic diseases contribute significantly to the growing demand of emergency medical care, and thus, on saturation of Emergency Departments. This is one of the most important current problems in healthcare systems worldwide. This work proposes an analytical model to calculate the theoretical throughput of a particular sanitary staff configuration in a Hospital Emergency Department, which is, the number of patients it can attend per unit time given its composition. The analytical model validation is based on data generated by simulation of the real system, based on an agent based model of the system, which makes it possible to take into account different valid sanitary staff configurations and different number of patients entering the emergency service. In fact, we aim to evaluate the response capacity of an ED, specifically of doctors, nurses, admission and triage personnel, who make up a specific sanitary staff configuration, for any possible configuration, according to the patient flow throughout the service. It would not be possible to test the different possible situations in the real system and this is the main reason why we obtain the necessary information about the system performance for the validation of the model using a simulator as a sensor of the real system. The theoretical throughput is a measure of the response capacity to patient's attention in the system and, moreover, it will be a reference in order to make possible a model for planning the entry of non-critical patients into the service by its relocation in the current input pattern, which is an immediate future goal in our current research. This research offers the availability of relevant knowledge to the managers of the Emergency Departments to make decisions to improve the quality of the service in anticipation of the expected growing demand of the service in the very near future.**

Keywords—Emergency Department (ED); Agent-Based Modeling and Simulation (ABMS); Decision Support Systems (DSS); Response Capacity; Lenght of Stay (LoS); Knowledge Discovery.

I. INTRODUCTION

The current research focuses on the field of modeling and simulation of a Hospital Emergency Department (ED) and, specifically, on the use of simulation as a source of data for the extraction of information. This information, finally, must provide us with an extensive knowledge of the behavior of the system in any situation.

We proceed with the main objective of providing a methodology that allows the managers of an ED to be able to make decisions to improve the quality of the service provided to patients who use the service.

With this objective in mind, in a previous paper, we explained the idea of characterizing the system through an analytical model based on the definition of a set of indexes, indicators of its attention capacity and its performance, given different possible scenarios [1]. The given model in [1] presents some limitations, since it does not take into account all possible combinations of the healthcare staff, as it's already mentioned in the referenced article. The generalization of this model is presented here.

Currently, given the growing demand for emergency medical care, mostly due to the progressive growth of aging, increased life expectancy and greater number of chronic diseases, the management of EDs is increasingly important. Particularly, how to manage the increasing number of patients entering into the service is one of the most important problems in EDs worldwide, because it requires a substantial amount of human and material resources, which unfortunately are often too limited, as well as a high degree of coordination between them [2][3]. A major consequence of the increase in patients entering the service is its saturation [4]. This results in an increase in the total time a patient spends in the service, from their entry to their discharge, called *Length of Stay* of patients in the service (*LoS*). This can produce a general discontent among patients for reasons such as being abandoned without receiving care, limited access to emergency care and an increasing patient mortality [5].

Some studies in the related literature try to analyze the factors that influence patients' long periods of stay of in the ED and its saturation [6] [7]. Others show that saturation and long waits increase the proportion of patients who leave the service without being seen by a doctor (LWBS) [8] [9]. The aim of some others is to reduce the *LoS,* and therefore, the total time the patient is waiting to be attended, or length of

waiting for patients *(LoW)*, and some of the solutions that have been found and have been implemented are called Fast Tracks [10] [11], or other measures known as See and Treat [12]. Finally, we highlight those references using simulation to test the effectiveness of the proposed measures for improvement in the *LoS* of patients in the ED [13] - [17].

The ED service is one of the most complex areas of the hospital due to its dynamism and variability over time. The operation of the system is the result of the interaction between the different elements of which it is composed, and all this makes it a complex real system.

Modeling and simulation of complex real systems, such as an ED, is one of the most powerful tools available for their description. Simulation provides a better understanding of their operation and of the activity of their elements, and it can help decision-making to establish strategies for an optimal system operation [18][19].

The final objective of modeling and simulation of a real system is to find additional knowledge about it. This can be achieved by inference processes on the variables of interest of the system in order to make predictions about the behavior of these variables under different conditions, based on information obtained from the generated data [20].

As a result of an intensive previous research, we have an ED simulator available, based on an Agent-Based Modeling (ABM) design of the system, which has been developed, verified and validated within our research group, the "High Performance Computing for Efficient Applications and Simulation" Research Group (HPC4EAS) of the Universitat Autònoma de Barcelona (UAB), in collaboration with the ED Staff Team of the Short Stay Unit of Hospital de Sabadell (one of the most important hospitals in Spain, which provides care service to a catchment area of 500,000 people, and attends 160,000 patients per year in the ED). The model describes the ED's behavior from the actions and interactions between agents, and between them and their physical environment. The input parameters that characterize each different scenario in the simulation of the real system are the healthcare staff configuration, the number and type of incoming patients each hour, and the period of time simulated. As output, given that the most widely used and accepted parameter in the literature as an indicator of the quality of service is the total *LoS* of patients in the service, each simulation provides data of this index of all patients in all locations in the ED. In addition, the simulator includes sensors to obtain fully temporalized information about the agents, in such a way that data on the number of patients per hour and location are also available for each iteration. The implementation of the simulator has been done with NetLogo, an agent-based simulation environment well-suited for modeling complex systems [21][22].

An initial application of the simulator, with interesting results, was carried out by analyzing the effects of different derivation policies over the ED performance, particularly by analyzing how these changes modify the *LoS* of patients in the service [23].

Another study in the same research line consisted of trying to find the optimal healthcare staff configuration to minimize the *LoS* of the patients in the service, taking into account a constraint related to the cost of the configurations and the amount of available resources [24].

There are a great number and variety of simulated agents, and different possible values for the input parameters in the simulator. This results in a large number of different possible scenarios to be simulated. Thus, the use of High Performance Computing (HPC) was necessary in both experiments, due to the high number of executions required and the amount of data to be computed.

The main purpose of these previous researches was to provide some understanding of specific variables affecting the normal system performance. This could support decision-making (DSS), aiding the administrators and heads of the ED to choose the policies that could permit them to achieve a better quality of service with the available human and technical resources.

Our current work tries to obtain further and different knowledge concerning the performance of the system. We propose a model for system characterization with respect to the sanitary staff available configuration in it. It is an analytical model based on a set of equations that allow us to obtain the necessary information to obtain knowledge regarding the theoretical capacity to patient care of the system with respect to its staff resources, given a specific staff configuration and according with the patient flow in the system.

The content of the paper is organized as follows: Section II presents the research objectives and methodology of the research; Section III briefly describes the ED process and the simulation model; Section IV presents the analytical model proposed; and the experimental results for model validation are showed in Section V. Finally, Section VI closes the paper with discussion and future work.

II. Research Objectives and Methodology

It is a fact that saturation of the ED service is mostly due to admission of patients with lower acuity level. Based on historical real data from the Hospital de Sabadell, these patients represent a high percentage of the admitted patients and most of them are non-critical (see Figure 4 in Section III.B). We hypothesized that a redistribution of these non-critical patients in the input pattern initially planned by historical data (Figure 1), can lead to an improvement in waiting times for all patients, and therefore, to an improvement in the quality of service from the point of view of the users of the service, as it could avoid long waiting times in the service.

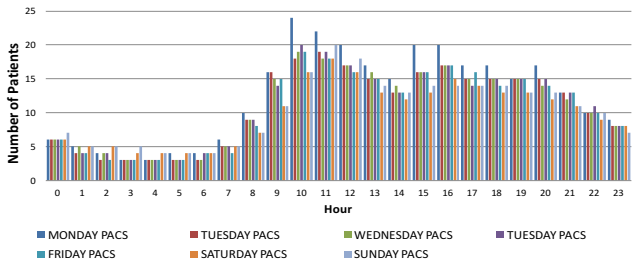

Figure 1. Input pattern of patients per hour and day of the week (historical data of 2014 of the Hospital de Sabadell).

In fact, the real starting point of this research was to understand the simulator as our main source of data. These data are the raw material for the analysis and they become information when we assign them some special meaning. When a model is found or designed, in order to interpret this information, and the model represents an added value, we refer to it as knowledge.

Moreover, simulation allows us to obtain data from situations, which cannot be proved in the real system, therefore any experimental limitation in the real system can be overcome through computer simulation. This idea suggested to us the hypothesis of the ability to gain knowledge about the ED service behavior from the data provided by the simulation of any possible reality.

From the analysis of the data from simulation, we can obtain information concerning patient's *LoS* in the service. The research we are conducting aims to improve the quality of service provided in a ED, trying to reduce the *LoS* of patients, through a model for scheduling the entry of non-critical patients into the service. The model will be based on the prediction of the *LoS* of patients in the ED by simulation. Simulation would also be the way to demonstrate the effectiveness of the scheduling model for patient admission, in which we are currently working on.

Specifically, the goal of the work presented in this paper is the first step on the way to the definition of this scheduling model. It consists of developing an analytical model to determine the *theoretical throughput* (*T_ThP*) of a particular healthcare staff configuration, which we define as the number of patients it can take care per unit time given its composition. It is a reference to measure the performance of the system and the capacity of the healthcare staff configuration to absorb the demand for the service, so it is an indicator of the response capacity of the system to patient attention.

It should be clarified that we propose a simplified model for the calculation of the system capacity, considering the system in a steady state. It is a continuous flow model, with regular admission and no queues. With this, we want to obtain, analytically, a reference value of the productivity of the system for its characterization in an ideal situation. This reference value will allow us to evaluate the effects on the behavior of the system against different measures through simulation. Specific changes in the input parameters of the simulator, in particular, referring to the patient input and the configuration of the sanitary staff, simulating different possible real situations, will modify the actual productivity of the system. The theoretical value obtained through the analytical model will be a reference to guide these changes.

The corresponding value for the *T_ThP* is an appropriate indicator for system characterization and it will indicate whether the considered healthcare staff configuration will generate endless queues for a specific scenario, or in another way, the number of patients attending the service is below its response capacity, and so the occupancy of the staff is not at its limit.

In the experimental results for the validation of the model in Section V, we conduct a sensitivity analysis on the effect of an increase or a decrease in the number of patients entering the service every hour, with respect to the theoretical value obtained as reference for the *T_ThP*. This analysis shows how the number of patients waiting to be attended in each phase of the process, which we call *Waiting Queue Length* (*WQL*), reaches endless values when the input for patients reaches and surpasses the obtained *T_ThP* with the model. It is also observed how the percentage of time in which the corresponding healthcare staff is attending or treating patients for each phase (occupancy) reaches 100% when this happens.

Once the system is characterized by this value, we can take it into account to act in order to avoid long waiting times through the admission scheduling of non-critical patients, and ultimately improve the *LoS* of all patients in the service.

The final aim of the complete research will be to obtain an input distribution of patients, which is as homogeneous as possible, so that the flow of patients in the service shall be in accordance with the response capacity of the system according to the healthcare staff resources at any time.

Moreover, the simulator will again be the main source of data for the model validation.

III. DESCRIPTION OF THE EMERGENCY DEPARTMENT OPERATION PROCESS

We divide this section into three subsections in which we describe the basic operation of the ED, the different types of patient and the functionality of the ED simulator.

A. Emergency Department Process

The operation of the ED is based on a process consisting of different steps or phases in which each patient is passing from their entry into the service until they are discharged, referred to another service or admitted to the hospital (Figure 2).

The ED is divided into different areas, which correspond with the different process phases:

- *Admissions Area*: Administrative staff carries out the registration of the patient's arrival and the reasons for their visit to the emergency service.
- *Triage Area*: Professional sanitary staff identifies the priority level with which the patient should be treated.

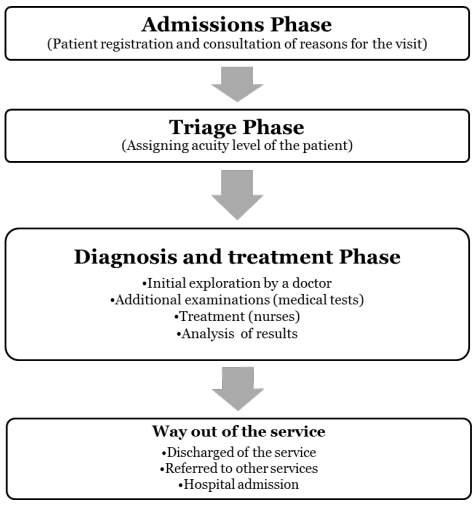

Figure 2. Operation of the Emergency department.

- *Diagnosis-Treatment Zone*: Healthcare staff (doctors, nurses and specialist technicians) try to identify the causes of the patient's health problem and, as far as possible, try to solve it. This area is in turn divided into different areas (medical room, nursing room, care boxes and X-ray laboratories).
- *Waiting Rooms*: Distributed in different zones of the ED, where patients wait to be treated at the different stages of the process.

B. Classification of patients

Real data from *Hospital de Sabadell* corroborate that the majority of patients attending the service are not critical patients and, therefore, they do not require immediate valuation or can be outpatients (Figure 3). If these non-critical patients had the possibility of getting information about when it is more advisable to go to the service, depending on the waiting time estimated for them, they would probably do it when the prevision for waits were lower. These are the patients suitable for a possible relocation in the current pattern of patients entering the service.

In the triage phase, patients are classified according to their acuity level and they are assigned a priority. The scale of priority and urgency to be applied in Spanish hospitals (Spanish Triage System) is based on the Andorran Triage Model (MAT) [25] (Figure 4).

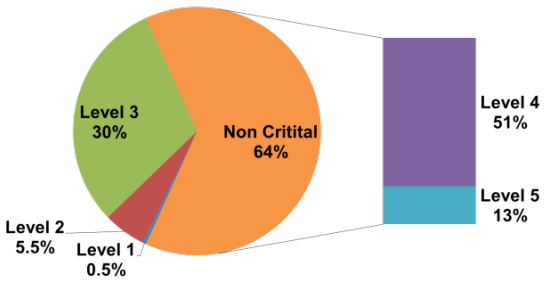

Figure 3. Percentage of patients by acuity level (historical data of 2014 of the Hospital de Sabadell).

TRIAGE	TYPE OF ATTENTIÓN	DESCRIPTION
LEVEL 1	REVIVAL	Extreme health condition life-threatening. It requires IMMEDIATE ATTENTION.
LEVEL 2	EMERGENCY	Health condition life-threatening. It requires IMMEDIATE ATTENTION. BUT NOT PRIORITY.
LEVEL 3	URGENCY	Acute condition but not life threatening. Requires NOT IMMEDIATE EVALUATION.
LEVEL 4	MINOR URGENCY	Acute condition, not life threatening. Requires DEFERRED VALUATION.
LEVEL 5	NOT URGENT	Symptomatic condition, not life threatening. DOESN'T REQUIRE URGENT ATTENTION. OUTPATIENT.

Figure 4. Classification of patients according to their level of urgency (Spanish Triage System).

C. Functionality of the Simulator

From the moment when the patient enters the service, the simulation runs according to the patient flow shown in Figure 5. The admission and triage phases are common to all patients entering the service, and there is a percentage, although low, of patients being referred to other services after the triage stage and also others who leave the service without being seen. After triage, patients with acuity level 1, 2 and 3 are treated separately from those with acuity level 4 and 5 for the diagnostic and treatment phase. In the simulation model, patients 1, 2 and 3 are treated in a specific area called Area A for diagnosis and treatment, and patients 4 and 5 are treated in a separate area identified as Area B. The admissions and triage phase share the same healthcare staff, but doctors and assistant nurses are different for Area A and B.

For our work, we are interested in tracking patients 4 and 5, those who are non-critical patients, and can be relocated in time for their arrival to service. So, we will consider all patients for admissions and triage phases, but only patients 4 and 5 (Area B) for diagnosis and treatment.

In the diagnostic and treatment phase, all patients generated by the system go through an initial medical exploration phase, which we will identify hereafter as *IE*. A percentage of them are directly discharged and leave the ED after the *IE* phase (showed by a continuous line in Figure 5). The rest remain in the ED and they go through a phase of complementary examinations and/or treatment carried out by technical staff and/or nurses. After this, they return to see the doctor, who analyzes the test and/or treatment results (we will use *AR* onwards to refer to this phase). Finally, they are discharged from the service (showed by a dashed line in Figure 5).

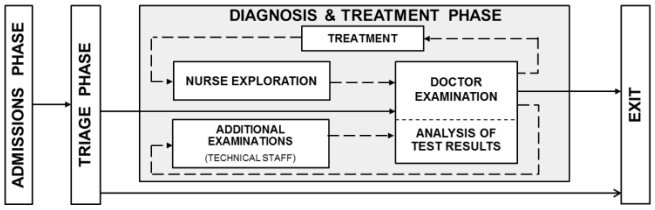

Figure 5. Patient flow in the Emergency Department.

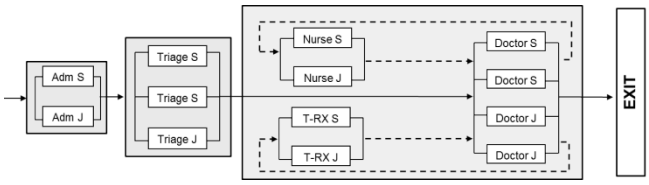

Figure 6. Sanitary staff working in parallel on each phase.

The simulator includes the following agents: patients, admissions staff, triage nurses, assistant nurses, doctors and radiology technicians. In the case of agents representing healthcare staff (all except patients), we consider two levels of experience (Junior/Senior) and all of them can work in parallel in each phase (Figure 6). The level of experience has an effect on the amount of time required for patient attention, which is different depending on their condition of junior or senior staff (hereafter *SS* and *JS*).

The actions and interactions between the involved agents at each process step result in changes of state of the agents, which ultimately result in the global operation of the system.

Each scenario of simulation is identified by an input healthcare staff configuration and a specific input of patients into the service, and the output of the simulation brings data concerning the number of attended patients, attention time and waiting time for each patient in all phases in their way through the service.

IV. ANALYTICAL MODEL

The quality of service, from the point of view of the user, is reflected in the time spent on patient attention and waiting times between different phases of the process. Moreover, from the point of view of service management, performance is directly related to the number of patients treated per unit time and an efficient use of resources.

We propose a model for system characterization, which should give us information and knowledge in order to make possible changes in the system to improve it. The model is based on the definition of a set of indicators of the quality of service, and a set of equations that allow us to measure some intrinsic characteristics of the system given a specific healthcare staff configuration, and the patient flow presented in Figure 5.

These equations will allow us to have information, and so knowledge, about the system capacity regarding its resources. We aim to use this knowledge to find an algorithm for the relocation of non-critical patients, modifying their current arrival pattern, such that their arrival at the service should be in accordance with the calculated system capacity.

A. Definition of indexes

As an indicator of the quality of service from the point of view of the user, we define an index called *Patient attention Time (PaT)* as the total time a patient is receiving attention throughout all stages in the service for a given configuration. This index is calculated from the summation of the values for the attention time in each stage, which are obtained from the simulator calibration, based on the corresponding values provided by the hospital:

$$PaT = \sum_{i=1}^{Stages} PaT_{stage\ i} \quad (1)$$

$PaT_{stage\ i}$ indicates the *Patient Attention Time in stage i*, and it is independent of the number of patients entering the service. Notice that PaT is not a fixed value for all patients, as it depends on the followed way by each patient (not all patients are required for additional examinations or receive some treatment).

Another parameter widely used and accepted in the literature as an indicator of the quality of service is the *Length of Stay (LoS).* It is defined as the total time a patient spends in the service. Unlike the previous one, the value of this index depends not only on the healthcare staff configuration, but also on the number and type of patients admitted to the service, as it includes the waiting time.

Finally, the *Length of Waiting (LoW)* is the total waiting time of a patient throughout the service. Note that,

$$LoS - PaT = LoW \text{ and always } PaT \leq LoS. \quad (2)$$

Moreover, the *Equivalent Patient attention Time* for stage *i* ($EpaT_{stage\ i}$) is defined as the attention time of a patient taking into account the possibility of working in parallel for the agents in that stage, and (3) shows how it is calculated:

$$EPaT_{stage\ i} = \frac{1}{\frac{SS_i}{PaT_{SS}^{\ i}} + \frac{JS_i}{PaT_{JS}^{\ i}}} \quad (3)$$

where SS_i and JS_i in (3) and (5) stand for the total number of senior/junior health workers in the stage *i* respectively, and the calculation is the corresponding one for parallelization on a pipeline model.

The slowest stage of the configuration will fix the speed at which patients can be attended in the service and also is the one which can saturate the system. It is, therefore, the inverse of the equivalent attention time of the slowest stage, which will determine the number of patients that a given configuration can treat per unit of time given its composition. We call this index *Theoretical Throughput (T_ThP),* which is the indicator we will use to measure the patient attention capacity of the configuration, that is, its response capacity for a specific situation. Expression (4) gives its calculation:

$$T\_ThP = \frac{1}{Max\ EPaT_i} \quad (4)$$

In fact, the *Theoretical Throughput* for a specify stage *i* will be obtained by the inverse of (3):

$$T\_ThP_{stage\ i} = \frac{SS_i}{PaT_{SS}^{i}} + \frac{JS_i}{PaT_{JS}^{i}} \quad (5)$$

B. Theoretical throughput for the diagnosis and treatment phase.

Unlike other stages of the process for a patient along his path through the ED, this is the most complex stage due to its non-linearity. All patients first go through an initial medical exploration (*IE*), which is their first contact with the doctor. There is a percentage p_1 of patients who require additional tests after the initial exploration phase with the doctor, and also a percentage p_2 of patients who require some treatment. Treatment is administered and controlled by assistant nurses. The return of these patients for the doctor's final diagnosis (after completing the complementary examinations requested by the doctor after his first contact with the patient (*AR*)) must be taken into account, as the time the doctor uses to see these patients again cannot be used to see new patients. The rest of patients will be discharged from service directly after their first contact with the doctor.

Figure 7 shows in detail patients' flow along this phase, in accordance with all these preliminary considerations.

The total number of assistant nurses, senior or junior, in the considered configuration is represented by *NS/NJ* respectively. The total number of doctors, also senior or junior, are represented by *DS/DJ*, and it is necessary to distinguish between:

- DS_{IE}/DJ_{IE}: Senior/Junior doctors attending patients in the Initial Exploration stage.
- DS_{AR}/DJ_{AR}: Senior/Junior doctors attending patients in the Analysis of Results stage.

We consider that doctors prioritize the attention of patients who have already gone through the *IE* (initial exploration stage), and therefore, these patients will be treated in the time the doctor is available for *AR* (analysis of results). This prevents endless queues on the return of patients from their requested complementary examination or treatment.

The *Theoretical Throughput (T_ThP)* has been defined as the number of patients which can be treated by the healthcare staff configuration working in each stage of the process, being so an indicator of the response capacity of each phase or stage. For its calculation in the diagnosis and treatment phase, it is necessary to consider the average attention time of each type of doctor depending on their experience (Junior or Senior), and depending on the type of care they are providing, either in the first step of initial exploration (*IE*), or in the second, consisting of the analysis of the results of a requested supplementary examination (*AR*). These times are known, determined by the calibration of the simulator, and denoted by PaT_i^j, which represents the average *Patient Attention Time* for a doctor type *i* doing *j*.

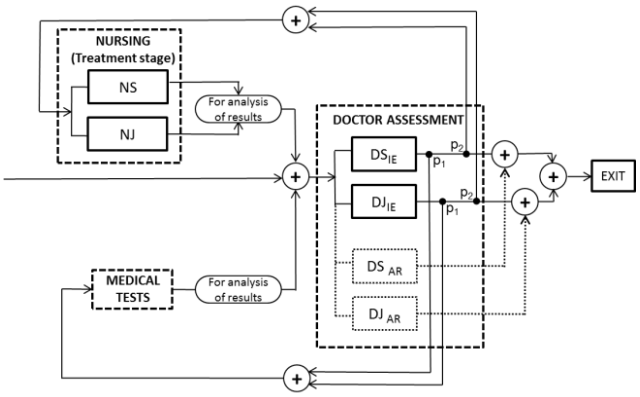

Figure 7. Patient flow in diagnosis & treatment phase.

Then we consider:

- PaT_{DS}^{IE}: Average attention time of a senior doctor (*DS*) in the Initial Exploration stage (*IE*).
- PaT_{DS}^{AR}: Average attention time of a senior doctor in the Analysis of Results stage (*AR*).
- PaT_{DJ}^{IE}: Average attention time of a junior doctor (*DJ*) in the Initial Exploration stage (*IE*).
- PaT_{DJ}^{AR}: Average attention time of a junior doctor in the Analysis of Results stage (*AR*).

Given these times, their inverse will give us the number of patients that each doctor can treat per unit time considered:

$$P_{DS}^{IE} = \frac{DS_{IE}}{PaT_{DS}^{IE}} = \text{Patients per minute for a } DS \text{ in } IE \text{ stage;}$$

$$P_{DJ}^{IE} = \frac{DJ_{IE}}{PaT_{DJ}^{IE}} = \text{Patients per minute for a } DJ \text{ in } IE \text{ stage;}$$

$$P_{DS}^{AR} = \frac{DS_{AR}}{PaT_{DS}^{AR}} = \text{Patients per minute for a } DS \text{ in } AR \text{ stage;}$$

$$P_{DJ}^{AR} = \frac{DJ_{AR}}{PaT_{DJ}^{AR}} = \text{Patients per minute for a } DJ \text{ in } AR \text{ stage.}$$

where DS_{IE}, DS_{AR}, DJ_{IE}, DJ_{AR} are unknown values.

From the historical real data provided by the *Hospital de Sabadell* we know that patients can go once, twice or more times for tests and/or treatment, and so see the doctor more than once (Figure 8). Anyway, for patients 4 and 5, the percentage of patients that require more than one test or treatment is very low.

There is a percentage p_1 of patients who, after their first contact with the doctor, require additional tests, and a percentage p_2 who require some treatment. Then, there is a percentage $1 - (p_1 + p_2)$ of patients who are discharged from the service directly after their initial exploration with the doctor, those who do not require any additional test nor any treatment.

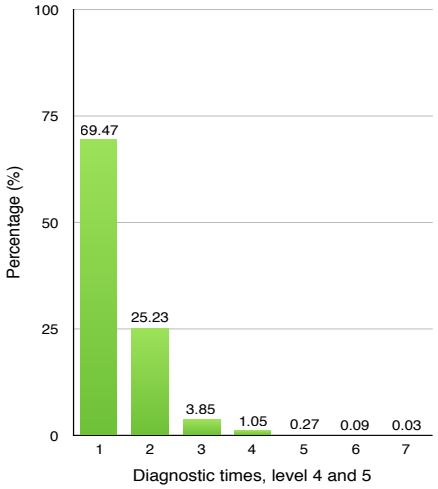

Figure 8. Percentage of number of diagnostic times (doctor care) for non-critical patients (Area B).

By observing the data represented in Figure 8, we can see that around 70% of patients 4 and 5 are discharged from the service directly after their initial exploration with a doctor. Therefore, only 30% of patients in Area B require some test or treatment ($p_1 + p_2$). Thus, given these percentages and the patient flow of Figure 7, we obtain the following relations of continuity:

$$P_{DS}^{IE} \cdot (p_1 + p_2) = P_{DS}^{AR} \quad (6)$$

$$P_{DJ}^{IE} \cdot (p_1 + p_2) = P_{DJ}^{AR} \quad (7)$$

$$DS_{IE} + DS_{AR} = DS \quad (8)$$

$$DJ_{IE} + DJ_{AR} = DJ \quad (9)$$

The solution of this linear system of equations gives us the values for DS_{IE}, DS_{AR}, DJ_{IE}, DJ_{AR}, and therefore, the values for P_{DS}^{IE}, P_{DS}^{AR}, P_{DJ}^{IE}, P_{DJ}^{AR}, for the considered configuration of doctors.

Now, we can obtain the *theoretical throughput* for the doctors' stage in the diagnosis and treatment phase by the summation of patients who have only been attended once by the doctor ($P_{only\ IE}$), those who have been required for additional testing (P_{Test}), and those who have gone to the nurses stage for some treatment (P_{Treat}), as shown in (10):

$$T\_ThP_{Doctors} = P_{only\ IE} + P_{Test} + P_{Treat} \quad (10)$$

where:

$$P_{only\ IE} = (P_{DS}^{IE} + P_{DJ}^{IE}) \cdot (1 - p_1 - p_2) \quad (11)$$

$$P_{Test} = (P_{DS}^{IE} + P_{DJ}^{IE}) \cdot p_1 \quad (12)$$

$$P_{Treat} = (P_{DS}^{IE} + P_{DJ}^{IE}) \cdot p_2 \quad (13)$$

When we introduce equations (11) to (13) on (10) we find:

$$T\_ThP_{Doctors\ stage} = P_{DS}^{IE} + P_{DJ}^{IE} \quad (14)$$

so,

$$T\_ThP_{Doctors\ stage} = \frac{DS_{IE}}{PaT_{DS}^{IE}} + \frac{DJ_{IE}}{PaT_{DJ}^{IE}} \quad (15)$$

Moreover, the *theoretical throughput* for the assistant nurses in the treatment stage, inside the diagnosis and treatment phase, will be calculated as shown in (5):

$$T\_ThP_{treatment\ stage} = \frac{NS}{PaT_{NS}} + \frac{NJ}{PaT_{NJ}} \quad (16)$$

Finally, the *theoretical throughput* for the diagnosis and treatment phase will be the lowest value of (15) and (16), and this value will be the indicator for the response capacity to patient attention in the ED, assuming that the admission and triage phases do not limit this value.

V. EXPERIMENTAL VALIDATION

Once we have defined the equations for the calculation of the theoretical throughput (*T_ThP*), we must validate them. For this validation we have used the simulator to see if the obtained values for the *T_ThP* for each stage in the ED process are in accordance with the generated data by the ED simulator. We consider a sufficient rate of patients entering into the service, the same each hour, to ensure that the system is running continuously and we assume the system is in a steady state, after a time of warm up.

We have used two different healthcare staff configurations (Staff I and II), and we only consider Area B for the diagnosis and treatment phase. The corresponding obtained values for the *T_ThP* calculated from the equations of the model are presented in Tables I and II, respectively.

TABLE I. THEORETICAL THROUGHPUT FOR EACH PHASE OF THE ED PROCESS CORRESPONDING TO STAFF I

		STAFF I				
		Healtcare Staff		*PaT* (minutes)		*T_ThP*
		Junior	Senior	Junior	Senior	(pat/hour)
ADMISSIONS PHASE		3	0	8.00	6.00	22.50
TRIAGE PHASE		1	2	12.00	8.00	20.00
DIAGNOSIS & TREATMENT	Nursing	5	7	30.00	27.00	25.56
	Doctors *IE*	5	2	23.89	21.74	14.68
	Doctors *AR*			19.17	15.25	

TABLE II. THEORETICAL THROUGHPUT FOR EACH PHASE OF THE ED PROCESS CORRESPONDING TO STAFF II

		STAFF II				
		Healtcare Staff		*PaT* (minutes)		*T_ThP*
		Junior	Senior	Junior	Senior	(pat/hour)
ADMISSIONS PHASE		1	1	8.00	6.00	17.50
TRIAGE PHASE		2	1	12.00	8.00	17.50
DIAGNOSIS & TREATMENT	Nursing	4	3	30.00	27.00	14.67
	Doctors *IE*	3	2	23.89	21.74	10.63
	Doctors *AR*			19.17	15.25	

The values for *PaT* in Tables I and II are the average values for each phase that result from the calibration of the simulator according to real data from the hospital. Moreover, it is important to point out that the simulator considers a random exponential distribution to model the real behavior of the *PaT*, depending on the type and age of patient.

We run the simulation for four different values for the number of patients entering the service per hour, around the theoretical value obtained as reference for the *T_ThP* for each phase from the equations of the model. Next, we conduct an analysis of the effect of the number of patients entering the service every hour, firstly on the percentage of time, which the corresponding healthcare staff spend on attending or treating patients (Occupancy) for each phase of the whole process, and secondly, on the number of patients waiting to be attended in each phase of the process, which we call *Waiting Queue Length* (*WQL*).

These are the indicators we use to validate our theoretical values. Therefore, we consider that the theoretical value obtained from the model is a good approach to the real throughput value, when the occupancy of the considered staff is below its maximum limit of capacity, and no queues are observed for this value, but they are generated when we add more patients per hour entering the service and the staff in this phase is at 100% of its capacity.

The analysis presented in the following sections shows how the *WQL* reaches endless values when the input for patients reaches and surpasses the obtained *T_ThP* with the model. The obtained results show how this situation inevitably leads to over-saturation of the system when we increase the simulation time. It is also observed how the occupancy of the corresponding staff in each stage reaches 100% when this happens.

A. Simulation results for admission phase.

We first go for the experiments for the validation of the *T_ThP* calculated value for the admissions phase. Once fixed, the input parameters for the configuration of the Staff I, and according to the results in Table I, we generate a constant and homogeneous patients input to ensure a steady state for the validation of the obtained values for the *T_ThP* for each phase. The results are shown in Figures 9 and 10.

The diagram in Figure 9 shows the results for the Staff I occupancy in the admissions phase for four different inputs of patients around the calculated *T_ThP*. We observe that the bar corresponding to the input of 21 patients per hour for admissions staff occupancy goes up to nearly 100% of occupancy, which is reached for 22 patients. This means that, with 22 or more patients, the admission phase has surpassed its limit of capacity, so this simulation result is in accordance with the *T_ThP* obtained with the analytical model for admissions phase in Table I. This first check validates this value.

On the other hand, Figure 10 shows the evolution on *WQL* with time, that is, the number of patients in the queue in the waiting room for this phase of the ED process depending on the number of simulated days, and for the

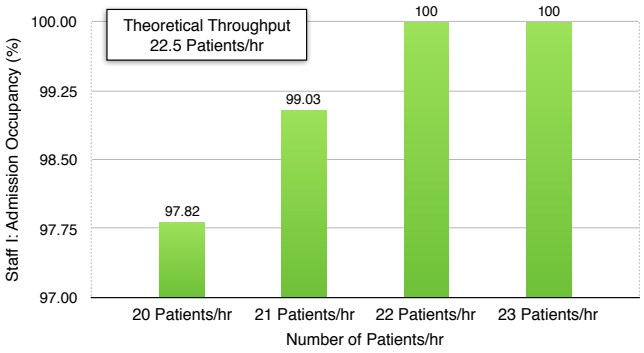

Figure 9. Occupancy percentage for admission phase (Staff I).

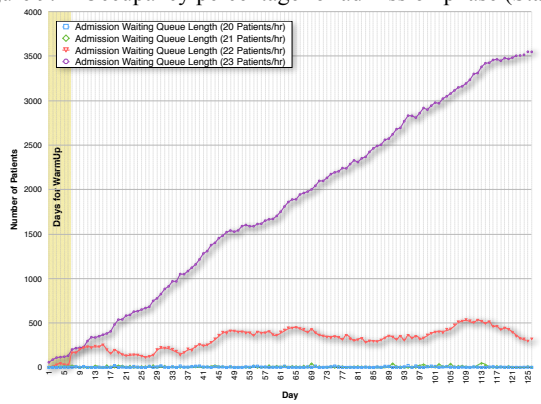

Figure 10. *WQL* evolution for admission phase (Staff I).

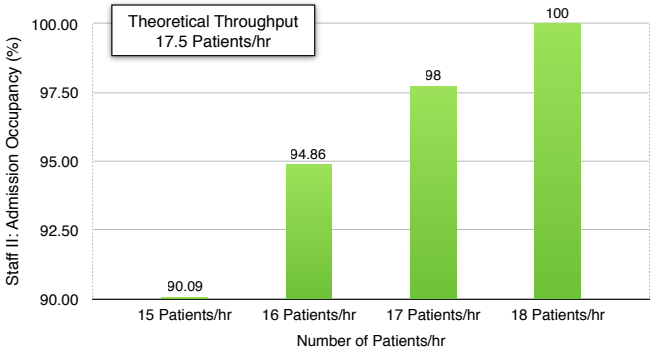

Figure 11. Occupancy percentage for admission phase (Staff II).

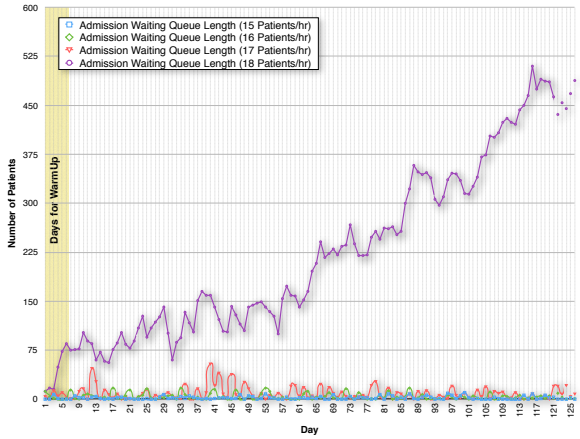

Figure 12. *WQL* evolution for admission phase (Staff II).

same four different input values for the number of patients entering the system.

The *WQL* is under control until the number of patients reaches and exceeds the obtained 22 patients for the *T_ThP* with the model, when it reaches endless values.

We proceed now with the validation of the admissions *T_ThP* value for the configuration in Staff II (see now Table II). Figures 11 and 12 show the results for the Staff II occupancy in the admissions phase and the *WQL* evolution again for four different inputs of patients around the calculated value for *T_ThP*.

When the input is 17 patients per hour, the admissions staff occupancy almost reaches its limit of capacity, and no important queues are generated. See how the bar of 17 patients/hr in the diagram in Figure 11 goes up to nearly 100% of occupancy, and temporal lines in Figure 12 for 17 or less patients per hour do not lead to saturation of the system, but only one more patient per hour entering the service produces endless queues. This simulation result is in accordance with the *T_ThP* obtained with the model (17.5 patients per hour) and so, it validates this value.

The fluctuations observed in the temporal lines in Figures 10 and 12 are due to the distribution used by the simulator to consider the variation of *PaT* depending on both the type and age of patients. The simulator uses an exponential distribution to model this fact, as a result of its calibration with the available real data from the hospital. These variations in the random values assigned to *PaT* for each generated patient can produce some queues, which appear anytime but, which the system can finally absorb if the number of patients entering the service per hour is below the system's capacity of attention.

Hereinafter, we proceed in the same way for validating the remaining values for *T_ThP* corresponding to the other stages: triage, doctors and nursing for treatment in the diagnosis and treatment phase.

B. Simulation results for triage phase.

The simulation results for the validation of the *T_ThP* value considering Staff I for the triage phase are shown in Figures 13 and 14. The bar chart of Figure 13 shows that the maximum attention capacity for this phase is 20 patients per hour, since it is for this value when 100% occupancy of the healthcare staff responsible for this stage is reached.

In Figure 14, we can observe the evolution of the *WQL* for the triage phase, again for four different inputs of patients. Endless queues are formed when 20 or more patients per hour enter the service, which is its limit of capacity. Meanwhile, there are no queues for values under the *T_ThP* calculated in Table I. This simulation result is again in accordance with the *T_ThP* obtained with the model, so it validates this value for the triage *T_ThP*.

Figures 15 and 16 show the *WQL* for the triage phase for Staff II and the corresponding staff occupancy respectively, for four different inputs of patients. When the input is 17 patients per hour, the triage staff occupancy nearly reaches the 100%, so it is almost at its limit of capacity. Only one patient more per hour collapses the system in this stage, as

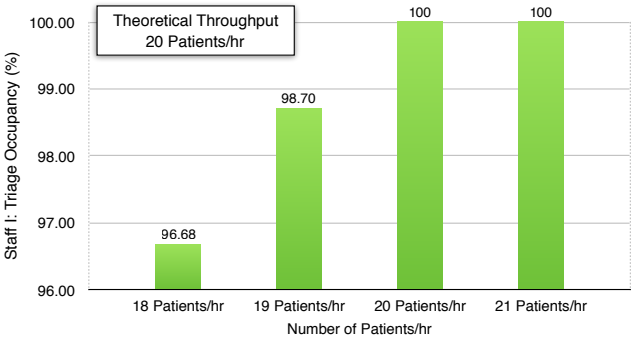

Figure 13. Occupancy percentage for triage phase (Staff I).

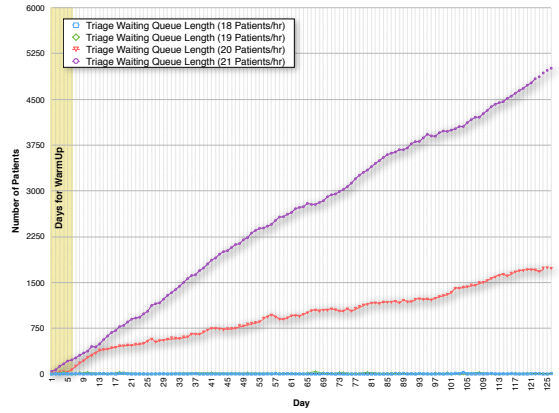

Figure 14. *WQL* evolution for triage phase (Staff I).

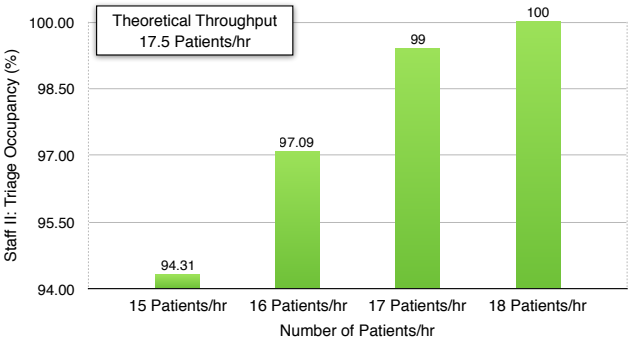

Figure 15. Occupancy percentage for triage phase (Staff II).

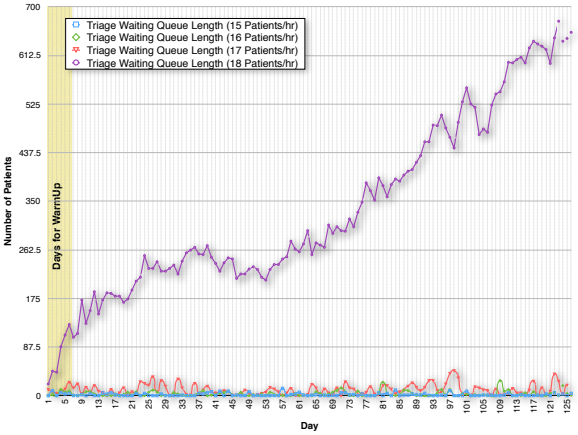

Figure 16. *WQL* evolution for triage phase (Staff II).

the temporal line shows in Figure 16 for 18 patients per hour entering the system.

This simulation result is in accordance with the *T_ThP* obtained with the analytical model (Table II), so it validates this value for the triage *T_ThP* for staff configuration II.

Once again the fluctuations observed in Figure 16 are due to the random exponential distribution used by the simulator to consider the variation of *PaT*.

C. *Simulation results for diagnosis and treatment phase.*

Here only patients 4 and 5 are considered, since we are analyzing the behavior in Area B, where non-critical patients are treated. According to data presented in Figure 8, the probability of these patients requiring some additional test or treatment has been fixed at 30%.

Figures 17 and 18 show the simulation experimental results for the doctor's stage of the diagnosis, considering the staff I configuration in Table I. In Figure 17, we can see the corresponding staff occupancy for this stage, for four different inputs of patients. Here, the *T_ThP* value obtained from the equations is 14.68 patients per hour. The simulation data shows how the occupancy for 14 patients per hour is almost at 100%, and also the analysis for the *WQL* shows how doctors are saturated when only one more patient per hour enters the service. Once again the *T_ThP* is at its limit of capacity and when this value is surpassed, the system collapses in this phase. These results are in accordance with the doctors' *T_ThP* obtained with the model and hence validate it.

In the same way, Figures 19 and 20 show the staff occupation rate and the *WQL* tendency for the doctors' stage, within the diagnosis and treatment phase for Staff II, again for four different inputs of patients. The probabilities for patients to require some additional test or treatment have also been fixed at 30%.

The *T_ThP* is between 10 and 11 patients, and we can see how when the input is of 11 patients per hour, medical staff occupancy reaches 100%. Therefore, this simulation result shows the system has surpassed its limit of capacity and this is in accordance with the *T_ThP* obtained with the model. Long and non-ending queues also collapse the service for 11 or more patients. Once again, this validates the obtained value for the Doctors *T_ThP* in this case.

Finally, we try to validate *T_ThP* for the for treatment stage, carried out by the assistant nurses inside the diagnosis and treatment phase. Figures 21 and 22 show the simulation results for this stage when we consider the specific configuration for the healthcare staff specified in Table I.

The obtained values by simulation for the occupation rate are in accordance with our theoretical value of 25.56 patients per hour (Figure 21), when the number of patients waiting for treatment grows dramatically and the *WQL* becomes very large (Figure 22). Also, with the staff II configuration, the obtained values for occupation in the nursing stage are in accordance with our theoretical value of 14.67 patients per hour (Figure 23), and again the number of patients waiting for attention grows dramatically (Figure 24).

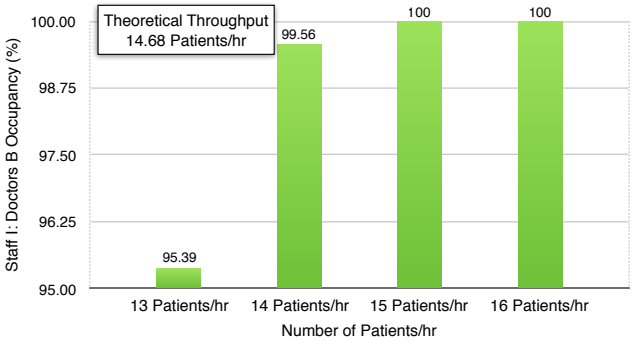

Figure 17. Occupancy for doctors phase (Staff I) in Area B.

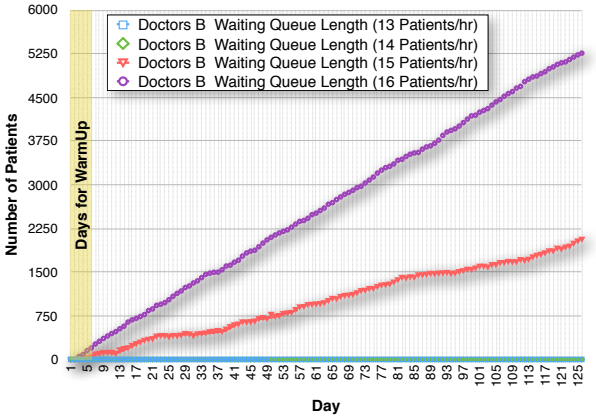

Figure 18. *WQL* evolution for doctors phase (Staff I) in Area B.

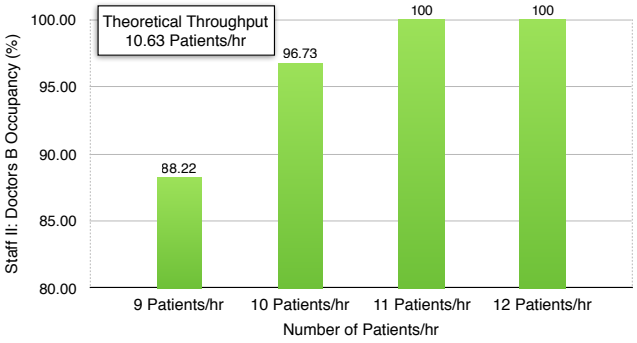

Figure 19. Occupancy for doctors phase (Staff II) in Area B.

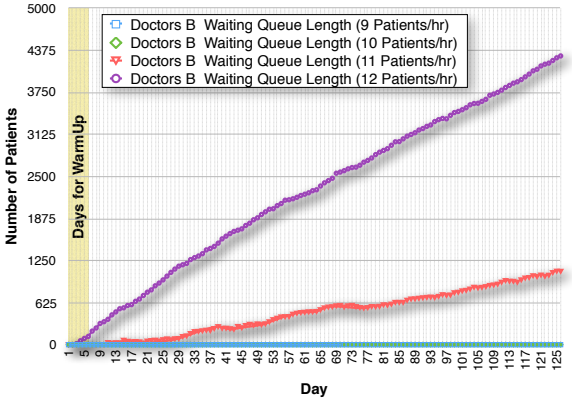

Figure 20. *WQL* evolution for doctors phase (Staff II) in Area B.

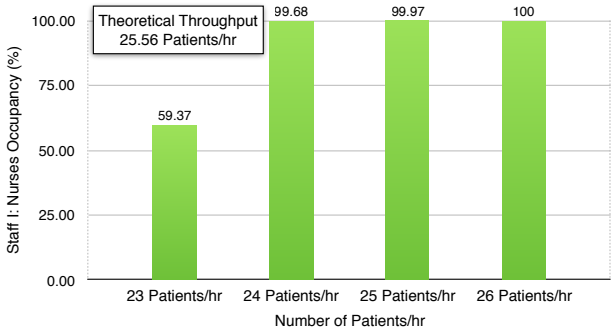

Figure 21. Ocupancy percentage for nursing phase (Staff I).

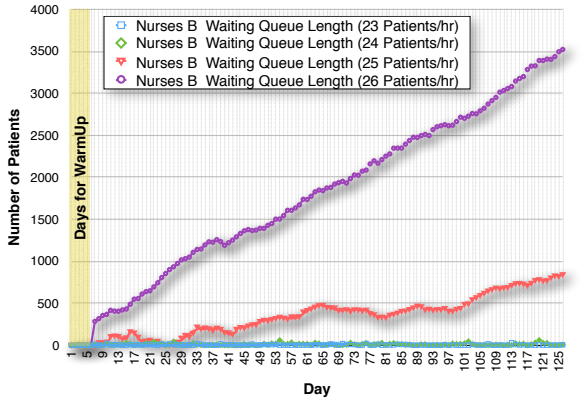

Figure 22. *WQL* evolution for nursing phase (Staff I).

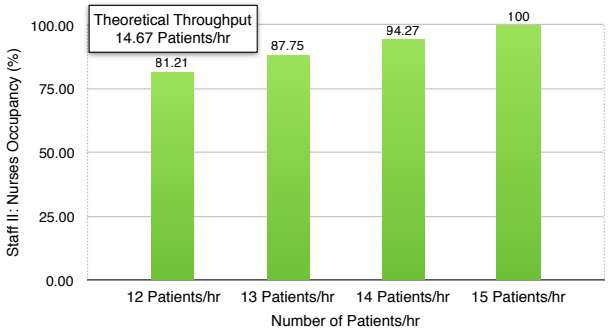

Figure 23. Occupancy percentage for nursing phase (Staff II).

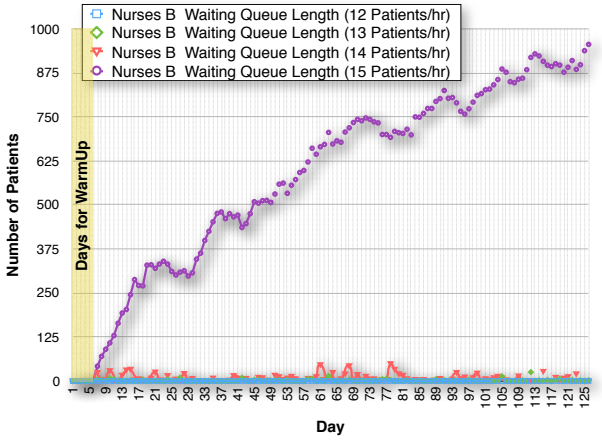

Figure 24. *WQL* evolution for nursing phase (Staff II).

In Figures 23 and 24 we can see how the simulation results for the nursing stage are once more in accordance with the model when we modify the staff parameters to Staff II configuration.

All the values of *T_ThP* for admission, triage, doctors and nurses have been validated, and they are in accordance with the simulation results with a very good approximation. The simulator is our sensor of the real system, so these results validate the proposed analytical model.

VI. Conclusion and Future Work

The main contribution presented in this paper is the ED's healthcare staff characterization, through its capacity, named theoretical throughput, which is the number of patients that the system should be able to absorb per unit time, given the staff composition.

We have defined an analytical model to determine the theoretical throughput of a particular healthcare staff configuration based on the number of admission staff, triage, assistant nurses, and doctors, and their respective attention times for patients.

The analytical model of equations to calculate the values of the *T_ThP* for admissions phase, triage phase, nursing and medical exploration stages in diagnosis and treatment phase, according to the actual patient flow in the ED process, has been validated. For the validation of the model we have used an ED simulator based on an agent based model of the system, as a sensor of the real system. Output data from simulation of different possible real situations have been analyzed to obtain the information for the model validation.

We have seen how the theoretical throughput is a reference to measure the performance of the system, and the capacity of the healthcare staff configuration to absorb the demand of the service, so it is an indicator of the response capacity of the system to patient attention.

The analytical model for the *T_ThP* calculation will give us information to relocate non-critical patients, so that the theoretical throughput will be the reference indicator for the redistribution of non-critical patients. The idea is to try to modify their current arrival according to system capacity at any time, which is our current research in progress. This relocation may improve the time patients stay in the service, and therefore the service quality.

Our future work will consist of designing a admission scheduling model for non-critical patients in the service, using the ED simulator for their *LoS* prediction.

The historical data provided by the hospital, the defined analytical model for the evaluation of the response capacity of the system, and the information obtained from the analysis of the data from simulation, will all enable the possibility of planning admission of non-critical patients into the service.

This proposed future model for relocation of patients will be efficient to the extent that a supposed “self-triage and recommendation system” is effective on patient entry, so that patient input curve gets flatter and approaches the value corresponding to the maximum capacity of the system, and therefore, an improvement in performance is expected.

A good relocation of non-critical patients and a significant improvement in the quality of service mean a

reduction on *LoS* of patients in the service, without removing patients, which in some cases, could make the reduction of under-utilized resources possible.

Finally, and more generally, our global proposal aims to improve the ED service, which is the main entry of patients in the healthcare system in relation to access, quality of service, user satisfaction and efficiency.

ACKNOWLEDGMENT

This publication is based upon work supported under contract TIN2014-53172-P, funded by the MINECO Spain.

REFERENCES

[1] E. Bruballa, M. Taboada, A. Wong, D. Rexachs, and E. Luque, "An Analytical model to evaluate the response capacity of emergency departments in extreme situations," The Seven International Conference on Advances in System Simulation (Simul 2015), pp. 12-16, 2015.

[2] F. Kadri, S. Chaabane, T. Berger, D. Trentesaux, C. Tahom, and Y. Sallez, "Modelling and management of the strain situations in hospital systems using ORCA approach," IEEE IESM, Rabat, Morocco, pp. 202-210, October 2013.

[3] F. Kadri, S. Chaabane, and C. Tahon, "A simulation-based decision support system to prevent and predict strain situations in emergency department systems," Simulation Modelling Practice and Theory, vol. 42, pp. 32-52, March 2014.

[4] A. Boyle, K. Beniuk, I. Higginson, and P. Atkinson, "Emergency Department Crowding: Time for interventions and policy evaluations," Emergency Medicine International, Article ID 838610, 2012.

[5] P.C. Sprivulis, J.A. Da Silva, I.G. Jacobs, A.R.L. Frazer, and G.A. Jelinek, "The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments," Medical Journal of Australia, vol. 184, num. 5, pp 208-212, 2006.

[6] Ph.Yoon, I.Steiner, and G. Reinhardt, "Analysis of factors influencing length of stay in the emergency department," Canadian Journal of Emergency Medicine (CJEM), vol. 5, issue 03, pp. 155-161, 2003.

[7] N.R. Hoot and D. Aronsky, "Systematic review of Emergency Department Crowding: causes, effects and solutions," Annals of Emergency Medicine, vol. 52, Issue 2, pp. 126-136, 2008.

[8] L.M. Stock, G.E. Bradley, R.J. Lewis, D.W. Baker, J. Sipsey, and C.D. Stevens, "Patients who leave emergency departments without being seen by a physician: magnitude o the problem in Los Angeles County," Annals of Emergency Medicine, vol. 23, pp. 294-298, 1994.

[9] C.M. Fernandes, A. Price, and J.M. Christenson, "Does reduced length of stay decrease the number of emergency department patients who leave without seeing a physician?," The Journal of Emergency Medicine, vol 15, pp. 397-399, 1997.

[10] M. Sanchez, A. Smally, R. Grant, and L. Jackobs, "Effects of a fast-track area on emergency department performance," The Journal of Emergency Medicin, vol.31, pp. 117-120, 2006.

[11] S. W. Rodi, M.V. Graw, and C.M. Orsini, "Evaluation of a fast track unit: alignment of resources and demand results in improved satisfaction and decreased length of stay for emergency department patients," Quality Management in Health Care, vol 15, pp. 163-170, 2006.

[12] R. Davies, ""See and Treat" or "See" and "Treat" in an Emergency Department," Proceedings of the Winter Simulation Conference, pp. 1519-1522, 2007.

[13] S. Samaha, W.S. Armel, and D.W. Starks, "The use of simulation to reduce the length of stay in an Emergency Department," Proceedings of the winter Simulation Conference, pp. 1907-1911, 2003.

[14] J. Wang, J. Li, K. Tussey, and K. Ross, "A simulation study to reduce length of stay in Emergency Department at a large community hospital," IIE Annual Conference. Proceedings. Institute of Industrial Engineers-Publisher, pp. 1-6, 2011.

[15] J. Wang, J. Li, K. Tussey, and K. Ross, "Reducing length of stay in emergency department: a simulation study at a community hospital," IEEE Transactions on Systems, Man, and Cybernetics- PART A: Systems and Humans, Vol. 42, No 6, pp. 1314-1322, 2012b.

[16] K.W. Tan, H.C. Lau, and F. Lee, "Improving patient length of stay in Emergency Department through dynamic queue management," Proceedings of the Winter Simulation Conference, December 2013.

[17] D.J. Medeiros, E. Swenson, and C. DeFlitch, "Improving patient flow in a hospital emergency department," Proceedings of the winter Simulation Conference, pp. 1526-1531, 2008.

[18] A. M. Mancilla, "Simulation: A tool for the study of real systems," Ingeniería y Desarrollo, Universidad del Norte, vol. 6, pp.104–112, 1999.

[19] J. Pavón, M. Arroyo, S. Hassan, and C. Sansores, "Simulation of social systems with software agents," CMPI-2006, Actas del Campus Multidisciplinar en Percepcion e Inteligencia, vol. 1, pp. 389-400, 2006.

[20] L. R. Izquierdo, J.M. Galán, J.I. Santos, and R. Del Olmo, "Modeling complex systems using agent-based simulation and system dynamics," Empiria: Revista de metodología de ciencias sociales, vol. 16, pp. 85–112, 2008.

[21] M. Taboada, E. Cabrera, M. L. Iglesias, F. Epelde, and E. Luque, "An agent-based decision support system for hospitals emergency departments," Procedia Computer Science, vol. 4, pp. 1870–1879, ICCS 2011.

[22] L. Zhengchun, E. Cabrera, D. Rexachs, and E. Luque, "A generalized agent-based model to simulate emergency departments," The Sixth International Conference on Advances in System Simulations, IARIA, pp. 65-70, Nice, France, October 2014.

[23] M. Taboada, E. Cabrera, and E. Luque, "Modeling, simulation and optimization of resources management in hospital emergency departments using the agent-based approach," Advances in Computational Modeling Research, pp. 1–31, 2013.

[24] E. Cabrera, M. Taboada, M. L. Iglesias, F. Epelde, and E. Luque, "Simulation optimization for healthcare emergency departments," Procedia Computer Science, vol. 9, ICCS, pp. 1464–1473, 2012.

[25] W. Soler, M. Gómez Muñoz, E. Bragulat, and A. Álvarez, "Triage: a key tool in emergency care," Anales del Sistema Sanitario de Navarra, Gobierno de Navarra, Departamento de Salud, vol. 33, supl.1, pp. 55-68, Pamplona 2010.