*Proceedings of the 2018 Winter Simulation Conference*
*M. Rabe, A.A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.*

# ANALYSING THE ED PATIENT FLOW MANAGEMENT PROBLEM BY USING ACCUMULATING PRIORITY QUEUES AND SIMULATION-BASED OPTIMIZATION

Marta Cildoz
Fermin Mallor

Institute of Smart Cities (ISC)
Public University of Navarre
Campus Arrosadia
Pamplona, 31006, SPAIN

Amaia Ibarra

Emergency Department
Navarre Hospital Compound
Irunlarrea-Str. 3
Pamplona, 31008, SPAIN

## ABSTRACT

This paper deals with the Emergency Department (ED) patient flow management problem. After triage, where acuity of patient's illness is assessed, each patient waits for treatment. Different care requires disparate resources, leading to different care-paths in the ED. During treatment patients can be in a diversity of treatment stages (waiting for the first consultation, doing clinical tests, waiting for a second consultation, etc.). Therefore, the selection of the next patient to be seen by a physician is not trivial, especially when there are different quality goals to be attained simultaneously. This research investigates disciplines based on accumulation of priority while waiting (APQ). Pure priority disciplines are independent of the quality goals set for the ED while the new ones can be optimized to achieve those goals as much as possible. The optimal APQ policies are obtained by using simulation-based optimization. Both types of policies are tested in a real case.

## 1 INTRODUCTION

Growth in utilization of emergency care is observed in high income countries. For instance, in the US, emergency admissions grew over 50 per cent from 1992 to 2006 (Schuur and Venkatesh 2012) and 9,3% from 2014 to 2017 in England (NHS England 2018) mainly due to its ageing population who are the main consumers of healthcare services. Some studies quantify that this factor itself can explain 40-50 per cent of the total growth (Tang et al. 2010; Strunk et al. 2006). This trend is expected to continue in the near future. Accordingly, emergency medicine represents a large share of health care spending, which could be as high as 10%, according to Lee et al. (2013). Nevertheless the expenditure in health care does not follow the demand growth pace, being stabilized in the last years around 10% of GDP (see Figure 1). Thus the mixture of a growing demand and a quite stable capacity of service lead to every time overcrowded emergency departments (ED) and makes operational health care management even more critical and important to guaranty the quality and universality of the public health care services. This paper uses a quantitative approach to improve the management of the patient flow in an ED.

EDs are especially difficult to manage: they evolve in a high stochastic environment due to the variability in the patient arrival rate, illness severity, and in general, health resources needed for treatment (material and human). This chaotic, stressful, and unpredictable environment makes difficult to find an optimal strategy for the patient flow management. Just after triage, in some EDs, patients are assigned to one of the physicians working at that shift, who decides the order in which patients receive the treatment, while in other EDs patients made up a single queue from which physicians pick up the next patient. In both situations the physicians have to manage the portfolio of pending patients taking into account their level of severity. In addition, pending patients can be in different stages of their treatment: waiting for the first consultation, waiting or doing clinical diagnostic tests, waiting for second consultation or waiting to

be discharged to a hospital ward. Thus, appropriate queue management policies need to consider a priority queue discipline for a multistage service.

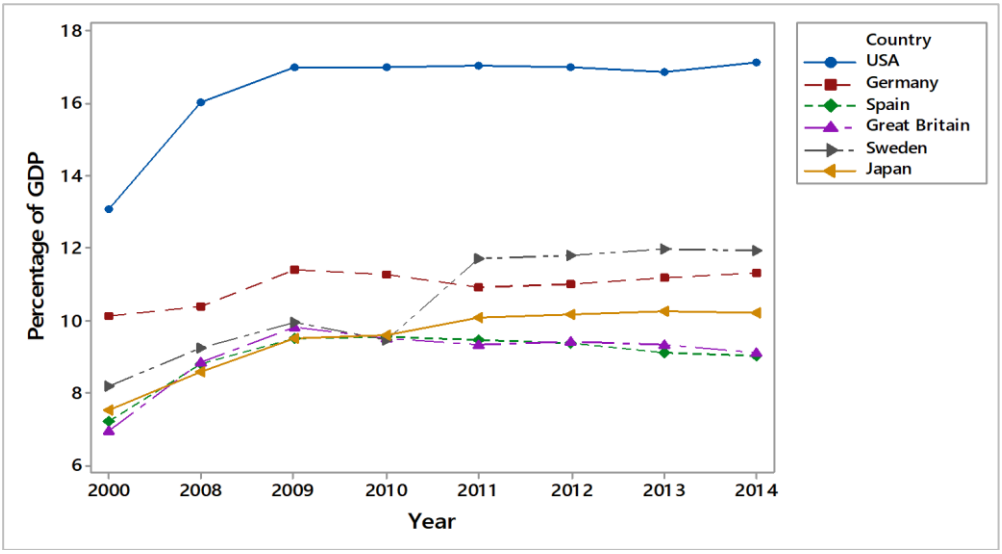


Figure 1: National expenditure in health care and increase in the demand of health care services (data from World Health Organization).

There are several objectives to be accomplished that guide the measurement of the patient flow management performance. One of the main ED performance measures is the arrival to provider time ("door to doc"), which is defined as the interval between the time a patient arrives at the ED and the time an attending physician sees the patient (Welch et al. 2011), but other objectives are also important as minimizing the length of stay (LoS) in the ED and minimizing the congestion (the number of patients simultaneously treated and present in the ED' facilities).

This paper uses a simulation model to evaluate the performance of different queue management policies. In particular the current policies, based on a priority discipline, are compared with the "accumulating priority queue" (APQ) strategy, term casted by Stanford et al. (2014). The APQ model tries to make more flexible the pure priority discipline by letting a patient with low priority that experiences a very long wait access to the physician even in the presence of patients with higher priority but with much less waiting time. This idea was first proposed by Kleinrock (1964) who let the patients from a given priority k (where $k \in \{1, 2, \dots, K\}$, and $K$ is the number of classes) accumulate "priority points" at a rate $\beta_k > 0$, where $\beta_k > \beta_j, for\ k < j$. The APQ model can be seen as a dynamic priority discipline in which patients of lower priority classes can overcome the priority of higher classes as their waiting time increases. Kleinrock (1964) obtained results about the mean waiting time before service which were extended by obtaining the waiting time distribution for each priority class in the single server and in the multi-server settings (Stanford et al. 2014; Sharif et al. 2014, respectively). All these models apply to the Markovian case (Poisson arrivals and exponentially distributed service times) and service provided in one step.

We adapt the APQ discipline to the case of multiple stages in the service and combine optimization with simulation to obtain the best values of the parameters defining the linear accumulating priority expressions for each type of patient in each service step.

The different management policies are tested in a real case, the Emergency Department of the Hospital Compound of Navarre (Spain). Analysis of historical data (more than 140,000 patients in a year) allowed us to construct a detailed discrete event simulation model, validated by the physicians, that includes all the main processes involved in the treatment of patients in the ED, as well as patient arrival

patterns accounting for intraday and intraweek seasonality, and service time distributions for all processes and consultations.

The rest of the paper is organized as follows. In section two, descriptions of an ED structure and the patient care-paths through it are provided. We also present the key performance indicators that are used to evaluate the quality of the healthcare service in the ED. In Section three we discuss patient flow management policies: the current based on priority disciplines, and the APQ policies. The simulation based optimization methodology to obtain the best AQP policy is also outlined. In Section four we present the case study and the results obtained from the application of the different management policies. We end the paper with a conclusion section.

## 2 PATIENT FLOW IN EMERGENCY DEPARTMENTS

In this section, it is presented the structure of an ED (demand and patients care-paths with different processes through the system) and the proposed Key Performance Indicators to assess performance and quality care.

### 2.1 Arrival of Patients

Patient arrivals to EDs vary over time, which affect workload assigned to the physicians. Our data analysis confirms results from other studies (e.g. Twomey et al. 2007) that observed intraday and intraweek seasonality in patient arrivals. Furthermore, the magnitude of these seasonal components also depends on the severity of the patients in such a way that the lower the severity the greater the intraday and intraweek seasonal component (Aguado-Correa et al. 2016) is. After the administrative registration process, the emergency patients are classified according to their level of severity. Depending on the hospital and country, this triage process usually uses one out of four ordinal ED triage scales (Twomey et al. 2007): the Australasian Triage Scale (ATS), the Canadian Triage Acuity Scale (CTAS), the Emergency Triage Scale (also known as Manchester Triage Scale) and the Emergency Severity Index (ESI). In our case study a rating system as the CTAS based on five priority levels of acuity is applied (Table 1: the *Access Time* is the upper limit for the arrival to provider time and the *Performance Level* is the minimum percentage of patients that should satisfy the access time requirement). The patient arrival pattern can be adequately modeled as a nonhomogeneous Poisson process for each type of patient (Luo et al. 2013), with intensity of arrivals depending on the patient acuity, the hour of the day, the day of the week and, in some cases, the season of the year.

Table 1: CTAS key performance indicators.

| Category | Classification | Access Time | Performance Level |
|---|---|---|---|
| 1 | Resuscitation | Immediate | 98% |
| 2 | Emergency | 15 min | 95% |
| 3 | Urgent | 30 min | 90% |
| 4 | Less urgent | 60 min | 85% |
| 5 | Not urgent | 120 min | 80% |

### 2.2 Patient Flow

Usually EDs organize the patient care into two different care circuits, one for the more critical patients (with severity index 1, 2 or 3) and other for less critical patients (with severity index 3, 4 or 5). Patients with acuity index 3 can be treated in both circuits. The staff (physicians, nurses and ancillaries) are assigned in every work shift to one of the two circuits. In this paper we pay attention to the patient flow management in the less critical patient circuit (although results could be also generalized to the more critical patient circuit). Figure 2 shows the flowchart of a patient through the ED. Patients arrive either by

own means (normal arrivals) or in ambulance. In a very short time they access to the examination room where a triage process classifies the patients according to their severity. After triage, some patients have to carry out some clinical tests required in the examination room. Then, all patients wait for the first consultation which can result in discharging the patient from the ED (to a hospital ward or to patient's home) or in requesting some additional tests. When the tests are carried out and their results obtained, the patients wait for a second consultation. The outcome of this second consultation for the patient can be to be discharged from the health system or the admission in a hospital ward. After concluding a consultation, a physician has to pick a pending patient up to provide a medical consultation, either her/his first one or her/his second one. The queue discipline implemented by the physician greatly influences the quality of service measures, which are discussed in the next section.

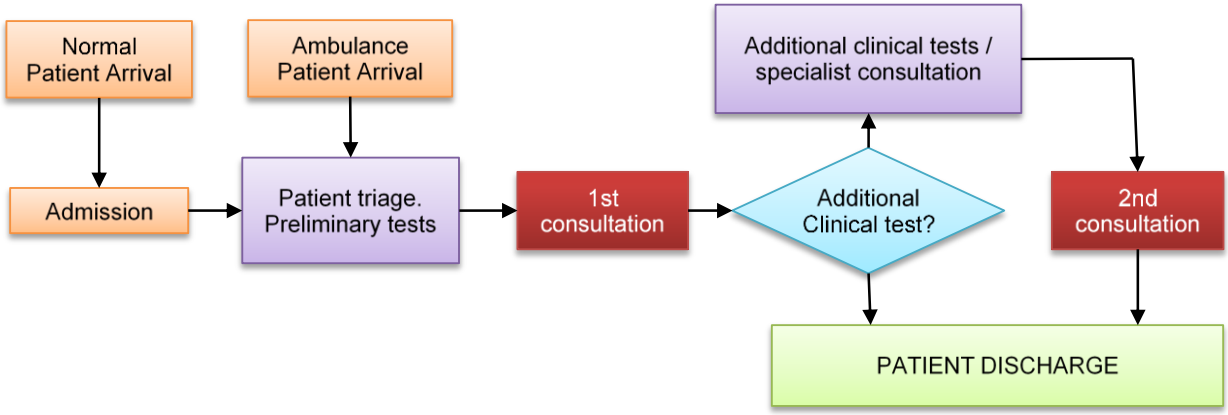


Figure 2: Patient flowchart in the ED.

### 2.3 Key Performance Indicators for Health Service Quality Measurement

Assuring quality care in EDs requires the development of indicators that are valid, relevant, and feasible (Lindsay et al. 2002). In this paper several indicators are used to measure performance with the ultimate aim of improving outcomes for ED patients. As we have said before, one of the highlighted ED performance measures is the arrival to provider time. It is of special importance in emergency healthcare services since a delay in the diagnostic evaluation by a qualified medical provider could cause health risk to patient, who might get worse. This time interval is used as one of the main criteria to optimize the EDs performance. Usually, all EDs define a maximum waiting time for each type of acuity level and set performance goals related with them, as it is exposed in Table 1. Nevertheless, there are other important measures which are influenced by the patient flow management policies: the arrival to discharge time, called "length of stay", is important also for the patient's quality perception on the received healthcare service. Both measures, time to first consultation and length of stay, will be measured for each class of patients determined in triage. Goals for each measure differ in each class: for example patients in class 1, the most urgent, should be immediately seen by the physicians, while not urgent patients can wait until 120 minutes (see Table 1 for CTAS scale). Finally, the last criterion considered is the "overcrowding", which affects the availability of resources, the increase of infection probability, the stress of the physicians and also the patient perception of quality. Overcrowding is measured as the number of patients that are simultaneously in the ED.

## 3 PATIENT FLOW MANAGEMENT POLICIES

A patient flow management policy is a rule that determines what patient will be the next attended by a physician after ending a consultation. The rule has to choose a patient from one of the different categories

that are obtained by combining all types of patients and the healthcare service stages (Figure 3). Next we present policies based on pure priority disciplines and on accumulating priority.

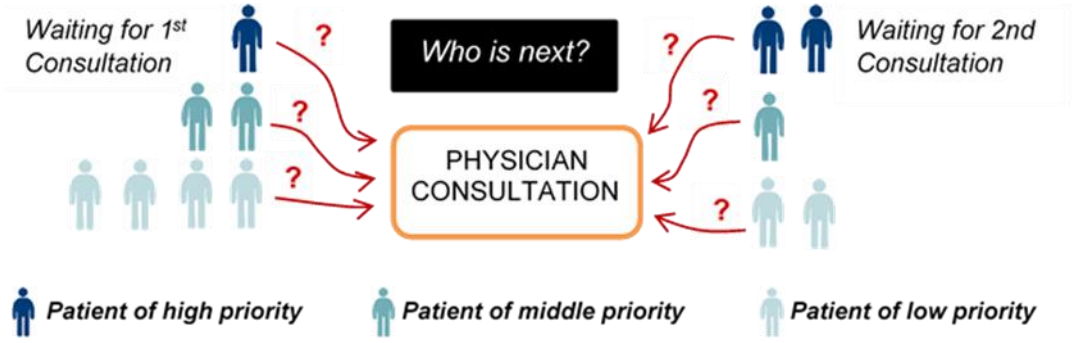


Figure 3: Physician consultation queue structure: several categories of patients in two different stages.

### 3.1 Simple Management Policies

The simplest queue disciplines are those based on pure priority rules. They are also the easiest to implement, which is very convenient in a dynamic and stressful environment like the ED, especially when physicians have to apply them. The priority order among the different illness acuity levels is straightforward after the triage process, however it is not so clear whether to prioritize first consultation over the second one or vice versa.

Priority disciplines are obtained by combining the priority rule for the patient type with the priority rule for the consultation. Taking into account the importance of the waiting time for the first consultation, the prioritization of this first consultation over the second leads to the 1st Consultation Priority Rule (named PR-1C). Under this discipline physicians always pick the next patient up in the following order: first consultation of high priority patients (1C-HP), first consultation of medium priority patients (1C-MP), first consultation of low priority patients (1C-LP) and then the corresponding second consultations, 2C-HP, 2C-MP and 2C-LP.

Other priority disciplines are possible. For example, giving more importance to the second consultation than to the first one, we obtain the queue discipline PR-2C. The strict prioritization of patients according to their illness acuity index provides the queue discipline PR-AI. Table 2 contains these three queue disciplines with full description of the order in which patients are picked up.

Table 2: Order of consultation of patients according to pure priority disciplines. The types of consultations are denoted by xC-YP, standing the xth consultation, x=1,2, of the patient with priority index Y, Y=H(High), M(Medium), L(Low).

| Discipline | 1st class | 2nd class | 3rd class | 4th class | 5th class | 6th class |
|---|---|---|---|---|---|---|
| *PR-1C* | 1C-HP | 1C-MP | 1C-LP | 2C-HP | 2C-MP | 2C-LP |
| *PR-2C* | 2C-HP | 2C-MP | 2C-LP | 1C-HP | 1C-MP | 1C-LP |
| *PR-AI* | 1C-HP | 2C-HP | 1C-MP | 2C-MP | 1C-LP | 2C-LP |

Each one of these priority disciplines is focused on achieving one specific objective. Discipline PR-1C tries to hierarchically minimize the access time to doctor by prioritizing all the first consultations. Discipline PR-2C hierarchically minimizes the number of patients in ED by discharging patients as soon as possible giving priority to all second consultations. Discipline PR-AI is focused in the access time to doctor and then LoS of patients, again hierarchically according to the illness acuity index.

### 3.2 Accumulating Priority Queue (APQ) Management Policies

The APQ management policy generalizes the pure priority queue discipline by setting a discipline based on priority points that patients of class $i$ accumulate at a rate $\beta_i$, where $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_k$ and $k$ is the number of different classes of patients. A class-$i$ customer arriving at time $t_0$ has accumulated $\beta_i(t - t_0)$ priority points, *PP(t)*, by time $t$. When the physician finishes a consultation the next patient to be seen is the one with more priority points. This model was studied by Stanford et al. (2014) in the single server case and Li and Stanford (2016) for the multi-server case, both cases with Poisson arrivals and identical exponential distributed service times. They obtained the waiting time distribution for each priority class, generalizing the results of Kleinrock (1964) who derived recursive expressions for the mean waiting time. As the authors already noted, the APQ model includes the FCFS model, obtained by setting $\beta_1 = \beta_2 = \cdots = \beta_k$, and the pure priority discipline, obtained by setting $\beta_i = M * \beta_{i+1}, i = 1, \ldots, k-1$ and M a sufficiently large value. Between both extremes of relationship among the set of beta parameters (equality and very large differences) it is possible to select appropriate values for them in order to weigh the waiting time allowing to overtake a higher priority patient.

To stress the capacity of the APQ policy to overtake patients with higher priority, in order to fulfill a time limit for the waiting time to the first consultation, we propose a modification of the APQ policy that takes into account these time limits. We name this modified policy as APQ with finite horizon policy and denoted it by APQ-h. The difference with the original APQ is that the accumulation of priority points at a constant rate finishes at the time limits for the consultation. From this moment no more priority is accumulated. This truncated APQ model is represented in Figure 4.

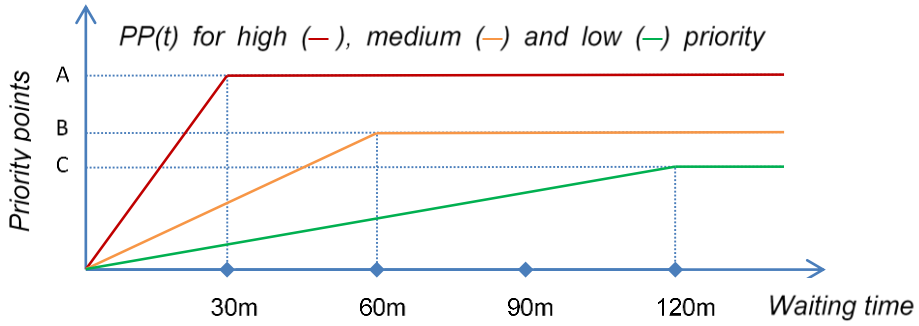


Figure 4: Accumulation of priority points with the APQ with finite horizon policy for patients according to their acuity level.

The parameters that define the APQ-h policy, given the time limits, are the maximum of priority that can be accumulated by each priority class (the values A, B and C in the Figure 4). It is reasonable to assume $A \geq B \geq C$. These values, together with the time limits determine the slopes $\beta_i's$ of the accumulating priority functions. When two patients have reached their waiting time limit then the queue discipline behaves like a pure priority discipline.

### 3.3 Determining Optimal APQ Management Policies

Determining the values of the APQ policy parameters that minimize objectives and/or fulfill constraints related with the KPI defined in Section 2.3 is not straightforward in a real setting because of its complexity: arrivals are non-homogeneous, there are non-Markovian service times and the ED patient flow of patients is made up of several stages. Simulation-based optimization (SBO) is a usual tool for analysis in the manufacturing context but not so much in healthcare system analysis, although it has already been used to find optimal assignment of resources in ED's (e.g. Ahmed and Alkhamis 2009) and to find optimal management policies at hospital departments, (e.g. Mallor et al. 2016; Azcárate et al.

2012; Mallor and Azcárate 2014, in the case of Intensive Care Units). In this research, we propose a simulation model to assess the performance of the ED ruled by APQ policies and a heuristic optimization method to find new promising APQ policies. From a modelling point of view, one of the most challenging tasks is the definition of the objective function that has to be optimized. Physicians try to optimize the three types of objectives described in Section 2.3: Minimizing the percentage of patients exceeding the time limit to first consultation ("door to doc" time), minimizing the Length of Stay and minimizing the work in progress (as indicator of overcrowding). These objectives may have different importance among them and among the different priority patients.

We introduce the following notation to define the multi-objective optimization problem:

- $i \equiv$ index denoting the type of patient according to the illness acuity index, $i = 1,2,3$ refer to patients of high, medium and low priority, respectively.
- $T_i \equiv$ time limit for accessing 1st consultation for patients of type $i$
- $W_i \equiv$ first consultation access time for a patient of type $i$
- $X_i \equiv \begin{cases} 0 \; if \; W_i \leq T_i \\ 1 \; if \; W_i > T_i \end{cases}$
- $P_i \equiv$ target for the maximum ratio of patients of type $i$ exceeding the time limit to first consultation
- $E(X_i) \equiv$ probability that a patient of type $i$ exceeds time limit $T_i$
- $E(LoS_i) \equiv$ expected length of stay in ED of a patient of type $i$
- $E(L) \equiv$ expected number of patients in the ED
- $\beta 1_i \equiv$ slope of the linear function accumulating priority for patient type $i$ in the 1st consultation
- $\beta 2_i \equiv$ slope of the linear function accumulating priority for patient type $i$ in the 2nd consultation

The decision variables of the optimization problem are the six slopes $\beta 1_i$, $\beta 2_i$, $i$=1,2,3. In case of a truncated linear function (policy APQ-h, represented in Figure 4), the slopes can be substituted by the maximum values $A$, $B$, $C$ as decision variables. The time limits $T_i$ and the ratio $P_i$ are parameters of the problem reflecting service quality goals, and the expectations $E(X_i)$, $E(LoS_i)$, $E(L)$, are the functions to be minimized.

Then the problem to find the optimal APQ management policy can be formulated as:

$$\underset{\beta 1_i \, \beta 2_i}{\text{minimize}} \{\mathrm{E(X_1)}, \mathrm{E(X_2)}, \mathrm{E(X_3)}, \mathrm{E(LoS_1)}, \mathrm{E(LoS_2)}, \mathrm{E(LoS_3)}, \mathrm{E(L)}\}.$$

We address this multi-objective problem by the weighted sum method, which provides a solution that reflects the manager preferences by translating them in the selection of weights. This method can also provide multiple solutions by varying the weights consistently. In addition, we introduce the service quality goals for the first consultation as constraints of the problem, minimizing the deviations to these goals in the objective function.

$$\underset{\beta 1_i \, \beta 2_i}{\text{minimize}} \; W_1(u_1\Delta_1 + u_2\,\Delta_2 + \, u_3\Delta_3) + \, W_2\big(v_1 E(LoS_1) + v_2\, E(LoS_2) + v_3\, E(LoS_3)\big) + W_3 E(L)$$

$$\begin{aligned} subject\ to \quad & E(X_i) - \Delta_i \leq P_i \qquad i = 1,2,3 \\ & \Delta_i \geq 0 \end{aligned} \tag{1}$$

Weights $\mathrm{W_1}, \mathrm{W_2}, \mathrm{W_3}$ express the different importance given by the manager to exceed the goals Pi, compared with the reduction in one time unit the expected LoS and the reduction in one person the average of the patients in the ED. On the other hand, the sets of weights $\{\mathrm{u_1}, \mathrm{u_2}, \mathrm{u_3}\}$ and $\{\mathrm{v_1}, \mathrm{v_2}, \mathrm{v_3}\}$

indicate the relative importance of achieving each objective in each type of patient. If the importance of types of patients is objective-independent then $u_i = v_i$.

Neither the objective function nor the constraints have an explicit expression in terms of the decision variables. They are stochastic functions that need to be evaluated by simulation. In this complex and stochastic optimization environment it is necessary to apply a SBO approach to solve the problem (1). Simulation and optimization are combined as follows. The optimization procedure determines a solution, that is, values for the slopes $\beta 1_i$, $\beta 2_i$, *i=1,2,3* that define an APQ policy. The ED is simulated under this policy and the KPI's recorded, which are used in the optimization procedure to evaluate the random objective and the fulfillment of the constraints. Using this information and its own search method, the optimization procedure decides the next APQ policy to be evaluated. This process continues until the stopping conditions of the optimization method are met.

The SBO technique was implemented in ARENA simulation software (Rockwell Automation, Version 15.0) and OptQuest optimization software (Laguna and Martí 2002).

## 4 CASE STUDY

This section reports the results of studies conducted to evaluate the previously proposed APQ management policies and their comparison with pure priority rules. Their efficacy is tested by using data coming from a real setting, the Emergency Department of the Hospital Compound of Navarre (HCN).

### 4.1 Description of the ED

The ED of the HCN is located in Pamplona and assists a population of half a million people with more than 140,000 annual users. The ED is staffed 24 hours per day with board-certified emergency physicians. In this study we focus on the care circuit for less critical patients (with severity index 3, 4 or 5) that has an average arrival rate of 11,26 patients per hour across the day (8:00-22:00). This arrival rate has a high intraday seasonality, as it is show in Figure 5. There are five physicians scheduled with a server rate of 2.60 patients per hour each one. Then, the average service utilization across the day is 81.38 % but the arrival rate is above the service rate for 3 hours (9:00-12:00). The maximum arrival rate peak is produced at 11:00 with a value of 123% the service rate (see Figure 5). Table 3 contains the quantitative description of the patient flow through the ED: percentages of each type of patient, probability distributions for durations of first and second consultations and the discharge probabilities after the first consultation.

Table 3: Percentage type of patient, parameters of the lognormal distribution for the consultation duration and probability of discharge after C1.

| Priority | % | Service time for first consultation (lognormal) | | Service time for second consultation (lognormal) | | Discharge probability after 1st consultation |
|---|---|---|---|---|---|---|
| | | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | |
| High | 40.63 | 2.89 | 0.45 | 2.29 | 0.45 | 0.361 |
| Medium | 32.3 | 2.71 | 0.45 | 2.12 | 0.45 | 0.513 |
| Low | 27.05 | 2.49 | 0.45 | 1.89 | 0.45 | 0.177 |

### 4.2 Simulation and Optimization Models

We built a discrete event simulation (DES) model representing the queue system that combines the patient flowchart described in Figure 2 and Section 2 and the queue disciplines described in Section 3 and Figure 3. This DES model is fully specified by the set of events (arrival of patients to the physician waiting room and end of physician consultation), the stochastic mechanism that generates them in the future, the consequences on the state variables when they occur and the queue discipline determining the order in which patients access to physicians. The hourly ratios of the non-homogeneous Poisson process

modelling each patient type arrival process and their service time distributions are estimated by data analysis of administrative records. This model was implemented in Arena Software and a 3D visualization was developed to facilitate validation by hospital experts.

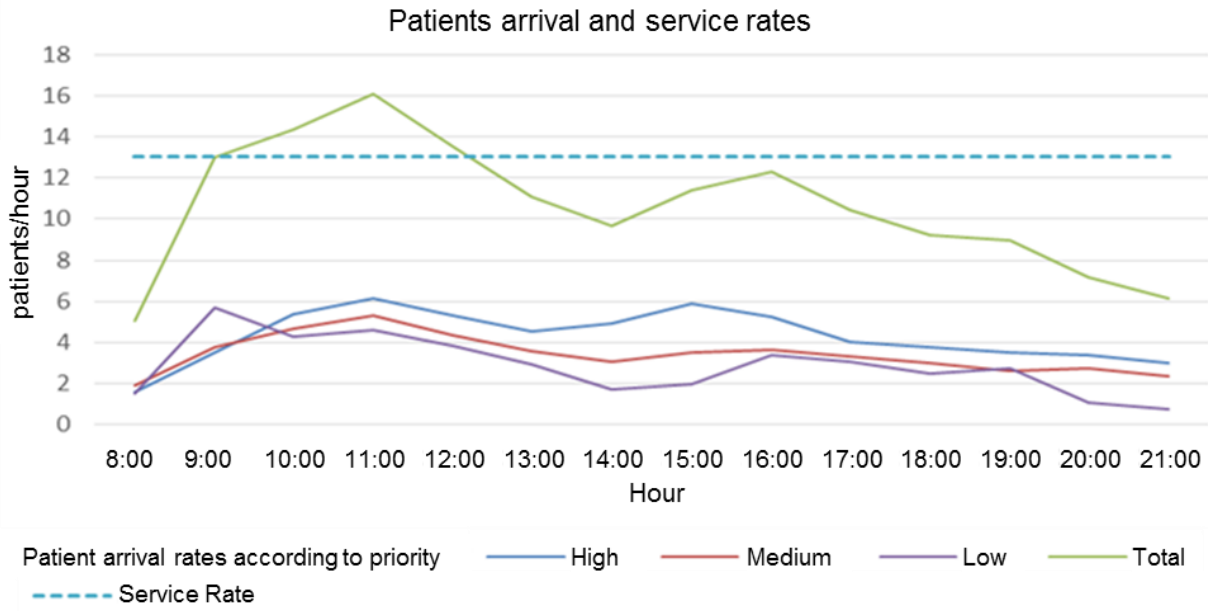


Figure 5: Arrival rate to the ED for each type of patient and the total. Aggregated service rate.

Pure priority disciplines, defined in Section 3.1, are evaluated with this simulation model to collect the estimations of the three KPI presented in Section 2.3. The parameters that define the APQ policies are obtained by solving the optimization problem (1). The assignment of values to all weighting parameters included in the objective function is the greatest challenge in the formulation of this optimization problem. The weighting parameters $v_1$ result from the comparison of the importance of aggregating one minute more to the waiting time among the types of patients. For example, physicians can consider that one hour of *LoS* of a patient with high priority is comparable with 90 minutes of *LoS* of a medium priority patient and two hours of a low patient priority. This assessment provides weights $v_1 = 2, v_2 = 1.5, v_3 = 1$. By reasoning similarly with the percentage of patients exceeding the access time limit to first consultation it is possible to determine the weighting parameters $u_i$. Weights $W_i$ compare the importance of different KPI's. To determine their values it is necessary to answer questions like the following: to assess the quality of the healthcare service, how many minutes of reduction in the *LoS* would be equivalent to a reduction of one percent the percentage of patients exceeding the time limit to access the first consultation? If the answer to this question were 15 minutes then the values of weights would be $W_1 = 15, W_2 = 1$ , for the objectives measured in percentage and minutes, respectively. Similarly, it can be determined the value for $W_3$.

Once the weights in the objective function and the ratio limits $P_i$ in the constraints are defined, the SBO approach is used to obtain the optimal values for the slopes $\beta 1_i, \beta 2_i$ in the accumulating functions.

### 4.3 Design of Experiments and Results

We carried out preliminary analysis by running the simulation model for 15,000 days, under different scenarios, defined by the management policies and the occupancy ratio. The KPI estimations were collected and graphically represented as a function of the number of simulated days in order to identify the stabilization point. As result of this analysis it was determined that 1,500 simulation days are enough to get good KPI estimations (see Figure 6).

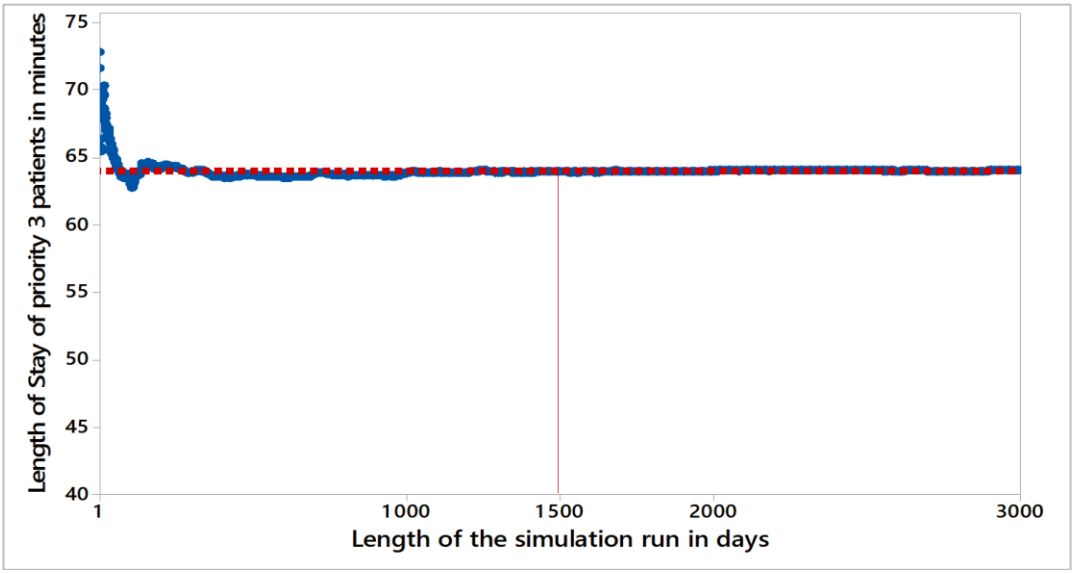


Figure 6: Length of stay of high priority patients and percentage of low priority patients above the time limit as a function of the number of simulated days.

We simulate the real ED with two sets of goals for the percentage of patients exceeding the time limit for the first consultation, (10, 15, 20) and (5, 10, 20) for $P_1$, $P_2$ and $P_3$, respectively. Patient arrival rates are changed to explore scenarios with higher and lower occupancy ratios (the current $\rho$ is increased and decreased in 5%). Furthermore, the influence of the seasonality in the arrival pattern is assessed by simulating scenarios with constant arrival rate. The representative APQ policy in each scenario is obtained by optimizing with the SBO methodology the objective function that describes the goals for the ED performance. By varying the weights in the objective function it can be investigated their influence in the resulting management policy.

As example of the obtained results Table 4 shows the KPI for pure priority disciplines and the APQ-h discipline. The simulated scenario has non-homogeneous Poisson arrivals as described in Figure 5 (ρ=76%) and values 0.05, 0.10 and 0.20 for ratios $P_1$, $P_2$ and $P_3$, respectively.

Table 4: KPI for priority disciplines (PR-AI, PR-1C and PR-2C) and APQ-h

| | **Discipline** | **PR-AI** | **PR-1C** | **PR-2C** | **APQ-h** |
|---|---|---|---|---|---|
| L, number of patients in the system. | | 20.3526 | 24.7961 | 16.8476 | 22.164 |
| Ratio of patients above the time limit | High (<0.05) | 0.0007 | 0.0008 | 0.0375 | 0.0191 |
| | Medium (<0.10) | 0.1421 | 0.0154 | **0.2532** | 0.0968 |
| | Low (<0.20) | **0.4205** | 0.0485 | **0.5224** | 0.1867 |
| Second consultation waiting time | High | 9.9462 | 89.7063 | 4.704 | 51.479 |
| | Medium | 61.253 | 186.06 | 4.68 | 77.0634 |
| | Low | 145.72 | 267.22 | 4.4347 | 236.87 |
| Objetive function value | | 232.9232 | 28.8693 | 481.0742 | **19.3627** |

Results show that APQ-h and PR-1C are able to achieve the goals for the percentages exceeding the time limit but not the others, but both obtain worse results for the waiting time. The APQ-h policy was obtained from OptQuest after applying an automatic stop criterion. The values for the maximum priority accumulated waiting for the first consultation is 297 (slope 9.9), 175 (slope 2.92) and 28 (slope 0.23),

respectively, for each type of patient. Then, the priority accumulated by a high priority patient in one minute equals the priority accumulated by a medium priority patient in 3.5 minutes and by a low priority patient in 43 minutes. The slopes for the second consultation are 1.14, 0.0315 and 0. It means that low priority patients are only seen by the physician when the ED is with no higher priority patients.

## 5 CONCLUSIONS

Applying appropriate policies for the patient flow management is important for the quality of the healthcare, as it is shown by the results of the simulation model. However, the uncertainty environment in which the ED evolves and the high pressure of the demand makes difficult for the physicians to know the best way to proceed. In consequence, for the sake of simplicity pure priority disciplines are usually applied. This is an easy way of handling the queue of patients and provides, in general, good results. Nevertheless, these policies are independent of the specific objectives set for the KPI and they apply always the same independently of the composition of the patient waiting queue.

In this paper we have investigated the use of APQ disciplines, which are adapted to the quality goals set by the ED managers and take into account the composition of the patient waiting queue. By applying a SBO methodology the APQ discipline that best achieves the quality goals is obtained. Then, the applied APQ policy is adaptive in the sense that the optimal parameters that defined it depend on the quality objectives and the patient mix.

The APQ-h policy obtained with the optimization problem determines that a patients waiting for first consultation (1C) are seen by the physician before those waiting for second consultation (2C) for each type of patient, as PR-1C dictates. However, this order can be reversed. For example, a high priority patient waiting for 2C more than 7 times the waiting time of a high severity patient waiting for the 1C is chosen in the first place. Simulation results show, in general, good behavior of the new type of management policies, being worthy to explore its implementation in practice. Moreover, it would also make sense to investigate a non-linear rate of accumulating priority taking into account the slack of a patient until exceeding the time limits.

## ACKNOWLEDGMENTS

This work was supported under grant MTM2016-77015-R (AEI, FEDER EU).

## REFERENCES

Aguado-Correa, F., M. Herrera-Carranza, and N. Padilla-Garrido. 2016. “Variability and Overcrowding Management: Ongoing Challenge for Spanish Hospital Emergency Departments”. *Journal of Health Management* 18(2):218–30.

Ahmed, M. A., and T. M. Alkhamis. 2009. “Simulation Optimization for an Emergency Department Healthcare Unit in Kuwait”. *European Journal of Operational Research* 198(3):936-942.

Azcárate, C., J. Barado, and F. Mallor. 2012. “Calibration of a Decision-Making Process in a Simulation Model By a Bicriteria Optimization Problem”. In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque et al., 782-791. Piscataway, New Jersey: IEEE.

Schuur, J. D., and A. K. Venkatesh. 2012. “The Growing Role of Emergency Departments in Hospital Admissions”. *New England Journal of Medicine* 367(5):389–91.

Kleinrock, L. 1964. “A Delay Dependent Queue Discipline”. *Naval Research Logistics Quarterly* 11(3–4):329–41.

Laguna, M., and R. Martí. 2003. “The OptQuest Callable Library”. In *Optimization Software Class Libraries. Operations Research/Computer Science Interfaces Series*, vol 18, edited by S. Voß and D.L. Woodruff, 193-218, Boston: Springer.

Lee, M. H., J. D. Schuur, and B. J. Zink. 2013. “Owning the Cost of Emergency Medicine: Beyond 2%”. *Annals of Emergency Medicine* 62(5):498–505.

Li, N., and D. A. Stanford. 2016. "Multi-Server Accumulating Priority Queues with Heterogeneous Servers". *European Journal of Operational Research* 252(3):866–78.
Lindsay, P., M. Schull, S. Bronskill, and G. Anderson. 2002. "The Development of Indicators to Measure the Quality of Clinical Care in Emergency Departments Following a Modified-Delphi Approach". *Academic Emergency Medicine* 9(11):1131–39.
Luo, W., J. Cao, M. Gallagher, and J. Wiles. 2013. "Estimating the Intensity of Ward Admission and Its Effect on Emergency Department Access Block". *Statistics in Medicine* 32(15):2681–94.
Mallor, F., and C. Azcárate. 2014. "Combining Optimization with Simulation to Obtain Credible Models for Intensive Care Units". *Annals of Operations Research* 221(1):255–71.
Mallor, F., C. Azcárate, and J. Barado. 2016. "Control Problems and Management Policies in Health Systems: Application to Intensive Care Units". *Flexible Services and Manufacturing Journal* 28(1–2):62–89.
National Audit Office. 2018. "Reducing Emergency Admissions". Report No. 833 Session 2017–19. Department of Health & Social Care, NHS England.
Rockwell Automation Technologies, Inc. 2016. "Arena Simulation Software."
Sharif, A. B., D. A. Stanford, P. Taylor, and I. Ziedins. 2014. "A Multi-Class Multi-Server Accumulating Priority Queue with Application to Health Care". *Operations Research for Health Care* 3(2):73–79.
Stanford, D. A., P. Taylor, and I. Ziedins. 2014. "Waiting Time Distributions in the Accumulating Priority Queue". *Queueing Systems* 77(3):297–330.
Strunk, B. C., P. B. Ginsburg, and M. I. Banker. 2006. "The Effect of Population Aging on Future Hospital Demand". *Health Affairs* 25(3):141-149
Tang, N., J. Stein, R. Y. Hsia, J. H. Maselli, and R. Gonzales. 2010. "Trends and Characteristics of US Emergency Department Visits, 1997-2007". *Jama* 304(6):664–70.
Twomey, M., L. A. Wallis, and J. E. Myers. 2007. "Limitations in Validating Emergency Department Triage Scales". *Emergency Medicine Journal* 24(7):477–79.
Welch, S. J., B. R. Asplin, S. Stone-Griffith, S. J. Davidson, J. Augustine, and J. Schuur. 2011. "Emergency Department Operational Metrics, Measures and Definitions: Results of the Second Performance Measures and Benchmarking Summit". *Annals of Emergency Medicine* 58(1):33–40.

## AUTHOR BIOGRAPHIES

**MARTA CILDOZ** studied industrial engineering at the Public University of Navarre, Spain. Currently, she is a Ph.D. student at the Institute of Smart Cities of the Public University of Navarre. Her research interests lie in the field of complex real problems simulation modelling. Her email address is marta.cildoz@unavarra.es.

**AMAIA IBARRA** is a physician at the ED of the Hospital of Navarre, Spain. Currently, she is a Ph.D. student at the department of statistics and operations research of the Public University of Navarre. Her research interests lie in the field of ED simulation modelling and in the study of clinical decision-making processes. His email address is amaia.ibarra.bolt@navarra.es.

**FERMIN MALLOR** is Full Professor of Statistics and Operations Research at the Public University of Navarre, Spain. He holds a PhD and MSc in Mathematics. He has been visiting researcher at the Missouri Science and Technology University, among others. His research interests include applications of simulation-optimization (classical and metaheuristics methods) in health, energy, logistics and production. His email address is mallor@unavarra.es.