Contents lists available at ScienceDirect

Operations Research for Health Care

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

Accumulating priority queues versus pure priority queues for managing patients in emergency departments

Marta Cildoz [a], Amaia Ibarra [b], Fermin Mallor [a,*]

[a] Institute of Smart Cities, Public University of Navarre, Campus Arrosadia, Pamplona, 31006, Spain
[b] Hospital Compound of Navarre, Irunlarrea, 3, Pamplona, 31008, Spain

ARTICLE INFO

Article history:
Received 31 October 2018
Accepted 21 October 2019
Available online 25 October 2019

Keywords:
Queue discipline
Accumulative priority queues
Patient flow
Emergency department
Simulation-based optimization

ABSTRACT

Improving the quality of healthcare in emergency departments (EDs) is at the forefront of many hospital managers' efforts, as they strive to plan and implement better patient flow strategies. In this paper, a new approach to manage the patient flow in EDs after triage is proposed. The new queue discipline, named accumulative priority queue with finite horizon and denoted by APQ-h, is an extension of the accumulative priority queue (APQ) discipline that considers not only the acuity level of patients and their waiting time but also the stage of the healthcare treatment. APQ disciplines have been studied in the literature from a queueing theory point of view, which requires assumptions rarely found in real EDs, such as homogeneity in the patient arrival pattern and only one service stage. The APQ-h discipline accumulates priority from the point of waiting for the first physician consultation until the moment the waiting time exceeds the upper time limit set to access the physician after the patient's arrival. A recent study shows that a management strategy of this type is applied in practice in several Canadian EDs. The main aim of this paper is to explore the implementation of APQ-h managing policies in a real ED. For this purpose, a simulation model replicating a real ED is developed. This simulation model is also used to obtain the optimal APQ type polices through a simulation-based optimization method that solves a multi-objective and stochastic optimization problem. Arrival to provider time and total waiting time in the ED are considered to be the key ED performance indicators. An extensive computational analysis shows the flexibility of the APQ-h and APQ discipline and their superiority over other pure priority disciplines in a real setting and in a variety of ED scenarios. In addition, no superiority over the APQ discipline is demonstrated.

1. Introduction

Growth in the utilization of emergency care is observed in high income countries. For instance, emergency admissions grew over 50% from 1992 to 2006 in the US [1] and 9.3% from 2014 to 2017 in England [2], mainly due to the ageing population, which encompasses the main consumers of healthcare services. Some studies quantify that this factor itself can explain 40%–50% of the total growth [3,4]. This trend is expected to continue in the near future. As a result of this growth, the National Center for Health Statistics [5] estimated 43.3 visits to emergency departments (ED) in the US per 100 persons in 2015, which equals a total of 136.9 million visits. Nevertheless, the capacity of healthcare services does not follow the demand growth pace, and it even decreases in some cases [5]. For example, regarding the number of hospital beds per 1000 habitants in the US, which was 4.5 per 1000 in 1980 and 2.5 per 1000 in 2014. Thus, the mixture of a growing demand and a fairly stable capacity of service leads to overcrowded EDs; approximately half of all EDs report operating near or above maximum capacity [6]. This restrictive environment makes operational health care management even more critical, and it is important to guaranty the quality and universality of public healthcare services.

However, EDs are especially difficult to manage; they evolve in a highly stochastic environment due to the variability in the patient arrival rate, illness severity, and, in general, the health resources needed for treatment (material and human) [7–9]. In this situation of resource scarcity, the grouping of patients according to their urgency to receive healthcare treatment is a strategy commonly used. Thus, upon arrival, patients undergo an initial assessment, i.e., triage, whose aim is to stratify them by illness acuity and prioritize them accordingly [10]. Examples of triage systems are the Emergency Severity Index (ESI); the Australasian Triage Scale (ATS), the Manchester Triage Scale (MTS), and the Canadian Triage Acuity Scale (CTAS). However, most triage systems do not provide explicit guidelines on how to manage the patient flow within and among the, usually five, assigned triage levels.

[*] Corresponding author.
E-mail address: mallor@unavarra.es (F. Mallor).

https://doi.org/10.1016/j.orhc.2019.100224

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%

Triage systems may include performance goals in terms of the percentage of patients who should have access to the physician consultation before certain time limits and should have a different time limit and a different percentage for each type of triage-level (see Table 1). ED managers and physicians, motivated by the achievement of such goals, follow pre-determined rules, such as FCFS, within the same triage level and strictly follow priority across different triage levels. Nevertheless, very often, especially in cases of overcrowding, they have to use their own discretion in making patient-routing decisions, as is mentioned in [11]. In this paper, by using patient-level ED visit data, the authors carried out an empirical study to understand how decision makers manage patients in the ED. They concluded that, generally, higher triage-level patients receive priority over low triage-level patients, but a lower triage-level patient who has waited longer can be prioritized over a higher triage-level patient who has waited less time. Then, they highlighted the need to consider not just the triage level but also the actual wait time in routing decisions. Therefore, the behaviour of these patient flow managers fits the so-called accumulative priority queue (APQ) policies, a term introduced by Stanford et al. [12]. Following this APQ strategy, patients accumulate “priority points” as they wait for treatment, and the patient with the most priority points is selected when a physician finishes a service. The accumulation rate of priority points depends on the patient’s triage level. In addition, in [11], it is also concluded that patients who have waited past the target set by the triage system (for example, 30 min for patients level 3 in the CTAS triage system; see Table 1) may not receive extra priority. The ED decision makers’ behaviour is described by a two piecewise linear concave marginal waiting cost function for each triage level, in which the break point is located around the target wait times. This important empirical observation suggests a modification of the classical APQ policy to define the new APQ-h (accumulative priority queue with a finite horizon) policy, which linearly accumulates priority points while the patient is waiting until the target wait time is reached, and then, no more priority points are accumulated.

The implementation of priority strategies in a real ED needs to consider not only the prioritization of patients to access to their first physician consultation but also the management of patients already in the process of being treated. After the first consultation with a physician, some patients are discharged from the ED, while others require some clinical tests and, once the results are obtained, have a second consultation with a physician. Thus, patient management should consider the following two components of the patient flow: first, the patients arriving from triage that must be served within time-deadlines and, second, the patients already being treated, both of which have a significant feedback constituent that produces operational congestion. Therefore, when a physician becomes idle, a decision has to be made regarding whether the next patient to be seen is having their first or for a second consultation; that is, managing the portfolio of pending patients must consider both the severity of the condition and the stage of their treatment.

The consideration of conflicting objectives is typical in the analysis of healthcare systems, as in all public services in which cost objectives compete with service quality objectives. Even in a case with fixed resources, as in the case of determining operative rules for optimally managing the ED patient flow, there are several conflicting objectives that guide the measurement of management performance. One of the main ED performance measures is the arrival to provider time (APT) (“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 [13]. Another important objective is minimizing the length of stay (LoS) in the ED. As was stated before, the upper limit for the APT is set for each type of patient (see Table 1) but can also be imposed onto the other performance measures; for example, EDs in hospitals in the United Kingdom should complete and discharge 98% of patients within 4 h, as it is mandated by the government (Mayhew and Smith [14]). The patient-flow management strategy should be selected to accomplish the goals and optimize the objectives set by the hospital direction board. One main characteristic of the APQ and APQ-h management policies is their capacity to represent very different dynamic priority rules by changing the value of the rates at which the different types of patients accumulate priority.

The main aim of this paper is to explore the implementation of APQ-h managing policy in a real ED framework that considers the acuity level of patients, the stage of treatment, the stochasticity of the ED and the different objectives set by managers. Specifically, the main contributions of this paper are as follows:

- The analysis of the ED patient flow management problem in a setting not previously considered in the literature to include the different acuity levels of the patients, several stages for treatment, the stochastic environment in which EDs evolve and different key performance indicators (KPIs).
- The proposal of the APQ-h policy to represent the real patient-routing decision making observed in empirical studies.
- The definition of a multi-objective and stochastic optimization problem to obtain the optimal APQ type policies which is solved by a simulation-based optimization method.
- Testing the performance of APQ-h (and APQ) policies by using a simulation model that reproduces the main features of a real ED, i.e., the stochasticity in arrivals, service times and paths thorough the ED.
- A sensitivity analysis to determine the influence of the optimal APQ-h and APQ policies on the structure and on the ED’s KPI of factors, such as the variability in the patient arrival pattern, the mix of patients, and the congestion level.
- Comparing the performance of the APQ-h and APQ with pure priority disciplines, to show their superiority. It also outperforms the priority rule used in the ED of the Hospital Compound of Navarre.
- No superiority of APQ-h over the APQ is demonstrated in the tested scenarios for the objective function proposed.

The rest of the paper is organized as follows. In Section 2, the related literature is reviewed. The characteristics of the ED patient flow and its KPIs are presented in Sections 3.1 and 3.2, respectively. Section 3.3 presents the management policies considered in this paper, the pure priority disciplines and the APQ-h and APQ disciplines, as well as simulation-based optimization methodology to determine the optimal management policy. Section 4 is focused on the case study, in which the main features of a real ED and the simulation model are described, and then, the optimal APQ-h or APQ policy is obtained and its performance is compared with the pure priority disciplines, including the currently followed by the majority of physicians in the ED studied. The last subsection includes a sensitivity analysis on the weights of the objective function. Finally, Section 5 presents the results and conclusions from an extended computational analysis carried out to test the pure priority disciplines and optimal APQ-h and APQ disciplines in a variety of EDs defined by different occupancy ratios, patient arrival patterns and mixes of patients. We end the paper with a conclusion section. Finally, Appendix additional numerical results.

Hereunder, Table 2 includes all acronyms used in this paper:

2. Related literature

The improvement of the performance of the ED has been addressed by many operations research studies in recent years, such as in Saghafian et al. [15], in which 350 papers dealing with the ED patient flow are reviewed. They distinguish the following three components of the ED patient flow: flow into, within, and out of the ED. The problem addressed in this paper deals with the patient flow optimization within the ED. In this specific context, Wiler et al. [10] reviews applications concerning patient flow and crowding in the ED.

In most healthcare settings with no appointment system, the queue discipline is either a first-in–first-out (FIFO) or a priority discipline, depending on the acuity of the patient's illness. This priority discipline applies in the ED, where, in general, patients with life-threatening injuries are treated before others. The use of a priority discipline with a FIFO rule inside each class of patients is almost generalized in the analysis of EDs by queueing models and/or discrete event simulations (see, for example, Taylor et al. [16], Haussman [17], Siddharthan and Jones [18], Laskowski et al. [19], Mokaddis et al. [20]). It is worth mentioning the paper by McQuarrie [21] that applies the shortest processing time rule, which is known to minimize waiting times. Although routinely applied in the manufacturing context, it is difficult to justify the use of this dispatching rule in EDs, given its unfairness to the more injured patients and the added difficulty of estimating the treatment times accurately. Nevertheless, this research raises the question of using other queue disciplines than the pure priority discipline to manage ED patient flow. The dispatching rules applied in manufacturing prioritize all jobs waiting for processing on a machine (the classic paper of Panwalkar and Iskander [22] presented a summary of 113 dispatching rules). The same idea can be applied in the ED patient flow management problem, i.e., whenever a physician has finished a patient's service, the dispatching rule selects the patient with the highest priority. Dispatching rules and other prioritizing policies to manage the patient flow in an ED are usually analysed by using queueing theory models or simulation models (or both in combination).

The paper by Armory et al. [23] provides a deep queueing-network view of patient flow in hospitals, with a special focus on EDs and the in wards patient flow, as the natural way for studying and improving its performance. They pointed out how the patient flow within the ED has been widely investigated, both academically (Hall et al. [24]; Saghafian, Austin and Traub, [15]; Zeltyn et al. [25]) and in practice (IHI, [26]; McHugh et al. [27]). Among all these studies, we highlight the paper of Huang et al. [28], which addresses many of the complexities of EDs that are often ignored in queueing models; this paper considers the patient triage level and the feedback of patients after the first consultation. They obtained an asymptotically optimal patient flow policy that is based on the $c\mu$ dispatching rule, which minimizes congestion costs subject to deadline constraints for the first consultation. Their analysis extended the results of Smith [29] and set the optimality of the known as the cμ rule, which prioritizes among the queues of the different categories of patients and then uses the FCFS discipline inside each queue. The waiting cost was assumed to be a linear function of the sojourn time. Later, the paper of Van Mieghem [30] shows that the generalized cμ rule minimizes the average waiting costs under the heavy-traffic asymptotic regime and the cumulative holding cost is a non-decreasing convex function. Mandelbaum and Stolyar [31] and Gurvich and Whitt [32] studied the queue-length version of the Generalized cμ-rule, in which the holding cost is a function of the queue length instead of the sojourn time. The aforementioned paper of Huang et al. [28] is the first to consider feedback and deadlines simultaneously; however, the need to assume a stationary heavy traffic and the use of diffusion approximations to obtain the results do not guarantee the optimality of the proposed control rules in a real setting (for example, it is necessary to assume that during the sojourn time of a patient within the ED, the various queue lengths do not change significantly, and the service duration is negligible relative to the queueing time).

Other types of queueing models developed to study the ED patient flow optimization problem without the need of asymptotic assumptions are those that specifically analyse the APQ strategies (Stanford et al. [12]). 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. In this way, they seek to overcome the drawback of pure disciplines that in periods of high demand, patients of the lowest priority can be "forgotten" in the system for long time periods. Kleinrock [33] obtained results about the mean waiting time before receiving service, which were extended by obtaining the waiting time distribution for each priority class in the single server and in the multi-server settings [12,34]. All these models assume Markovian distributions (Poisson arrivals and exponentially distributed service times), and in addition, there is only one stage for the service without feedback.

The ED healthcare process can be represented by a queueing system with feedback to model the patients who need clinical tests after the first assessment and need to return for a second consultation. [11] models the ED in which patients are waiting to see a physician as a multi-class queueing system and investigates how decision makers choose which patient is the next to be seen by an available physician. They obtained strong evidence of the practical use of a sophisticated prioritization behaviour that is consistent with the APQ-h discipline and that, consequently, supports the research carried out in this paper. Nevertheless, they only consider the prioritization to the first consultation without addressing the feedback of patients already in the process of being treated.

A related research is exposed in the paper of Ferrand et al. [35], where the patient flow management problem is analysed by using a simulation model that reproduces a real life setting that includes different acuity levels, and the stochastic environment. They conclude that dynamic priority queues outperform other approaches based on different implementations of fast tracks for low priority patients. The main difference between the Ferrand et al.'s model [35] and ours is that we consider deadline constraints for the first consultation, whose fulfilment becomes an important goal in addition to the minimization of the LoS (the only one considered in Ferrand et al. [35]). As a consequence, we do not assume the policy of prioritizing treatment over the first consultation as they do, and the management problem is addressed from a bi-objective point of view. In Zayas-Caban et al. [36] the prioritization of treatment is also criticized in a patient management problem focused on maximizing the profit when a reward is obtained from patients that complete the treatment and there could be abandonment of patients before the treatment is complete.

Table 2
Acronyms used in this paper.

Acronym	Definition
General terms	
ED	Emergency Department
KPI	Key performance indicator
CTAS	Canadian Triage Acuity Scale
OFV	Objective function value
SBO	Simulation based optimization
DES	Discrete event simulation model
Management policies	
APQ	Accumulative priority queue management policy
APQ-h	Accumulative priority queue with a finite horizon management policy (an extension of the normal APQ)
PP	Priority points
β_{1i}	Rate at which patients of class i who are waiting for the first consultation accumulate PP
β_{2i}	Rate at which patients of class i who are waiting for the second consultation accumulate PP
PR	Pure priority rule
FCFS	First come first served management policy
FIFO	First in first out management policy
PR-1C	1st Consultation pure priority rule
PR-2C	2nd Consultation pure priority rule
PR-AI	The acuity index pure priority rule
PR-HN	The rule which is used by the majority of the medical stuff in the HCN
Key performance indicators	
APT	Arrival to provider time (“door to doc”)
LoS	Length of stay
TWT	Total waiting time
Classes of patients	
1C	Patients waiting for the first consultation
2C	Patients waiting for the second consultation
HP	High-priority patients
MP	Medium-priority patients
LP	Low-priority patients
1C-HP	High-priority patients waiting for the first consultation
1C-MP	Medium-priority patients waiting for the first consultation
1C-LP	Low-priority patients waiting for the first consultation
2C-HP	High-priority patients waiting for the second consultation
2C-MP	Medium-priority patients waiting for the second consultation
2C-LP	Low-priority patients waiting for the second consultation
P3	Priority 3 patients
P4	Priority 4 patients
P5	Priority 5 patients
Scenario factors and levels	
F1	ED congestion level. It is the average occupation rate, ρ, $f_1 = \{90\%, 95\%\}$
F2	Arrival ($\lambda(t)$) seasonality, $f_2 = \{T0, Tu, Tp\}$
F3	Mix of patients, $f_3 = \{B0, B3, B4, B5\}$
T0	Constant arrival rate of patients $\lambda(\mathrm{t})$
Tu	Triangular pattern for the arrival rate $\lambda(\mathrm{t})$, with a peak at 11:30 a.m. and a ratio of $(\lambda_{\max} - \lambda_{\min})/\lambda_{\min} = 0.5$. It extends the triangular shape across the entire time range
Tp	Triangular pattern for the arrival rate $\lambda(\mathrm{t})$, with a peak at 11:30 a.m. and a ratio of $(\lambda_{\max} - \lambda_{\min})/\lambda_{\min} = 0.5$. It only applies the triangular shape in the time range [10:00, 13:00], with the arrival rate out of this range being constant
B0	Equilibrated. Balanced distribution among all types of patients (1/3 of P3, 1/3 of P4 and 1/3 of P5)
B3	Biased mix of patients towards priority 3 patients (50% of P3, 25% of P4 and 25% of P5)
B4	Biased mix of patients towards priority 4 patients (25% of P3, 50% of P4 and 25% of P5)
B5	Biased mix of patients towards priority 5 patients (25% of P3, 25% of P4 and 50% of P5)

3. Prioritization of patients in emergency departments

3.1. Patient routing

Fig. 1 shows a flowchart of a patient being processed through an ED. Patients arrive either by their own means (normal arrivals) or in an ambulance, and in the first case, the administrative registration process must be carried out. In a very short time, patients gain access to the examination room, where a triage process classifies the patients according to their severity. Usually, EDs organize the patient care into two different care circuits, one for the more critical patients and another for less critical patients. In this paper, we pay attention to the patient flow management in the less critical patient circuit which has dedicated physicians, nurses and ancillaries that are not shared with the most critical patient circuit.

Depending on the hospital and country, the triage process usually uses one of the four ordinal ED triage scales [37] mentioned in the Introduction section. Without a loss of generality, we consider that the triage classifies ED patients on 5 acuity levels, as is the case of CTAS (Table 1: Access time is the upper limit for the arrival to provider time, and performance level is the minimum percentage of patients that should satisfy the access time requirement).

After triage, all patients wait in a queue for the first consultation (red arrow in Fig. 1), in which a physician is needed to evaluate them. This first consultation can result in discharging the patient from the ED (to a hospital ward or to the patient’s home) or in ordering some clinical tests, such as blood tests, X-ray, scan, specialist’s consultation, etc. Once the tests and complementary diagnosis are carried out and their results are ready, the patient re-enters the queue (blue broken line arrow in Fig. 1) and waits for a second consultation with the ED physician to be reviewed before being discharged from the ED.

After concluding a consultation, a physician has to choose a pending patient from the queue to provide a medical consultation. This queue is formed by patients of different priorities, and within these priority categories, patients can be classified into one of the following two categories: new patients who have arrived just after triage or patients who have re-entered the queue for being re-evaluated. The queue discipline implemented by the physician greatly influences the quality of the service measures, which are discussed in the next section.

3.2. Key performance indicators

Assuring quality care in the ED requires the development of indicators that are valid, relevant, and feasible [38]. Welch et al. [39] and Welch et al. [13] list various metrics by which ED performance can be measured, such as the arrival to provider time (APT, or "door-to-doc" time). This important time interval is widely used in emergency healthcare services, since many illnesses are time-dependent, and a delay in the diagnostic evaluation by a qualified medical provider could be a health risk for the patient. Most EDs define a maximum waiting time for each acuity level and set performance goals related to them, as is explained in Table 1's CTAS; for example, class 1 patients, the most urgent, should be immediately seen by the physician, while nonurgent patients can wait up to 120 min. The ratio of patients whose APT exceeds the time limit is considered a KPI.

There are also other important measures influenced by the patient flow management policies, such as the arrival to discharge time, called the "length of stay" (LoS), which has an impact on the patient's quality perception of the received healthcare service. The LoS depends directly on the treatment needed. It will be, in general, much greater for patients who need additional diagnostic tests and a second consultation than for patients who are discharged after the first consultation. The waiting times for the first and the second consultation of a patient with acuity level i are denoted by τ_i and ν_i, respectively. Thus, the total waiting time (TWT) for a patient of acuity level i is τ_i when only one consultation is needed and $\tau_i + \nu_i$ when two consultations are needed. Because the queue discipline implemented to manage the physician waiting room directly impacts those waiting times, the TWT is considered a KPI in this study.

Other KPIs could be considered, such as measuring the "overcrowding" level, which affects the availability of resources and causes an increase in the infection probability, the physician's stress level, waiting times, LoS [40], medical error probability [41], and the patient's perception of quality. Overcrowding occurs when demand exceeds available capacity, i.e., when there is no space left to meet the timely needs of the next patient requiring emergency care; however, according to [42], "No measure is universally applicable as a marker of overcrowding and should be used with caution when comparing performance between institutions". One scoring system that has become a national standard in the United States is the National ED Overcrowding Scale ("NEDOCS", http://www.nedocs.org), whose elements include total patients in the ED, as well as the waiting time of the longest admitted patient, among others. Other studies, such as that of Weiss et al. [43], which found, using multivariate regression analysis, that the combination of patients in the waiting room and the total registered patients was a better model than the NEDOCS score for quantifying paediatric ED overcrowding. Little's formula relates the average number of patients in the waiting room with the average waiting time. Thus, aiming at the minimization of the TWT implies reducing the number of patients in the ED's waiting room, which is a main contributor to overcrowding.

3.3. Patient flow management policies

A patient flow management policy is a rule that determines which patient will be the attended next by a physician when he/she becomes available (after ending a consultation). The implemented policies should be designed to achieve good ED performance, which is assessed by a set of KPIs, such as those defined in the previous section.

As has been mentioned, the physician's queue of pending patients is formed by several categories of patients, which are defined by both the illness acuity level and the healthcare service stage. The patients waiting in the physician consultation waiting room can be in one of two stages, i.e., waiting for the first consultation stage, denoted by **1C**, or waiting for re-evaluation after having some medical test, denoted by **2C**. Without loss of generality, we assume that the patients are classified into three different levels of priority according to their illness acuity as follows: high priority, denoted by **HP**; medium priority, denoted by **MP**; and low priority, denoted by **LP**. Subsequent analysis can be readily adapted to any number of acuity level categories. Therefore, the patients in the physician waiting room can be classified in one of the six categories represented in Fig. 2, which are denoted by **1C-HP**, **1C-MP**, **1C-LP** (high-, medium-, and low-priority patients waiting for the first consultation) and **2C-HP**, **2C-MP**, **2C-LP** (high-, medium-, and low-priority patients waiting for the second consultation).

In the next subsections, policies based on pure priority disciplines and on accumulating priority queues are described.

3.3.1. Pure priority rules

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 such as the ED, especially when physicians have to apply them. A pure priority discipline defines the total order among the categories of patients and chooses the first patient in the non-empty highest priority category. This total order has to be compatible with the partial order induced by the different illness acuity levels in each process stage. That is, in the total order 1C-HP < 1C-MP < 1C-LP and 2C-HP < 2C-MP < 2C-LP; however, this order can be reversed between different consultations, that is, 1C-MP < 2C-HP could be possible. There are 20 different pure priority disciplines satisfying this partial-ordering condition.

In this paper, we consider the four more meaningful pure priority disciplines, named PR-1C, PR-2C, PR-AI, and PR-HN. Table 3 contains a full description of the order in which each category of patients is chosen according to each one of these four pure priority disciplines. The 1st consultation pure priority (PR-1C) rule always prioritizes a first consultation over a second one; thus, the order among categories is as follows: 1C-HP, 1C-MP, 1C-LP, 2C-HP, 2C-MP, 2C-LP. The 2nd consultation pure priority (PR-2C) rule always prioritizes the second consultation over the first one; thus, the order among categories is as follows: 2C-HP, 2C-MP, 2C-LP, 1C-HP, 1C-MP, 1C-LP. The acuity index pure priority (PR-AI) rule prioritizes the patients according to their illness acuity index, and within each priority, it prioritizes the 1st consultation over the 2nd consultation; thus, the order among categories is as follows: 1C-HP, 2C-HP, 1C-MP, 2C-MP, 1C-LP, 2C-LP. Finally, PR-HN is the one that is generally followed by the majority of physicians in the HCN, which combines the PR-AI for HP patients with the PR-2C for the MP and LP patients.

Each one of these priority disciplines is focused on achieving a different objective. Discipline PR-1C attempts to hierarchically minimize the APT by prioritizing all the first consultations. Discipline PR-2C hierarchically minimizes the number of patients in the ED by discharging patients as soon as possible, giving priority

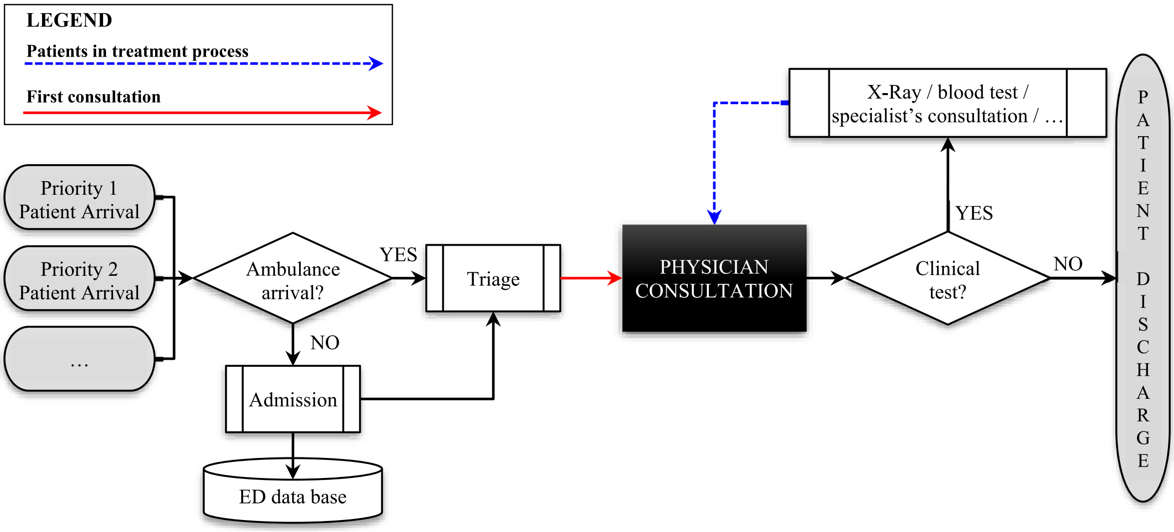

Fig. 1. Patient flowchart in the ED. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

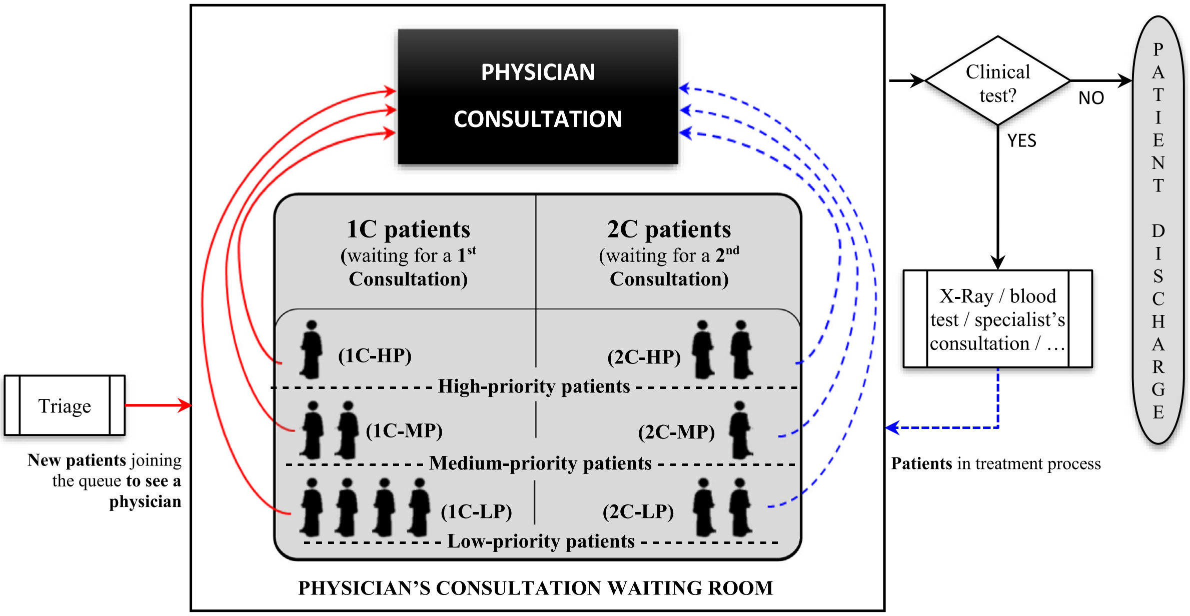
PHYSICIAN CONSULTATION
1C patients (waiting for a 1st Consultation)
2C patients (waiting for a 2nd Consultation)
(1C-HP)
(2C-HP)
High-priority patients
(1C-MP)
(2C-MP)
Medium-priority patients
(1C-LP)
(2C-LP)
Low-priority patients
PHYSICIAN'S CONSULTATION WAITING ROOM
Triage
New patients joining the queue to see a physician
Clinical test?
NO
YES
X-Ray / blood test / specialist's consultation / …
Patients in treatment process
PATIENT DISCHARGE

Fig. 2. Physician consultation queue structure: different priority categories of patients in two different stages.

Table 3
Ordering induced according to the types of patients by several pure priority disciplines.

Discipline	Order induced in the patient categories					
	1st	2nd	3rd	4th	5th	6th
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
PR-HN	1C-HP	2C-HP	2C-MP	2C-LP	1C-MP	1C-LP

to all second consultations to minimize their waiting time in the system. Discipline PR-AI focuses on providing the best possible treatment to the higher priority patients according to the acuity index, assuring the APT limits first and then minimizing the TWT in the ED.

3.3.2. Accumulation priority queues

The APQ management policy generalizes the pure priority queue discipline by setting a discipline based on priority points (PP) that patients of class i accumulate at a rate β_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 $\beta_i(t - t_0)$ PP by time t. When the physician finishes a consultation, the next patient to be seen is the one with the highest PP. Clearly, the APQ model includes the FCFS discipline, obtained by setting $\beta_1 = \beta_2 = \cdots = \beta_k$, and the pure priority disciplines, obtained by setting $\beta_i = M * \beta_{i+1}, i = 1, \ldots, k - 1$ and M to a sufficiently large value. Between both extremes of relationships among the set of beta parameters (equality and very large differences), it is possible to select appropriate values for them to weigh the waiting time, which allows them to overtake a higher priority patient.

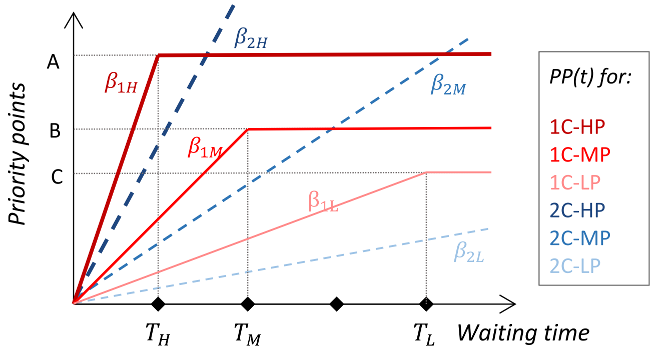

Fig. 3. Accumulation of priority points with the APQ-h policy for patients classified in three acuity levels.

In this study, we also propose a modification of the APQ policy that takes into account the time limit targets for each priority. It is motivated by the empirical study of Ding et al. [11], already described in the Introduction section, that analyses the patient routing behaviours of ED decision makers in four EDs using CTAS in Canada. They found that the behaviour of the routing decision makers is best fit by a piece-wise linear concave marginal waiting cost function for each triage level, in which marginal waiting cost has a significantly positive slope below the point where the slope changes and is nearly constant above the CTAS triage-level target wait times. We name this APQ modified policy as the *APQ with finite horizon* policy and denote it with APQ-h. Therefore, the difference between this new policy and the original APQ is that the accumulation of PP at a constant rate finishes at the waiting time limits for the first consultation. From then on, no more priority is accumulated which remains at the maximum value that can be attained in for patients waiting for the first consultation. However, as there is no waiting time target for the second consultation, the limitation of PP does not apply for patients waiting for the second consultation. This truncated APQ model is represented in Fig. 3.

Therefore, an APQ-h discipline, as well as an APQ discipline, is determined by the vector $\boldsymbol{\beta}$ of the slopes at which the different categories of patients accumulate PP. In our setting with 6 categories of patients, $\boldsymbol{\beta} = (\beta_{1H}, \beta_{1M}, \beta_{1L}, \beta_{2H}, \beta_{2M}, \beta_{2L})$, and in the case of the APQ-h discipline, the parameters associated with the first consultation can be replaced by parameters A, B and C, denoting the maximum PP accumulated at the time limits for the first consultation (see Fig. 3).

3.4. Determination of the optimal APQ-h discipline

In this subsection, we address the problem of finding the optimal values for the vector of parameters $\boldsymbol{\beta}$ that determines the APQ-h (or APQ) discipline with best performance according with the KPIs defined in Section 3.2. The problem is multi-objective and of a stochastic nature. The necessary notation to define the optimization problem is:

i ≡ index denoting the class of the patient according to the illness acuity index, $i = 1, 2, 3$ refer to patients of high, medium and low priority, respectively.

$\bar{\lambda}_i$ ≡ average arrival rate of patients of class i.

T_i ≡ APT limit (1st consultation time limit) for patients of priority i.

τ_i ≡ waiting time for the 1st consultation for a patient of priority i.

α_i ≡ probability that a priority i patient is discharged after the first consultation.

ν_i ≡ waiting time for the 2nd consultation for a patient of priority i.

$$X_i \equiv \begin{cases} 0 \text{ } if \ \ \tau_i \le T_i \\ 1 \text{ } if \ \ \tau_i > T_i \end{cases}$$

$E(X_i)$ ≡ ratio of patients of priority i exceeding the APT limit, T_i.

P_i ≡ target for the ratio of patients of priority i exceeding their APT limit.

$E(\tau_i)$ ≡ expected TWT for priority i patients who only need one consultation.

$E(\tau_i + \nu_i)$ ≡ expected TWT for priority i patients needing two consultations with a physician.

$E(TWT_i) = \alpha_i \mathrm{E}(\tau_i) + (1 - \alpha_i)\mathrm{E}(\tau_i + \nu_i)$ ≡ expected TWT for a patient of class i.

β_{1i} ≡ slope of the linear accumulating priority function for priority i waiting for the 1st consultation.

β_{2i} ≡ slope of the linear accumulating priority function for priority i waiting for the 2nd consultation.

The decision variables of the optimization problem are the slopes β_{1i}, β_{2i}. The time limits T_i and the ratios P_i are the parameters of the problem reflecting the service quality goals, and the expectations $(E(X_i) - P_i)^+$, $E(\tau_i)$, and $E(\tau_i + \nu_i)$ are the functions to be minimized.

Then, the problem of finding the optimal APQ-h (or APQ) management policy, particularized to the case with three types of patients and two consultations, can be formulated as follows in (1):

$$\min_{\boldsymbol{\beta}} \left\{ \begin{matrix} (E(X_1) - P_1)^+, (E(X_2) - P_2)^+, (E(X_3) - P_3)^+, \ \mathrm{E}(\tau_1), \ \mathrm{E}(\tau_2), \ \mathrm{E}(\tau_3), \\ \mathrm{E}(\tau_1 + \nu_1), \ \mathrm{E}(\tau_2 + \nu_2), \ \mathrm{E}(\tau_3 + \nu_3) \end{matrix} \right\} \tag{1}$$

We address this multi-objective problem by the weighted sum method, as follows:
Problem [P1]

$$\min_{\boldsymbol{\beta}} W \left(u_1\bar{\lambda}_1\Delta_1 + u_2\bar{\lambda}_2\Delta_2 + u_3\bar{\lambda}_3\Delta_3\right) + \left(v_1\bar{\lambda}_1 E(TWT_1)\right.$$
$$\left. + v_2\bar{\lambda}_2 E(TWT_2) + v_3\bar{\lambda}_3 E(TWT_3)\right) \tag{2}$$

where $\Delta_i = (E(X_i) - P_i)^+ = \max\{(E(X_i) - P_i), 0\}$.

The weight W expresses the importance of exceeding the goals P_i compared to reducing a time unit of the total waiting time. The sets of weights $\{u_1, u_2, u_3\}$ and $\{v_1, v_2, v_3\}$ indicate the relative importance of achieving each objective in each type of patient. We will consider that the importance of the types of patients is objective-independent and then $u_i = v_i$. Moreover, each patient category is weighted according to their average arrival rate $\bar{\lambda}_i$.

The objective function has no explicit expression in terms of the decision variables. It is a stochastic function that needs to be evaluated by simulation. Therefore, a simulation based optimization (SBO) methodology is used to solve the optimization problem [P1]. SBO is a tool typically used for analysis in the manufacturing context but has not been used often in healthcare system analysis, although it has already been used to find the optimal assignment of resources in EDs (e.g., [44]) and to find optimal management policies for hospital departments, (e.g., [45–47], in the case of intensive care units).

The rationale of the SBO methodology is as follows: the optimization procedure proposes values for the slopes β_{1i}, β_{2i}, that define an APQ-h (or APQ) policy, which is the input for the simulation model. The ED is simulated under this APQ-h (or APQ) policy and the outputs – KPIs – are recorded and used to evaluate the random objective. Then, the optimization procedure uses this information and the history of the solutions already evaluated

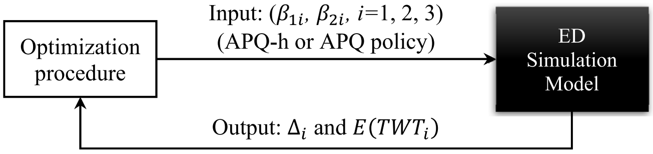
Optimization procedure
Input: (β_{1i}, β_{2i}, i=1, 2, 3)
(APQ-h or APQ policy)
ED Simulation Model
Output: Δ_i and $E(TWT_i)$

Fig. 4. Simulation based optimization approach.

to decide the next solution – APQ-h (or APQ) policy – to be assessed by the simulation model. This process continues until the stopping conditions of the optimization method are met (see Fig. 4).

4. Case study

4.1. Description of the emergency department

The efficacy of the APQ-h and APQ management policies, as well as their comparison with the pure priority rules, are tested by using a simulation model that represents the ED of the Hospital Compound of Navarre (HCN). Moreover, the management policy currently followed by the majority of physicians in the HCN's ED, especially in days of severe overcrowding, does not achieve the goals set by the ED managers, and therefore better management policies should be investigated. 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 h per day with board-certified emergency physicians. As in many other EDs, it organizes the patient care into two different care circuits: one for the more critical patients, i.e., circuit B, and another for less critical patients, i.e., circuit A. In this study, we focus on care circuit A, which has its own staff that is not shared with the circuit B and treats patients of priorities 3, 4 and 5.

In the studied care circuit A, there are five exploration rooms and a senior physician in each exploration room. The patient routing within the care circuit is the same as that described in Fig. 1, with patients of priorities 3 (P3), 4 (P4), and 5 (P5) arriving to the ED system, which correspond to the high, medium, and low priorities in Sections 2 and 3. The next subsection shows the data analysis for adequately modelling the emergency department (patients, medical stuff, care paths, etc.).

Data analysis

The hospital administration provided the electronic records of all patients who visited the ED in the period from 2014–2016. The records contain the arrival time, number of physician consultations, medical test requests, description of the illness and acuity, among others. These data facilitate the estimation of the arrival pattern for each type of patient and the patient path through the ED, including the probability of discharge after the first consultation. However, the duration of the physician consultation was not available in the hospital database and had to be recorded in situ and complemented by eliciting the opinions of physicians and other staff.

The patient arrivals are modelled as a nonhomogeneous Poisson process (NHPP) for each type of patient [48], with the intensity of arrivals $\lambda_i(t)$ depending on the patient class i, and the hour of the day t (there is even a different pattern depending on the day of the week). This seasonality, also observed in other studies (e.g., [37]), depends on the acuity level of the patients in such a way that the lower the acuity, the greater the intraday and intraweek seasonal component [7] is. Fig. 5 shows the arrival rates per hour for the three types of patients of circuit A across the three types of days (holidays, day after a holiday and a normal work day). The average arrival rate of patients across the day (8:00–21:00) and week is 12.17 patients per hour.

The average service utilization across the day is 90.8%, but the arrival rate is above the service rate for 3 h (10:00–13:00). The maximum arrival rate peak occurs at the hourly interval from 11:00–12:00, with a value of 129.87% of the service rate (see Fig. 5). Table 4 contains the quantitative description of the patient flow through the ED, including the probability distributions for first and second consultations service time and the discharge probabilities (α_i) after the first consultation of each priority i patient. Both consultations' service times follow a lognormal distribution with different location parameter value (μ) and the same scale value (σ), which leads to a different expected duration.

The service rate of each type of day is calculated from the estimated service time for each patient type and the mix of patients of each type of day, which is slightly different from one to another (the second column of Table 4 shows the percentages of patients of each priority on days after a holiday). In this study, we will focus on the most adverse day, the days after a holiday (generally Mondays), in which a service rate of 2.66 patients per hour and physician (13.30 in total, since there are five physicians scheduled all day) is obtained.

4.2. Simulation model

A discrete event simulation (DES) model is built to assess the performance of the ED under different queue disciplines and under different working and demand pressure conditions. The events simulated are as follows:

- *Arrival of a new patient* to the ED with properties such as the priority level and the number of consultations needed. The registration and triage process times are very small, and patients never queue; then, in the simulation model, these times are neglected. Therefore, if any of the physicians are idle at the patient arrival time, the patient enters the first consultation; however, if all physicians are occupied, then the new patient joins the queue in the waiting room.
- *End of a physician consultation* The patient is then discharged or exits the ED to begin the complementary diagnostic tests. The physician begins a new consultation if there are any patients waiting.
- *Re-entry of a patient to the physicians' waiting room* after medical test are carried out, and the results are ready. At this moment, the patient joins the queue in the waiting room, or the second consultation begins, if there is an available physician.

The arrival of patients is simulated by sampling from the NHPP and the duration of the physician consultation from the lognormal distribution. The selection of the next patient to be seen by a physician is simulated by following the rules of the queue discipline that is implemented in the simulation model. The diagnostic tests, specialist consultation, etc. are performed outside the limits of the ED and are not the responsibility of the ED physicians. Therefore, we keep this part of the hospital and care process out of the limits of our model. The stochastic delay that this additional tests suppose is randomly simulated following a triangular distribution (30, 60, 90) minutes. The simulated priority queue system model is represented in Fig. 6.

The design of the simulation model is flexible enough to modify the mix of patients, the seasonality of the arrivals, the level of congestion and the variability of the service times to create many different representative scenarios of other hospitals' EDs. Thus, the robustness of a queue discipline can be investigated by assessing its performance in a wide range of ED scenarios (see Section 5).

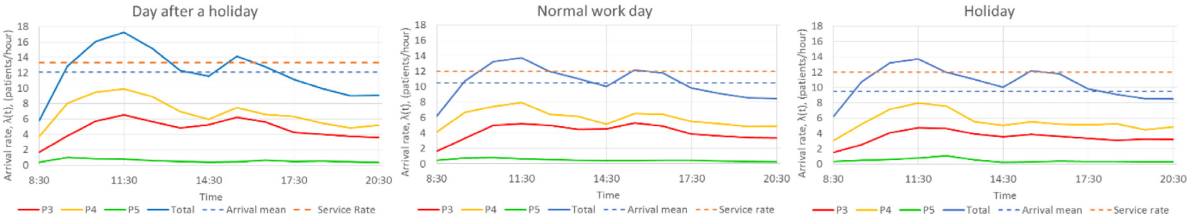

Fig. 5. Arrival rates of patients, total and according to priority, and service rates for each type of day.

Table 4
Percentage type of patient (day after a holiday), parameters of the lognormal distribution for the consultation duration and discharge probability after C1.

Priority i	$\%_i$	Service time (min) for the first consultation (S_{1i}): lognormal		Service time (min) for the second consultation (S_{2i}): lognormal		Discharge probability after the 1st consultation (α_i)
		μ_{1i}	σ_{1i}	μ_{2i}	σ_{2i}	
3	38.76	2.89	0.45	2.29	0.45	0.361
4	56.56	2.71	0.45	2.12	0.45	0.513
5	4.67	2.49	0.45	1.89	0.45	0.177

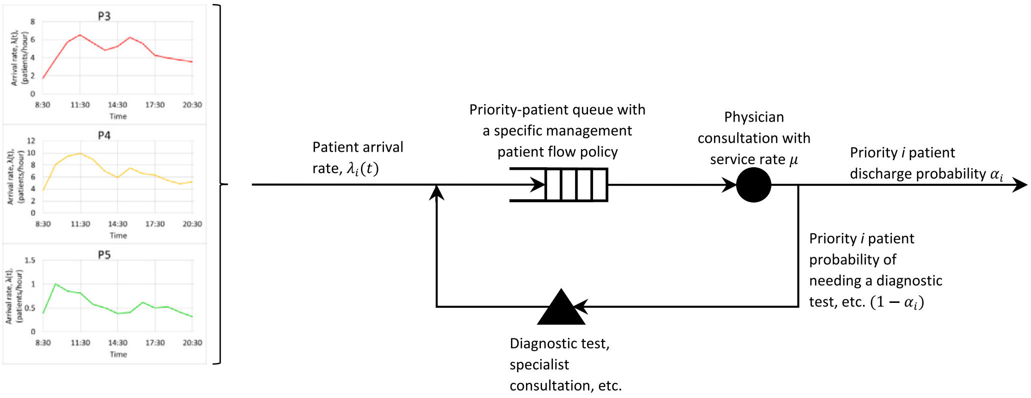
P3
P4
P5
Arrival rate, λ(t), (patients/hour)
8:30
11:30
14:30
17:30
20:30
Time
Patient arrival rate, $\lambda_i(t)$
Priority-patient queue with a specific management patient flow policy
Physician consultation with service rate μ
Priority i patient discharge probability α_i
Priority i patient probability of needing a diagnostic test, etc. $(1 - \alpha_i)$
Diagnostic test, specialist consultation, etc.

Fig. 6. ED priority queue model.

The simulation model was developed using the Arena Simulation software from Rockwell Automation [49], and 3D visualization was developed to help with the validation by ED professionals. The ED professionals were shown the KPIs in real time, such as the queue length and level of occupation and judged them by their experience before accepting the model as valid. A video of the simulation model can be seen at http://www.unavarra.es/quphs/proyectos, or downloaded specifically at [50].

Moreover, several equivalence tests have been performed to determine whether the means of simulated patients' LoS are close enough to be considered equivalent to the means of historical patients' LoS in the ED. The range of acceptable values for the difference was fixed in ten minutes (less than 5% of average length of stay). This equivalence test was carried out for the six groups of patients defined by priority and number of consultations. At a significance level of 0.05 we could claim the equivalence of both LoS means, from the historical data and from the simulation data, in all six groups.

To determine the simulation run length necessary to accurately estimate the KPI, a preliminary analysis was carried out by running the simulation model for 15,000 days. The KPI estimations were collected and graphically represented as a function of the number of simulated days to identify the stabilization point. As result of this analysis, it was determined that 2000 simulation days are enough to obtain good KPI estimations (see Fig. 7).

4.3. Optimal prioritization policies

The problem [P1] is to find the optimal APQ-h policy to manage the patient flow of an ED such as the HCN's described in 4.1. In particular, we consider the management of the healthcare circuit of lower priority patients (levels 3, 4, and 5). The values of the parameters included in the objective function of the instance of [P1] that is solved are in Table 5.

In addition, the values to weigh the importance between both terms in the objective function and the relative importance of achieving each objective in each type of patient were determined by the ED physician co-author of this article, following a discussion with her colleagues. Specifically, the objective independence of the weights for each patient priority was set, that is, $u_i = v_i$. Additionally, the objectives for priority 3 patients were set to be twice as important as for priority 5 patients, and priority 4 patients were 50% more important than priority 5 patients. Therefore, the priority weights were adjusted as follows: $u_3 = 2u_5$, $u_4 = 1.5u_5$, and $u_5 = 1$. Finally, a weight of W = 5 was assigned when the patients exceeded the time limit, which is expressed in percentage ($ratio \times 100$). The time unit is expressed in half-hours. Therefore, the increment of 1% in the patients that exceed the first consultation time limit is equivalent in the objective function to an increment of 2.5 h in the total waiting time.

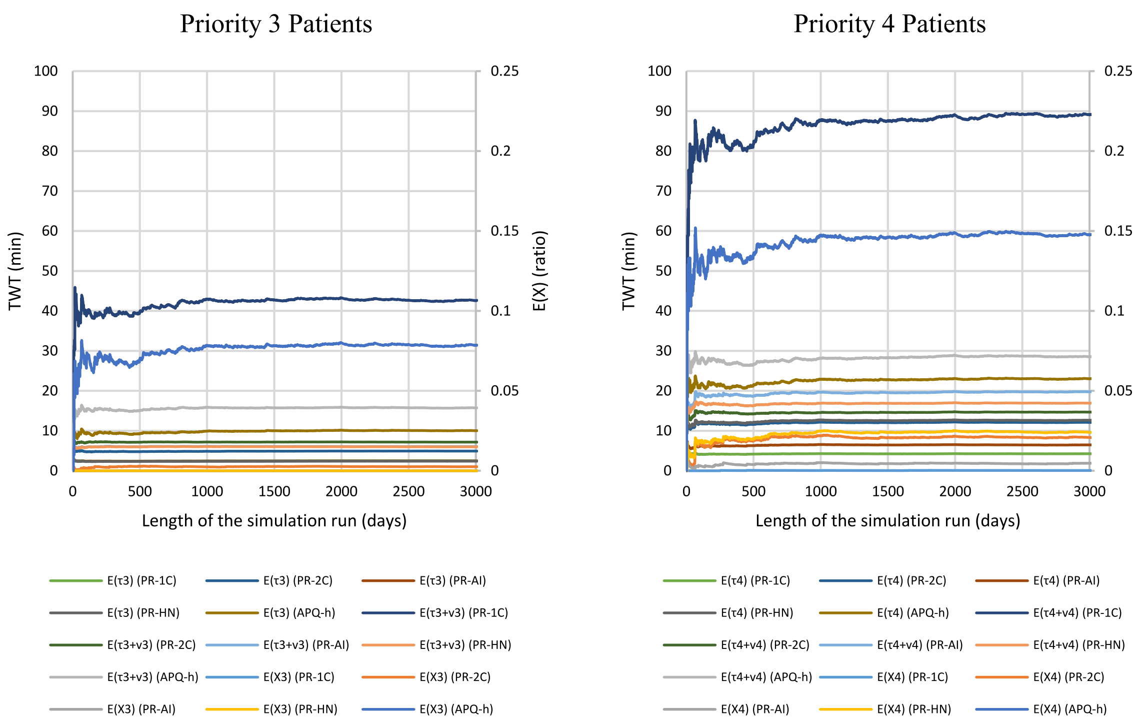

Fig. 7. Estimation of the KPI values as a function of the number of simulated days.

Table 5
Parameters of the instance solved.

Parameter	Value
Maximum APT limit (time unit: half-hours)	
T_3	1 (30 min)
T_4	2 (60 min)
T_5	4 (120 min)
Maximum ratio of patients exceeding the APT limit	
P_3	0.10
P_4	0.15
P_5	0.20
Average hourly arrival rate (patients/hour)	
$\overline{\lambda_3}$	4.68
$\overline{\lambda_4}$	6.83
$\overline{\lambda_5}$	0.56
Probability of being discharged after the first consultation	
α_3	0.36
α_4	0.51
α_5	0.18

The parameters β are required to sum to 10 to facilitate the comparison of the results among the different scenarios and avoid multiple optimal solutions. The SBO technique described in 3.4 was implemented in the ARENA simulation software (Rockwell Automotion [49], Version 15), which is a suitable software to implement discrete event simulation models and OptQuest optimization software, which is based on the scatter search metaheuristic, as proposed by Laguna and Martí [51].

The optimal values for the maximum priority accumulated by patients waiting for the first consultation are 51.801 ($\beta_{13} = 1.7$), 33.288 ($\beta_{14} = 0.5548$) and 16.86 ($\beta_{15} = 0.1405$), for patients of priority 3, 4 and 5, respectively. Then, the priority accumulated by a priority 3 patient in almost 20 min equals the priority accumulated by a priority 4 patient in 60 min and by a priority 5 patient in almost three hours. The slopes for the second consultation are 7.5775, 0.0005 and 0, which means that priority 5 patients are only seen by the physician for a re-evaluation when the ED no longer has higher priority patients. Therefore, the relative importance that should be given to the different stages of the care process is not the same for all priorities, that is, the dominance relation between β_{1i} and β_{2i} is not always the same ($\beta_{13} < \beta_{23}$ while $\beta_{14} > \beta_{24}$).

The optimization process was also applied to determine the optimal value of the APQ parameters, but no significant difference was found in the ED performance (KPI values) when using the optimal APQ-h. Thus, there is no practical difference in applying any of both queue disciplines.

The simulation results for the KPI for pure priority disciplines and the optimal APQ and APQ-h disciplines are shown in Table 6. Disciplines APQ-h and APQ, and PR-1C are able to achieve the goals for the probability of patients exceeding the time limit but they are not achieved by the other disciplines, including the currently used one. It should be noted that particularly, the currently used pure priority rule in the HCN's ED, PR-HN, has their KPIs out of control and are considerably improved with the optimized new policy APQ-h (and also by the APQ).

The APQ and APQ-h policies' KPIs $E(X_i), i \in \{3, 4, 5\}$ fall within the limits; however, as is expected, the PR-1C is the only pure priority policy whose KPIs $E(X_i)$ are also within the boundaries. This policy focuses on assisting patients waiting for the first consultation, which leads to better values for $E(X_i)$ – further from the limits – while the rest of the performance KPIs and, consequently, the objective function value are significantly worse.

Display and interpretation of the simulation results using star graphs

In this subsection, the ED performance when using pure disciplines to manage the patient flow of an ED such as the HCN is compared with the ED performance when using the optimal APQ-h and APQ, combining the two factors taken into account. The simulation results for the KPIs and the value of the objective function are displayed by using a star graphs (see Fig. 8). The upper vertical axis (OFV) represents the objective function value

Table 6
KPI for pure priority disciplines and APQ and APQ-h.

Discipline			PR-AI	PR-1C	PR-2C	PR-HN	APQ & APQ-h
Ratio of patients exceeding the time limit P_i. ($P_3 = 0.1, P_4 = 0.15, P_5 = 0.2$)		$E\,(X_3)$	<0.001	<0.001	0.004	<0.001	0.061
		$E\,(X_4)$	0.111	0.016	**0.300**	**0.306**	0.148
		$E\,(X_5)$	**0.457**	0.066	**0.452**	**0.453**	0.199
Total waiting time	Patients who need a single consultation	E (τ_3)	2.838	2.885	5.458	2.747	9.552
		E (τ_4)	21.474	9.683	44.564	45.420	26.494
		E (τ_5)	171.137	28.978	166.449	168.034	54.699
	Patients who need medical tests (2 consultations)	E $(\tau_3 + \nu_3)$	7.596	65.038	8.021	7.394	18.816
		E $(\tau_4 + \nu_4)$	129.149	155.640	47.341	51.224	113.371
		E $(\tau_5 + \nu_5)$	235.253	226.350	169.732	172.291	223.557
Objective function value			103.723	44.511	858.719	879.589	31.990

while the first three axes clockwise show the ratios of patients exceeding the APT limit for each priority ($E\,(X_3)$, $E\,(X_4)$, and $E\,(X_5)$). The values below the upper limits for each priority objective are in black and those above them are in red.

The next three axes clockwise ($E\,(\tau_3)\,, E\,(\tau_4)$, and $E\,(\tau_5)$) represent the expected TWT in the ED system of patients who only need a single consultation with the physicians, while the last three axes ($E\,(\tau_3 + \nu_3)\,, E\,(\tau_4 + \nu_4)$, and $E\,(\tau_5 + \nu_5)$) display the expected TWT in the ED system of patients who are not discharged after the first consultation. Because the goal is to minimize all KPIs and the objective function, the nearer each value is to the centre of the chart, the better the performance is.

Fig. 8 displays and compares the KPIs obtained by the analysed queue disciplines in the real HCN scenario. The star plot on the right displays the KPIs for the PR-HN (the rule which is used by the majority of the medical stuff in the HCN) and the optimal APQ-h and APQ, while the star plot on the left displays those for the pure priority rules PR-1C, PR-2C, and PR-AI. The APQ-h and APQ policies provides results for $E[X_i]$ that are lower than but close to the P_i boundaries, which produces both no penalties and room to improve the results in the other KPIs. The discipline PR-1C respects the P_i limits, while the PR-AI, PR-2C, and PR-HN policies do not (PR-AI: $E\,(X_5) = 0.457 > 0.2$; PR-2C: $E\,(X_4) = 0.300 > 0.15, E\,(X_5) = 0.452 > 0.2$; PR-HN: $E\,(X_4) = 0.306 > 0.15, E\,(X_5) = 0.453 > 0.2$).

However, because PR-1C prioritizes the first consultation, the $E[\tau_i + \nu_i]$ values are worse than those obtained by APQ-h and APQ (65.04 vs. 18.82, 155.64 vs. 113.37, and 226.35 vs. 223.557 for $E[\tau_3 + \nu_3], E\,[\tau_4 + \nu_4]$, and $E\,[\tau_5 + \nu_5]$), respectively. As a consequence, the APQ type policies obtains a better global performance, as measured by the value of the objective function.

4.4. Sensitivity analysis for the criteria's importance

In this section, we analyse the robustness of the optimal solutions to the APQ type policies' parameters when the weight W in (2) is varied. The weight fixed by the physicians (W = 5) – whose main objective is to achieve the performance level for the APT limits – is considered to be one of the extreme values for the range of studied values. From that point, the weight is being reduced until a minimum of 0.2 is reached (at this point the worsening of 1% in the number of patients that exceed the time limit for first consultation is equivalent to increase the total waiting time in 6 min).

W is varied in the range [0.2, 5] and the following two different optimal solutions are found: one is optimal for W varying from 0.2 to 0.3 and the other is optimal for W varying from 0.4 to 5. The solution for W $\geq$ 0.4, provides $\Delta_i = 0, i = 3, 4, 5$, and then the first part of the objective is fully minimized with a value of zero. The solution is that mentioned in the previous section (the optimal solution for the case study). Fig. 9 shows the KPI values for each solution through the interval of values [0.2, 0.7] with steps of 0.1 (note that there is no change from 0.4 onwards).

However, from W $=$ 0.2 to W $=$ 0.3, $\Delta_5 > 0$ due to the domination of the objective function by the minimization of the TWT (objective 2). The TWT of priority 4 patients is significantly reduced ($E\,(\tau_4)$ is reduced from 26.494 to 23.776 and $E\,(\tau_4 + \nu_4)$ from 113.371 to 86.179), which represents 57% of the patients. The priority 3 patients' waiting time is almost the same, while the waiting time of priority 5 patients worsens. The optimal solution in this case is $\beta_{13} = 1.359$, $\beta_{14} = 0.6580$, $\beta_{15} = 0$, $\beta_{23} = 8.0051$, $\beta_{24} = 0.0010$, and $\beta_{15} = 0$. Contrary to the previous solution, in this case, priority 4 patients who are waiting for their second consultation have a greater accumulating priority rate than all priority 5 patients, who are only assisted if there are no other patients in the ED.

5. Extended simulation study to a general set of ED scenarios

5.1. Selection of scenarios

In this section, an extended analysis of the performance and a comparison between pure priority disciplines and optimal APQ-h and APQ disciplines are carried out in different ED scenarios. The set of scenarios is designed from the HCN ED, which is described in Section 4, by varying the average occupancy rate, the pattern of the intraday seasonality, and the composition of the mix of patients. Specifically, we consider the following values for the abovementioned factors:

- **ED congestion level** (named as factor 1 and denoted by **F1**): The average occupation rates ρ of 90% and 95% are considered. The number of physicians is maintained while the patient arrival rate is modified accordingly.
- **Arrival seasonality (F2):** Three arrival seasonality patterns are considered, ranging from no seasonality (constant arrival rate of patients, denoted as *T0*) to a maximum hourly seasonality, which is described by two different triangular patterns for the arrival rate $\lambda(t)$, both with a peak at 11:30 a.m. and a ratio of $(\lambda_{max} - \lambda_{min})/\lambda_{min} = 0.5$. The first triangular pattern, denoted as Tu, extends the triangular shape across the entire time range, while the second triangular patterns, denoted as Tp, only applies the triangular shape in the time range [10:00, 13:00], with the arrival rate out of this range being constant (see Fig. 10). As a consequence, each one of the three seasonality patterns have different values for λ_{max}.
- **Mix of patients (F3):** Four different mixes of patients are considered as follows: balanced distribution among all types of patients (1/3 of P3, 1/3 of P4 and 1/3 of P5) and a biased mix towards each priority (50% of P3, 25% of P4 and 25% of P5; 25% of P3, 50% of P4 and 25% of P5; and 25% of P3, 25% of P4 and 50% of P5). These scenarios are denoted by B0, B3, B4 and B5, respectively.

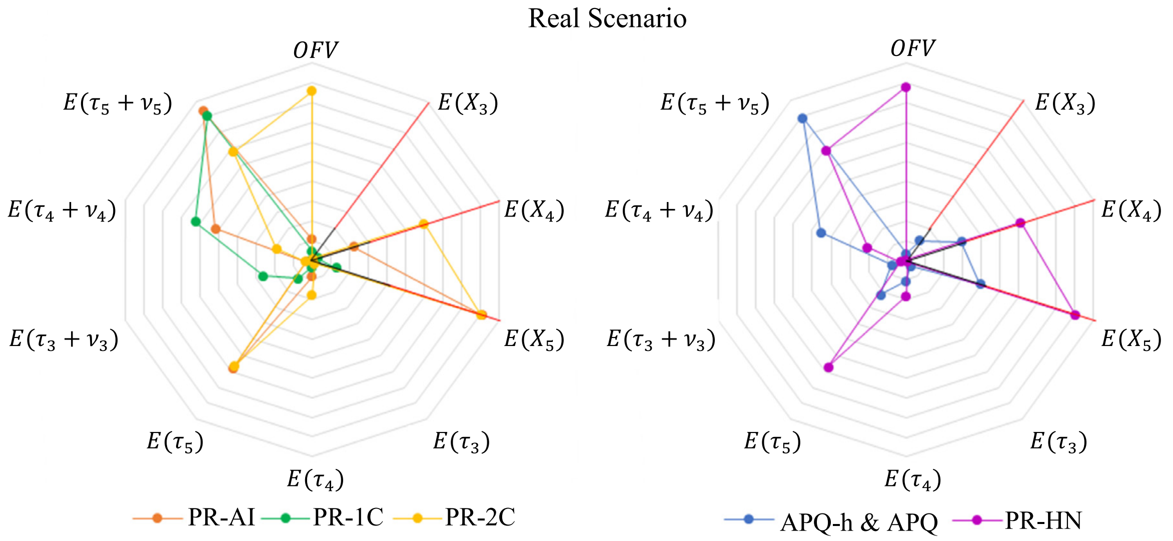

Fig. 8. Star plot of the simulated results of the real ED scenario of the HCN.

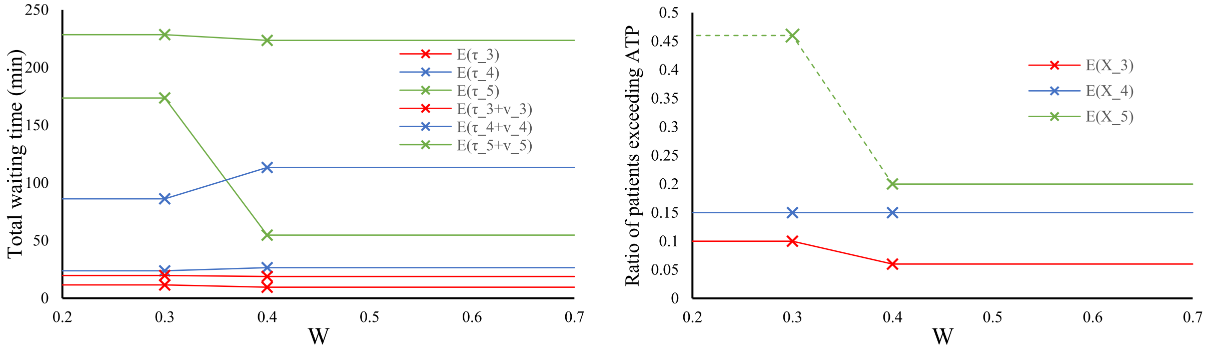

Fig. 9. Outcomes of the optimal solution for different objective functions (W ranging from 0.2 to 0.7 as there is no change from 0.4 onwards): total waiting time in the system (left graph) and ratio of patients exceeding the time limit for the first consultation (right graph). The latter represents the values that does not achieve the target for the ratio (above the limit) in dashed line and those that does (equal or below the limit) in solid line. The crosses indicates the change-points.

Each scenario is denoted by a vector (f_1, f_2, f_3), where f_i is the level of factor F_i, $i = 1, 2, 3$, and $f_1 \in 90\%, 95\%$, $f_2 \in T0, Tu, Tp$, and $f_3 \in B0, B3, B4, B5$. Then, a total of 24 scenarios will be analysed with the simulation model.

5.2. Analysis of the results

Influence of the demand factors on the ED performance. The results from the simulation of the different ED scenarios show the influence of the demand factors (quantity, seasonality and typology) on the ED performance. This influence is visualized in Table 7, which displays three selected sets of ED scenarios results by disclosing the value for every KPI considered: the policy applied is in first column, the objective function value is in second column, the time target objective values for P3, P4, and P5 priority patients are in columns third ($E(X_3)$), fourth ($E(X_4)$) and fifth ($E(X_5)$) respectively, the APT for P3, P4, and P5 are in sixth ($E(\tau_5)$), seventh ($E(\tau_4)$), and eighth ($E(\tau_5)$) columns and the TWT for P3, P4, and P5 patients who need two consultations are in ninth $E(\tau_3 + \nu_3)$), tenth ($E(\tau_4 + \nu_4)$), and eleventh ($E(\tau_5 + \nu_5)$) respectively. The last two columns are the description of each scenario and the improvement of the objective function value with respect to the best Pure Priority Rule. The first set of scenarios are of type (90%, Tu, -); that is, they differ in the mix of patients. The scenarios in the second set are of type (95%, Tu, -); that is, they only differ from the scenarios in the first set in the congestion level, that is, of 95%. Finally, the scenarios in third set are of type (95%, -, B4); that is, they differ in the seasonality pattern for the arrivals. The following observations can be extracted from this table and, in general, from all scenarios results:

- The results are very sensitive to the increase in the occupancy ratio from 90% to 95% (the KPIs worsen from the first row, 90%, to the second row, 95%).
- The mix of patients also has a large influence; the higher the severity of patients who represent the maximum percentage in the mix of patients, the worse the performance is (as observed in the first and second rows).
- The seasonality also influences the performance. The best results are observed in the case of homogeneous arrivals, and the worst results are observed in the case of a triangular pattern extended throughout the day.

APQ-h versus the pure priority disciplines: performance comparison

The simulations of the different ED scenarios ruled with pure priorities and APQ-h and APQ disciplines produced results that highlight the very different behaviours of all of them; while each pure priority focused on the achievement of a specific subset of KPIs, disregarding the others, the APQ-h and APQ policies are able to balance all the KPIs according to their relative importance

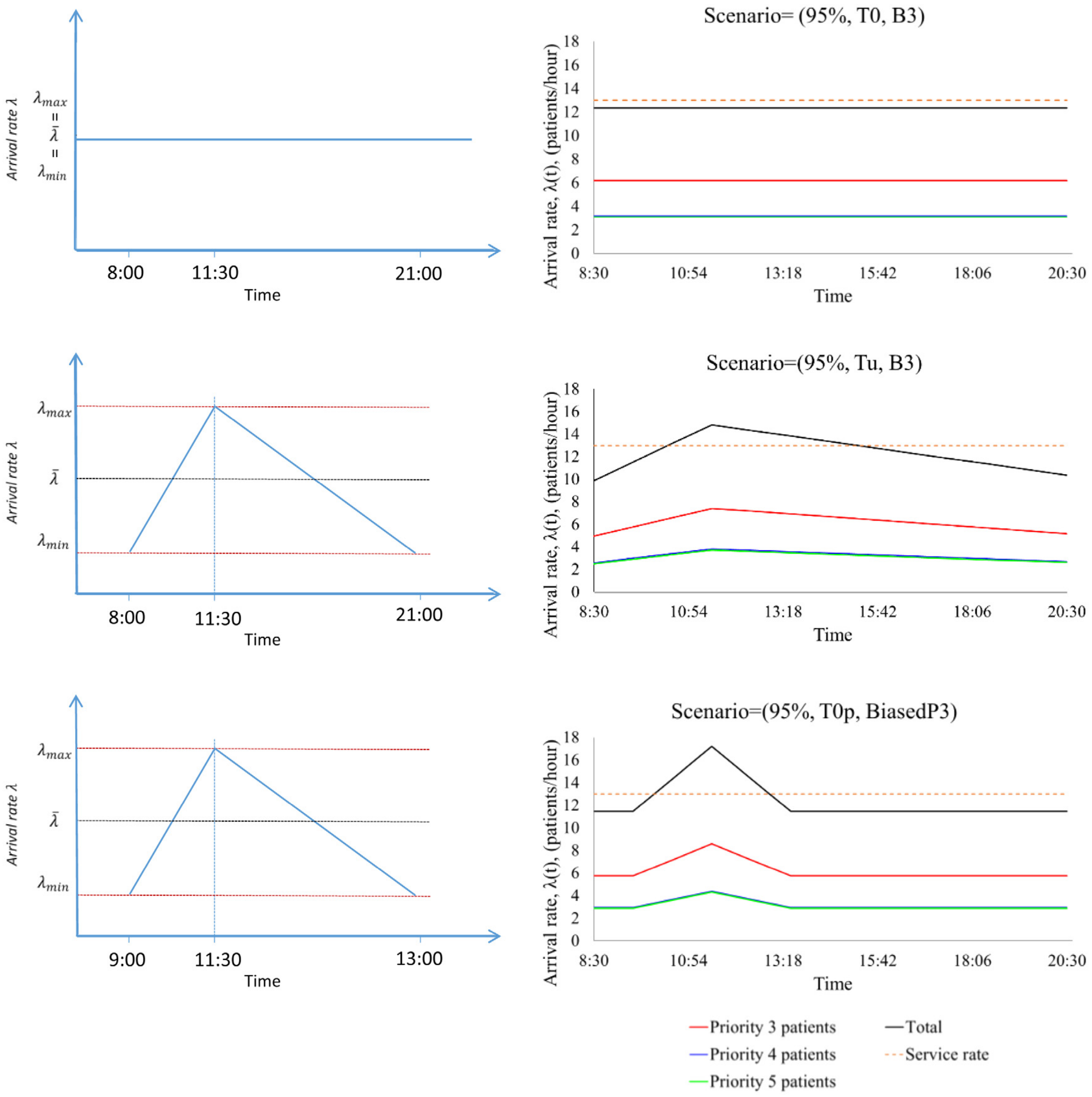

Fig. 10. The three patterns for the seasonality of the arrivals: from top to down scenarios T0, Tu and Tp, respectively.

expressed through the weights in the objective function. This general statement is graphically visualized in Fig. 11, in which the simulated results of all the ED scenarios are represented in the same star plot for each queue discipline.

The PR-2C policy prioritizes the minimization of the TWT for patients who need medical diagnostic tests. The shape created in Fig. 11 by the KPIs associated with the PR-2C policies is graphically shifted to the right and down, as this pure discipline focuses on discharging patients waiting for the second consultation, that is, on $E(\tau_i + \nu_i)$. In cases of a high congestion and a high percentage of high priority patients, disregarding the first consultation produces the non-fulfilment of the time limits for the APT and, therefore, the results in positive values for the ratio Δ_i.

The PR-1C policy prioritizes the minimization of the TWT for patients who need a single consultation. Opposite to PR-2C, the shape created in Fig. 11 by the KPIs associated with the PR-2C policies is graphically shifted to the top left, as this pure discipline only pays attention to the APT limit target, ignoring the waiting time for the second consultation, ν_i. Therefore, $\Delta_i = 0$ and the $E(\tau_i)$ values are small but the $E(\tau_i + \nu_i)$ values are large.

The PR-AI policy prioritizes the minimization of the APT and TWT for the highest priority patients. The shape created in Fig. 11 by the KPIs associated with the PR-AI policies is a mixture of the previous ones, i.e., the best results in all KPIs for the highest priority patients and the worst for the lowest priority patients.

The PR-HN policy prioritizes the minimization of the APT and TWT for the highest priority patients and the minimization of the TWT for patients of other priorities who need medical diagnostic tests. The shape created in Fig. 11 by the KPIs associated with the PR-HN policies is similar to the shape created by the PR-2C policies. The difference is that the values for the highest priority KPIs, $E(\tau_3 + \nu_3)$ and $E(\tau_3)$, are nearer to the centre of the chart.

The optimal APQ-h and APQ policies produce balanced results. The optimal APQ-h and APQ policies obtain worse results than PR-1C for $E(\tau_i)$, although they achieve $\Delta_i = 0$ and better results for $E(\tau_i + \nu_i)$. The opposite is concluded when compared with PR-2C, i.e., the results are better for $E(\tau_i)$ and Δ_i but worse for $E(\tau_i + \nu_i)$. When compared with PR-AI, the results are worse for the highest priority patients but better for the lower priority patients. Therefore, the shapes of the star plots associated with the APQ-h and APQ policies are more centred and close to the central point than the shapes of the other policies, meaning more balanced results are obtained.

Table 7
KPIs for three selected sets of ED scenarios.

Queue discipline	Obj	$E(X_3)$	$E(X_4)$	$E(X_5)$	$E(\tau_3)$	$E(\tau_4)$	$E(\tau_5)$	$E(\tau_3+\upsilon_3)$	$E(\tau_4+\upsilon_4)$	$E(\tau_5+\upsilon_5)$	Scenario	APQ-h & APQ improvement
(90%, Tu, -)												
PR-AI	160.97	0.00	<0.01	0.06	2.21	5.25	28.90	5.40	15.26	112.07		
PR-1C	230.70	0.00	<0.01	<0.01	2.24	3.63	8.47	27.51	54.68	112.30	F1: 90%;	
PR-2C	107.42	<0.01	0.01	0.15	4.29	9.58	51.13	6.43	11.92	53.84	F2:Tu;	0%
PR-HN	115.53	0.00	0.01	0.16	2.17	10.05	51.71	5.38	13.87	56.25	**F3: B0**	
APQ-h & APQ	107.59	<0.01	0.01	0.15	3.75	9.72	51.30	6.09	12.06	54.69		
PR-AI	179.01	0.00	<0.01	<0.01	2.04	3.53	14.73	4.66	8.36	95.07		
PR-1C	219.59	0.00	0.00	<0.01	2.07	2.82	6.63	22.23	36.82	95.35	F1: 90%;	
PR-2C	103.02	<0.01	<0.01	0.08	4.00	6.50	37.97	5.99	8.68	40.73	F2:Tu;	<1%
PR-HN	108.58	0.00	<0.01	0.08	1.98	6.85	38.26	4.61	9.76	42.19	**F3: B5**	
APQ-h & APQ	102.96	<0.01	<0.01	0.08	3.31	6.65	38.05	5.44	8.68	41.34		
PR-AI	155.07	0.00	<0.01	0.12	2.12	5.12	41.71	4.84	19.70	122.97		
PR-1C	235.59	0.00	<0.01	<0.01	2.15	3.89	10.02	27.59	59.45	122.63	F1: 90%;	
PR-2C	110.37	<0.01	0.01	0.20	3.90	10.40	61.22	5.97	12.79	64.26	F2:Tu;	0%
PR-HN	124.77	0.00	0.01	0.20	2.09	10.64	61.57	4.87	13.98	65.85	**F3: B4**	
APQ-h & APQ	110.94	0.00	0.01	0.20	2.93	10.56	61.15	5.76	13.39	64.77		
PR-AI	158.37	<0.01	0.03	0.13	2.53	9.97	45.46	7.41	27.57	126.60		
PR-1C	256.74	<0.01	<0.01	<0.01	2.59	4.79	9.89	36.30	78.57	126.67	F1: 90%;	
PR-2C	244.57	<0.01	0.06	0.21	5.36	15.74	64.96	7.59	18.37	67.47	F2:Tu;	20%
PR-HN	251.56	<0.01	0.07	0.21	2.49	16.73	66.37	7.32	23.52	72.77	**F3: B3**	
APQ-h & APQ	126.33	0.02	0.09	0.20	6.77	17.49	57.00	10.71	24.16	62.41		
(95%, Tu, -)												
PR-AI	247.60	0.00	<0.01	0.13	2.46	6.44	45.17	6.08	19.69	169.50		
PR-1C	370.02	0.00	<0.01	<0.01	2.51	4.28	11.19	41.67	86.75	169.89	F1: 95%;	
PR-2C	1693.08	<0.01	0.02	0.27	4.94	12.10	78.88	7.24	14.77	82.25	F2:Tu;	27%
PR-HN	1668.13	0.00	0.02	0.28	2.40	12.59	79.66	6.04	17.01	85.12	**F3: B0**	
APQ-h & APQ	180.28	0.07	0.14	0.20	9.96	22.49	60.53	14.64	28.44	66.53		
PR-AI	282.31	0.00	<0.01	0.02	2.26	4.02	21.80	5.23	9.98	145.47		
PR-1C	352.40	0.00	0.00	<0.01	2.30	3.21	8.47	33.03	57.85	145.94	F1: 95%;	
PR-2C	155.47	<0.01	<0.01	0.17	4.58	7.75	58.18	6.76	10.17	61.02	F2:Tu;	0%
PR-HN	153.97	0.00	<0.01	0.17	2.21	8.13	58.57	5.17	11.51	62.75	**F3: B5**	
APQ-h & APQ	155.46	<0.01	<0.01	0.17	3.81	7.81	58.28	6.13	10.13	61.81		
PR-AI	435.44	<0.01	<0.01	0.21	2.37	6.36	66.02	5.47	27.30	184.93		
PR-1C	378.37	<0.01	<0.01	0.01	2.41	4.66	13.97	41.53	93.64	185.01	F1: 95%;	
PR-2C	2220.07	<0.01	0.03	0.33	4.48	13.52	95.78	6.76	16.22	99.20	F2:Tu;	48%
PR-HN	2196.88	0.00	0.03	0.33	2.32	13.87	96.25	5.45	17.70	100.96	**F3: B4**	
APQ-h & APQ	195.82	0.02	0.14	0.20	7.51	24.39	63.95	13.59	31.18	79.45		
PR-AI	522.48	<0.01	0.05	0.22	2.85	13.04	69.98	8.66	37.68	183.51		
PR-1C	395.08	<0.01	<0.01	<0.01	2.89	5.74	13.51	53.15	118.20	183.64	F1: 95%;	
PR-2C	2144.24	<0.01	0.10	0.33	6.22	20.84	97.85	8.65	23.82	99.89	F2:Tu;	46%
PR-HN	2152.75	<0.01	0.11	0.33	2.77	22.18	99.53	8.49	30.11	106.13	**F3: B3**	
APQ-h & APQ	213.26	0.05	0.15	0.20	8.40	26.42	63.46	14.93	33.36	126.48		
(95%, -, B4)												
PR-AI	187.94	<0.01	<0.01	<0.01	2.15	3.63	14.07	4.88	8.68	94.22		
PR-1C	226.95	<0.01	0	<0.01	2.17	2.95	6.70	20.76	34.71	94.53	F1: 95%;	
PR-2C	108.41	<0.01	<0.01	0.07	4.17	6.81	37.24	6.25	9.08	40.10	**F2:T0;**	0%
PR-HN	107.13	0.00	<0.01	0.07	2.06	7.09	37.62	4.81	10.13	41.64	F3: B5	
APQ-h & APQ	108.30	<0.01	<0.01	0.07	3.46	6.90	37.37	5.64	9.06	40.76		
PR-AI	216.66	0.00	<0.01	0.18	2.35	6.34	58.32	5.36	26.05	164.97		
PR-1C	336.36	0.00	<0.01	0.01	2.40	4.65	13.88	37.17	81.93	164.92	F1: 95%;	
PR-2C	1544.95	<0.01	0.03	0.29	4.41	13.17	84.17	6.61	15.80	87.77	**F2:Tp;**	21%
PR-HN	1552.25	0.00	0.03	0.29	2.29	13.45	84.71	5.33	17.18	89.57	F3: B4	
APQ-h & APQ	171.90	0.02	0.11	0.20	7.69	21.04	61.43	13.00	26.68	68.54		
PR-AI	435.44	<0.01	<0.01	0.21	2.37	6.36	66.02	5.47	27.30	184.93		
PR-1C	378.37	<0.01	<0.01	0.01	2.41	4.66	13.97	41.53	93.64	185.01	F1: 95%;	
PR-2C	2220.07	<0.01	0.03	0.33	4.48	13.52	95.78	6.76	16.22	99.20	**F2:Tu;**	48%
PR-HN	2196.88	0.00	0.03	0.33	2.32	13.87	96.25	5.45	17.70	100.96	F3: B4	
APQ-h & APQ	195.82	0.02	0.14	0.20	7.51	24.39	63.95	13.59	31.18	79.45		

Table 8 compares the values of the objective function obtained for each discipline in each of the 24 ED scenarios analysed and quantifies the improvement obtained by the APQ-h and APQ policies with respect to the best pure priority policy. Table 10 of Appendix displays the detailed results for each scenario by disclosing the value for every KPI considered. The last columns of Table 8 show the optimal APQ-h and APQ policy in terms of the values of β_i. Generally, the highest improvements are obtained in the toughest environments for the ED, that is, the 95% occupation level, seasonality in the patient arrivals and high percentages of high severity patients. For low occupation levels (90%), almost no improvement is achieved, as the optimal APQ-h and APQ policies are very similar to the best pure priority rule solution (PR-2C) – giving more importance to the second consultation than to the first one, as it is easy to obtain the time target limit. For example, in the first row scenario, the slopes are 0.0396, 0.0001, and 0 for first consultations (β_{13}, β_{14}, and β_{15} respectively), and 5.7720, 4.1227, and 0.0656 for second consultations (β_{23}, β_{24}, and

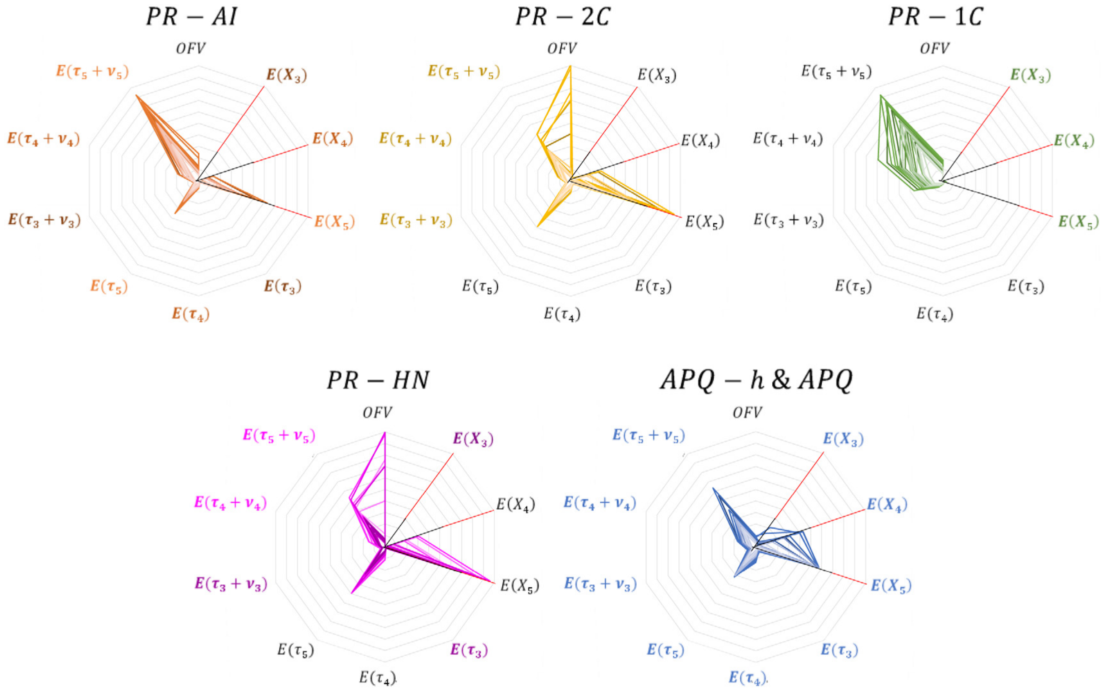

Fig. 11. Representation of the scenarios KPIs ruled by the PR-2C, PR-AI, PR-1C, PR-HN, APQ-h and APQ policies.

β_{25}respectively), and the objective function value obtained for the APQ-h and APQ policy are almost the same as for the PR-2C policy.

The similarity between the optimal APQ-h and APQ policies and one pure priority policy is observed in other ED scenarios. However, having no strict priority imposed by the APQ-h (or APQ) policy can enormously affect the KPI results. This fact is illustrated in Table 9, which contains the KPI results and APQ-h optimal solution for the (95%, Tu, B0) ED scenario. This optimal policy favours the ED discharge of patients by assigning larger slopes to patients waiting for the second consultation, i.e., $((\beta_{23} = 9.1343) > (\beta_{24} = 0.2565) > (\beta_{25} = 0.2414) > (\beta_{13} = 0.2402) > (\beta_{14} = 0.1274) > (\beta_{15} = 0.0002))$. However, even if the order of attending patients seems to be similar to the PR-2C discipline, the flexibility of the APQ type policies allows P3 and P4 patients who have been waiting for their first consultation for a long time to overtake patients who have been waiting for less time for their second consultation. As a result, the PR-2C policy does not fulfil $\Delta_i = 0$, but the APQ-h and APQ policies do, and the objective function value is 1693.08 for PR-2C and 180.28 for the APQ-h and APQ policy.

6. Conclusions

In this paper, the performance of the APQ-h and APQ policies in a real ED setting that considers the stochasticity in the arrivals of patients and the different stages of the health care process has been investigated. Therefore, this study extends the theoretical results obtained in studies [12,52] that assumed a homogeneous Poisson process for the arrivals and only one stage for the patient treatment. The results show that ED performance is better when it is managed with the APQ or APQ-h policy than with other priority policies. This observation supports the use of any of both policies in practice to manage the ED patient flow, in fact the APQ-h is already followed in some EDs, as reported in [11]. We identify that managers of the ED hospitals included in Ding et al.'s study [11] apply a structure for queue discipline that is equal to the APQ-h. This policy that might seem counterintuitive because the patients stop accumulating priority points, in fact, states that from that moment on the patients waiting for first consultation are selected by priority and a FIFO rule within each category, which is a rule widely used to manage EDs as it is mentioned in the introduction (see, for example, Taylor et al. [16], Haussman [17], Siddharthan and Jones [18], Laskowski et al. [19], Mokaddis et al. [20]). Any patient of high priority having reached the time limit for the first consultation will never be overtaken by other patient of lower priority waiting for the first consultation, independently of their respective waiting times. Only patients having waited for a very long time for their second consultation could overtake such high priority patient. We use a simulation-based optimization methodology to obtain the optimal APQ-h and APQ policies that are superior compared to other pure priorities disciplines, especially when high congestion and non-stationary ED environments are considered. Moreover, in the case study of the ED of the HCN, the use of APQ-h and APQ significantly outperforms the current priority rule, PR-HN, whose obtained KPIs were out of control.

However, the analysis also shows that in not very congested ED scenarios, with a time-regular affluence of patients, the application of APQ-h or APQ has no advantage over the best pure priority policy. In these cases, it is recommended to apply the pure priority discipline because it is easier to implement and is very convenient in a dynamic and stressful environment such as the ED, especially when physicians have to apply them. Furthermore, pure priority disciplines require less information, only requiring the type of patient and the stage of healthcare process but not the recording of the waiting time for each patient, as it is necessary with the APQ-h and APQ policies.

The analysis of the ED performance was carried out by considering several KPIs related to the APT and the waiting time for consultations. Other specific KPIs could be considered to assess the performance of the ED under different patient flow management policies. Nevertheless, the application of pure priority rules is goal- and objective-independent, and therefore, ED performance will remain unchanged. The computational analysis carried out in this paper shows that these rules can be optimal

Table 8
Summary of the objective value of each scenario with the different queue disciplines and the improvement of the optimal APQ-h and APQ with respect to the best pure priority rule.

Scenario			Management policy						APQ-h solution					
F1	F2	F3	PR-AI	PR-1C	PR-2C	PR-HN	APQ-h & APQ	Improvement	β_{13}	β_{23}	β_{14}	β_{24}	β_{15}	β_{25}
0.9	T0	B0	106	145	75	81	75	0%	0.0396	5.7720	0.0001	4.1227	0.0000	0.0656
		B5	113	137	71	75	71	0%	0.0368	4.9629	0.0001	4.9437	0.0000	0.0565
		B4	103	145	76	81	76	0%	0.0443	4.9691	0.0001	4.9539	0.0000	0.0326
		B3	108	161	82	88	82	0%	0.0601	7.8318	0.0001	2.0637	0.0000	0.0443
	Tp	B0	143	205	98	106	98	0%	0.0608	5.7806	0.0001	4.1199	0.0000	0.0386
		B5	158	196	94	99	94	<1%	0.0376	5.3776	0.0001	4.5459	0.0000	0.0388
		B4	139	208	101	108	101	0%	0.0608	4.9572	0.0001	4.9453	0.0000	0.0366
		B3	144	230	108	115	108	0%	0.0601	8.8210	0.0001	1.0723	0.0000	0.0465
	Tu	B0	161	231	107	116	108	0%	0.0602	5.3336	0.0001	4.5612	0.0000	0.0449
		B5	179	220	103	109	103	<1%	0.0407	5.7440	0.0001	4.1624	0.0000	0.0528
		B4	155	236	110	125	111	0%	3.4125	3.9649	0.0013	2.1180	0.0000	0.5033
		B3	158	257	245	252	126	20%	0.1202	9.5409	0.0967	0.1213	0.0001	0.1208
0.95	T0	B0	171	240	114	114	114	0%	0.0479	5.7642	0.0001	4.1353	0.0000	0.0525
		B5	188	227	108	107	108	0%	0.0456	5.9277	0.0001	3.9886	0.0000	0.0380
		B4	164	244	117	117	117	0%	0.0653	9.7665	0.0001	0.0999	0.0000	0.0682
		B3	163	254	119	122	120	0%	0.0603	9.8036	0.0001	0.0681	0.0000	0.0679
	Tp	B0	225	330	895	879	159	29%	0.1211	9.4163	0.0825	0.2582	0.0001	0.1218
		B5	253	315	142	141	142	0%	0.0433	4.9817	0.0001	4.9343	0.0000	0.0406
		B4	217	336	1545	1552	172	21%	0.1217	5.0678	0.0726	4.6173	0.0001	0.1205
		B3	212	349	1522	1538	185	13%	1.3123	5.8177	0.4861	2.3816	0.0007	0.0016
	Tu	B0	248	370	1693	1668	180	27%	0.2402	9.1343	0.1274	0.2565	0.0002	0.2414
		B5	282	352	155	154	155	0%	0.0378	5.3497	0.0001	4.5821	0.0000	0.0303
		B4	435	378	2220	2197	196	48%	0.2401	7.0783	0.126	2.3531	0.0002	0.2023
		B3	522	395	2144	2153	213	46%	1.7532	5.1798	0.1021	2.9624	0.0024	0.0001

Table 9
KPI results and APQ-h optimal solution for the (96%, Tu, B0) ED scenario.

Scenario	Queue discipline	Obj	$E(X_3)$	$E(X_4)$	$E(X_5)$	$E(\tau_3)$	$E(\tau_4)$	$E(\tau_5)$	$E(\tau_3+\upsilon_3)$	$E(\tau_4+\upsilon_4)$	$E(\tau_5+\upsilon_5)$	Slopes obtained by solving the optimization problem
F1: 95%; F2:Tu; F3: B0	PR-AI	247.60	0.00	<0.01	0.13	2.46	6.44	45.17	6.08	19.69	169.50	$\beta_{13}=0.2402$
	PR-1C	370.02	0.00	<0.01	<0.01	2.51	4.28	11.19	41.67	86.75	169.89	$\boldsymbol{\beta_{23}}=$ **9.1343**
	PR-2C	1693.08	<0.01	0.02	**0.27**	4.94	12.10	78.88	7.24	14.77	82.25	$\beta_{14}=0.1274$
	PR-HN	1668.13	0.00	0.02	**0.28**	2.40	12.59	79.66	6.04	17.01	85.12	$\boldsymbol{\beta_{24}}=$ **0.2565**
	APQ-h & APQ	180.28	0.07	0.14	0.20	9.96	22.49	60.53	14.64	28.44	66.53	$\beta_{15}=0.0002$
												$\boldsymbol{\beta_{25}}=$ **0.2414**

for certain objectives but provide very bad values for others. For example, in non-congested EDs, the PR-2C policy works better than the rest of the pure priority rules, while in very congested EDs, the PR-1C policy works better, especially when great importance is given to avoiding exceeding the APT limit. However, by definition, optimal APQ-h and APQ policies are objective- and goal-dependent, which provides them with a flexibility that managers can use to adapt them to achieve specific objectives or to obtain solutions that balance all of them.

The introduced APQ-h discipline is a modification of the APQ discipline, justified by the previous empirical study [11]. In our computational study, we optimized the parameters of both types of policies to compare them and to determine which one is better or in which ED scenarios one outperforms the other, but in all tested scenarios both policies produced the same results. Differences in KPI values were in decimals, which is attributable to the non-exact optimization procedure and the evaluation of the KPIs by simulation. Therefore, given that no practical differences between them have been found in any of the analysed scenarios with the considered objective functions and both of them have been found to be superior to the other pure priority policies, any of both modalities of the APQ policies could be recommended to be implemented for the management of the ED patient flow. However, although we have not found scenarios in which they differ, their structure is somewhat different, and it is possible that in some situations or under different objective functions, one of the two disciplines will surpass the other. This issue remains to be investigated. Moreover, it would also make sense to investigate a non-linear rate for accumulating priority by taking into account the slack of a patient until the APT is exceeded. Finally, the problem has been solved using commercial optimization software, which, in some scenarios, has shown a slow convergence to the *optimal* solution. Therefore, treating the problem from a multi-objective point of view and developing an efficient optimization algorithm to estimate the Pareto frontier remains an objective for future research.

Acknowledgement

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

Appendix

Table 10 displays each scenario detailed results by disclosing the value for every KPI considered: the policy applied is in first column, the objective function value is in second column, the time target objectives values for P3, P4, and P5 priority patients are in columns third ($E(X_3)$), fourth ($E(X_4)$) and fifth ($E(X_5)$) respectively, the APT for P3, P4, and P5 are in sixth ($E(\tau_5)$), seventh ($E(\tau_4)$), and eighth ($E(\tau_5)$) columns and the TWT for P3, P4, and

Table 10
Summary of the objective and KPI values of each scenario with the different queue disciplines and the improvement of the optimal APQ-h & APQ with respect to the best pure priority rule.

Queue discipline	Obj	$E(X_3)$	$E(X_4)$	$E(X_5)$	$E(\tau_3)$	$E(\tau_4)$	$E(\tau_5)$	$E(\tau_3+\upsilon_3)$	$E(\tau_4+\upsilon_4)$	$E(\tau_5+\upsilon_5)$	Scenario	APQ-h & APQ improvement
PR-AI	106.15	0.00	<0.01	0.02	2.06	4.44	17.99	4.88	12.08	68.84		
PR-1C	145.03	0.00	<0.01	<0.01	2.10	3.24	6.52	18.01	33.27	68.69	F1: 90%;	
PR-2C	75.27	<0.01	<0.01	0.06	3.87	7.80	30.90	5.81	10.01	33.50	F2:T0;	0%
PR-HN	80.72	0.00	0.01	0.06	2.00	8.18	31.25	4.88	11.61	35.61	F3: B0	
APQ-h & APQ	75.28	<0.01	<0.01	0.06	3.47	7.79	30.99	5.63	9.90	34.23		
PR-AI	107.83	<0.01	0.01	0.05	2.34	7.71	26.39	6.59	19.91	79.42		
PR-1C	161.07	<0.01	<0.01	<0.01	2.38	4.16	7.39	22.88	46.64	78.84	F1: 90%;	
PR-2C	82.13	<0.01	0.03	0.10	4.71	11.85	38.40	6.78	14.22	41.17	F2:T0;	0%
PR-HN	88.00	<0.01	0.04	0.10	2.30	12.73	39.29	6.46	18.17	45.72	F3: B3	
APQ-h & APQ	82.44	<0.01	0.03	0.10	4.21	12.03	38.55	6.46	14.86	42.66		
PR-AI	103.00	0.00	<0.01	0.05	1.97	4.30	24.53	4.47	14.45	75.83		
PR-1C	145.25	0.00	<0.01	<0.01	1.99	3.40	7.51	17.25	35.15	75.19	F1: 90%;	
PR-2C	76.10	<0.01	<0.01	0.09	3.51	8.11	36.10	5.41	10.34	38.43	F2:T0;	0%
PR-HN	81.32	0.00	0.01	0.09	1.92	8.33	36.36	4.44	11.31	39.82	F3: B4	
APQ-h & APQ	76.11	<0.01	<0.01	0.09	3.25	8.10	35.99	5.37	10.20	39.02		
PR-AI	113.35	0.00	<0.01	<0.01	1.89	3.09	9.91	4.29	7.29	57.48		
PR-1C	136.58	<0.01	0.00	<0.01	1.91	2.55	5.29	14.77	22.92	57.67	F1: 90%;	
PR-2C	70.90	<0.01	<0.01	0.03	3.58	5.58	23.45	5.44	7.63	25.97	F2:T0;	0%
PR-HN	74.75	0.00	<0.01	0.03	1.86	5.84	23.78	4.23	8.54	27.22	F3: B5	
APQ-h & APQ	70.90	<0.01	<0.01	0.03	3.04	5.67	23.52	4.98	7.63	26.51		
PR-AI	170.52	0.00	<0.01	0.05	2.31	5.44	27.84	5.65	15.23	112.20		
PR-1C	239.86	0.00	<0.01	<0.01	2.37	3.81	8.53	26.20	53.05	112.32	F1: 95%;	
PR-2C	113.90	<0.01	<0.01	0.14	4.51	9.93	49.71	6.72	12.34	53.22	F2:T0;	0%
PR-HN	113.72	0.00	0.01	0.14	2.27	10.36	50.22	5.56	14.30	55.79	F3: B0	
APQ-h & APQ	113.95	<0.01	<0.01	0.14	4.05	9.88	49.88	6.45	12.27	54.18		
PR-AI	162.56	<0.01	0.02	0.11	2.65	9.78	41.85	7.65	26.42	122.19		
PR-1C	253.61	<0.01	<0.01	<0.01	2.70	4.94	9.79	32.77	72.76	121.96	F1: 95%;	
PR-2C	119.22	<0.01	0.05	0.18	5.52	15.32	61.29	7.83	18.18	63.26	F2:T0;	0%
PR-HN	122.38	<0.01	0.06	0.19	2.59	16.43	62.45	7.55	23.25¡	68.74¡	F3: B3	
APQ-h & APQ	119.65	<0.01	0.05	0.18	4.82	15.60	61.64	7.22	20.12	64.85		
PR-AI	164.32	0.00	<0.01	0.10	2.22	5.29	39.62	5.10	19.67	123.38		
PR-1C	244.05	<0.01	<0.01	<0.01	2.27	4.03	10.22	26.21	57.31	122.91	F1: 95%;	
PR-2C	116.62	<0.01	0.01	0.18	4.06	10.65	59.17	6.28	13.16	63.08	F2:T0;	0%
PR-HN	117.40	0.00	0.01	0.18	2.20	10.94	59.72	5.08	14.40¡	64.75¡	F3: B4	
APQ-h & APQ	116.87	<0.01	0.01	0.18	3.34	10.77	59.44	5.54	13.91	63.53		
PR-AI	187.94	<0.01	<0.01	<0.01	2.15	3.63	14.07	4.88	8.68	94.22		
PR-1C	226.95	<0.01	0	<0.01	2.17	2.95	6.70	20.76	34.71	94.53	F1: 95%;	
PR-2C	108.41	<0.01	<0.01	0.07	4.17	6.81	37.24	6.25	9.08	40.10	F2:T0;	0%
PR-HN	107.13	0.00	<0.01	0.07	2.06	7.09	37.62	4.81	10.13	41.64	F3: B5	
APQ-h & APQ	108.30	<0.01	<0.01	0.07	3.46	6.90	37.37	5.64	9.06	40.76		
PR-AI	160.97	0.00	<0.01	0.06	2.21	5.25	28.90	5.40	15.26	112.07		
PR-1C	230.70	0.00	<0.01	<0.01	2.24	3.63	8.47	27.51	54.68	112.30	F1: 90%;	
PR-2C	107.42	<0.01	0.01	0.15	4.29	9.58	51.13	6.43	11.92	53.84	F2:Tu;	0%
PR-HN	115.53	0.00	0.01	0.16	2.17	10.05	51.71	5.38	13.87	56.25	F3: B0	
APQ-h & APQ	107.59	<0.01	0.01	0.15	3.75	9.72	51.30	6.09	12.06	54.69		
PR-AI	158.37	<0.01	0.03	0.13	2.53	9.97	45.46	7.41	27.57	126.60		
PR-1C	256.74	<0.01	<0.01	<0.01	2.59	4.79	9.89	36.30	78.57	126.67	F1: 90%;	
PR-2C	244.57	<0.01	0.06	0.21	5.36	15.74	64.96	7.59	18.37	67.47	F2:Tu;	20%
PR-HN	251.56	<0.01	0.07	0.21	2.49	16.73	66.37	7.32	23.52	72.77	F3: B3	
APQ-h & APQ	126.33	0.02	0.09	0.20	6.77	17.49	57.00	10.71	24.16	62.41		
PR-AI	155.07	0.00	<0.01	0.12	2.12	5.12	41.71	4.84	19.70	122.97		
PR-1C	235.59	0.00	<0.01	<0.01	2.15	3.89	10.02	27.59	59.45	122.63	F1: 90%;	
PR-2C	110.37	<0.01	0.01	0.20	3.90	10.40	61.22	5.97	12.79	64.26	F2:Tu;	0%
PR-HN	124.77	0.00	0.01	0.20	2.09	10.64	61.57	4.87	13.98	65.85	F3: B4	
APQ-h & APQ	110.94	0.00	0.01	0.20	2.93	10.56	61.15	5.76	13.39	64.77		
PR-AI	179.01	0.00	<0.01	<0.01	2.04	3.53	14.73	4.66	8.36	95.07		
PR-1C	219.59	0.00	0.00	<0.01	2.07	2.82	6.63	22.23	36.82	95.35	F1: 90%;	
PR-2C	103.02	<0.01	<0.01	0.08	4.00	6.50	37.97	5.99	8.68	40.73	F2:Tu;	<1%
PR-HN	108.58	0.00	<0.01	0.08	1.98	6.85	38.26	4.61	9.76	42.19	F3: B5	
APQ-h & APQ	102.96	<0.01	<0.01	0.08	3.31	6.65	38.05	5.44	8.68	41.34		
PR-AI	247.60	0.00	<0.01	0.13	2.46	6.44	45.17	6.08	19.69	169.50		
PR-1C	370.02	0.00	<0.01	<0.01	2.51	4.28	11.19	41.67	86.75	169.89	F1: 95%;	
PR-2C	1693.08	<0.01	0.02	0.27	4.94	12.10	78.88	7.24	14.77	82.25	F2:Tu;	27%
PR-HN	1668.13	0.00	0.02	0.28	2.40	12.59	79.66	6.04	17.01	85.12	F3: B0	
APQ-h & APQ	180.28	0.07	0.14	0.20	9.96	22.49	60.53	14.64	28.44	66.53		

(*continued on next page*)

Table 10 (*continued*).

Queue discipline	Obj	$E(X_3)$	$E(X_4)$	$E(X_5)$	$E(\tau_3)$	$E(\tau_4)$	$E(\tau_5)$	$E(\tau_3+\nu_3)$	$E(\tau_4+\nu_4)$	$E(\tau_5+\nu_5)$	Scenario	APQ-h & APQ improvement
PR-AI	522.48	<0.01	0.05	0.22	2.85	13.04	69.98	8.66	37.68	183.51		
PR-1C	395.08	<0.01	<0.01	<0.01	2.89	5.74	13.51	53.15	118.20	183.64	F1: 95%;	
PR-2C	2144.24	<0.01	0.10	0.33	6.22	20.84	97.85	8.65	23.82	99.89	F2:Tu;	46%
PR-HN	2152.75	<0.01	0.11	0.33	2.77	22.18	99.53	8.49	30.11	106.13	F3: B3	
APQ-h & APQ	213.26	0.05	0.15	0.20	8.40	26.42	63.46	14.93	33.36	126.48		
PR-AI	435.44	<0.01	<0.01	0.21	2.37	6.36	66.02	5.47	27.30	184.93		
PR-1C	378.37	<0.01	<0.01	0.01	2.41	4.66	13.97	41.53	93.64	185.01	F1: 95%;	
PR-2C	2220.07	<0.01	0.03	0.33	4.48	13.52	95.78	6.76	16.22	99.20	F2:Tu;	48%
PR-HN	2196.88	0.00	0.03	0.33	2.32	13.87	96.25	5.45	17.70	100.96	F3: B4	
APQ-h & APQ	195.82	0.02	0.14	0.20	7.51	24.39	63.95	13.59	31.18	79.45		
PR-AI	282.31	0.00	<0.01	0.02	2.26	4.02	21.80	5.23	9.98	145.47		
PR-1C	352.40	0.00	0.00	<0.01	2.30	3.21	8.47	33.03	57.85	145.94	F1: 95%;	
PR-2C	155.47	<0.01	<0.01	0.17	4.58	7.75	58.18	6.76	10.17	61.02	F2:Tu;	0%
PR-HN	153.97	0.00	<0.01	0.17	2.21	8.13	58.57	5.17	11.51	62.75	F3: B5	
APQ-h & APQ	155.46	<0.01	<0.01	0.17	3.81	7.81	58.28	6.13	10.13	61.81		
PR-AI	143.36	<0.01	<0.01	0.05	2.17	5.24	26.33	5.27	14.82	97.16		
PR-1C	204.91	0	<0.01	<0.01	2.20	3.61	8.68	25.79	48.01	97.34	F1: 90%;	
PR-2C	98.30	<0.01	0.01	0.13	4.16	9.35	45.18	6.23	11.62	47.68	F2:Tp;	0%
PR-HN	105.92	0.00	0.01	0.13	2.10	9.79	45.52	5.23	13.55	49.98	F3: B0	
APQ-h & APQ	98.42	<0.01	0.01	0.13	3.63	9.43	45.16	5.92	11.70	48.59		
PR-AI	144.13	<0.01	0.03	0.12	2.51	10.10	40.70	7.30	26.31	111.21		
PR-1C	229.65	<0.01	<0.01	<0.01	2.56	4.85	10.12	33.33	68.43	111.05	F1: 90%;	
PR-2C	107.58	<0.01	0.06	0.18	5.23	15.26	57.27	7.37	17.90	59.70	F2:Tp;	0%
PR-HN	115.07	0.00	0.07	0.19	2.47	16.38	58.33	7.22	22.68	64.77	F3: B3	
APQ-h & APQ	108.06	<0.01	0.06	0.18	4.62	15.47	57.62	7.00	18.98	61.28		
PR-AI	139.19	0.00	<0.01	0.10	2.05	5.12	36.84	4.72	18.84	106.74		
PR-1C	208.15	0.00	<0.01	<0.01	2.10	3.86	10.24	25.17	52.17	106.48	F1: 90%;	
PR-2C	100.68	<0.01	0.01	0.17	3.79	10.02	53.02	5.81	12.37	56.22	F2:Tp;	0%
PR-HN	107.52	0.00	0.01	0.17	2.04	10.27	53.40	4.74	13.52	57.74	F3: B4	
APQ-h & APQ	100.84	0.00	0.01	0.17	3.38	10.07	53.09	5.69	12.38	56.82		
PR-AI	157.97	0.00	<0.01	<0.01	2.00	3.44	14.25	4.59	8.27	82.61		
PR-1C	195.89	0.00	0.00	<0.01	2.03	2.77	6.76	21.46	34.10	82.85	F1: 90%;	
PR-2C	94.14	<0.01	<0.01	0.06	3.92	6.41	33.87	5.84	8.58	36.49	F2:Tp;	<1%
PR-HN	98.99	0.00	<0.01	0.06	1.95	6.66	34.28	4.54	9.60	37.90	F3: B5	
APQ-h & APQ	94.05	<0.01	<0.01	0.06	3.29	6.45	34.01	5.33	8.52	37.09		
PR-AI	225.26	0.00	<0.01	0.11	2.43	6.47	40.28	6.01	19.19	151.76		
PR-1C	329.99	0.00	<0.01	<0.01	2.49	4.31	11.22	37.43	75.17	152.09	F1: 95%;	
PR-2C	894.61	<0.01	0.02	0.24	4.83	11.81	69.95	7.11	14.42	73.48	F2:Tp;	29%
PR-HN	879.47	0.00	0.02	0.24	2.39	12.34	70.55	5.92	16.64	76.28	F3: B0	
APQ-h & APQ	159.74	<0.01	0.09	0.19	6.24	18.37	59.93	9.91	24.16	65.86		
PR-AI	212.05	<0.01	0.05	0.19	2.80	12.86	61.11	8.45	34.61	163.66		
PR-1C	349.20	<0.01	<0.01	<0.01	2.89	5.76	13.36	46.74	101.97	163.77	F1: 95%;	
PR-2C	1521.67	0.01	0.09	0.29	6.09	19.81	86.10	8.51	22.76	87.91	F2:Tp;	13%
PR-HN	1538.46	<0.01	0.10	0.29	2.77	21.06	87.60	8.37	28.66	94.17	F3: B3	
APQ-h & APQ	185.39	0.01	0.15	0.20	6.23	24.94	62.74	11.88	30.66	109.67		
PR-AI	216.66	0.00	<0.01	0.18	2.35	6.34	58.32	5.36	26.05	164.97		
PR-1C	336.36	0.00	<0.01	0.01	2.40	4.65	13.88	37.17	81.93	164.92	F1: 95%;	
PR-2C	1544.95	<0.01	0.03	0.29	4.41	13.17	84.17	6.61	15.80	87.77	F2:Tp;	21%
PR-HN	1552.25	0.00	0.03	0.29	2.29	13.45	84.71	5.33	17.18	89.57	F3: B4	
APQ-h & APQ	171.90	0.02	0.11	0.20	7.69	21.04	61.43	13.00	26.68	68.54		
PR-AI	252.86	0.00	<0.01	0.02	2.26	3.97	20.41	5.14	9.84	129.08		
PR-1C	315.32	0.00	<0.01	<0.01	2.28	3.19	8.62	30.71	51.30	129.42	F1: 95%;	
PR-2C	142.01	<0.01	<0.01	0.14	4.42	7.53	52.07	6.63	9.94	55.01	F2:Tp;	0%
PR-HN	140.50	0.00	<0.01	0.14	2.19	7.91	52.50	5.10	11.29	56.67	F3: B5	
APQ-h & APQ	142.01	<0.01	<0.01	0.14	3.71	7.65	52.19	6.03	9.95	55.69		

P5 patients who need two consultations are in ninth ($E(\tau_3+\nu_3)$), tenth ($E(\tau_4+\nu_4)$), and eleventh ($\mathrm{E}(\tau_5+\nu_5)$) respectively. The last two columns are the description of each scenario and the improvement of the objective function value with respect to the best Pure Priority Rule.

References

[1] J.D. Schuur, A.K. Venkatesh, The growing role of emergency departments in hospital admissions, N. Engl. J. Med. 367 (2012) 389–391, http://dx.doi.org/10.1056/NEJMp1206519.

[2] NHS England, reducing emergency admissions, 2018, https://www.nao.org.uk/wp-content/uploads/2018/02/Reducing-emergency-admissions-Summary.pdf.

[3] N. Tang, J. Stein, R.Y. Hsia, J.H. Maselli, R. Gonzales, Trends and characteristics of US emergency department visits, 1997-2007, JAMA 304 (2010) 664, http://dx.doi.org/10.1001/jama.2010.1112.

[4] B.C. Strunk, P.B. Ginsburg, M.I. Banker, The effect of population aging on future hospital demand, Health Aff. 25 (2006) http://dx.doi.org/10.1377/hlthaff.25.w141.

[5] National center for health statistics, health, United States, 2016: With chartbook on long-term trends in health, Cent. Dis. Control (2017) 314–317, http://www.ncbi.nlm.nih.gov/pubmed/28910066 (accessed July 25, 2018).

[6] M. McHugh, K. Van Dyke, M. McClelland, D. Moss, Improving Patient Flow and Reducing Emergency Department Crowding: A Guide for Hospitals, Rockville, MD, 2011. https://hsrc.himmelfarb.gwu.edu/cgi/viewcontent.cgi?referer=https://scholar.google.es/&httpsredir=1&article=1041&context=sphhs_policy_facpubs (accessed July 25, 2018).

[7] F. Aguado-Correa, M. Herrera-Carranza, N. Padilla-Garrido, Variability and overcrowding management, J. Health Manag. 18 (2016) 218–230, http://dx.doi.org/10.1177/0972063416637697.

[8] S. Peiró, J. Librero, M. Ridao, E. Bernal-Delgado, Variabilidad en la utilización de los servicios de urgencias hospitalarios del Sistema Nacional de Salud, Gac. Sanit. 24 (2010) 6–12, http://dx.doi.org/10.1016/j.gaceta.2009.06.008.

[9] M.E.H. Ong, K.K. Ho, T.P. Tan, S.K. Koh, Z. Almuthar, J. Overton, S.H. Lim, Using demand analysis and system status management for predicting ED attendances and rostering, Am. J. Emerg. Med. 27 (2009) 16–22, http://dx.doi.org/10.1016/j.ajem.2008.01.032.

[10] J.L. Wiler, R.T. Griffey, T. Olsen, Review of modeling approaches for emergency department patient flow and crowding research, Acad. Emerg. Med. 18 (2011) 1371–1379, http://dx.doi.org/10.1111/j.1553-2712.2011.01135.x.

[11] Y. Ding, E. Park, M. Nagarajan, E. Grafstein, Patient prioritization in emergency department triage systems: An empirical study of Canadian triage and acuity scale (CTAS), SSRN Electron. J. (2018) http://dx.doi.org/10.2139/ssrn.2843932.

[12] D.A. Stanford, P. Taylor, I. Ziedins, Waiting time distributions in the accumulating priority queue, Queueing Syst. 77 (2014) 297–330, http://dx.doi.org/10.1007/s11134-013-9382-6.

[13] S.J. Welch, B.R. Asplin, S. Stone-Griffith, S.J. Davidson, J. Augustine, J. Schuur, Emergency department operational metrics, measures and definitions: Results of the second performance measures and benchmarking summit, Ann. Emerg. Med. 58 (2011) 33–40, http://dx.doi.org/10.1016/j.annemergmed.2010.08.040.

[14] L. Mayhew, D. Smith, Using queuing theory to analyse the Government's 4-h completion time target in Accident and Emergency departments, Health Care Manag. Sci. 11 (2008) 11–21, http://dx.doi.org/10.1007/s10729-007-9033-8.

[15] S. Saghafian, G. Austin, S.J. Traub, Operations research/management contributions to emergency department patient flow optimization: Review and research prospects, IIE Trans. Healthc. Syst. Eng. 5 (2015) 101–123, http://dx.doi.org/10.1080/19488300.2015.1017676.

[16] T.H. Taylor, A.M.C. Jennings, D.A. Nightingale, B. Barber, D. Leivers, M. Styles, J. Magner, A study of anaesthetic emergency work. Paper 1: The method of study and introduction of queuing theory, Br. J. Anaesth. 41 (1969) 70–75, https://pdfs.semanticscholar.org/8098/a9ec0f728d721a614dc70fa3d621e69826d2.pdf (accessed July 27, 2018).

[17] R.K. Dieter Haussmann, Waiting time as an index of quality of nursing care, Heal. Serv. Res. 5 (1970) 92–105, http://www.ncbi.nlm.nih.gov/pubmed/5482376 (accessed July 1, 2019).

[18] K. Siddharthan, W.J. Jones, J.A. Johnson, A priority queuing model to reduce waiting times in emergency care, Int. J. Health Care Qual. Assur. 9 (1996) 10–16, http://dx.doi.org/10.1108/09526869610124993.

[19] M. Laskowski, R.D. McLeod, M.R. Friesen, B.W. Podaima, A.S. Alfa, Models of emergency departments for reducing patient waiting times, PLoS One 4 (2009) e6127, http://dx.doi.org/10.1371/journal.pone.0006127.

[20] G.S. Mokaddis, I.A. Ismail, S.A. Metwally, K.M. Metry, Response times for health care system, J. Appl. Math. Bioinform. 1 (2011) 131–146.

[21] D.G. McQuarrie, Hospitalization utilization levels. The application of queuing. Theory to a controversial medical economic problem, Minn. Med. 66 (1983) 679–686.

[22] S.S. Panwalkar, W. Iskander, A survey of scheduling rules, Oper. Res. 25 (2008) 45–61, http://dx.doi.org/10.1287/opre.25.1.45.

[23] M. Armony, S. Israelit, A. Mandelbaum, Y.N. Marmor, Y. Tseytlin, G.B. Yom-Tov, On patient flow in hospitals: A data-based queueing-science perspective, Stoch. Syst. 5 (2015) 146–194, http://dx.doi.org/10.1287/14-SSY153.

[24] R. Hall, D. Belson, P. Murali, M. Dessouky, Modeling patient flows through the health care system, in: R.W. Hall (Ed.), Int. Ser. Oper. Res. Manag. Sci., Springer, Boston, MA, 2013, pp. 3–42, http://dx.doi.org/10.1007/978-1-4614-9512-3_1.

[25] S. Zeltyn, T. Lauterman, D. Schwartz, K. Moskovitch, S. Tzafrir, F. Basis, Y.N. Marmor, A. Mandelbaum, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, A. Shtub, Simulation-based models of emergency departments: Operational, tactical and strategic staffing, ACM Trans. Model. Comput. Simul. 21 (2011) 1–25, http://dx.doi.org/10.1145/2000494.2000497.

[26] Institute for healthcare improvement (IHI), patient first: Efficient patient flow management impact on the ED, 2011, http://www.ihi.org/resources/Pages/ImprovementStories/PatientFirstEfficientPatientFlowManagementED.aspx (accessed February 1, 2019).

[27] M. Mchugh, K. Van Dyke, M. Mcclelland, D. Moss, Improving patient flow and reducing emergency department crowding: A guide for hospitals, 2011, https://www.ahrq.gov/sites/default/files/publications/files/ptflowguide.pdf (accessed February 1, 2019).

[28] J. Huang, B. Carmeli, A. Mandelbaum, Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback, Oper. Res. 63 (2015) 892–908, http://dx.doi.org/10.1287/opre.2015.1389.

[29] W.E. Smith, Various optimizers for single-stage production, Nav. Res. Logist. Q. 3 (1956) 59–66, http://dx.doi.org/10.1002/nav.3800030106.

[30] J.A. van Mieghem, Dynamic scheduling with convex delay costs: The generalized $c| mu$ rule, Ann. Appl. Probab. 5 (2007) 809–833, http://dx.doi.org/10.1214/aoap/1177004706.

[31] A. Mandelbaum, A.L. Stolyar, Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule, Oper. Res. 52 (2004) 836–855, http://dx.doi.org/10.1287/opre.1040.0152.

[32] I. Gurvich, W. Whitt, Scheduling flexible servers with convex delay costs in many-server service systems, Manuf. Serv. Oper. Manag. 11 (2008) 237–253, http://dx.doi.org/10.1287/msom.1070.0211.

[33] L. Kleinrock, A delay dependent queue discipline, Nav. Res. Logist. Q. 11 (1964) 329–341, http://dx.doi.org/10.1002/nav.3800110306.

[34] A. Bin Sharif, D.A. Stanford, P. Taylor, I. Ziedins, A multi-class multi-server accumulating priority queue with application to health care, Oper. Res. Heal. Care 3 (2014) 73–79, http://dx.doi.org/10.1016/j.orhc.2014.01.002.

[35] Y.B. Ferrand, M.J. Magazine, U.S. Rao, T.F. Glass, Managing responsiveness in the emergency department: Comparing dynamic priority queue with fast track, J. Oper. Manag. 58–59 (2018) 15–26, http://dx.doi.org/10.1016/j.jom.2018.03.001.

[36] G. Zayas-Caban, J. Xie, L.V. Green, M.E. Lewis, Policies for physician allocation to triage and treatment in emergency departments, IISE Trans. Healthc. Syst. Eng. (2019) 1–15, http://dx.doi.org/10.1080/24725579.2019.1620384.

[37] M. Twomey, L.A. Wallis, J.E. Myers, Limitations in validating emergency department triage scales, Emerg. Med. J. 24 (2007) 477–479, http://dx.doi.org/10.1136/emj.2007.046383.

[38] P. Lindsay, The development of indicators to measure the quality of clinical Care in emergency departments following a modified-delphi approach, Acad. Emerg. Med. 9 (2002) 1131–1139, http://dx.doi.org/10.1197/aemj.9.11.1131.

[39] S. Welch, J. Augustine, C.A. Camargo, C. Reese, Emergency department performance measures and benchmarking summit, Acad. Emerg. Med. 13 (2006) 1074–1080, http://dx.doi.org/10.1197/j.aem.2006.05.026.

[40] L.I. Horwitz, J. Green, E.H. Bradley, US emergency department performance on wait time and length of visit, Ann. Emerg. Med. 55 (2010) 133–141, http://dx.doi.org/10.1016/j.annemergmed.2009.07.023.

[41] E.B. Kulstad, R. Sikka, R.T. Sweis, K.M. Kelley, K.H. Rzechula, ED overcrowding is associated with an increased frequency of medication errors, Am. J. Emerg. Med. 28 (2010) 304–309, http://dx.doi.org/10.1016/j.ajem.2008.12.014.

[42] J.S. Olshaker, N.K. Rathlev, Emergency department overcrowding and ambulance diversion: The impact and potential solutions of extended boarding of admitted patients in the emergency department, J. Emerg. Med. 30 (2006) 351–356, http://dx.doi.org/10.1016/j.jemermed.2005.05.023.

[43] S.J. Weiss, A.A. Ernst, M.R. Sills, B.J. Quinn, A. Johnson, T.G. Nick, Development of a novel measure of overcrowding in a pediatric emergency department, Pediatr. Emerg. Care 23 (2007) 641–645, http://dx.doi.org/10.1097/PEC.0b013e31814a69e2.

[44] M.A. Ahmed, T.M. Alkhamis, Simulation optimization for an emergency department healthcare unit in Kuwait, European J. Oper. Res. 198 (2009) 936–942, http://dx.doi.org/10.1016/j.ejor.2008.10.025.

[45] F. Mallor, C. Azcárate, J. Barado, Control problems and management policies in health systems: application to intensive care units, Flex. Serv. Manuf. J. 28 (2016) 62–89, http://dx.doi.org/10.1007/s10696-014-9209-8.

[46] C.M. Macal, M.J. North, N. Collier, V.M. Dukic, D.S. Lauderdale, M.Z. David, R.S. Daum, P. Shumm, J.A. Evans, J.R. Wilder, D.T. Wegener, Modeling the spread of community-associated MRSA, in: Proc. Title Proc. 2012 Winter Simul. Conf., IEEE, 2012, pp. 1–12, http://dx.doi.org/10.1109/WSC.2012.6465271.

[47] F. Mallor, C. Azcárate, Combining optimization with simulation to obtain credible models for intensive care units, Ann. Oper. Res. 221 (2014) 255–271, http://dx.doi.org/10.1007/s10479-011-1035-8.

[48] W. Luo, J. Cao, M. Gallagher, J. Wiles, Estimating the intensity of ward admission and its effect on emergency department access block, Stat. Med. 32 (2013) 2681–2694, http://dx.doi.org/10.1002/sim.5684.

[49] I. Rockwell, Automation Technologies, Arena, 2016.

[50] Q-UPHS: quantitative methods for uplifting the performance of health services, 2019, http://www.unavarra.es/digitalAssets/233/233798_100000proyecto05_video01.mp4.

[51] M. Laguna, R. Martí, The optquest callable library, in: Optim. Softw. Cl. Libr., Kluwer Academic Publishers, Boston, 2005, pp. 193–218, http://dx.doi.org/10.1007/0-306-48126-x_7.

[52] N. Li, D.A. Stanford, Multi-server accumulating priority queues with heterogeneous servers, European J. Oper. Res. 252 (2016) 866–878, http://dx.doi.org/10.1016/j.ejor.2016.02.010.