Omega 117 (2023) 102844

Contents lists available at ScienceDirect

Omega

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

Real-time resource allocation in the emergency department: A case study☆

Davide Duma[a,1], Roberto Aringhieri[b,*]

[a] Dipartimento di Matematica "Felice Casorati", Università degli Studi di Pavia, via Adolfo Ferrata 5, I-27100, Pavia, Italy
[b] Dipartimento di Informatica, Università degli Studi di Torino, Corso Svizzera 185, I-10149, Torino (Italy)

ARTICLE INFO

Article history:
Received 8 November 2021
Accepted 23 January 2023
Available online 26 January 2023

This manuscript was processed by Associate Editor Kuhn.

Keywords:
Emergency department
Patient flow
Online allocation
Simulation
Process mining

ABSTRACT

Overcrowding is a phenomenon that affects Emergency Departments (EDs) worldwide determining a harmful impact on the healthcare provided. Because of the wide variety of different patient paths, in the literature the ED processes are usually modeled making significant assumptions, and neglecting fundamental aspects. Such assumptions could make sense for strategic or tactical decisions but nowadays the objective most frequently required by practitioners is the optimization of already available resources. To deal with this problem, we need to act at the operational level, with a particular attention to resource allocation in real time since arrivals and activities to be performed are known only over time. In this paper we present the case study of an Italian ED to investigate if an online allocation algorithm based on prioritization combined with a prediction tool can improve the ED performance, alleviating the overcrowding. Then, we propose several policies for the online allocation of the ED resources, which take into account the real-time state of the ED and the prediction of the next activities provided by an ad hoc process mining model. The proposed approach is validated and analyzed on the case study through a fine-grained simulation model. Results suggest that if the decisions of allocating resources for the execution of activities take into consideration the probably subsequent activities, then there is a room for improvement for the average door-to-doctor time, ED Length-of-Stay, and resource utilization.

1. Introduction

An emergency department (ED) is a hospital unit specialized for providing 24-hour emergency care to unscheduled patients. The ED staff has the task of performing initial treatments for a broad spectrum of pathologies and injuries with different urgency. Such treatments require the execution of several activities using shared (and usually scarce) human and equipment resources, which need to be coordinated to deal with a patient flow that varies for volume and patient characteristics throughout the day, along the week, and according to the seasons. These aspects make resource management complex, which can have a direct impact on patients' health.

The management of the EDs is a problem destined to grow in complexity because of an increasing demand without a corresponding increase in resources. The reason is the aging population ofhigh-income countries, since the main users of this health service are elderly people. For instance, admissions in EDs grew over 50% from 1992 to 2006 in the United States [1], while in Italy people aged 65 or older are destined to increase by 33% in the next 30 years [2].

A phenomenon that affects the EDs worldwide is the overcrowding, which nowadays reaches crisis proportions [3,4]. Excessive number of patients waiting or treated in the hallways, long patient waiting times, ambulances diverted, and patients that leave without being seen (LWBS) are all effects of the ED overcrowding [5]. Consequently, this phenomenon could have dangerous consequences on the safety of patients, as it has been shown that as ED crowding increases, waiting times and medical errors also increase. Then the overcrowding is the cause of worse patient outcomes [6]. Some studies found a correlation between the mortality rate of ED patients when the ED is overcrowded. In [7] the incidence of patients with delayed resuscitation efforts in crowded days has been estimated through an odds ratio of 2.0, with an even higher incidence of mortality. In [8] an overall inpatient mortality has been lowered from 1.5% to 1.3% in correspondence of a decrease of the Door-To-Doctor Time (DTDT) and ED Length-of-Stay (EDLOS). Other undesirable consequences are a stronger stress

☆ This manuscript was processed by Associate Editor Prof. Benjamin Lev.

* Corresponding author.
E-mail addresses: davide.duma@unipv.it (D. Duma), roberto.aringhieri@unito.it (R. Aringhieri).

[1] This work has been done when the author was Ph.D. student at the Computer Science Department of the University of Turin.

https://doi.org/10.1016/j.omega.2023.102844
0305-0483/

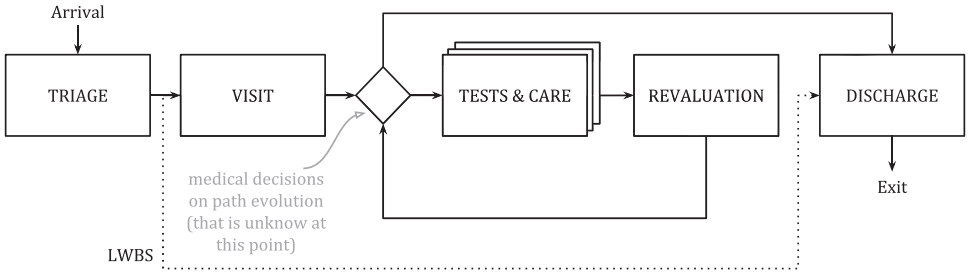

Fig. 1. Flowchart of a general patient path.

among the ED staff [9] and higher costs because of the decreased productivity [10].

In the real world, such a problem is usually addressed providing a real-time measure (e.g., the NEDOCS) that estimates the level of crowding. However, the limits of such measures have been shown in [11], besides the fact that they are just used as an alarm bell for exceptional and unsystematic interventions while an issue such as the ED overcrowding should be addressed with structural solutions.

One of the most widely used methodologies in the OR&MS literature for modeling the ED process is simulation [12–14], which allows scenario analysis for the improvement of the performance taking into account stochastic aspects of the ED environment. Although most of the solutions try to determine at a strategic-tactical level how many resources need to be added to eliminate bottlenecks [15,16], the medical literature (as well as the experience of the case study that we present in this paper afterwards) suggests that usually there are not enough public funds to put into practice an increase in resources [10,17]. Therefore, the decision maker needs some supports for optimizing the existing human and equipment resources. This has led in recent years to an increase in research work at the operational level, focusing almost exclusively on the scheduling and allocation of human resources [18,19].

In this paper we deal with the real-time resource allocation, that is the online decision problem arising at the online operational level since arrivals and activities to be performed are known only over time. These sources of uncertainty require the adoption of policies for assigning the available human and equipment resources to a dynamic list of patients requiring to be treated. To this scope, online allocation algorithms could be used to allocate the ED resources in real-time in such a way to deal with uncertainty. This type of algorithms lies within the more general class of multi-stage optimization algorithms, for which online optimization is a suitable methodology to deal with both dinamicity and uncertainty, through sequential decisions to be taken over time [20]. Furthermore, we can exploit updated information about the current use of resources and patients waiting or undergoing treatments, with a lookahead approach. However, such an approach requires a fine-grained knowledge of the patient paths.

We illustrate a flowchart representing the general path of a patient within an ED in Fig. 1. Generally, such a framework differs among different EDs only for the set of the *Tests & Care* activities (exams, therapies, and observations), whose competence could be of the ED or another ward (e.g., specialist visits are usually performed in the corresponding ward). After the arrival, the patient is triaged by a nurse and they waits for the admission, which starts with the first medical visit. From that point the patient enters within an activity cycle organized as follows: after the visit, the physician can prescribe a sequence of *Tests & Care* activities to be executed before a revaluation, until the physician decides that the patient can be discharged or hospitalized.

One of the main factors of uncertainty of such an emergency pathway is the lack of knowledge regarding the sequence of the activities to which the patient will undergo: the ED staff knows only a part of the next activities of their treatment after the first visit (or re-evaluation visit). Since different activities involve different resources, this makes more complex their allocation. The missing knowledge of the subsequent activities could be given by a prediction tool capable to provide more information, such as a classification of the patients in accordance with the predicted future need of activities and resources. Further, the prediction tool could support the decision process as depicted in the Figs. 2 and 3, which retraces the same framework in Fig. 1. For the sake of simplicity, we do not consider urgency codes in this example but they are taken into account in the proposed approach.

In Fig. 2, every patient is represented with a certain number of circles marked with the same ID number in order to show the possible waiting in multiple queues simultaneously: each circle indicates that the patient is waiting for an activity, that could be a visit or a *Tests & Care* activity. In the absence of additional information, all patients who are waiting for a visit (i.e., the first medical visit or a revaluation visit) are indistinguishable. Conversely, the order in which they will be visited could have a different impact on the resource bottlenecks.

In Fig. 3, we show how a generic prediction tool could classify the patients that are waiting for the (first or revaluation) visit on the basis of their characteristics and their already performed activities. For each cluster of patients (grouped in a dashed circle), the tool could provide several indices such as probability of the need of a certain activity in the next *Test & Care* sequence or the probability to be discharged at the end of that visit. Such probabilities could be taken into account to avoid or limit the bottlenecks on the resources. For instance, if two patients with non-urgent codes and similar waiting times are in the queue for the medical visit, and one of them has a high probability to require a resource with a long queue, then it should not make sense to visit them firstly, since they will continue to wait for a long time after the visit. Conversely, if a patient that is occupying a critical resource (e.g., a stretcher) has a high probability to be discharged after the next visit, it could be useful to give them the priority in order to release such a resource.

The effectiveness of the standard process discovery approaches to mine a process model that is adequate with respect to these needs is investigated in [21]. Because of its limitations, an ad hoc process discovery algorithm called *Hybrid Activity Forest* (HAF) has been proposed in [22]. The HAF consists of a set of *Hybrid Activity Trees* (HATs), each one of them is associated to a subset of patient characteristics (e.g., main symptom, age, urgency, sex, arrival hour, ...) and it is capable to make accurate predictions on the basis of the already performed activities at the ED. The information contained by a HAT could be useful in the real-time resource allocation, since it allows the estimation of the probability of subsequent activities to be performed by a patient in the next minutes or hours.

The aim of this work is to investigate if an online allocation approach combined with a prediction tool can improve the ED perfor-

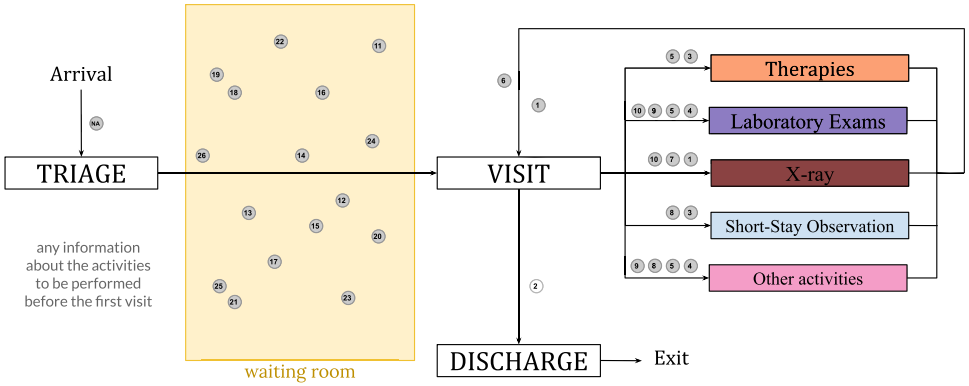

Fig. 2. Patient flow in a general ED without prediction about the next activities to be performed.

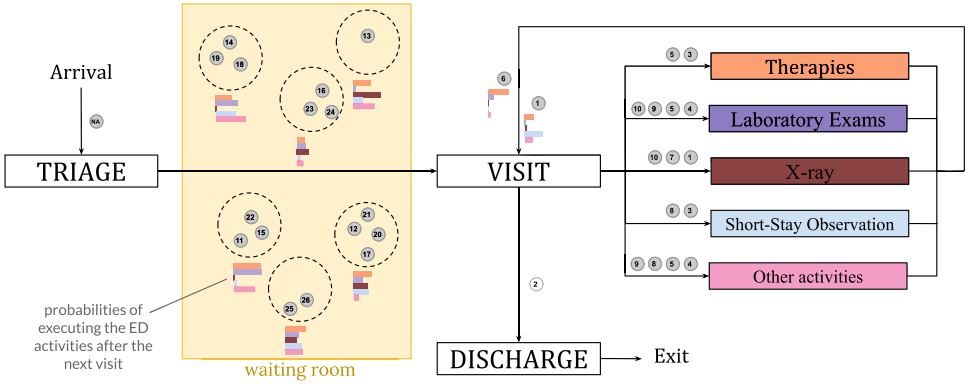

Fig. 3. Patient flow in a general ED exploiting the information provided by a prediction tool.

mance, alleviating the inherent phenomenon of the ED overcrowding. In this paper, we present a real case study of an Italian ED and we propose an online allocation framework for the resource allocation in real time, which takes into account the current state of the ED and the prediction of the next activities provided by the HAF. We analyze the impact of an online algorithm based on such a framework, embedding it within a fine-grained simulation model. Our approach falls in the general class of online algorithms with lookahead formalized in [23,24], in which the instance revelation rule that provides deterministic information previews in addition to the partial available information is given by a prediction model. To the best of our knowledge, an analysis of online approaches based on prediction for the real-time resource allocation of the ED has never been addressed in literature.

We would like to remark that the fine granularity of our simulation model is due to the need of replicating the activities of the patients in accordance with their pathways and the interdependence between activities and the resources utilization. To model the patient flow through the ED in such a way to analyze the impact of an online algorithm, we use a Discrete Event Simulation (DES) methodology as suggested in [23,25], in which events occur at a particular time instant and marks a change of the state in the system. However, we use the Agent-Based Simulation (ABS) semantics to model straightforwardly the pathways of patients and the tasks of the human resources having a behavior that is not representable by a simple resource pool, such as the time for handover, the assignment of the same resource to a patient to ensure continuity in the treatment, or a limited availability in certain phases of the work shift.

This paper is organized as follows. A literature review about the existing OR&MS approaches to deal with the improvement of the ED performance at the operational level is presented in Section 2. In Section 3 we introduce the case study of an Italian ED, reporting its organization and describing the available data. In Section 4 we resume our work presented in [21,22] in which process mining techniques are exploited to obtain a model capable to predict the evolution of the patient paths. In Section 5 we propose an online allocation framework that provides several policies to operate on different possible bottlenecks of the patient flow. The simulation model that embed such policies is described in Section 6. In Section 7 we provide a quantitative analysis to prove the effectiveness of the proposed online allocation framework. Finally, Section 8 closes the paper.

In this paper, we used the so-called *singular they*, which is an epicene (gender-neutral) third-person pronoun, and its derivative forms.

2. Literature survey

In this section we present a literature review about the OR&MS studies that deal with the problem of improving the ED performance. After an overview of the main optimization methods used to address the real-time allocation of the ED resources, we focus our attention on the simulation methodologies used to model the flow of patients in the ED. Furthermore, the novelties of our research with respect to the existing studies are highlighted.

In recent years, the lack of funds in EDs has led to an increase in research on how to optimize the resources already available, which takes place mainly at the operational level. A first type of attempt lies at the Emergency Medical System (EMS) level, where patients are transported by ambulances to the EDs of a certain geographical area choosing the destination with the aim of balancing the workload among the EDs and reducing waiting times [26–29]. The second type of attempt concerns the resource man-

agement within the ED. Here we can identify two classes of decision problems, that is (i) the shifts of the human resources and (ii) the resource allocation.

The first class can be generally placed at the offline operational level, since decisions about the number of physicians [30–34], nurses [32–35], and other medical health personnel [34] and their shifts are taken at least few days before the actual execution of the activities.

In the second class of problems, equipment resources are fixed, while human resources have been already organized in shift and need to be allocated. This class can be divided into two subclasses, that is the offline and online optimization of the ED resource allocation. While the former deals with the a priori assignment of resources to certain areas of the ED or categories of patients, in accordance with fluctuation in demands throughout the week or the year [36], the latter deals with the real-time resource allocation as patients arrive and require the same limited resource, taking decision based on the current state of the ED system.

Most of the research about the real-time ED resource allocation focus on the dynamic patient-physician assignment process, which has a crucial role. The trade-off between patient satisfaction and efficiency is analyzed in [37] through a queuing system with stochastic arrival times and visit durations to evaluate the impact of patient-centered and facility-centered policies for general walk-in clinics, including EDs. In [38] a queueing theory framework is introduced to evaluate different policies to allocate physicians to patients, using deadlines to model the maximum DTDT and feedback to balance the access to the visit room to different classes of patients. In [39] authors design a DES model to evaluate the performance of different queue management policies for the access to the visit room, that is giving priority to patients with the higher urgency codes and/or those that have to undergo the first visit or the revaluation visit, which have different probabilities to be discharged after such visits. Optimization is used to find the best value of parameters that determine a prioritization, with the aim of minimizing the DTDT of non-urgent patients. The study is extended in [40], providing a scenario analysis that proves the domination of the accumulating priority queues over pure priority queues with respect to several Key Performance Indices (KPIs). The allocation of the physician to patients is also optimized in [41], where a DES model is used to find the optimal parameter settings of the allocation rule with the aim of limiting the physician workload and minimizing both the DTDT and the EDLOS. A greedy heuristic based on priority queue and a general variable neighborhood search for the scheduling of the patients with different urgencies to physician is proposed in [42]. For the same problem, in [43] an integer linear program is formulated and evaluated on an ED modeled through a queueing network, taking into account the expected residual consultation times of patients waiting for a visit, which are estimated from real data.

Although the dynamic patient-physician assignment could be one of the most impactful online decision problems, interviews with ED staff of the presented case study revealed that there may also be other activities and resources that can congest the patient flow. Some examples could be the exhaustion of the SSO beds, which implies the occupation of the beds used for the medical visits, or the increase of patients in the waiting room and corridors of the ED, which must be monitored periodically by nurses, slowing the performance of other activities, such as blood tests or therapies.

Nevertheless, to the best of our knowledge the unique work that addresses the problem of the real-time allocation of both human and equipment resources of an ED is [44]. The authors propose an online scheduling of the beds and 6 different tasks (medical assessment, vital signs & electrocardiogram, take bloods, pathology test, administer treatment, review & discharge) based on updated information available every time a patient arrives at the triage or ends a task. The ED system is modeled as a two-layer assignment and sequencing problem, considering the dependency between them: the patient-bed assignment is modeled as a parallel machine scheduling environment with machine groups, then the task-resource allocation is modeled as a flexible job shop accordingly. Because of the lack of data about activity records, authors develop five possible patient paths, under the assumption that they are deterministic since patients arrive.

In this paper we propose an online allocation approach based on a dynamic prioritization of queues for the real-time allocation of physicians, nurses, a X-ray technician, three different exam machines, beds (divided into visit rooms and short-stay observation units), and stretchers, which are used for executing more than 7.800 possible patient paths involving 17 different ED activities, that is replicated from real data. The proposed dynamic prioritization consists on a greedy algorithm that tries to alleviate the effect of possible bottlenecks, considering the resources availability on the short term. The idea beside the algorithm is to determine a prioritization depending also on the probability that patients will require some resources in the next minutes or hours. Such a probability can be estimated through the knowledge about the possible path evolutions provided by a prediction model.

Another novelty of our method lies in considering as a factor of uncertainty not only the arrival time of patients and their type, but also the evolution of their pathway over time, that is what happens in the reality. We take into account the interdependence between their events, which is predicted and optimized exploiting the knowledge provided by a process mining model, that is the first attempt in the context of the real-time ED management.

Due to inherent uncertainty of ED processes, simulation models are widely used for supporting decision making [12–14]. An intertwining of different levels of causal dependencies should be considered when modeling such a complex system, in order to mimic the volatile behaviors of the EDs with the challenging objective of optimizing their patient flow on the basis of selected KPIs from both a patient and an efficiency perspective [45]. Furthermore, its explainability and the possibility of visualizing the impact of decisions on entire ED system is appreciated by practitioners [46].

DES is the most common methodology in the literature of the patient flow optimization at the ED. DES models are generally used to its appropriateness in replicating the ED processes and guarantee the reliability of the results [47], considering events (patient arrivals, activity executions, change of the staff shift, etc.) that occur over time in accordance with theoretical or empirical distributions. ABS is alternatively used to DES in studies that focus on the interactions between independent agents, since it allows the modeling of particular patient pathways [14] or the organizational behavior of the ED staff when performing activities that involve patients or other healthcare personnel [13].

Prior simulation models often introduce assumptions to deal with their limits at different levels. While DES models often consider human resources as non-interacting pools, that involves a significant simplification from the point of view of efficiency, ABS models suffer from the difficulty of calibrating at the macroscopic level the result of the interactions between the agents modeled at the microscopic level. Furthermore, because of the wide variety of different patient paths within the ED process and the missing of data or tools to mine them, in the above studies the ED processes (or a part of them) are modeled making significant assumptions and neglecting fundamental aspects, such as the interdependence between activities and accordingly the access to resources. Such assumptions could make sense for strategic or tactical decisions, but could lead to misleading results when it comes to managing a complex process in real time.

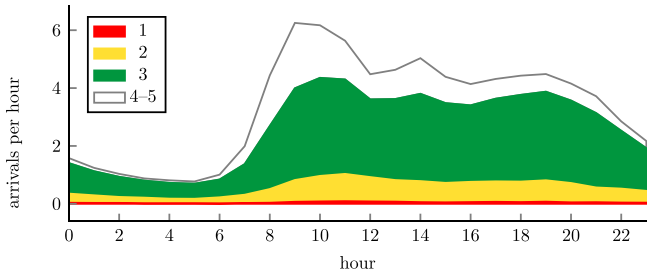

Fig. 4. Patient accesses during the day divided for urgency code.

Then, we introduce a novel DES model that replicates with a high-granularity the possible patient paths and simultaneous access to multiple resources. In addition, we adopt the semantics of ABS to model some management behaviors of the medical staff that can lead to an increase in overcrowding under specific circumstances, which have been defined and modeled in accordance with the suggestions of the ED staff, as reported in Section 6.

3. The case study

As case study for showing the applicability and the effectiveness of the proposed approach, we used the data collected by the ED of Ospedale Sant'Antonio Abate di Cantù, in Italy. For all 88 272 in 2013–2015, the ED recorded accesses information about main symptom (between 31 options), urgency code (from 1 to 5, similar to those of the Emergency Severity Index adopted for the triage in the United States [48]), personal data, arrival mode (autonomous or by ambulance), and a set of timestamps about the activities executed within the ED (completion of triage, exam reporting time, start of short-stay observation, etc.). Further information have been provided by the head of the ED about the number of beds (4) and Short-Stay Observation (SSO) units (5) and their location in the visit rooms (ordinary rooms, shock room, and minor codes ambulatory (MCA)), the number of stretchers (10) the management settings for the specialist visits and exams (whose competence may be the ED or other wards with precise scheduling rules and timetables for access). The human resources vary during the shifts of the day and between weekdays and weekend as follows: 1–3 physician(s), 4–6 nurses, and 0–1 technician.

The accesses have different fluctuations throughout the day, among the days of the week, and among the seasons, but also among the urgency classes. Among all these fluctuations, those that have the higher impact occur in different time windows of the same the day, with a peak of non-urgent patients in the business hours (8:00–20:00), as shown in Fig. 4. In such hours we can observe different arrival rates: for instance, the average number of accesses in the time window 9:00–10:00 is approximately three times that of 7:00–8:00, and the 50% more than 11:00–12:00. Even with an adequate staff scheduling, a very variable number of care requests must be met during different hours of the day and in within short time limits. A Markov model of the arrival process of this ED has been discussed in a previous work [49].

3.1. Organization of the emergency department

After the triage and the waiting for the admission, the patient is visited in one of the visit rooms by a physician. The shock room is reserved to patients with urgency code 1 and certain symptoms (e.g., myocardial infarction, stroke, polytrauma, etc.) that require a particular equipment (e.g., defibrillator, surgical table, anesthesia machine, etc.). The MCA ambulatory is instead used from 8:00 to 16:00 in the weekdays for the treatment of the large fraction of patients with emergency code 4–5, who usually require few and short treatments in addition to the first visit. In all the other cases the ordinary visit rooms are used.

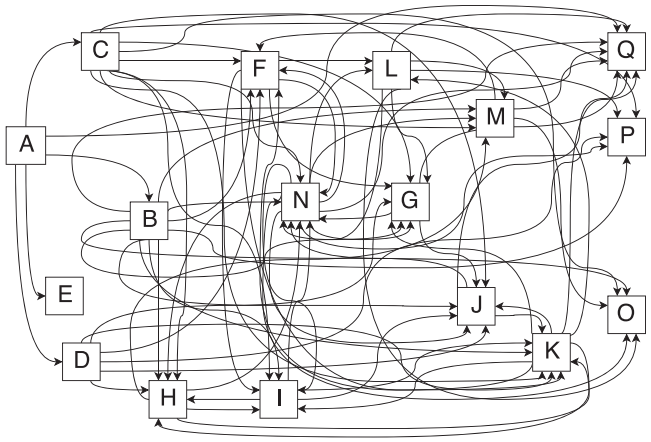

Fig. 5. Example of a spaghetti process model mined through a standard approach (Heuristic Miner). Nodes with letters refer to the activities listed in Table 1, while arcs indicate the possible sequences of activities. Such a process model is not simple and memoryless.

At the end of the medical visit, the physician can prescribe therapies, tests or observations. All therapies, the X-ray exams and the SSO are competence of the ED, while specialist visits, computer tomographies (CTs) and ecographies are executed in other wards according to a preferential track with respect to other patients. Furthermore, from a management point of view it is important to highlight two different ways of using SSO: the first have a fixed duration due to the monitoring of the patient, the second is a temporary occupation of the unit waiting for the patient to be hospitalized when there are no beds in the ward of their destination. We notice that they both require an SSO unit and the supervision of nurses and physicians.

After completing the treatments prescribed, the patient is revalued again by the same physician (or another that made the handover with the previous one), which establish if continuing with other treatments or discharging the patient, which could go home or be transferred to another ward or hospital. At any time, the patient can also decide to interrupt their path and be discharged on their own responsibility, both before the visit (LWBS) and during the treatments.

Table 1 summarizes all the activities for the treatment of a patient within the ED. The first and the second column indicate an identifier for each activity and its description, respectively. Then, we classify the activities into 5 classes called *Triage, Visit, Tests & Care, Revaluation,* and *Discharge*. In the fourth column the competence of the activity (ED or extra-ED) is indicated. An estimation of the average durations is reported in the fifth column, that is an information provided by the ED head for all the activities such that the start and/or the end timestamps are not available in the ED data-set, otherwise indicated with "◎". The last nine columns indicate the main human and equipment resources necessary for performing the activities (note that nurses and stretchers could be also occupied by patients waiting for an activity).

4. Prediction model

In this section we summarize the features of the prediction model based on the HAF introduced in [22], which is obtained developing an ad hoc process discovery approach for the same case study addressed in this paper. Nevertheless, we would like to remark that this approach can be easily generalized to other EDs.

Table 1
Activities in a patient path, average durations and resources involved ($\mathcal{P}$ = physician, $\mathcal{N}$ = nurse, $\mathcal{X}$ = X-ray technician, $\mathcal{E}$ = extra-ED human resource(s), $\mathcal{B}_V$ = visit room bed, $\mathcal{B}_O$ = SSO unit, $\mathcal{S}$ = stretcher, $\mathcal{M}_X$ = X-ray machine, $\mathcal{M}_O$ = other extra-ED machine). Symbol $\surd$ indicates the need of a resource for all patients and $\surd^*$ for patient with codes 1–2 or with walking difficulties, while # indicates resources shared simultaneously with other patients during the execution of the activity.

id	description	activityclass	comp.	avg dur. (min)	resources								
					$\mathcal{P}$	$\mathcal{N}$	$\mathcal{X}$	$\mathcal{E}$	$\mathcal{B}_V$	$\mathcal{B}_O$	$\mathcal{S}$	$\mathcal{M}_X$	$\mathcal{M}_O$
A	Triage	Triage	ED	5		$\surd$					$\surd^*$		
B	Medical Visit	Visit	ED	15	$\surd$	$\surd$			$\surd$				
C	Shock-Room	Visit	ED	15	$\surd$	$\surd$			$\surd$				
D	MCA Visit	Visit	ED	15	$\surd$	$\surd$			$\surd$				
E	Pediatric Fast-Track	Discharge	extra	1		$\surd$		$\surd$					
F	Therapy	Tests & Care	ED	2		$\surd$					$\surd^*$		
G	Laboratory Exams	Tests & Care	ED	3		$\surd$					$\surd^*$		
H	X-Ray Exams	Tests & Care	ED	3			$\surd$				$\surd^*$	$\surd$	
I	Computed Tomography	Tests & Care	extra	10		$\surd^*$		$\surd$			$\surd^*$		$\surd$
J	Ecography	Tests & Care	extra	15		$\surd^*$		$\surd$			$\surd^*$		$\surd$
K	Specialist Visit	Tests & Care	extra	15		$\surd^*$		$\surd$			$\surd^*$		
L	SSO	Tests & Care	ED	◎	#	#				$\surd$			
M	Pre-hosp. SSO	Tests & Care	ED	◎	#	#				$\surd$			
N	Revaluation Visit	Revaluation	ED	10	$\surd$	$\surd$			$\surd$				
O	Hospitalization	Discharge	ED	1	$\surd$								
P	Discharge (Ordinary)	Discharge	ED	1	$\surd$								
Q	Interruption	Discharge	-	0									

As depicted in Fig. 5, the ED process has the characteristics of a *spaghetti process* that is an unstructured process in which the huge variety of sequences of events affects the trade-off between the main quality criteria of the discovered process model, that is simplicity, precision, generality and fitness [50]. Since the lack in the process mining literature of an algorithm suitable for our purpose [21], an ad hoc process discovery algorithm has been proposed in [22]. For each group of patients with enough similar paths, that is they have a consistent probability to perform the same subsequence of activities after the first visit, the algorithm builds a HAT, which is a tree that represents activities as nodes and arcs indicate the consecutiveness between activities with weights equal to the frequency, and special leaves are (small) graphs that generalize infrequent behaviors.

From a non trivial pre-processing of the ED data-set described in Section 3, the corresponding event log has been generated. The HAF takes in input the event log and two parameters: the clustering gain ratio $g \in (0, 1)$ and the pruning threshold $\ell \in \mathbb{N}$. The former is a parameter that is used to cluster the patients in groups having similar pathways in accordance with their characteristics (main symptom, urgency code, age, etc.) and it defines how much the behavior of the patient paths of different clusters should differ. Then, for each cluster a HAT is built according to the pruning threshold parameter ℓ, which indicates the minimum level of statistical significance to represent the consecutiveness between two activities in a certain point of the pathway. Fixed a cluster of patients, such a parameter allows us to determine a tree structure with all the sub-pathway of those patients from the first visit to certain points, for which we have enough historical data to make prediction without overfitting the process model. The remaining sub-pathways (i.e., the last subsequence of activity with a number of occurrences less than ℓ) represents infrequent behaviors that could be considered for purposes such as the generation of new instances when modeling the ED process (e.g., for simulation) but that are not consistent for prediction. An example of HAT is reported in Fig. 6: square nodes indicate the most frequent activities (S is the triage plus the first visit and X is the discharge), while pentagon nodes represent a graph that generalizes infrequent path evolutions (as shown for the leaf at the end of the sequence SGGH).

The HAF can be used as a prediction model, associating to each patient the HAT corresponding to their cluster. Each time the patient undergoes an activity, they moves accordingly on that HAT, and the probability of the occurrence of a certain activity in the future can be efficiently computed using the weights on the arcs of corresponding the sub-tree. Once a pruning threshold ℓ has been set, from a statistical perspective it is reasonable to predict the future activities of patients, whose pathway is represented by square nodes through the tree, while the arrival on a pentagonal node indicates that we are in correspondence of an infrequent pathway that we can not predict precisely.

From an implementation point of view, the HAF can be converted to a table, readable in input by the simulation model presented in Section 6 and accessible by the online allocation algorithm proposed in Section 5. Each row represents a different combination of a cluster and a past activity sequence while each column represents all the possible next activities. For each pair "row and column" are reported two probability values, that is the probabilities that a patient will undergo the activity in the column (i) before the next revaluation visit, and (ii) before the end of their path in the ED. Several examples of these probabilities are shown in Table 7 of the work illustrated in [22].

5. Online resource allocation

In this section we present a general framework capable of representing different policies for the ED real-time resource allocation. Such policies are managed by a *decision maker*, which is able to consult in real time the activities to be performed and the waiting time of each patient. The decision maker operates on the list of patients waiting for the execution of an activity. This means that patients doing some activities (e.g., medical visit, laboratory and X-ray exams, specialist visit, ...) are not in the list and they will be re-inserted as soon as their current activity finishes. On the basis of all this information, every time a resource becomes available, the decision maker consults the list of waiting patients ranked in real time: the first of the list is the next to be served in accordance with a fixed policy among those that we present in this section as follows. In Section 5.1 we define a baseline policy that broadly replicates the allocation used in the real case study, expressing with quantitative criteria the practical sense of the ED staff in taking decisions. Then, in Section 5.2 we propose an online allocation framework for the real-time resource allocation, which exploits the knowledge acquired by the prediction model presented in Section 4 providing five different policies that differs for the definition of a *prioritization score*.

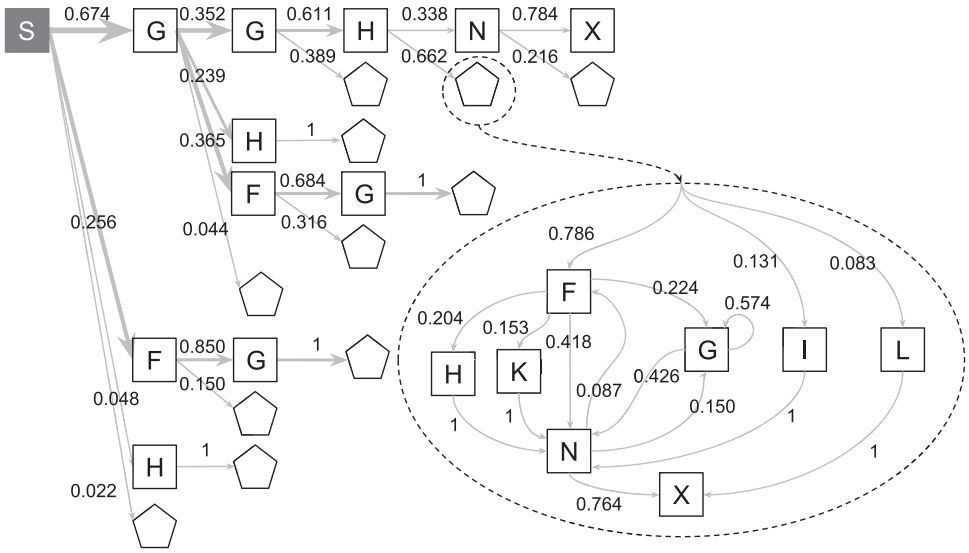

Fig. 6. HAT of the cluster of patients with dyspnea arriving on a weekday by ambulance, obtained fixing $g = 0.2$ and $\ell = 30$.

5.1. The baseline policy

After several interviews with the medical staff, we have compiled a list of criteria to lead the allocation of the ED resources. The main criterion is the urgency code c. Patients with urgency code $c = 1$ have always the priority on patients with $c > 1$. Patients with urgency code $c = 2$ and $c = 3$ have priority on patients with $c > 2$ and $c > 3$, respectively. Finally, patients with code $c = 4$ and $c = 5$ (the so called *minor codes*) have the lowest priority even if they are usually treated without any difference.

Unfortunately, the strict application of these rules would result in a possible *starvation* situation for less urgent codes in the more crowded hours of the day: in those cases, such constraints could be relaxed allowing patients with lower priority to be treated before the others on the basis of the common sense of the ED staff.

In order to replicate such a common sense as a systematic rule within the baseline policy, we assign to each patient a priority index Φ that is equal to $\min\{c, 4\}$ (e.g., the initial urgency code) at the moment of their insertion in the list of patients waiting for activities managed by the decision-maker. The decision-maker checks and updates the priority of patients in real time as follows:

- if $c \geq 4$ and the next activity belongs to the visit class, then Φ is set to 3 after t_Φ minutes elapsed from the insertion in the list;
- if $c = 3$ and the next activity belongs to the visit class, then Φ is set to 2 after $2t_\Phi$ minutes elapsed from the insertion in the list;
- if $c \leq 2$, then $\Phi = c$ all the time.

The issue of estimating the value of t_Φ is addressed in Section 7.1. The decision-maker allocates the available resources to the patients waiting for an activity in increasing order of Φ. At each step, the subset of patients S_Φ with priority Φ is selected. If two or more patients with urgency code $c \leq 2$ are in S_Φ, the resources are assigned with an extension of the *First Consultation Priority Rule*, that is to perform before the execution of (i) shock-room visits, (ii) first visits, (iii) revaluation visits, and (iv) other activities. Otherwise for $c \geq 3$, the most promoted activities are (i) revaluation visits, (ii) first or MCA visits, and (iii) other activities. If two patients in S_Φ need to undergo the same activity, the decision-maker selects the patient with the higher waiting time.

5.2. An online allocation framework

Our online allocation framework is based on a greedy algorithm capable to allocate a resources in accordance with a patient prioritization based on two levels, an upper level ensuring fairness and a lower that focuses on efficiency. This prioritization is used for all the ED activities and is continuously updated with respect to the time spent in the queue and the activities to be performed, which are revealed over time.

As well as the baseline policy reported in Section 5.1, at the first level we define a priority among patients that is based on their urgency codes c and waiting times w since their last insertion in the decision maker list $\mathcal{Q}$, that is at the completion time of the previous activity. For each patient $p \in \mathcal{Q}$, the following *waiting index* $\Psi_p \in \mathbb{N}_0$ is computed:

$$\Psi_p = \left\lfloor \frac{w}{m_c} \right\rfloor, \tag{1}$$

where m_c is a constant time fixed to normalize the waiting times with respect to the urgency, with $m_1 < m_2 < m_3 < m_4 = m_5$. At this point, the subsets S_ψ, $\psi = 0, 1, \ldots$, of patients p such that $\Psi_p = \psi$, are considered in decreasing order of ψ in order to ensure a fair allocation of resources among patients with similar waiting times with respect to their urgency in order to guarantee their safety. Then, patients within each set S_ψ are prioritized at the lower level.

At the second level, a policy Π is fixed in order to avoid bottleneck caused by a possible critical resource. We check the availability of that resource in real time and we try to optimize its utilization to lower the level of crowding. The idea is to promote the execution of those activities involving patients that will occupy critical resources through a *prioritization score* $\Sigma_\Pi \in [0, 1]$. Then, for patients p with the same *waiting index* the value of Σ_Π^p establishes their priorities promoting the activities of those with the higher values of the score. We notice that the prioritization score is used to prioritize patients waiting for all activities. Differently from the baseline policy, there is not a priority among types of activities to be respected (as in the extended *First Consultation Priority Rule*): beyond urgencies and waiting times, which are considered in both cases, the proposed online algorithm promotes the allocation of resources only among patients with different scores. The rationale behind this choice is to promote a more flexible allocation of resources, in order to deal with the critical ones. In addition, we

remark that the only resource shared between the medical visits (first and revaluations) and the other activities is the nurse: if such a resource is not a bottleneck then the patients waiting for the first visit and the other patients can be seen as in different queues, otherwise nurses can be assigned to both the type of patients, which is different from the current policy of the case study.

While for a part of patients in $\mathcal{Q}$ one or more subsequent activities after the next one are known, they are uncertain for the rest of patients. In fact, the trace of a patients is not known at the beginning of the process of care but its evolution is revealed over time: at the end of the first or the revaluation visit, we are aware of what will be the next activities of the patient until the next revaluation visit, as shown in Fig. 1. Therefore, when patients are waiting for an activity of the *Visit* or *Revaluation* class, we could estimate the probability of undergoing a certain activity after the next visit computing its frequency on the HAT of their cluster, as reported in Section 4. To this end, we use for each cluster of patients:

- a HAT, called $\mathbb{H}_{check}$, with minimum absolute frequency ℓ on the tree edges sufficiently high to have statistical relevance is used to check if probability of the evolutions of a certain path can be estimated: let $\mathcal{P}^S$ and $\mathcal{P}^F$ be respectively the set of patients for which such checking is successful or has failed;
- a HAT, called $\mathbb{H}_{comp}$, with $\ell = 1$ is used to compute frequencies of next activity for patients in $\mathcal{P}^S$;
- a function $\mathbb{P} : \mathbb{A} \to [0, 1]$ that given a certain activity $Y \in \mathbb{A}$ of a patient p gives the relative frequency $\mathbb{P}(Y)$ of the occurrence of Y, that is computed using $\mathbb{H}_{comp}$ if $p \in \mathcal{P}^S$, or it is set equal to the the frequency of the past cases of patients in $\mathcal{P}^F$ during the simulation, otherwise.

The rationale behind the different computation of $\mathbb{P}(Y)$ when $p \in \mathcal{P}^S$ or $p \in \mathcal{P}^F$ is that in the former case we have a sufficient number of cases belonging to the same cluster in historical data to estimate the further activities of such a patients based on their subsequence of already undergone activities, while in the latter case we are not able to exploit the knowledge of the prediction tool. For this reason, when $p \in \mathcal{P}^F$ we estimate the required probability as follows

$$\mathbb{P}(Y) = \frac{|\{p' \in \mathcal{P}^F_{past} : p' \text{ underwent to } Y\}|}{|\mathcal{P}^F_{past}|}$$

where $\mathcal{P}^F_{past}$ is the set of patients in $\mathcal{P}^F$ that preceded the patient p, regardless of their cluster. Let $X, Y \in \mathbb{A}$ be two activities among those reported in Table 1. We extend the function $\mathbb{P}$ as follows:

- $\mathbb{P}(X \vee Y)$ indicates the probability that either X, Y, or both occur;
- $\mathbb{P}(X < Y)$ indicates the probability that both activities X and Y occur, with X appearing at least once before the first occurrence of Y.

In accordance with the analysis of the ED staff, we identified 5 different policies Π determining 5 different *prioritization scores* Σ_Π at the lower level. Each score is based on the probability of the occurrence of a certain activity $\mathcal{A}_p$, which is 0 or 1 when the next activity is not in the *Visit* or *Revaluation* class, otherwise it is estimated using the function $\mathbb{P}$.

When $\mathcal{A}_p \notin \{B, C, D, N\}$, the five *prioritization scores* are defined as follows.

Σ_E **– *Exit score***

The rationale of this policy is to foster those patients whose probability to be discharged after the first or revaluation visit in order to reduce the EDLOS and the number of patients in the hallways and in the SSO area of the ED.

Observe that such patients slow the work of the nurses because they increase the time dedicated to the supervision task. Furthermore, the exit of patients that occupy a stretcher or a SSO unit allows the allocation of that resource to another patient.

Patients have a priority that is equal to the probability to exit after the next activity (i.e., activities B, C, D, N in Table 1).

The score of the patient p is defined as follows:

$$\Sigma^p_E = \mathbb{P}(O \vee P \vee Q). \tag{2}$$

Σ_X **– *Extra score*** Some activities (i.e., activities H, I, J, K in Table 1) can not be performed always but only during a period of time, typically from the early morning to the late evening.

The rationale of this policy is therefore to avoid a patient being forced to wait for such activities all night long. This can be obtained by promoting those patients requiring such activities before their closing time.

This can result in a significantly reduction of the EDLOS.

Conversely, if a patients needs one of these activities during the closing time, the ED can treat them without haste but without stopping their care.

The score of the patient p is defined as follows:

$$\Sigma^p_X = \begin{cases} \max\{\mathbb{P}(H \vee I \vee J \vee K), \mu\} & \text{if } 0 \le f_{end} - t \le \delta, \\ \min\{1 - \mathbb{P}(H \vee I \vee J \vee K), \mu\} & \text{if } f_{start} - t > 0, \\ \mu & \text{otherwise,} \end{cases} \tag{3}$$

where t is the instant in which the decision maker is allocating the resources, f_{start} and f_{end} are the open and closing time of the extra-ED activities, δ is a parameter set to indicate how much time before the closing time we want to promote the activities of patients needing activities H, I, J and K, and $\mu \in (0, 1)$ is a parameter set to fix a default score for patients for which the estimated probability is not sufficiently high/low to establish a discriminant in access to the resources. We remark that when the extra-ED resources are accessible and the time left until closing is greater then δ, all patients have the same default score μ, then the prioritization is given only by the upper level (i.e., Ψ_p).

Σ_O **– *Observation score*** SSO units (beds) are a critical resource due to their limited amount; further, when such resources are all busy, patients have to be observed on stretchers in the hallways or visit rooms, slowing down the other ED activities.

Furthermore, activities L (ordinary SSO) and M (pre-hospitalization SSO) are different: the former is a period of time on which the patients need to be observed by the medical staff, the latter is just a temporary accommodation waiting for a bed in a ward of the hospital.

For this reason, the ending of the activity L depends on the starting time, while the ending of the activity M depends on exogenous factors.

The rationale of this policy is to manage patients that have to undergo the activity L, trying to start it as soon as possible when the number u of free SSO units is over a certain threshold u^+, and to promote the activities of other patients when u is under a critical threshold u^-.

The resulting score is defined as follows:

$$\Sigma^p_O = \begin{cases} \mathbb{P}(L) & \text{if } u \ge u^+, \\ 1 - \mathbb{P}(L) & \text{if } u \le u^-, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Σ_S **– *Stretcher score*** The rationale of this policy is to promote the release of stretchers by patients that need to undergo a generic activity X having an higher probability to be discharged or to occupy a SSO unit after that. The score is defined as follows:

$$\Sigma^p_S = \begin{cases} \mathbb{P}(L \vee M \vee O \vee P \vee Q) & \text{if } p \in \mathcal{P}_{str}, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\mathcal{P}_{str}$ is the set of patients that occupy a stretcher.

Σ_H – ***Hospitalization score*** We are considering patients needing a SSO unit waiting for their hospitalization because beds are not available in the destination ward.

Such patients could be treated without haste but without stopping because of they have a usually long EDLOS, most of which is caused by the impossibility to be hospitalized at the end of their paths (independently on the ED efficiency). Once the physician has established and booked the hospitalization, usually a series of activities are performed in preparation for treatment in the destination ward. Apparently, there is any reason to perform these activities in priority over other patients (except for the urgency already addressed at the higher prioritization level), conversely such a choice could worsen the EDLOS of other patients.

The rationale of this policy is to promote those patients having lower probability of needing a pre-hospitalization in the SSO units.

The score is defined as follows:

$$\Sigma_H^p = \begin{cases} 1 - \mathbb{P}(M < N) & \text{if } p \notin \mathcal{P}_{str}, \\ 1 & \text{otherwise.} \end{cases} \tag{6}$$

We exclude patients that are occupying a stretcher because they are a critical resource.

Conversely, after the (first or revaluation) visit the activities of a patient until the next revaluation visit are known. In these cases, we are able to assign more appropriate score values based on the known activities that a patient has to undergo until the next revaluation visit. Let Λ be the set of the next known activities, then the corresponding scores are trivially computed as follows:

$$\Sigma_E^p = 0, \tag{7}$$

$$\Sigma_X^p = \begin{cases} 1 & \text{if } 0 \le f_{end} - t \le \delta \text{ and } \Lambda \cap \{H, I, J, K\} \neq \emptyset, \\ 0 & \text{if } f_{start} - t > 0 \text{ and } \Lambda \cap \{H, I, J, K\} \neq \emptyset, \\ \mu & \text{otherwise,} \end{cases} \tag{8}$$

$$\Sigma_O^p = \begin{cases} 1 & \text{if } u \ge u^+ \text{ and } L \in \Lambda, \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

$$\Sigma_S^p = \begin{cases} 1 & \text{if } p \in \mathcal{P}_{str} \text{ and } \Lambda \cap \{L, M, O, P, Q\} \neq \emptyset, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

$$\Sigma_H^p = \begin{cases} 0 & \text{if } p \notin \mathcal{P}_{str} \text{ and } M \notin \Lambda, \\ 1 & \text{otherwise.} \end{cases} \tag{11}$$

Fixed a policy and the parameters of its score (that is a choice that could be taken after performing a quantitative analysis as that presented in Section 7.2), a summary of our online approach is formally described in Algorithm 1. This procedure is invoked every time a resource is released or a new patient is added in the waiting list $\mathcal{Q}$. The available resources at that moment are defined through the multiset (R, m), where $m : R \to \mathbb{N}_0$ indicates for each resource set R, that is those listed in Table 1, the amount available. In the first part of the algorithm (lines 3–7) all the *waiting indices* Ψ_p and the *prioritization scores* are updated. In the second part (lines 8–24) the algorithm scans the subsets S_ψ in decreasing order of ψ, that is from that with the higher normalized waiting times to that with the lower one. Then, within each S_ψ, patients are scanned in decreasing order of Σ_Π^p. For each patient p', we define the multiset (R, m') (line 13), where $m' : R \to \{0, 1\}$ indicates the resources necessary for the execution of the next activity $\mathcal{A}_{p'}$. Furthermore, if the next activity is a first visit (line 14) one of the available physician $d_{p'}$ is selected randomly and is candidate to be assigned to p' (if $\mathcal{D} = \emptyset$ then $d_{p'}$ is a dummy physician that will fail the allocation check at line 18), otherwise if the next activity is a revaluation visit (line 16) $d_{p'}$ is defined as the physician previously assigned during the first visit (or a substitute one if the work shift is changed). The allocation is made if all available resources are sufficient and, in the case of a revaluation visit, if the assigned physician is available (lines 18–20). Finally the list and the available resources are updated (lines 21–23).

Algorithm 1: Online resource allocation.

Input: Policy Π; set of all waiting patients $\mathcal{Q}$; multiset of all available resources (R, m); set of all available physicians $\mathcal{D}$.

1 **begin**
2 **foreach** $p \in \mathcal{Q}$ **do**
3 compute Ψ_p as in Equation (1);
4 **if** $\mathcal{A}_p \in \{B, C, D, N\}$ **then** compute Σ_Π^p as in Equations (2)-(6);
5 **else** compute Σ_Π^p as in Equations (7)-(11);
6 **end**
7 $\psi_{\max} = \max_{p \in \mathcal{Q}} \Psi_p$;
8 **foreach** $\psi = \psi_{\max}, \psi_{\max} - 1, \ldots, 0$ **do**
9 $\mathcal{S}_\psi = \{p \in \mathcal{Q} : \Psi_p = \psi\}$; $K \leftarrow |\mathcal{S}_\psi|$;
10 **foreach** $k = 1, \ldots, K$ **do**
11 $p' \leftarrow p$ with the kth highest value of Σ_Π^p in $\mathcal{S}_\psi$;
12 $(R, m') \leftarrow$ multiset of resources required by p';
13 **if** $\mathcal{A}_{p'} \in \{B, C, D\}$ **then** $d_{p'} \leftarrow randsel(\mathcal{D})$;
14 **else**
15 **if** $\mathcal{A}_{p'} = N$ **then** $d_{p'} \leftarrow$ physician previously assigned to p for the first visit;
16 **end**
17 **if** $m' \le m$ *and* $(\mathcal{A}_{p'} \notin \{B, C, D, N\}$ *or* $d_{p'} \in \mathcal{D})$ **then**
18 **if** $\mathcal{A}_{p'} \in \{B, C, D\}$ **then**
19 assign $d_{p'}$ to p';
20 $\mathcal{Q} \leftarrow \mathcal{Q} \setminus \{p'\}$; $m \leftarrow m - m'$;
21 allocate to p' the required resources (eventually including $d_{p'}$);
22 **if** $\mathcal{A}_{p'} \in \{B, C, D, N\}$ **then** $\mathcal{D} \leftarrow \mathcal{D} \setminus \{d_{p'}\}$;
23 **end**
24 **end**
25 **end**
26 **end**

For the sake of simplicity, in Algorithm 1 we consider all physicians and nurses eligible for all activities in which these types of resources are required. Actually, the allocation algorithm allocates only physicians and nurses that have the competence of a certain activity (e.g., only triage nurses can perform A and only the physician in the MCA can perform D), as described in Section 6. Similarly, beds are distinguished and allocated in accordance with the type of visit room in which the visit is performed.

6. The simulation model

In order to evaluate the impact of our approach, we need a tool that is able to replicate the possible (uncertain) behaviors of the ED under different situations, and to embed the policies presented in Section 5 as subroutines. In [25], the authors propose a framework based on DES to address the issue of modeling a complex real world system in which several stochastic processes are involved. In such an environment, online algorithms can be tested on a large number of different instances of the problem in order to provide a quantitative analysis based on the average case, which

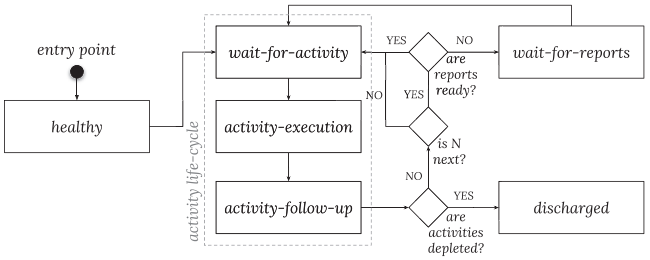

Fig. 7. Statechart of the agent *Patient*.

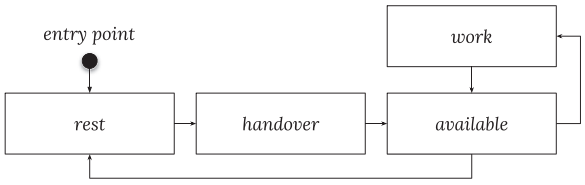

Fig. 8. Statechart of the agent *Physician*.

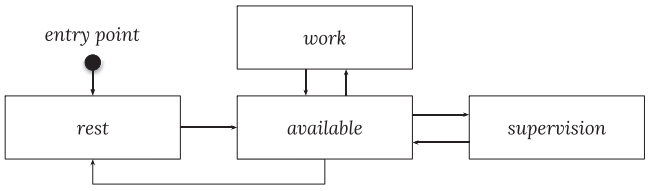

Fig. 9. Statechart of the agent *Nurse*.

can be used to estimate a set of performance indices for the decision support [51].

Although we model the ED process using a DES approach, we use an ABS semantics because it allows the tracking of the behavior of each agent acting in the simulated environment [52], and a set of rules describes the agent behavior and its interaction [53]. Therefore, we illustrate the proposed model through the description of each agent. The rules modeling the agent behavior are represented by a statechart. The first agent type is the patient, whose statechart reproduces the general pathway structure, in accordance with the diagram in Fig. 1. Then, three types of agent describes the human resources of the ED, whose statecharts represent the work shift and the execution of their tasks. Finally, the fifth agent implements the decision maker and its possible policies described in Section 5.

Patient. The patient population is reproduced from the event log: an agent is created for each access to the ED from the data-set and relevant information for the replication of its path (i.e., urgency code c, trace, arrival time, and several activity durations) are collected as agent attributes. Although in some particular cases the patient's urgency code could change during their stay at the ED, for the lack of simplicity we always consider the urgency code assigned during the triage. Each agent progresses in their path within the ED in accordance with its trace, following the statechart shown in Fig. 7.

The agent is on the *healthy* state until the arrival time. Then it moves on the *wait-for-activity* state – the first of the 3 states representing the general life-cycle of each activity – sending a message to the *decision-maker* and waiting for its reply indicating the allocation of the needed resources. The second state is the *activity-execution*, which has duration defined by the estimations provided by the ED staff or mined by the ED data-set, and depending on the activity type as reported in Table 1. After a timeout, the agent passes on the *activity-follow-up* state, which represents a period of inactivity after the visit or a therapy of urgent patients (duration is set to 0 in the other cases).

The follow-up duration depends on the activity X and on the urgency code, then it is implemented through a triangular distribution of minimum 0, modal value ζ_c^X and maximum $2\zeta_c^X$. The value of ζ_c^X is mined from the data-set except for particular cases in which it can be assessed with a good approximation (e.g., patients that leave the ED after that activity) and computing the average value for each urgency code c.

At the end of an activity life-cycle, the agent can be in three different situations. The first is the depletion of activities of the trace, then it passes on the *discharged* state. The second situation happens when the patient needs a revaluation visit but all the medical reports concerning their previous treatment are not available (e.g., the X-ray technician does not send the report): in such a case the patient lies in the *wait-for-report* state until all the reports are available. The last situation includes all the other cases, for which the patient passes on the *wait-for-activity* state and is ready to undergo the next activity of the trace.

While the actual sequence of activities of a patient is determined by the trace in the event log and is simulated as described above, the timing is a consequence of the allocation policy. For this purpose, several information is continuously updated to keep track of (i) the activities already executed, (ii) the current waiting time for the next activities, (iii) the next known activities (i.e., all those until the next first/revaluation visit), and (iv) a pointer to the HAT used to predict the activities after the next first/revaluation visit.

Physician. Each physician is represented by an agent with an attribute that indicates its competence (visit rooms, MCA, or SSO) and with a schedule defining its working shift. First and revaluation visits are performed by the agents with competence visit rooms or MCA, depending on the type of visits that it executes, while the SSO competence indicates the supervision of the patients that occupies the SSO units. The agent passes between the *rest* and *available* states in accordance with a schedule which define the start and the end of the working shift. When the agent is available and receives a message by the *decision-maker* indicating a task and its duration, it goes on the *work* state, on which it stay until the expiration of a timeout. Furthermore, at the beginning of the shift, the *handover* state models a certain time λ_{beg} of inactivity due to the receipt of medical records. Conversely, λ_{end} minutes before the shift end, the agent can be assigned only to urgent patients with $c \leq 2$, or taken over previously, as commonly happens in reality. The statechart is reported in Fig. 8.

Nurse. The agent is implemented as well as the physician. An attributed indicates the competence (triage, SSO, MCA or general). Triage and SSO competences indicate that such agents could execute only the tasks of triage and supervision of the SSO units, respectively. The MCA and general competence includes several tasks, such as first and revaluation visits (supporting the physician), therapies, test collection for examination, and assistance in other exams or specialist visits for patients with walking difficulties. The MCA and general competence are mutually exclusive, that is the nurse take care of the patients doing the first visit in the MCA or in the general visit room. The moving time of the nurse from the execution place of an activity to another one is considered by adding a time d_{move} to each nurse task. Furthermore, agents have an additional task on the supervision of patients waiting in the triage waiting room and corridor, which is executed each τ minutes and have duration equals to $n_p^{\mathcal{A}} d_{sup} / n_{nur}^{\mathcal{A}}$, where d_{sup} is the average duration for assisting a patient during the supervision task, while $n_p^{\mathcal{A}}$ and $n_{nur}^{\mathcal{A}}$ are the number of patients and nurses in the considered area of competence $\mathcal{A}$, that is waiting room, ED hallways or SSO units. The statechart is shown in Fig. 9.

X-ray technician. The agent is implemented similarly to the other medical agents, but having only competence on the X-ray

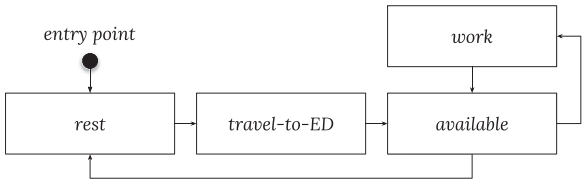

Fig. 10. Statechart of the agent *X-ray technician*.

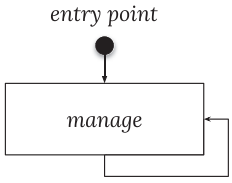

Fig. 11. Statechart of the agent *Decision-maker*.

scan. Since at nighttime no technician is working in the ED, we model the *on demand* technician availability for patients with code $c = 1$ by adding a *travel-to-ED* state representing the travel of 20 min reaching the ED. The statechart is illustrated in Fig. 10.

Decision-maker. When a patient needs to undergo an activity, the agent is informed by a message and such a request is inserted in a prioritized queue recording the patient ID, the request timestamp, the set of resources needed, the urgency code c and a score.

The agent scans the queue in real time and chooses the patients to whom to assign the resources available at that moment. Such decisions are taken in accordance with the online resource allocation policies reported in Section 5, that is the *baseline policy* or Algorithm 1 with a fixed policy Π. Then, the agent updates the set of the free resources and send a message to the agents representing the patient and the human resources involved in the activity.

The statechart of the agent is composed of a single state, and a transition fired at each unit of time in order to update information and to take decisions, as shown in Fig. 11.

The ABS semantic allows us to model the continuity of the care process because it is possible to identify individual resources (i.e., single physician and nurses) simulating their interactions: the same physician (or the one receiving the medical records during the handover) is always assigned to a patient for the activities that follow its first medical visit, that is revaluation visits and discharge.

The simulation model is implemented in such a way to be sensitive to overcrowding. At the increasing of the number of patients waiting in some areas, the nurses with competence on such areas are busier in the supervision task, then less time is dedicated to the other activities and this feeds even more the level of crowding. Furthermore, the occupation of certain resources restricts the use of others related to them (e.g., physicians need nurses to perform visits) and increases the occupation of other ones (e.g., stretchers or SSO units are not released until a physician is not assigned for the last revaluation visit before the discharge). Another important aspect represented by the model is the behavior of the human resources during the beginning and the ending of their shift, which are critical periods that cause a bottleneck in the patient flow.

7. Quantitative analysis

In this section we perform a quantitative analysis to study the impact of the online allocation approaches presented in Section 5 on several performance indices taking into account the perspectives of patients with different urgency. To this purpose we replicate the arrival process and all the characteristics of the patients during the whole year 2016, and recorded by the ED of Ospedale Sant'Antonio Abate of Cantù described in Section 3. The time horizon of the simulation is one year, of which the first 7 days are used as warming up period, while statistics are collected on the remaining 359 days (2016 is a leap day).

The proposed simulation model has been implemented using *AnyLogic 7.2*. All the results reported in this section are the average values over 30 simulation runs starting from different initial conditions randomly generated. The average time required for a single simulation run over the whole time horizon is 17.4 s.

7.1. Validation

The activity durations are replicated from data when they are available (i.e., for first visits of urgent patients, specialist visits, and SSO), otherwise we use the average durations suggested by the ED staff. We introduced further time parameters, as summarized in Table 2. Their values result from tuning of the simulation model in such a way to obtain the maximum fitness of the DTDT and the EDLOS with respect to the real data, that is for the values reported in the second column of the table.

The real case statistics and the results of the validation test are compared in Table 3. While the DTDTs of the simulation model are very close to the real case, EDLOS present more significant deviations for non-urgent patients. Such differences are due to the huge complexity of the ED system and the very high level of detail in modeling the patient behavior, which is not common in the literature [19]. Furthermore, they are justified by a large lack of information and noise in the ED data-set, especially for some non-urgent patients (e.g., patients whose EDLOS includes also the night in the ED waiting for a treatment).

We would remark that the main purpose of our analysis is to evaluate the sensitiveness of the model to online allocation approaches. Conversely, we are not interested in the classical what-if analysis for which a better fitness could be required. For these reasons, we can state that our simulation model is enough representative of the ED with respect to the purposes of the quantitative analysis.

Table 2
Parameters ranged during the model validation.

Param.	Value (min)	Definition
λ_{beg}	15	physician handover duration
λ_{end}	30	time at the end of physician shift with no new patients assigned
d_{move}	1	nurse moving time
d_{sup}	1	duration of the nurse supervision task per patient
τ	30	time period of the supervision task
t_{Φ}	35	time for the re-prioritization (see Section 5.1)

7.2. Results

In this section, we report the results of the quantitative analysis performed in order to evaluate the sensitiveness of the model to the online allocation approaches proposed in Section 5. The setting of the model validated in the previous section is the baseline policy to which our online approaches are compared.

The parameter for the normalized waiting time defined in Eq. (1) are the following:

$$m_1 = 1 \text{ min}, \ m_2 = 10 \text{ min}, \ m_3 = 120 \text{ min}, \ m_4 = m_5 = 180 \text{ min}. \quad (12)$$

For instance, it means that 1 min of waiting time for a patient with $c = 2$ are equivalent to 12 min and 18 min of waiting time for patients with $c = 3$ and $c = 4$, respectively.

The prediction exploited by the proposed online algorithm through the five different scores are obtained from the 2013–2015

Table 3
Model validation (patients with urgency code $c = 1$ are visited without waiting).

	DTDTs (min)					**EDLOSs** (h)				
urgency code	1	2	3	4 – 5	overall	1	2	3	4 – 5	overall
real case	-	21.6	86.7	81.5	70.1	14.9	7.7	5.0	2.8	5.2
model (baseline)	-	18.6	82.1	79.9	66.7	14.5	8.0	6.3	3.7	6.2

Table 4
Comparing online policies: DTDTs and EDLOSs (first column reports the score parameters).

	DTDTs (min)					**EDLOSs** (h)				
urgency code	1	2	3	4 – 5	overall	1	2	3	4 – 5	overall
baseline	-	18.6	82.1	79.9	66.7	14.5	8.0	6.3	3.7	6.2
Σ_E	-	11.2	41.4	54.8	37.5	15.6	9.3	4.7	2.8	5.4
Σ_X ($\delta = 120$ min, $\mu = \frac{1}{2}$)	-	11.2	43.2	59.7	39.6	15.8	9.5	3.9	2.0	4.9
Σ_O ($u^- = 3$, $u^+ = 7$)	-	12.1	54.0	70.7	48.1	15.6	9.3	4.1	2.1	4.9
Σ_S	-	11.9	51.3	69.1	46.3	15.7	9.3	4.0	2.2	4.9
Σ_H	-	10.9	40.0	58.1	37.4	15.8	9.6	4.0	2.0	4.9

data-set. We use this data-set and a HAF generated a priori as explained in Section 4. During the simulation, such a HAF is used to estimate the scores defined in Section 5.2, in accordance with the patient cluster that is identified at the triage. The setting of the ad hoc process mining approach (described in Section 4) is $g = 0.2$ and $\ell = 30$, which generates a HAF with 18 clusters and 18 different HATs.

DTDT and EDLOS are the indices evaluated to analyze the ED performance of the proposed online approaches. We also analyze the DTDTs in 6 different 4-hours frames of the day (frame $F = 1$ is 0:00–4:00, frame $F = 2$ is 4:00–8:00, etc) because of the high variability of demand throughout the day, as depicted in Fig. 4. We also identify a set of 46 *overcrowding days*, which are the days having a maximum length of the admission queue (i.e., patients waiting for the first visit) longer than 10 patients in the baseline.

The results reported in Table 4 prove the effectiveness of the proposed online allocation algorithm regardless of the policy and the score adopted. DTDTs are significantly reduced for each urgency class, especially for the most numerous class $c = 3$ saving the 51% of the time. This result supports the rationale behind the idea of prioritizing patients in a unique queue discussed in Section 5.2: although the activities of the *Visit* and the *Revaluation* classes have the precedence on those of the *Test & Care class* in the baseline policy, the DTDTs are decreased by the online algorithm that relaxes this greedy rule in exchange for a more flexible resource allocation that results advantageous over time.

The impact on the EDLOS provide a trade-off: urgent patients ($c = 1, 2$) stay in the ED from 60 to 90 min more compared to the baseline configuration, while the EDLOS of the other patients decreases from 55 to 145 min. The explanation behind the increase of the time spent by urgent patients before the discharge could be in the differences between the prioritization rule defined by the priority index Φ in the baseline and the first level of prioritization of Algorithm 1 defined in Eq. (1), which depend on the implicit choices of t_Φ by the ED staff and the setting of $m_1, \ldots, m_5$ in the proposed approach, respectively. In particular, the second choice seems to have a propensity to promote the execution of the activities (except for first visits) of non-urgent patients with respect to the common sense of the ED staff in the real case. However, the DTDT is the most important index when dealing with life-threatening patients, that is approximately 0 for $c = 1$ (i.e., patients with this code are visited and have an immediate first treatment as soon as they arrive at the ED as well as in the simulation model), while for $c = 2$ the DTDT decreases from 35% to 41% compared to the baseline. Further parameter variation experiments (e.g., by varying the values of m_c, $c = 1, \ldots, 5$) can provide a different balance of average DTDTs and EDLOSs among the urgency classes to meet any time limits established by (national or regional) guidelines or any preference established by the decision maker, as discussed in Section 7.3.

On average, both DTDTs and EDLOS have a significant reduction up to 44% and 21% respectively, that is maximized using the *hospitalization score* Σ_H. This result leads us to think that hospitalizations, which depend on the availability of beds by the other hospital wards, represent the most important bottleneck. Therefore an optimized management of SSO based on the concept of lookahead offers the largest margin of improvement of the performance of the case study.

In Table 5 we compare the results of the baseline with our online allocation algorithm using the five proposed scores, focusing on DTDTs of patients arriving in the frame $F = 3$ (8:00-12:00), which is that with the high number of accesses. In the first part of Table 5, average DTDTs over all days are reported showing an even stronger impact of the proposed approach in peak hours: DTDTs of patients with $c \leq 3$ are reduced up to the 71% using the *exit score* Σ_E. In the second part of Table 5, we focus only on the *overcrowding days*. The proposed solutions seem to be very effective to alleviate the overcrowding, preserving its effect compared to other days. The *extra score* Σ_X has the best performance on average in the most crowded days. This result suggests to promote the execution of activities that surely or probably precede specialist visits or exams with a closing hour when the level of crowding is high. Consequently, the policy defined by Σ_X increases the number of patients that undergo these activities before closure, avoiding spending the night in the ED.

The variation of the average DTDT during the six frames of the day is illustrated in Fig. 12. The impact of the online allocation algorithm is evident in almost all the frames, especially in the peak hours (frames $F = 3$ and $F = 4$), and in the overcrowding days. Scores Σ_X and Σ_H give similar DTDTs. The former is slightly better in the central hours of the day, that is when the patients needing an extra-ED activity and have a higher priority. A side-effect can be observed after the closing time of those activities, when the accumulated work to promote those patients causes an increasing of the DTDTs and better performance can be obtained using Σ_H. In general, the online approach seems to be really effective when dealing with the peaks of demand, and flattening the average DTDT throughout the day.

In order to evaluate the impact of the proposed online approaches on the queues and on the resources, we introduce an additional set of performance indices in Table 6. The impact of the different scores on such indices can demonstrate if the prediction

Table 5
DTDTs (min) in the frame from 8:00 to 12:00 (first column reportsthe score parameters).

	All days				Overcrowding days			
urgency code	2	3	4 – 5	overall	2	3	4 – 5	overall
baseline	25.2	95.9	101.4	83.7	41.8	205.2	158.5	156.0
Σ_E	8.8	26.9	61.6	34.6	14.7	69.3	144.7	83.3
Σ_X ($\delta = 120$ min, $\mu = \frac{1}{2}$)	9.3	28.3	64.5	36.3	12.0	57.6	135.9	74.4
Σ_O ($u^- = 3$, $u^+ = 7$)	10.3	38.0	72.2	43.7	13.4	72.2	140.1	82.8
Σ_S	10.2	35.0	70.2	41.5	13.3	70.5	140.6	82.2
Σ_H	9.4	28.4	64.5	36.4	11.9	57.9	136.9	74.8

Table 6
Performance indices.

Indices	Definition
q	average number of patients/min in the pre-admission queue
r	average number of patients/min under treatment (post-admission)
u_{str}	stretcher utilization
γ	number of *overcrowding days*

Table 7
Impact on queues and resource utilization (first column reports the score parameters).

	All days			Overcr. days, F=3			
	q	r	u_{str}	q	r	u_{str}	γ
baseline	2.8	15.9	53.2%	5.9	21.0	63.8%	46
Σ_E	1.6	13.8	54.0%	3.4	18.1	59.6%	27
Σ_X ($\delta = 120$ min, $\mu = \frac{1}{2}$)	1.7	12.4	51.6%	3.1	15.8	56.2%	28
Σ_O ($u^- = 3$, $u^+ = 7$)	2.0	12.6	51.0%	3.8	16.3	55.8%	30
Σ_S	1.9	12.5	50.3%	3.8	16.0	54.0%	31
Σ_H	1.6	12.5	52.2%	3.1	15.9	56.4%	28

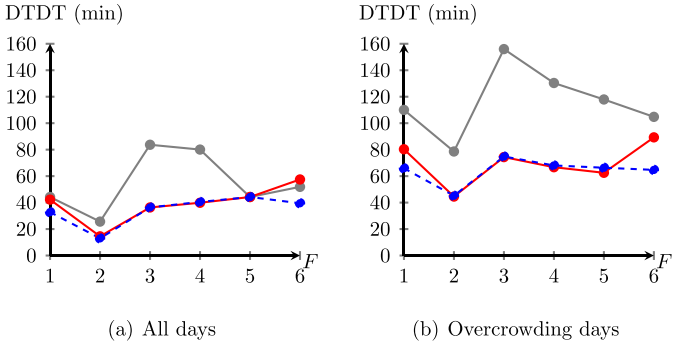

Fig. 12. Average DTDTs of patients in the frames $F = 1, \ldots, 6$ of the day (baseline in grey, Σ_X in red, Σ_H in blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

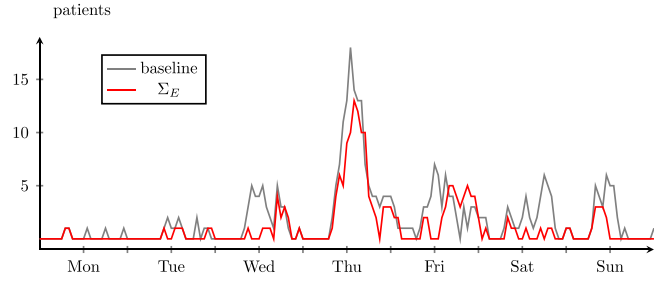

Fig. 13. Length of the pre-admission waiting list in the ED during the week 1–7 April.

performed using the HAF is effective in predicting the occupation of the resources that are a bottleneck for the patient flow. Furthermore, the indices q and r provide information that could represent the perception of crowding by the ED staff, indicating how many patients on average are in the pre-admission waiting room (q) and in the rooms and the corridors dedicated to treatment (r).

In Table 7 we compare the different scores on the indices introduced in Table 6 with the baseline, reporting the average values over all the year in the first part, and a focus on the frame $F = 3$ in the more crowded days in the second part. As expected, the *exit score* Σ_E is able to reduce the queue in the pre-admission waiting room, but a counter-intuitive aspect can be observed in correspondence of the overcrowding days, where the *extra score* Σ_X and the *hospital score* Σ_H have lower values of q. Such indices decrease also the average number of patients in the rest of the ED, reducing the sense of crowding. Finally, the stretcher utilization is minimized by the *stretcher score* Σ_S, whose impact is equivalent to have an extra stretcher when the ED is overcrowded. Finally, in the last column (γ) it is shown that the use of the proposed online allocation algorithm can reduce the days of overcrowding.

Figs. 13 and 14 report respectively the trend of the number of patients in the waiting list queue and concurrently under treatment in the ED during a week. The week has been selected in such a way that the baseline has one overcrowding day (Thursday). Except for a limited period of time, the *exit score* Σ_E dominates the baseline showing the capability of reducing the level of crowding in the ED. When the *exit score* Σ_E performs slightly worse than the baseline in one of the two figures, we would remark that the sum of the number of patients in pre-admission queue and concurrently under treatment is less than the same value for the baseline.

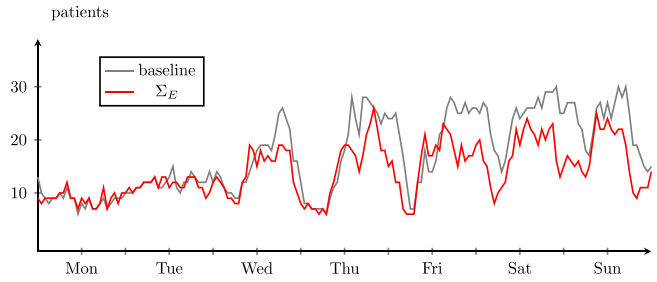

Fig. 14. Number of patients concurrently under treatment (between admission and discharge) in the ED during the week 1–7 April.

7.3. General insights and practical considerations

Results reported in the quantitative analysis offer various general insights, which go beyond the case study. First, we can state that if the decisions of allocating resources for the execution of activities take into consideration the probably subsequent activities and the necessary resources, then there is a significant room for improvement for the DTDT, the EDLOS, and the resource utilization.

Due to the high number of significant differences between the considered ED environments and a lack of publicly available data, a fair comparison between our approach and those in the literature is not possible. However, we are able to provide general insights that are in common with several prior works. According to Cildoz et al. [39] the *First Consultation Priority Rule*, included in our baseline (as described in Section 5.1) and commonly used in practice, is far from being the best solution. Although different KPIs has been defined, the common suggestion is that this practical rule should be replaced by policies that take into account the characteristics of patients in the queue and the probability of their discharge after the visit, in order to improve the patient flow. This fact further motivates the need of a decision support system. Another interesting analogy can be observed with [43] regarding the performance improvement when giving a higher priority to patients who are estimated to occupy fewer resources (i.e., only physicians in [43] vs. all the mentioned ones in this paper). The authors find a reduction of the overall DTDT and EDLOS up to 58% and 8%, respectively, compared to the *First Come First Served* policy. In addition this known insight, our analysis shows that if we consider the probability of further events in addition to the discharge (i.e., need of other activities) and take into account also a prioritization for patients waiting for other resources (e.g., X-ray machine), then the use of other resources can be improved and both DTDTs and EDLOSs can be further decreased.

Results in Section 7.2 indicate that the hospitalization score provides the best performances on ordinary days, but the overall average DTDT can be further improved using the extra-score in the *overcrowding day*. However, this comparison should not be considered as a general insight, since the characteristics of the ED environment and the prediction tool could lead to the preference of different policies. To determine the best policy to be adopted, a quantitative analysis like that proposed above should be performed. After identifying the resources that could cause crowding in the specific ED, new policies or a combination of those proposed in this paper could be further investigated. Furthermore, an accurate choice of the values of parameters m_c used to normalize the waiting time of patients with different urgency codes is strongly recommended to guarantee an adequate level of safety and fairness, in compliance with medical guidelines and local regulations. For instance, (national or regional) guidelines could indicate a time limit for the DTDT, the EDLOS, or the time spent in a SSO unit. If the performance indices obtained from the adoption of a certain policy do not satisfy one or more of these requirements, parameters m_c should be tuned to obtain a different trade-off between different emergency codes, or between different phases of the path, that is by setting values of m_c that change according to the activity for which the patient is waiting. In our case study, the ED staff has given greater importance to the DTDT (decreased by the online allocation algorithm) of patients with urgency code $c = 2$ than their EDLOS (increased by the online allocation algorithm). In fact, the conditions of these patients are usually stabilized during the first medical visit (otherwise their code is changed to $c = 1$) and after the visit they are constantly monitored by a nurse, which assist them even during treatments that do not require the same for non-urgent patients (e.g., they are accompanied to the ward where they have to undergo the specialist visit) as indicated in Table 1.

From an implementation point of view, at the ED of Cantù, as well as in most of EDs, the real-time resource allocation task could be supported by a software tool that reports for each patient the timestamps of the previous events (e.g., arrival time, visit completion time, etc.) and the activity in which they is currently undergoing or waiting. Such information would be integrated within a decision support tool integrating the HAF and Algorithm 1, with the aim of automating the procedure and suggesting to which patient to allocate the available resources. The final decision on every particular case would however be left to the decision maker.

8. Conclusions

The aim of this work is to investigate if an online allocation approach combined with a prediction tool can improve the ED performance, alleviating the inherent phenomenon of the overcrowding. We proposed an online allocation framework for the ED resource allocation, which takes into account the real-time state of the ED and the prediction provided by an ad hoc process mining model. We analyzed the impact of an online allocation algorithm based on such a framework, embedding them within a fine-grained simulation model that reproduces the operative context and the patient flow of an Italian ED.

Using simple policies that exploit prediction provided by the ad hoc process mining model, we are able to reduce significantly the duration of the process of care, and to have a less crowded environment in which the medical staff can work better also from a qualitative point of view. Since the proposed scores are designed to have an impact on the queues or on the resources that we want to optimize, the quantitative analysis proves the effectiveness of a prediction tool in combinations an online allocation algorithm based on prioritization for the real-time ED resource allocation in order to improve the patient flow and alleviating the ED overcrowding. The policies acting on bottlenecks caused by hospitalizations and activities performed in other wards (i.e., those defined by the hospitalization score and the extra score) result the most effective for the case study.

A general insight provided by this study is that if the decisions of allocating resources for the execution of activities take into consideration the probably subsequent activities and the necessary resources, then there is a significant room for improvement for the DTDT, the EDLOS and the resource utilization.

Possible extensions of this work are twofold, that is in terms of further scenario analysis and methodological. One of the main features of the process discovery approach proposed in [22] is the ability of representing the path of a certain group of patients. Exploiting such a feature, the simulation model proposed in this paper is suitable to perform several other scenario analysis, such as those regarding the impact of self-referred patients that does not need emergency care but they access the ED as a faster alternative to primary care. Such analysis would allow us to have a quantitative estimate of how much the wrong organization of the primary

care system can penalize the work in the emergency department. Furthermore, the lack of data has led us to make assumptions on the durations of several activities, with some of them being supposed to be deterministic: a quantitative analysis based on different probability distributions would allow to investigate the impact in the stochasticity of durations on the dynamic resource allocation and the study of online allocation approaches supported by prediction models based also on durations. From a methodological perspective, the online allocation algorithm could be generalized by defining different values of the standardization parameters m_c depending on the next activity to be executed with the aim of reducing waiting times before most important activities. In addition, it would be important to study how to combine different policies and their prioritization scores in such a way to obtain additional improvements.

Data availability

The authors do not have permission to share data.

CRediT authorship contribution statement

Davide Duma: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Roberto Aringhieri:** Methodology, Supervision, Data curation, Writing – original draft.

Acknowledgments

The authors wish to thank the anonymous reviewers for their valuable comments which significantly improved the quality of the article.

References

[1] Schuur J, Venkatesh A. The growing role of emergency departments in hospital admissions. N Engl J Med 2012;367(5):391–3.
[2] ISTAT, Futuro della popolazione: meno residenti, più anziani, famiglie più piccole, 2019. https://demo.istat.it/.
[3] Paul S, Reddy M, Deflitch C. A systematic review of simulation studies investigating emergency department overcrowding. Simulation 2010;86(8–9):559–71.
[4] Di Laura D, D'Angiolella L, Mantovani L, Squassabia G, Clemente F, Santalucia I, Improta G, Triassi M. Efficiency measures of emergency departments: an Italian systematic literature review. BMJ Open Qual 2021;10(3).
[5] Hwang U, Concato J. Care in the emergency department: how crowded is overcrowded? Acad Emerg Med 2004;11(10):1097–101.
[6] Filippatos G, Evridiki K. The effect of emergency department crowding on patient outcomes. Health Sci J 2015;9(1):1–6.
[7] Hong KJ, Shin SD, Song KJ, Cha WC, Cho JS. Association between ED crowding and delay in resuscitation effort. Am J Emerg Med 2013;31(3):509–15.
[8] Claret P-G, Bobbia X, Olive S, Demattei C, Yan J, Cohendy R, et al. The impact of emergency department segmentation and nursing staffing increase on inpatient mortality and management times. BMC Health Serv Res 2016;16(279).
[9] Cildoz M, Ibarra A, Mallor F. Coping with stress in emergency department physicians through improved patient-flow management. Socioecon Plann Sci 2020;71.
[10] Derlet R, Richards J. Overcrowding in the nation's emergency departments: complex causes and disturbing effects. Ann Emerg Med 2000;35(1):63–8.
[11] Hoot N, Zhou C, Jones I, Aronsky D. Measuring and forecasting emergency department crowding in real time. Ann Emerg Med 2007;49(6):747–55.
[12] Gul M, Guneri A. A comprehensive review of emergency department simulation applications for normal and disaster conditions. Comput Ind Eng 2015;83:327–44.
[13] Ahsan K, Alam M, Morel D, Karim M. Emergency department resource optimisation for improved performance: a review. J Ind Eng Int 2019;15:253–66.
[14] Yousefi M, Yousefi M, Fogliatto F. Simulation-based optimization methods applied in hospital emergency departments: a systematic review. Simulation 2020;96(10):791–806.
[15] Kuo Y-H, Leung J, Graham C. Simulation with data scarcity: Developing a simulation model of a hospital emergency department. In: Proceedings - Winter simulation conference; 2012. p. 1–12.
[16] Kenny E, Hassanzadeh H, Khanna S, Boyle J, Louise S. Patient flow simulation using historically informed synthetic data. Stud Health Technol Inform 2021;276:32–7.
[17] Derlet R. Overcrowding in emergency departments: increased demand and decreased capacity. Ann Emerg Med 2002;39(4):430–2.
[18] Aboueljinane L, Sahin E, Jemai Z. A review on simulation models applied to emergency medical service operations. Comput Ind Eng 2013;66:734–50.
[19] Salmon A, Rachuba S, Briscoe S, Pitt M. A structured literature review of simulation modelling applied to emergency departments: current patterns and emerging trends. Oper Res Health Care 2018;19:1–13.
[20] Bakker H, Dunke F, Nickel S. A structuring review on multi-stage optimization under uncertainty: aligning concepts from theory and practice. Omega 2020;96:102080.
[21] Duma D, Aringhieri R. Mining the patient flow through an emergency department to deal with overcrowding. In: Health care systems engineering. ICHCSE 2017. In: Springer proceedings in mathematics and statistics, Vol. 210. Springer, Cham; 2017. p. 49–59.
[22] Duma D, Aringhieri R. An ad hoc process mining approach to discover patient paths of an emergency department. Flexible Serv Manuf J 2020;32:6–34.
[23] Dunke F, Nickel S. A general modeling approach to online optimization with lookahead. Omega 2016;63:134–53.
[24] Dunke F, Nickel S. Online optimization with gradual look-ahead. Oper Res 2021;21(4):2489–523.
[25] Dunke F, Nickel S. Evaluating the quality of online optimization algorithms by discrete event simulation. Cent Eur J Oper Res 2017;25(4):831–58.
[26] Leo G, Lodi A, Tubertini P, Di Martino M. Emergency department management in Lazio, Italy. Omega 2016;58:128–38.
[27] Aringhieri R, Bocca S, Casciaro L, Duma D. A simulation and online optimization approach for the real-time management of ambulances. In: 2018 Winter simulation conference (WSC); 2018. p. 2554–65.
[28] Acuna JA, Zayas-Castro JK, Charkhgard H. Ambulance allocation optimization model for the overcrowding problem in US emergency departments: a case study in Florida. Socioecon Plann Sci 2020;71:100747.
[29] Li M, Carter A, Goldstein J, Hawco T, Jensen J, Vanberkel P. Determining ambulance destinations when facing offload delays using a Markov decision process. Omega 2021;101:102251.
[30] Cildoz M, Mallor F, Mateo P. A GRASP-based algorithm for solving the emergency room physician scheduling problem. Appl Soft Comput 2021;103.
[31] Zaerpour F, Bijvank M, Ouyang H, Sun Z. Scheduling of physicians with time–varying productivity levels in emergency departments. Prod Oper Manage 2021.
[32] Koyuncu M, Araz OM, Zeger W, Damien P. A simulation model for optimizing staffing in the emergency department. In: Cappanera P, Li J, Matta A, Sahin E, Vandaele NJ, Visintin F, editors. Health care systems engineering. Springer International Publishing; 2017. p. 201–8.
[33] Aringhieri R, Bonetta G, Duma D. Reducing overcrowding at the emergency department through a different physician and nurse shift organisation: a case study. In: Daniele P, Scrimali L, editors. New trends in emerging complex real life problems. AIRO Springer series, Vol. 1. Springer Nature; 2018. p. 43–53.
[34] Apornak A, Raissi S, Keramati A, Khalili-Damghani K. Optimizing human resource cost of an emergency hospital using multi-objective Bat algorithm. Int J Healthc Manag 2021;14(3):873–9.
[35] Yeh J-Y, Lin W-S. Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department. Expert Syst Appl 2007;32(4):1073–83.
[36] Feng Y-Y, Wu I-C, Chen T-L. Stochastic resource allocation in emergency departments with a multi-objective simulation optimization algorithm. Health Care Manag Sci 2017;20(1):55–75.
[37] Pazoki M, Samarghandi H. Regulating patient care in walk-in clinics. Omega 2021;99:102200.
[38] Huang J, Carmeli B, Mandelbaum A. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. Oper Res 2015;63(4):892–908.
[39] Cildoz M, Mallor F, Ibarra A. Analysing the ED patient flow management problem by using accumulating priority queues and simulation-based optimization. In: Proceedings - 2018 winter simulation conference; 2019. p. 2107–18.
[40] Cildoz M, Ibarra A, Mallor F. Accumulating priority queues versus pure priority queues for managing patients in emergency departments. Oper Res Health Care 2019;23.
[41] Vanbrabant L, Braekers K, Ramaekers K. Improving emergency department performance by revising the patient–physician assignment process. Flexible Serv Manuf J 2020.
[42] Alves de Queiroz T, Iori M, Kramer A, Kuo Y-H. Scheduling of patients in emergency departments with a variable neighborhood search. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), LNCS, Vol. 12559; 2021. p. 138–51.
[43] He S, Sim M, Zhang M. Data-driven patient scheduling in emergency departments: a hybrid robust-stochastic approach. Manage Sci 2019;65(9):4123–40.
[44] Luscombe R, Kozan E. Dynamic resource allocation to improve emergency department efficiency in real time. Eur J Oper Res 2016;255(2):593–603.
[45] Abo-Hamad W, Arisha A. Simulation-based framework to improve patient experience in an emergency department. Eur J Oper Res 2013;224(1):154–66.
[46] Fitzgerald J, Dadich A. Using visual analytics to improve hospital scheduling and patient flow. J Theor Appl ElectronCommerce Res 2009;4(2):20–30.
[47] De Santis A, Giovannelli T, Lucidi S, Messedaglia M, Roma M. A simulation-based optimization approach for the calibration of a discrete event simulation model of an emergency department. Ann Oper Res 2022.
[48] N. Gilboy, T. Tanabe, D. Travers, A. Rosenau, Emergency severity index (ESI): a triage tool for emergency departments, Emergency severity index implementation handbook, 2012 Edition, AHRQ Publication No. 12-0014(2011).

[49] Ballarini P, Duma D, Horváth A, Aringhieri R. Petri nets validation of Markovian models of emergency department arrivals. In: Application and theory of Petri nets and concurrency, Vol. 12152. Springer International Publishing; 2020. p. 219–38.
[50] Buijs J, van Dongen B, van der Aalst WMP. Quality dimensions in process discovery: the importance of fitness, precision, generalization and simplicity. Int J Coop Inform Syst 2014;23(1) 1440001/1–39.
[51] Vanbrabant L, Braekers K, Ramaekers K, Van Nieuwenhuyse I. Simulation of emergency department operations: acomprehensive review of KPIs and operational improvements. Comput Ind Eng 2019;131:356–81.
[52] Gilbert N. Agent-based models. Quantitative applications in the social sciences, volume 153. SAGE Publications, Inc; 2008.
[53] Gilbert N, Terna P. How to build and use agent-based models in social science. Mind Soc 2000:57–72.