Flexible Services and Manufacturing Journal (2020) 32:6–34
https://doi.org/10.1007/s10696-018-9330-1

An ad hoc process mining approach to discover patient paths of an Emergency Department

Davide Duma[1] · **Roberto Aringhieri**[1]

Published online: 7 December 2018

Abstract
The Emergency Department (ED) management presents a really high complexity due to the admissions of patients with a wide variety of diseases and different urgency, which require the execution of different activities involving human and medical resources. This can have an impact on ED overcrowding that may affect the quality and access of health care. In this paper we propose an ad hoc process mining approach to discover the paths of the patients served by an ED. Our aim is to obtain a process model capable (1) to replicate properly the possible patient paths, and (2) to predict the next activities in the view of a possible application to online optimisation. To prove its effectiveness, we apply our ad hoc approach to a real case study.

Keywords Emergency Department · Overcrowding · Process mining · Patient flow

1 Introduction

An Emergency Department (ED) is a medical treatment facility inside of a hospital or in other primary care centre and is specialised in emergency medicine providing a treatment to unplanned patients, that is patients who present without scheduling.

The ED operates 24 h a day, providing initial treatment for a broad spectrum of illnesses and injuries with different urgency. Such treatments require the execution of different activities, such as visits, exams, therapies and intensive observations. Therefore human and medical resources need to be coordinated in order to efficiently manage the patient flow, which varies over time for volume and characteristics.

A phenomenon that affects EDs all over the world reaching crisis proportions is the overcrowding (Paul et al. 2010). It is manifested through an excessive

✉ Roberto Aringhieri
roberto.aringhieri@unito.it

Davide Duma
davide.duma@unito.it

[1] Dipartimento di Informatica, Università degli Studi di Torino, Corso Svizzera 185, 10149 Turin, Italy

number of patients in the ED, long patient waiting times and patients Leaving Without Being Seen (LWBS); sometimes patients being treated in hallways and ambulances are diverted (Hwang and Concato 2004). Consequently, the ED overcrowding has a harmful impact on the health care: when the crowding level raises, the rate of medical errors increases and there are delays in treatments, that is a risk to patient safety. Not only overcrowding represents a lowering of the patient outcomes, but it also entails an increase in costs (George and Evridiki 2015) because of the decreased productivity. Moreover, the ED overcrowding causes stress among the ED staff, patient dissatisfaction and episodes of violence (Derlet and Richards 2000; Cildoz et al. 2017).

The Emergency Care Pathway (ECP) was introduced by Aringhieri et al. (2017) formalising, from an operational research perspective, the idea of emergency health care delivery systems (Calvello et al. 2013). The ED overcrowding can be addressed in different points of the ECP and, in particular, into two phases: (1) the ambulance rescue performed by the Emergency Medical Service (EMS) and (2) the treatment in the ED. The former is performed only by a part of patients because the other arrives at the ED with their own means.

Regarding the first phase, Aringhieri et al. (2017) suggested to analyse the interplay between the EMS and the network of EDs operating on a given area at the system level. The analysis of a simple EMS dispatching policy (Aringhieri et al. 2018), based on the real-time workload of the EDs, showed that there is room to improve the efficiency of the ED network reducing the patient waiting time. Further, such an improvement is more significant as soon as the percentage of the patients transported by the EMS increases.

We focus in the second phase, that is the management of the ED patient flow. Simulation is widely used to test what-if scenarios to deal with overcrowding (Paul et al. 2010), analysing the use of different resources, setting or policy within the care planning process. Although most of the solutions proposed in literature foresee the use of new additional resources, often the resources available to departments are scarce and there is no economic possibility of new investments (Derlet and Richards 2000; Derlet 2002). Then human and equipment resources available should be used as efficiently as possible optimising existing resources and processes. For this reason, research addressing short-term decision problems are increasing in the recent years (Abouelji nane et al. 2013). Placing in the perspective to alleviate the ED overcrowding without changing the ED resources and settings, there are two way to act: (1) changing the human resources planning (Yeh and Lin 2007; Fitzgerald and Dadich 2009; Sinreich et al. 2012) or (2) adopting different policies in the allocation of the human and equipment resources (Kuo et al. 2012; Luscombe and Kozan 2016; Feng et al. 2017; Koyuncu et al. 2017).

Because of the wide variety of different patient paths within the ED process and the missing of data or tools to mine them, strong assumptions and simplifications are usually made, neglecting fundamental aspects, such as the interdependence between activities and accordingly the access to resources. Actually, the greatest effort in modelling the ED behaviour is to replicate such different paths. Moreover, in order to implement online optimisation algorithms to deal

with overcrowding to intervening on bottlenecks, models capable of making predictions on the patient paths evolution would be useful.

As reported by Rebuge and Ferreira (2012), the analysis of the care processes in health care organisations is a challenging task due the highly dynamic, complex, ad hoc, and multi-disciplinary nature of such processes. Process Mining is a promising approach to improve their understanding through the analysis of the data recorded in health care information systems (Mans et al. 2013a). However, not all process mining techniques perform well in capturing the complex and ad hoc nature of clinical workflows (Rebuge and Ferreira 2012). In literature there are several process mining approaches that use specialised data-mining algorithms to extract knowledge from dataset, creating a process model that takes into account dependency, order and frequency of events, but also decision criteria and durations. After presenting a review of the process mining in health literature, Partington et al. (2015) report about a case study in which process mining techniques are applied to the administrative and clinical data of the patients suffering from chest pain symptoms in four hospitals in South Australia. Nowadays huge amounts of data are collected by EDs, recording diagnosis and treatments of patients. Process Mining can exploit such data and provide an accurate view on health care processes (Basole et al. 2015; Rojas et al. 2017a, b; Abo-Hamad 2017; Alvarez et al. 2018), ensuring their understanding in order to generate benefits associated with efficiency (Rojas et al. 2016).

In Duma and Aringhieri (2017) we applied several process discovery techniques from the literature for a real case study. We tried to model the ED from a control flow perspective and to identify the path of each patient on the basis of the only information known at the access of the patient. We shown that standard process discovery approaches could be not able to provide models adequate to our aims in terms of simplicity and precision. This because the ED process we would mine 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 simplicity and precision discovering the process, as discussed in Duma and Aringhieri (2017).

In this paper we propose a new framework to mine an ED process model based on ad hoc process discovery tools. Our purpose is to obtain simple and precise process model capable to replicate the large variety of the paths and to predict the use of the ED resources by each patient on the basis of the only information known at the access of the patient. We apply our new framework to a real case study arising at *Ospedale Sant'Antonio Abate di Cantù*, Italy. The paper is structured as follows. The case study is reported in Sect. 2 describing the population of the patients and the ED organisation, also providing a simple retrospective analysis. After describing how to pre-process our datasets, in Sect. 3 we report the results of a mining based on standard approaches in order to justify the need of an ad hoc mining solution to develop a proper model for the ED under consideration. The conformance of the discovered model is then discussed in Sect. 4 testing its replicability and its robustness over a new dataset. Finally, Sect. 5 closes the paper discussing the importance of process mining and reporting new research directions.

2 The case study under consideration

We present a real case study concerning the ED sited at *Ospedale Sant'Antonio Abate di Cantù*, which is a medium size hospital in the region of Lombardy, Italy. The ED serves about 30,000 patients per year.

The resources available within the ED are: 4 beds for the medical visits placed in 3 different visit rooms, in addition to one bed within the shock-room and another one in the Minor Codes Ambulatory (MCA), one X-ray machine, 5 Short-Stay Observation (SSO) units (beds), 10 stretchers and 10 wheelchairs to transport patients with walking difficulties. The medical staff is composed of 4–6 nurses and 1–3 physician(s), depending on the time of day and the day of week, in addition to the X-ray technician.

Thanks to the collaboration with the ED, we have information concerning all the 88,272 accesses made in the years 2013–2015. Such information is available on a dataset extracted from the Hospital Information System, and it contains records about personal data of the patients, their diagnosis, their arrival times, and the activities executed (e.g., X-ray, blood exams,...).

2.1 Patient population

From the personal data and the diagnosis available in the ED dataset, we can present an overview of the patient population of the case study. Such data is sex (male 52.7% or female 47.3%) and age of the patient, the urgency code (1–5, in descending order of urgency), the main symptom (undefined 35.2%, trauma 30.7%, abdominal pain 6.9%, temperature 4.5%, chest pain 3.8%, dyspnea 3.4%, and other 25 options), timestamps and resources used during the activities, and type of discharge (ordinary 82.0%, hospitalisation 10.6% and abandonment 7.4%). Information about the type of access (autonomously 79.9% or with a rescue vehicle 20.1%) is also provided in the dataset.

The patient population is quite uniformly distributed across the different ages, with slight peaks for the age groups 5–9 and 35–54. To motivate this fact we compared the access frequencies of the 5-year age classes with the demographic distribution. As shown in Fig. 1, the almost uniform distribution of accesses among the

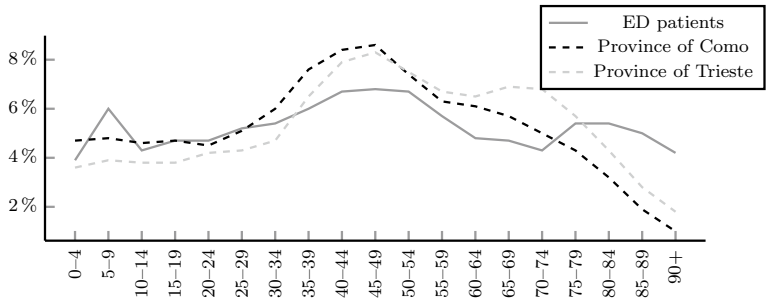

Fig. 1 Comparison between territorial and patient age distributions

age classes is due to the balance between the lower percentage of children and older people in the territorial area and the higher percentage of adults, which have a lower number of accesses per person. For the comparison, we used ISTAT data about 2014 in the province of Como, in which Cantù is located, observing that Lombardy Region and Italian territory have very similar distributions, but there are areas with a different age distribution, such as the Province of Trieste, for which we expect a different ED demand. This because in addition to a greater number of accesses, older patients have urgency codes 1–2 more frequently (30.0% of cases for patients over 65 years old against 12.1% for under 64) and consequently they have higher Emergency Department Length-Of-Stay (EDLOS), as we will see below.

2.2 Organisation of the Emergency Department

A patient is interviewed and registered as soon as possible by a triage-nurse on his/her arrival in the ED, recording personal data, the main symptom and the urgency code from 1 (most urgent) to 5 (less urgent), in accordance with Table 1.

After the triage, the patient is visited in one of the visit rooms by a physician. Certain patients are visited in other special rooms such as the shock-room, which is properly equipped for severely urgent interventions, and the MCA, provided by the ED from Monday to Friday in the time slot 8:00–16:00 for adult patients with low urgency codes and good ambulation ability.

After a medical visit, the physician can prescribe therapies, tests or observations. Therapies are various but always performed by a nurse and identified in the same way within the dataset. Tests could be laboratory tests, which are performed by a nurse, X-ray examinations, performed by a X-ray technician with the assistance of a nurse for urgent or motor-impaired patients, or other investigations that are not competence of the ED, that could be a Computerised Tomography (CT), an ecography or a specialist visit. Then, there are two different SSO, both requiring a SSO bed unit and the supervision of nurses and physicians: the first is the ordinary SSO for medical reason, while the second is the pre-hospitalisation SSO, that is when patient need to be hospitalised but a bed is not yet available within the assigned hospital ward.

After examinations, treatments and specialist visits, the patient is revalued again by a physician of the ED, which establish how to continue the treatments, the need of hospitalisation or the discharge for patients needing non-urgent investigations.

Table 1 Urgency codes: description and frequency over 2013–2015

Number	Colour	Description	Frequency (%)
1	Red	Immediate danger of death	1.5
2	Yellow	Need of a timely medical visit	15.8
3	Green	Need of treatments or investigations	61.6
4	Blue	Symptoms that could be treated as primary care	13.8
5	White		7.3

There are different ways in which a patient can leave the ED and/or be discharged. The first one is before the triage, when the patient can leave without a visit (LWBS). Another possibility is after the triage in the case of a non-urgent patients under 18 years old, which are under competence of the paediatric department and, from the ED point of view, is a discharge. Further, during tests and treatments the patient has the right to interrupt the care. Finally, after all the necessary visits and investigation patients can be discharged or hospitalised.

In Table 2 we summarise all the activities that could be performed by a patient within the ED, in accordance with the suggestions of the ED staff of the case study collected in several interviews about the ED management system and the content of a dataset. The first and the second columns indicate respectively an identifier for each activity and its description. Then, we classify the activities into 5 classes called *Triage*, *Visit*, *Tests & Care*, *Revaluation* and *Discharge*. In the fourth column the activities that are competence of the ED are indicated with a mark. Finally, in the last column the timestamps available in the dataset records are indicated, that is the start time t_S, the prescription or request time t_P, the report time t_R and the end time t_E.

A unique *Triage* activity is defined for the homonym class, which consist of the triage and registration procedures. After the triage, the patient should have a first visit, which is usually the *Medical Visit* performed in one of the visit rooms. Alternatively, the first visit can be performed in the *Shock-Room*, that is an adequately equipped room for some urgent patients, or in the MCA for patients with urgency code 4–5 (*MCA Visit*). Instead, the *Revaluation* refers to the successive visits after some tests or treatments, performed generally in the same visit room by the same physician. A *Therapy* activity is recorded each time a nurse provide a general care

Table 2 Activities in a patient path

Id	Description	Class	ED comp.	Timestamps
A	Triage	Triage	✓	t_E
B	Medical visit	Visit	✓	t_E
C	Shock-room	Visit	✓	t_E
D	MCA visit	Visit	✓	t_E
E	Paediatric fast-track	Discharge		t_P
F	Therapy	Tests & Care	✓	t_P, t_E
G	Laboratory exams	Tests & Care	✓	t_P, t_R
H	X-ray exams	Tests & Care	✓	t_P, t_R
I	Computed tomography (CT)	Tests & Care		t_P, t_R
J	Ecography	Tests & Care		t_P, t_R
K	Specialist visit	Tests & Care		t_P, t_R
L	Short-Stay Observation (SSO)	Tests & Care	✓	t_S, t_E
M	Pre-hospitalisation SSO	Tests & Care	✓	t_S, t_E
N	Revaluation visit	Revaluation	✓	t_E
O	Hospitalisation	Discharge	✓	t_E
P	Discharge (ordinary)	Discharge	✓	t_E
Q	Interruption	Discharge		t_E

treatment (under the prescription of the physician) to the patient, such as taking a drug or medicate a wound. Activities G–J are exams consisting of an execution (in which the patient is involved), and of a reporting performed by a technical staff. The *Laboratory Exams* and the *X-Ray Exams* are competence of the ED: the formers consist in blood collections made by a physician, while the latters are X-ray scans executed by the X-ray technician. The *CT* and the *Ecography* are instead performed in an ambulatory that serves several hospital departments, but the ED has the highest priority. Similarly, the *Specialist Visit* is usually performed by a specialist physician in another department of the hospital, giving priority to ED patients. The *SSO* and the *Pre-hospitalisation SSO* are both observations made in the SSO units under the supervision of a physician and a nurse: the former is performed to ensure the settling of the medical conditions, while the latter is a temporary stay awaiting for the release of a bed in a specific hospital department for the hospitalisation. Finally, the patient discharging from the ED is performed in four different ways: beginning a *Paediatric Fast-Track* in the paedriatric department, performing a *Hospitalisation* in the same or through a transfer to another hospital, going home with an ordinary *Discharge*, or in deciding the *Interruption* of the process of care, that is the LWBS or the refusal of treatments and/or hospitalisation.

Figure 2 depicts a general patient path: after the triage, a Visit class activity is always provided except for a LWBS patient. Then the patient can be discharged or continue with a sequence of Tests & Care class activities, that is always followed by a revaluation visit, after which the patient can be discharged or go on with other Tests & Care class activities.

2.3 Retrospective analysis

The ED of Cantù performed a retrospective analysis using the NEDOCS in the aftermath of several management changes, such as the introduction of the MCA or a new staff rostering. In addition to inadequacy of this and other similar measures, proved by Hoot et al. (2007), the NEDOCS is a one-dimensional index that expresses the request of several resources and therefore is not useful to identify bottlenecks. Furthermore, the analysis performed by the ED of Cantù has been affected by the lack of several information that has been dealt with approximations. For all these reasons, we omit the NEDOCS results, focusing on a brief retrospective analysis that describes the variability of demand over time. The results

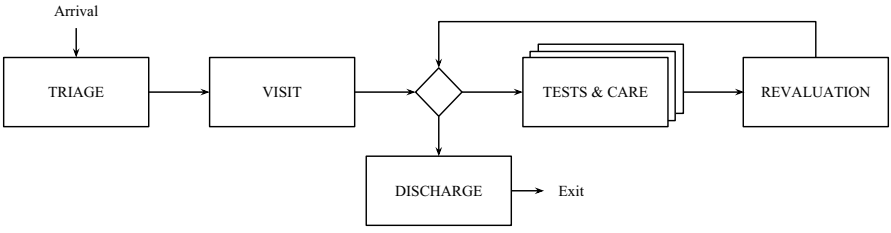

Fig. 2 A general path for a patient within the ED

reported in this section are obtained by a statistical analysis of the ED dataset regarding the 88,272 accesses during the years 2013–2015.

The accesses have different fluctuations over the day, among the days of the week and among the seasons, but also among the urgency classes. The higher arrival rate fluctuations occur during the business hours of the day, as shown Fig. 3a, especially for the minor codes, which usually go to the ED instead of relying on primary care. For the same reason, a higher number of non-urgent arrivals has been registered on Monday, as shown in 3b. Conversely, the urgency class 1 has the highest coefficient of variation among the different months of the week, because of medical and epidemiological reasons that causes more arrivals in winter. Nevertheless, from Fig. 3c a uniform workload over the year (except for August) could be deducted, in fact, the workload do not depend directly of the number of accesses. Then, we report in Fig. 3d the average number of patients concurrently treated (including all the activities between the first visit and the discharge), that is a more consistent indicator with respect to the ED staff perception.

The statistics in Table 3 justify this fact, indeed more urgent patients have a longer average EDLOS. Such a difference is due to the higher frequency of SSO for patients with urgency codes 1 and 2, caused by a higher percentage of hospitalisations. The average waiting times confirm us that the priority among urgency codes is respected. Finally, lower urgency codes also have an higher rate of LWBS patients, while the percentage of patients Leaving After Being Seen (LABS), that is patients leaving after the medical visit but without finishing the treatment, is similar for all the urgency classes.

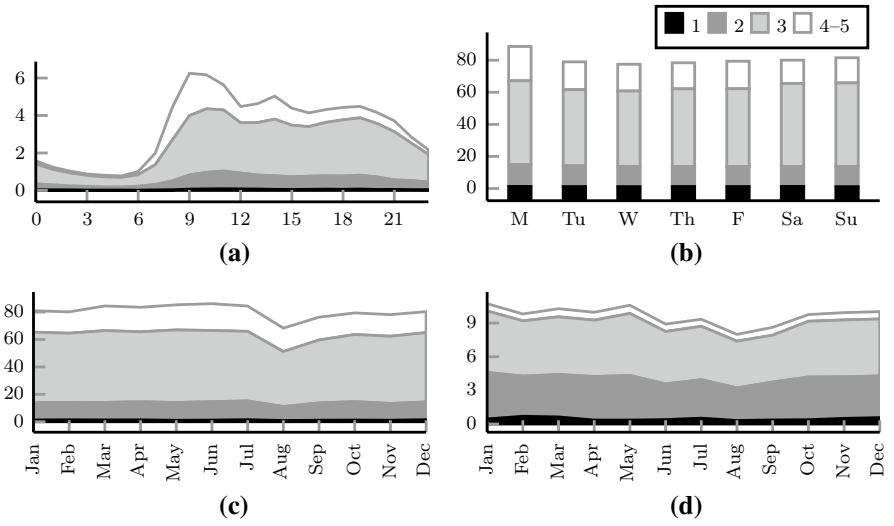

Fig. 3 Patient accesses classified by urgency code. **a** over the day (patients per hour), **b** over the week (patients per day), **c** over the year (patients per day) and **d** number of patients concurrently treated

Table 3 Waiting times, LWBS and statistics on the treatment of patients

Urgency code	Average wait time (min)	Percentage of LWBS (%)	Average EDLOS (h)	Perc. of SSO (%)	Average SSO duration (h)	Perc. of hosp. (%)
1	15	0.2	9	28.4	19	60.3
2	34	0.5	7	19.5	20	28.9
3	65	3.3	2.5	5.4	19	7.8
4–5	68	11.1	1	0.5	17	1.1

3 Process discovery

After reporting how to pre-process the huge amount of available data, in this section we report the results of a process mining based on standard approaches in order to justify the need of an ad hoc process mining solution to develop a proper model for the ED under consideration. Our aim is to have a model capable (1) to replicate properly the possible patient paths, and (2) to predict the next activities and the required resources of patients on the basis of their characteristics and their activities performed until that moment.

3.1 Pre-processing

In order to use discovery mining techniques, we need to pre-process the ED database to create an event log, which consists of a set of traces (i.e. temporally ordered sequences of events of a single case), their multiplicity and other information about the single events, such as timestamps and/or durations, resources, case attributes and event attributes. In our case, the events correspond to the activities concerning the patient treatments recorded for each access within the ED of Cantù's database, while each trace identifies a patient path. An example of the ED database records from our case study is shown in Table 4, where two accesses are reported.

The event log has been generated taking into account the accesses of the 3-years period from 2013 to 2015. Each case of the event log consists in an access and events consist in activities, which has been classified into 17 event classes corresponding to the same number of activities reported in Table 2. An example of the resulting event log is shown in Table 5, where we report the rows of the event log corresponding two the two instances of accesses taken into account in Table 4.

Because of the control flow perspective that we are taking into account, we need to estimate the start time and the end time of each activity involving a patient. For instance, tests after blood collection are not part of the activity in this sense, because the patient can continue with the execution of other activities while the blood sample is analysed and reported. However, we have to take into account several noise factors that may be present in the dataset (Mans et al. 2013b; Suriadi et al. 2017). A list of the noise factors that we are dealing with is the following:

$\mathfrak{N}_0$– *missing timestamps*	for activities of the classes A–K and N–P one or both start and end timestamps are not available;
$\mathfrak{N}_1$– *timely execution*	urgent patients' activities are performed without worrying about the registration of the information at the exact moment, consequently triage or shock-room activities could refer to a later time;
$\mathfrak{N}_2$– *forgetfulness in recording therapies*	therapies are sometimes recorded during the discharge instead of the actual execution time, because they are activities that could be performed on the fly;
$\mathfrak{N}_3$– *multiple recording*	for technical reasons, two ore more records can refer to the same event for the event classes G–J, that is when more examinations are performed through a unique collection, scan or specialist visit;
$\mathfrak{N}_4$– *fake or missing revaluation visit*	sometimes the revaluation record can refer to the passage of the medical record between two physicians for the change of work shift, while other times a revaluation visit could be performed without to be recorded if the patient is discharged at once (but from ED suggestions we know that always a revaluation is performed between tests and discharge);
$\mathfrak{N}_5$– *fake medical visit*	paediatric visits are performed in the Paediatric Department but that are activities also recorded in the dataset of the ED.
$\mathfrak{N}_6$– *tests reported after discharge*	activities (such as non-urgent investigation) are included within the patient path but could be analysed and reported after the patient discharge.

Table 4 Example of two ED database records corresponding to two different accesses: all the information about the personal data and the activities of a patients are contained in a unique row, leaving the cells not applicable blank.

Id	Code	Symptom	Sex	Age	registr_time	vis_room	visit_end	xray1_pres	xray1_rep	...	exit_time
007776	2	Urological	M	67	8/4/13 16:59	Shock-C	8/4/13 17:25			...	8/4/13 17:31
007777	3	Trauma	F	39	8/4/13 17:22	Visit-A	8/4/13 18:22	8/4/13 18:22	8/4/13 19:08	...	8/4/13 19:16

Table 5 Example of event log corresponding to the activities of two patients

act_name	timestamp1	timestamp2	id_patient	Code	Symptom	Sex	Age	arrival_type
Triage		8/4/13 16:59	007776	2	Urological	M	67	Self
Triage		8/4/13 17:22	007777	3	Trauma	F	39	Self
Shock-room		8/4/13 17:25	007776	2	Urological	M	67	Self
Discharge		8/4/13 17:31	007776	2	Urological	M	67	Self
Visit		8/4/13 18:22	007777	3	Trauma	F	39	Self
X-ray	8/4/13 18:22	8/4/13 19:08	007777	3	Trauma	F	39	Self
Discharge		8/4/13 19:16	007777	3	Trauma	F	39	Self

Columns *timestamp1* and *timestamp2* contain timestamps that are interpreted depending on the activity: the former can refer to the start or the prescription time, while the latter is used for the report or the ending time

In Fig. 4 two examples of traces with noise are reported, with the corresponding timestamps available (black dots) and several missing useful timestamps (white dots): we can estimate the missing start time subtracting the average service time and/or reporting time in accordance with the directions of the ED staff, as reported in Table 6. A noise of type $\mathfrak{N}_6$ can be observed in trace 1: actually all activities finish before the discharge, but if we take into account the end times, we have the wrong trace ABGNPH. Trace 2 contains both noise phenomena $\mathfrak{N}_1$ and $\mathfrak{N}_2$. The former occurs when the shock-room visit is registered after the actual end because the urgency of treating the patient has the priority on the recording. The latter is due to the incorrect time of insertion of the therapy execution, whose recording is made during the final check at the discharge. In this case is not possible to know exactly the moment in which the activities C and F have been performed, so we approximate the end time of the shock-room visit with the timestamps of the dataset, while for the therapy execution we suppose that the start time is immediately after the prescription by the physician.

The pre-processing algorithm has been implemented as follows:

1. Start time and end time of each activity are estimated in accordance with Table 6 (noise $\mathfrak{N}_0$).
2. A sorting time $\bar{t}$ is fixed for each activity in order to avoid overlapping of activities (because of $\mathfrak{N}_0$); we chose the more reliable time, that is $\bar{t} = t_S$ for activities F, L and M, $\bar{t} = t_E$ for the other ones.

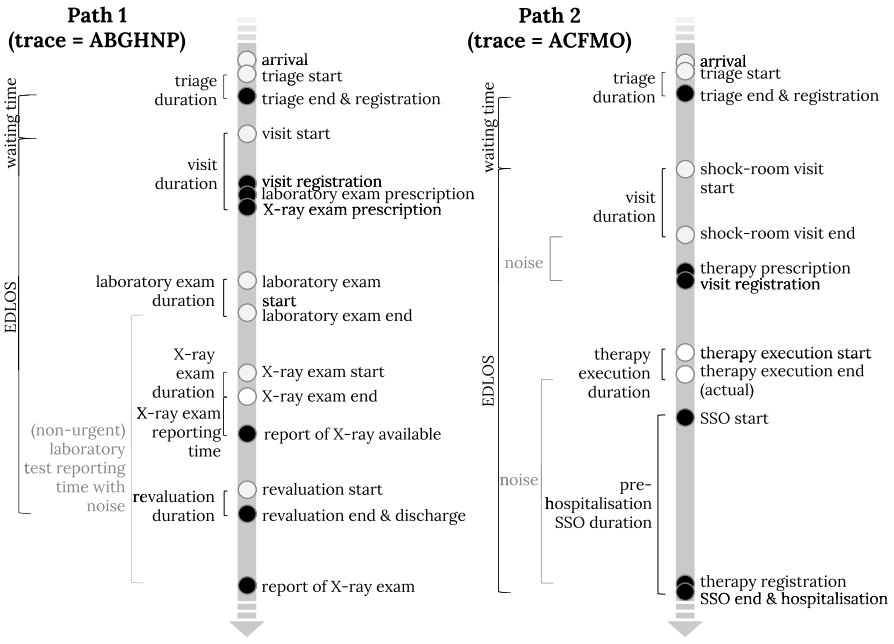

Fig. 4 Example of the activities for two different paths with the corresponding timestamps (black dots) and other significant times (white dots)

Table 6 Average duration of the activities according to the ED staff and estimation of the missing timestamps

Event class	Activity duration d	Reporting duration r	Start time t_S	End time t_E	Sorting time $\bar{t}$
A	5 min	n.a.	$t_E - d$	Available	t_E
B	15 min	n.a.	$t_E - d$	Available	t_E
C	15 min	n.a.	$t_E - t_E^{\text{triage}}$	Available	t_E
D	15 min	n.a.	$t_E - d$	Available	t_E
E	0 min	n.a.	t_E	Available	t_E
F	2 min	n.a.	$t_E - d$	Available	t_S
G	3 min	15 min	$t_R - r - d$	$t_R - r$	t_E
H	3 min	30 min	$t_R - r - d$	$t_R - r$	t_E
I	10 min	45 min	$t_R - r - d$	$t_R - r$	t_E
J	15 min	45 min	$t_R - r - d$	$t_R - r$	t_E
K	15 min	n.a.	$t_E^{\text{last before K}}$	t_R	t_E
L	Available	n.a.	Available	Available	t_S
M	Available	n.a.	Available	Available	t_S
N	10 min	n.a.	$t_E - d$	Available	t_E
O	1 min	n.a.	$t_E - d$	Available	t_E
P	1 min	n.a.	$t_E - d$	Available	t_E
Q	0 min	n.a.	t_E	Available	t_E

3. If activity E occurs, all the other activity are removed, except the triage (noise $\mathfrak{N}_5$).
4. The activities of the same path are sorted in chronological order of $\bar{t}$ composing the trace.
5. For each trace, let $\bar{t}_{\text{exit}}$ be the sorting time of the discharge (one among activities O, P and Q) and let $\tau > 0$ be a parameter denoting the amount of time before the discharge in which the forget recording of therapies is remedied. If $\bar{t}_{\text{exit}} - \bar{t}_F < \tau$, then $\bar{t}_F = \max\{\bar{t}_F, t_R^F + 1 \text{ min}\}$, where t_R^F is the prescription time of that therapy (noise $\mathfrak{N}_2$).
6. For each trace, let $\bar{t}_Y$ be the sorting time of a certain Tests & Care class activity. If $\bar{t}_Y > \bar{t}_{\text{exit}}$, then $\bar{t}_Y$ is fixed 1 min before the first revaluation visit after the prescription time of that activity (noise $\mathfrak{N}_6$).
7. For each activity of each trace:
   - if it precedes the triage time, then it is moved 1 min after the triage time (noise $\mathfrak{N}_1$);
   - if it is not a triage and it precedes the visit time, then it is moved 1 min after the visit time (noise $\mathfrak{N}_1$).
8. For each trace, if there is no revaluation visit between a Tests & Care activity and the discharge, then a fake revaluation visit is inserted a minute before the discharge (noise $\mathfrak{N}_4$).
9. For each trace, consecutive Tests & Care activities of the same type such that the time between them is less than δ are merged keeping the start time of the first one and the end time of the last one (noise $\mathfrak{N}_3$).

In our pre-processing, parameters τ and δ have been fixed equal to 10 and 30 min, respectively. The derived event log is composed of 475,870 events concerning 88,272 cases. The execution time required by the pre-processing procedure implemented in C++ is 26.4 s for the whole dataset. Excluding LWBS and the paediatric fast-tracks, corresponding to the trivial traces AQ and AE, the remaining 66,551 cases generated 7868 different traces of length ranging in [3, 31], with an average value of 5.5. For instance, the traces resulting from the rows of the event log in Table 5 are ACP and ABHP for the patients with id 007776 and 007777, respectively. The high number of different traces with a low frequency is partially caused by medical reasons (i.e. patients need very different treatments), but also by noise phenomena $\mathfrak{N}_0$–$\mathfrak{N}_6$ that have not been relieved completely.

3.2 Standard process discovery

We report a summary of the analysis of process discovery techniques from the literature. The models and the results presented in this section are similar to those discussed in Duma and Aringhieri (2017) but they differs from the use of the event log obtained by the pre-processing procedure described in Sect. 3.1.

In addition to the requirement of computational efficiency, not always found testing standard approaches, four main quality criteria of the process discovery algorithms have be assessed (Buijs et al. 2014): fitness, precision, generality and simplicity. Fitness indicates how much of the observed behaviour is captured by the process model, that is how many traces of the mined event log can be replied on it. The precision points out if behaviour completely unrelated to what was seen in the event log are allowed by the model. The generality is the capacity of the model to generate different sequences of activities with respect to the observations in the log. Finally, the simplicity is the easiness in understanding the process using the mined model.

The huge number of traces suggests the use of discovery techniques that deal with low frequent behaviour and noise. We focus on two different process miners, the *HeuristicMiner* (HM) (Weijters and Ribeiro 2011) and the *Inductive Miner–infrequent* (IMi) (Leemans et al. 2014), both based on the control-flow perspective.

The HM takes into account the order and the causal dependencies among the events within a trace, generating a model that uses the Heuristic Net notation, which is flexible because it can be easily converted in other notations, for instance a Petri Net. The IMi is an extension of the *Inductive Miner* (IM), that is a divide-and-conquer approach based on dividing the events into disjoint sets taking into account their consecutiveness within traces, then the event log is split into sub-logs using these sets. The IMi uses the same approach but filters a fixed percentage of traces representing infrequent behaviour to create a PN. Both the techniques require low computational time, that is an important requirement due to the dimension of our event log. On the contrary, the two approaches perform differently with respect to the quality criteria.

The process models $\mathcal{H}$ and $\mathcal{I}$ mined by the event log using the HM and the IMi are shown in Figs. 5 and 6. Such process discovery techniques are provided by ProM 6.6, an open source pluggable tool (Van Der Aalst et al. 2009).

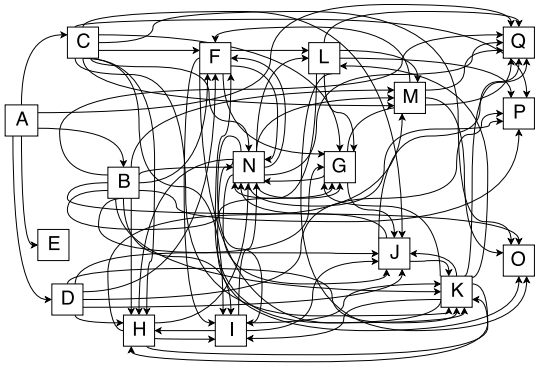

Fig. 5 Process model mined with the HM: model $\mathcal{H}$ (heuristic net)

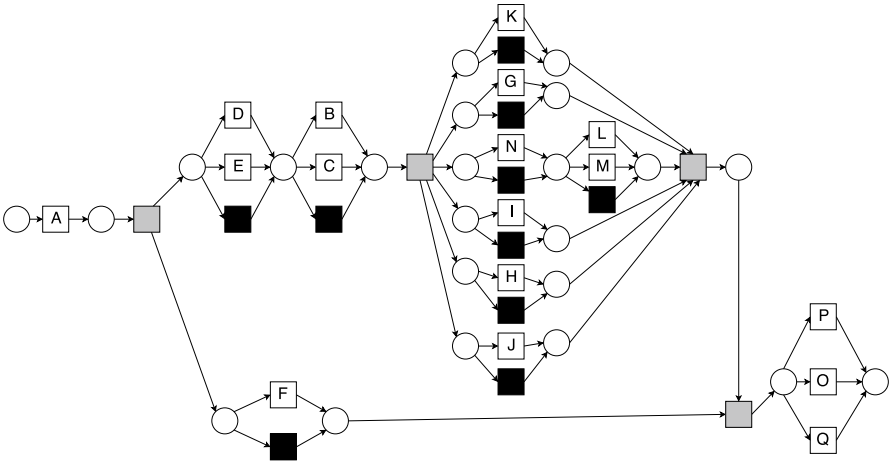

Fig. 6 Process model mined with the IMi: model $\mathcal{I}$ (Petri net)

The model $\mathcal{H}$ has been generated varying the parameters *dependency* and *relative-to-best* of the HM in such a way to reach the best fitness, that is an index of the capacity to reproduce the behaviour recorded in the event log, equal to 64%. The obtained model $\mathcal{H}$, as well all the other generated varying the parameters, is a so-called *Spaghetti process* that is not sufficient simple to understand the whole process. In addition to the problem of non-simplicity, the model is not adequate to predict the evolution of the route because it has no memory regarding the activities already performed.

The model $\mathcal{I}$ has been obtained varying the noise parameter of the IMi in order to have a good precision, avoiding or limiting infrequent behaviour. However, we observed very slight deviations among the models ranging the noise percentage, that has been fixed to 20%. Contrariwise to $\mathcal{H}$, this model is very simple but not precise: the parallelisms among activities allowed by $\mathcal{I}$ (represented by the grey boxes in the Fig. 6) imply additional behaviour that is not present in the event log, for instance

traces with two Visit class activities are allowed by the model but not in reality. At the same time, there is an insufficient fitness, because the model $\mathcal{I}$ do not allow to replicate behaviour present in the event log, such as the execution of two or more event of the same event class (e.g. multiply therapies or multiply X-ray exams).

Other standard approaches have been tested in a preliminary analysis without satisfying our requirements. For instance, we tried to use the Fuzzy Miner, which is a discovery algorithm based on significance and correlation. This approach has been applied in Abo-Hamad (2017) for an ED case study to show the main highway paths for patients to gain insights into bottlenecks and resource utilisation. However the Fuzzy Miner is not suitable to our purpose, because the level of granularity necessary to implement a process model to analyse resource allocation policies is very high, then varying the parameters of such an algorithm we deal with the same trade-off between precision and simplicity founded for the models $\mathcal{H}$ and $\mathcal{I}$.

3.3 Ad hoc process discovery model

Starting from the observations in Sect. 3.2, we would like to design a model with a better compromise between fitness, precision, generalisation and simplicity. A way to obtain a simple but precise process model is to use a tree-structure that allows us to follow the possible different evolutions of the paths. However the huge variability of the traces would generate a model of huge dimensions, that is not good from the simplicity point of view. This issue could be addressed through a clustering of the patients with respect to their characteristics, such as symptoms and urgency. Indeed the treatment of patients with illness or injuries belonging to different medical specialty and very various even within the same specialty. Such a classification should identify classes of patients in such a way to reduce as much as possible the dimension of the trees, and to group patients with different characteristics in order to guarantee their statistical relevance.

An example of the process model that we would propose is shown in Fig. 7. Each node represents an activity executed after all the activities indicated by the ancestor nodes, while the arrows indicate that a certain activity can be performed after

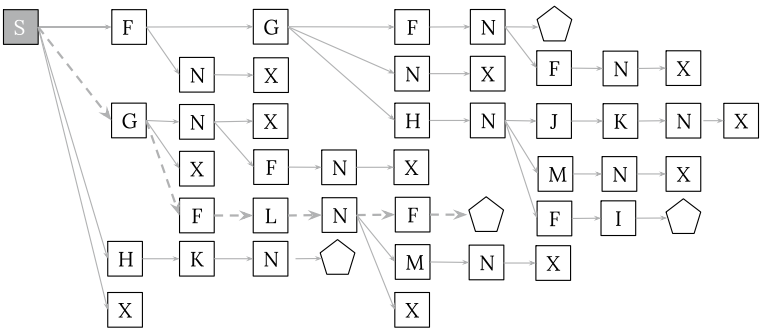

Fig. 7 Example of process model with a tree-structure. Dashed edges highlight the possible path SGFLNF

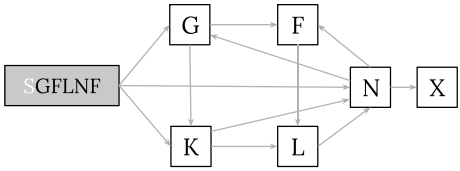

Fig. 8 Example of a sub-process model with a graph-structure

another one. The presence of one or more edges from a node indicates that one and only one of them have to be crossed, representing a sort of XOR condition. Therefore, branches represent the different path evolutions after the execution of the node from which they start.

The tree-structure allows us to keep track of the path previously done, which is a way to have memory of the past activities (unlike model $\mathcal{H}$) and to predict what could happen in the future. The labelling of edges with frequencies allows us to estimate, in a computationally efficient way, the probability that a certain event will occur from a certain point on wards. However a model mined from the event log with these rules would replicate all but only the paths in the data, leading to an over-fitting—that does not satisfy the generalisation requirement—and generating a high number of nodes. To overcome these limitations, we summarise infrequent branches with graphs, in which we do not keep track of the past activities.

A possible path is highlighted with dashed edges in Fig. 7, whose trace starts with a node labelled with G and followed by other tree nodes labelled with F, L, N and F respectively. In this case, the branch ends with a pentagonal box indicating that the model continues with a graph similar to that depicted in Fig. 8.

Before introducing an ad hoc algorithm for the process discovery of the real case study, we imposed a process structure based on the framework in Fig. 2 drawn together with the staff and consistent with the previously obtained models.

Excluding the cases of LWBS and the paediatric fast-track, which are trivial and not interesting for the process discovery, each path begins with the activity A (triage) followed by an activity of the Visit class, that is B, C or D. Then, the patient performs a sub-process that we call Investigations Process (IP), consisting of a number $n \geq 0$ of activity sequences of the Tests & Care class, that is F–M activities, at the end of each there is always an activity N (revaluation visit). Finally, at the end of the IP, the path ends with a Discharge class activity, that is E, O, P or Q.

We are interested in studying the evolution of the path inside the IP, that is the sub-process that differentiates the paths and should be predicted in order to optimise the resource allocation. Indeed, there are two moments of the path in which the prediction make sense, that is before a Visit class activity or before the revaluation visit. After these activities, the physician decides if the patient can be discharged or if a set of Tests & Care class activities is necessary. Such set is partially ordered, because some activities must be performed in a certain sequence (e.g. X-rays could be necessary before the specialist visit at the orthopaedist ward), while other activities that do not impact on others can be performed in different orders. Of course the latters include all the exams, that is activities G–J, while we assume in general that the formers need to be executed in the order registered in the event log because

of the impossibility to go specifically from the data. From our perspective, traces with the same activities and two or more consecutive activities G–J with different order identify the same path, even if in the records they are executed in different way because of management decisions. For this reason, we define a unique order of those activities that can be performed in any order, that is $G \prec H \prec I \prec J$, where $\prec$ indicates that the former activity precedes the latter.

3.3.1 Phase 1: patient clustering with decision tree

We use the Decision Tree (DT) learning approach of the data mining to predict the first sequence of Tests & Care class activities before the revaluation visit, possibly null in case of discharge immediately after the visit. To this aim, the label is expressed as a string in which characters identify the activities of the sub-trace between the first visit (excluded) and the first revaluation visit (included), using the only character X if no activities are performed in the IP. The attribute are all the information known at the triage: sex, age, arrival mode (with an ambulance or autonomously), main symptom, urgency code, time-dependence (yes/no referred to urgent patient with specific symptoms), arrival day (Monday–Sunday), type of arrival day (weekday or weekend), month of arrival (January–December), arrival time slot (60 min period) and type of first medical visit (ordinary, shock-room or MCA).

The DT approach requires the following parameters. We use the criterion called "gain ratio", that is used to reduce a bias towards multi-valued attributes by taking the number and size of branches into account when choosing an attribute. We fixed a confidence equal to 0.25 and imposed a minimum leaf size equal to the 1% of the whole patient population of the event log. Finally, we set the minimal gain parameter to 0.25 and to 0.2 in such a way to obtain two different DTs of different size, with number of leaves equal to 9 (Fig. 9) and 18 (Fig. 10), respectively.

We denote with $\{C_i\}_{i=1,\dots,9}$ and $\{C'_i\}_{i=1,\dots,18}$ the clusters obtained in correspondence of the leaves of the two DTs, which are two different partitions of the set of all patients that all the visited patients. Observe that $C_i = C'_i$, for $i = 1, \dots, 7$, $C_8 = C'_8 \cup C'_9$, and $C_9 = C'_{10} \cup \cdots \cup C'_{18}$. The clusters obtained through the data

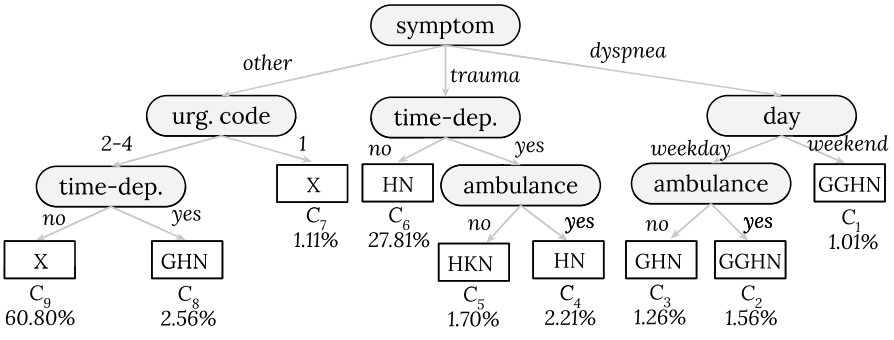

Fig. 9 Decision tree with gain parameter set to 0.25 and obtained clusters $C_1 - C_9$

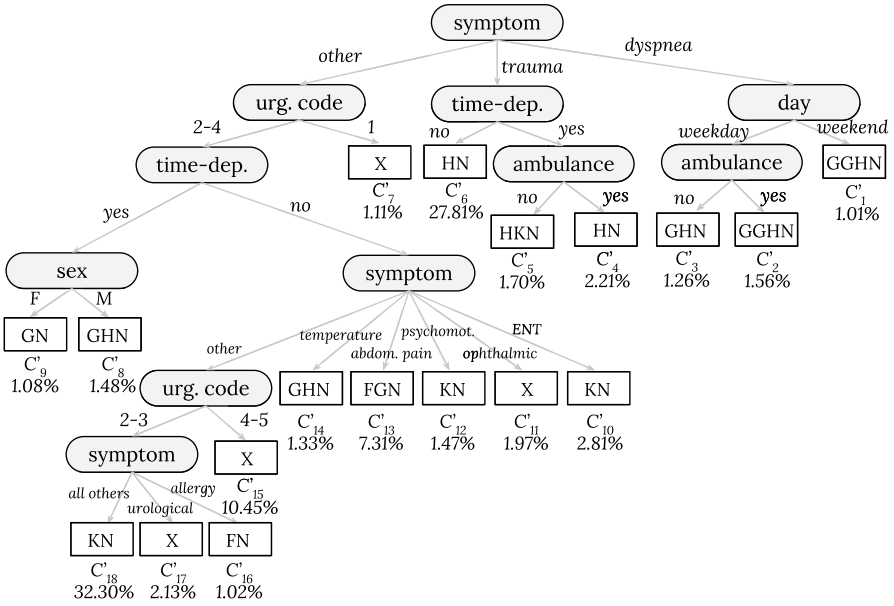

Fig. 10 Decision tree with gain parameter set to 0.2 and obtained clusters $C'_1 - C'_{18}$

mining allow us to reduce the number of such paths for each subset of patients and to group patients that have similar frequencies to follow a certain path. The DT has been applied using RapidMiner Studio 7.1.

3.3.2 Phase 2: process modelling

For each cluster defined in the first phase of our ad hoc approach, we model the behaviour of its patients, that is the possible patient paths. To this end, we use a notation that we call *Hybrid Activity Tree (HAT)*, that is a graph $\mathbb{G} = (\mathbb{A}, \mathbb{T})$, where $\mathbb{A}$ is a set of nodes labelled with the ED activities (those in Table 2) and $\mathbb{T}$ is a set of oriented edges indicating possible transitions between nodes and labelled with a weight $f \in (0, 1]$ equal to the relative frequency of that transition. We remark that different nodes can be labelled with the same activity: each of them represents the execution of such an activity after the execution of different activity sequences.

Globally, the HAT represents all the possible paths in the IP phase as a tree, in which the root node S has $m > 0$ child nodes representing the m first possible activities that can be performed after the medical visit (activities B–D), each of them has a number $m_i \geq 0$ of child nodes representing the second activity, and so on, until reaching a leaf node. This node always represents a general Discharge class activity, labelled with X, or the starting node of a graph, called Sub-Tree Activity Graph (STAG), which is used to model infrequent behaviour (indicated with a pentagon in Fig. 7). A STAG is a graph to model infrequent paths having the first part of the sequence in common, which consist in the sequence of nodes from the root to the node that connects the tree to the STAG. Also within the STAG, an edge indicates

that a certain activity can be performed after another one, but unlike what happen for the tree nodes, at most a node within a STAG can labelled with a certain activity. Therefore, a node can have more incoming edges representing after which activities that one can be performed.

The proposed process discovery approach takes into account a certain cluster C, focusing on the IP of the path and using a parameter ℓ that indicate the minimum absolute frequency required for considering a certain transition sufficiently significant. Starting from the dataset of all patients of the cluster C, the *Hybrid Activity Tree Miner (HATM)* is built as follows:

1. Let $\mathbb{G}_C$ be the HAT of the cluster C, initially equal to $(\{S\}, \emptyset)$, where S is a node denoting the start of the IP. Let $\wp$ indicate the node on which we are positioned.
2. For each trace Ψ of cluster C with the uniformed notation introduced in the pre-processing phase, let $\Sigma = (\sigma_1, \ldots, \sigma_m)$ be its sub-trace corresponding to the IP and let $\wp$ be positioned on the root node S.
3. For each activity σ_i, for i from 1 to m, we check if exists a transition from $\wp$ labelled with σ_i. If it exists we increase of one the weight of the edge connecting the two nodes, otherwise we add a node with label σ_i and a transition from $\wp$ to the new node.
4. If $i < m$, we set $\wp$ on the existing or new node with label σ_i and we go to step 3. Otherwise, if exists other traces in C, we go to step 2. An iteration of steps 2–4 is depicted in Fig. 11.
5. We set $\wp$ on S and, for each outgoing edge $e \in \mathbb{T}$, we check if its frequency $f_e \geq \ell$. In positive case, we iterate the check for each son node, otherwise we mark that node.
6. For each marked node of $\mathbb{G}_C$ we prune the sub-tree τ in its correspondence and we connect the tree in that point with a STAG γ built in such a way that:
   - if exists at least one node labelled with a certain activity in τ, then a unique node is inserted in γ with that label;

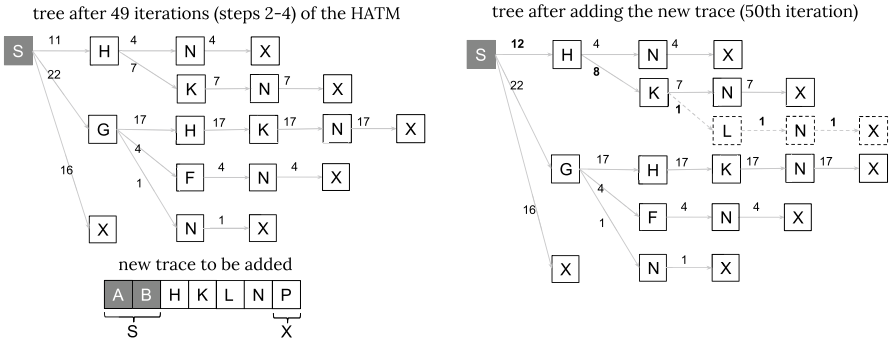

Fig. 11 The example shows how the new trace ABHKLNP is added to the tree by the HATM during the 50-th iteration: after converting the trace into SHKLNX, it is added following the common initial path SHK (increasing the edge labels) and adding the new branch LNX

- if exists at least one edge between from one of the nodes with label L to one of the nodes with label L' in τ, then a unique edge with the same direction is inserted in γ between the node with the label L and the one with label L';
- weights of edges in γ are computed as sum of all the weights on edge in τ having same labels to the connected nodes;

7. for each node of $\mathbb{G}_C$ that is not part of a STAG, if two ore more STAGs are connected to that node, then they are merged and weights on edges are summed. An example is depicted in Fig. 12.

The HATM is a process discovery algorithm that guarantee the 100% of fitness. Since patient paths are added one by one to the tree and to the STAG, each trace of the generating event log is full replicable in the discovered model.

We call *Hybrid Activity Forest (HAF)* a set of HATs that model the behaviour of different clusters $C_1, \ldots, C_l$ of patients. Let $\Gamma = \{C_1, \ldots, C_9\}$ and $\Gamma' = \{C'_1, \ldots, C_{18}\}$ be the sets of the partitions obtained through the two clusterings performed in the phase 1 of our approach. We generate 6 different HAFs taking into account Γ or Γ' and fixing $\ell \in \{1, 30, 100\}$.

Table 7 reports the main characteristics of the mined process models using the HATM implemented in C++. Fixing $\ell = 1$, pure tree models are obtained, which are over-fitted models able to replicate all but only the traces of the event log. These models allow us to have always memory of the activities previously performed. The pure tree models provide a high number of nodes, that is not good to understand the behaviour of the process, but could be used without problems of computational efficiency because of the tree structure, which avoid cycles and allows a simple calculation of frequency of a certain event. Observe that the average number of pure tree nodes in a single HAT is an index of the simplicity of the process model.

More generally, models generated with higher values of ℓ have a higher percentage of traces of the mined event log that are replicable in the STAGs and a lower number of nodes on the tree, which allows us to better understand the main path

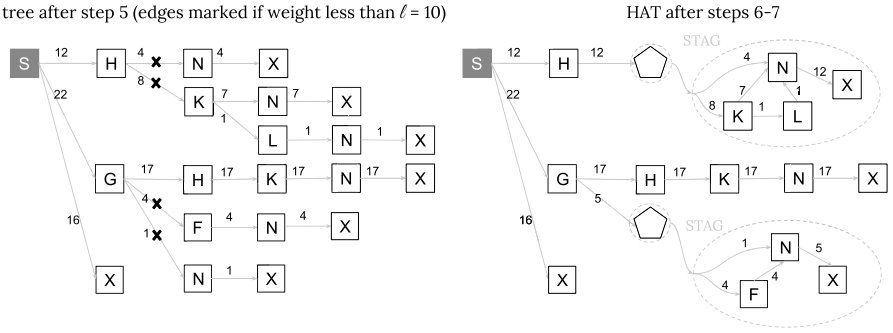

Fig. 12 The example shows how the tree obtained at the end of step 4 is pruned and how the STAGs are created from the pruned subtrees. For instance, the 3 edges N→X of the top subtree are merged in a single edge N→X of the STAG with a label 12, which is equal to the sum of their weights

Table 7 Characteristics of the HAFs using different clusters and values of ℓ

Name	Clustering	ℓ	Average number of pure tree nodes in a single HAT	Mined traces totally replicated on tree nodes		Comp. time (s)
				(Number)	(Percentage) (%)	
$\mathcal{F}_1$	Γ	1	5311	66,551	100.0	3.8
$\mathcal{F}'_1$	Γ'	1	2884	66,551	100.0	3.6
$\mathcal{F}_{30}$	Γ	30	55	55,956	84.1	3.7
$\mathcal{F}'_{30}$	Γ'	30	35	52,982	79.5	3.3
$\mathcal{F}_{100}$	Γ	100	24	51,186	76.9	3.7
$\mathcal{F}'_{100}$	Γ'	100	14	46,804	70.3	3.3

executed by the patients of the clusters. A slightly improvement is given using the clustering Γ instead of Γ'. However, lower dimensions of the tree mean also less precision and more generalisation. The HATM required always less than 4 s of computational time for each parameters combination. Figures 13, 14, 15 and 16 show the differences of using different values of the parameter ℓ, for two clusters that are equals for both clustering Γ and Γ'.

In Figs. 13 and 14 two different models are discovered for the paths of patients with dyspnea arrived at the ED in a weekday with their own means, in which thicker arrows indicates transitions with higher absolute frequencies. In this case the value $\ell = 100$ (Fig. 14) is too high to have a significant process model, because of the low number of patients in this cluster (1041 patients). The result is similar to that obtained for the Heuristic Net $\mathcal{H}$, but in this case we have two different simpler graphs denoted with a pentagon: one for patient that execute the activity G and one for all the others (explicated in the figure). On the contrary, for $\ell = 30$ (Fig. 13) the most common paths or the initial parts of them are easy deductible and different

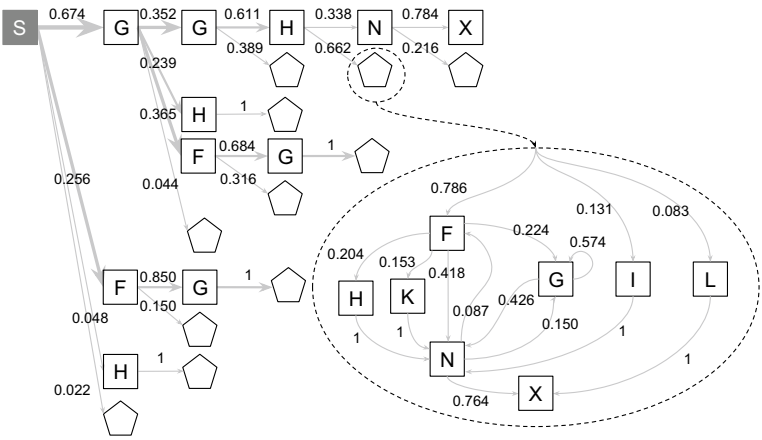

Fig. 13 HAT of cluster $C_2 = C'_2$ fixing $\ell = 30$

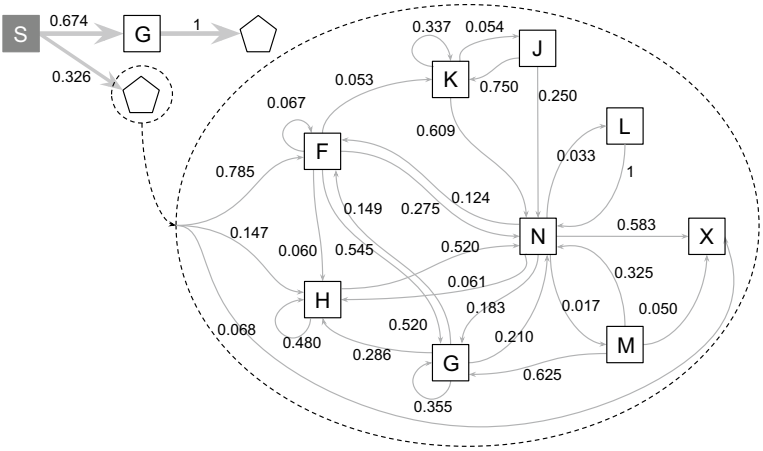

Fig. 14 HAT of cluster $C_2 = C_2'$ fixing $\ell = 100$

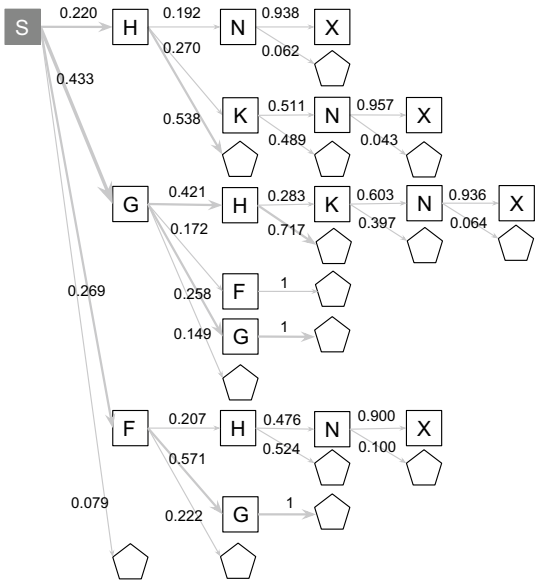

Fig. 15 HAT of cluster $C_4 = C_4'$ fixing $\ell = 30$

frequencies can be observed in different path evolutions. In this case, we have a high number of STAGs, but they are simpler. The same observations can be made for the cluster C_4, that is time-dependent trauma patients arrived by an ambulance (Fig. 15 for $\ell = 30$ and Fig. 16 for $\ell = 100$).

Models mined fixing $\ell = 30$ give us also information useful to make prediction of the next activities of a patient when he/her is waiting for a visit. We report an example of the type of prediction that can be made. Let us suppose to have 3 patients with the same urgency, π_1 of the cluster C_2 and π_3 and π_4 of the cluster C_4, which are waiting for a revaluation visit, occupying a scarce resource (e.g. a stretcher). Let us

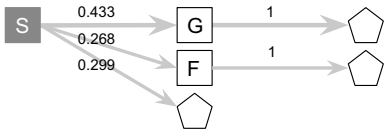

Fig. 16 HAT of cluster $C_4 = C'_4$ fixing $\ell = 100$

suppose they performed the activity sequences ABGGH, ABH and ABFH, respectively. This means that the π_1 is positioned before the unique node labelled with N in Fig. 13, π_2 is before the node labelled with N at the top of Fig. 15, and π_3 is on the other node with the same label on the bottom of the same model. In order to release stretchers as soon as possible, the model suggests to visit π_2 because the frequency of the discharge after the activity N is equal to 0.938, which estimates a higher probability of discharge compared to π_1 (0.784) or π_3 (0.900).

The discovered models could be used as follows. A HAT with $l = 30$ or $l = 100$ can be used by a simulation model to keep track of a patient path during the execution of its activities. Until that path is on the tree part of the HAT, it means that the historical data guarantees statistical relevance, then predictions about the further activities can be made starting from the same node of the correspondent HAT with $l = 1$ because of the greater precision of such a model.

4 Conformance checking

The quality of a process model is usually assessed using four quality criteria, that is fitness, precision, generality and simplicity (Buijs et al. 2014). The fitness and the simplicity of the HAFs are already discussed in Sect. 3.3 along the development of the proposed process models: the former is always at the maximum values (100%) while the latter depends on the clustering and the parameter ℓ.

In Sect. 4.1 we complete the conformance checking by evaluating the generality and the precision of the discovered HAFs. Further, we strengthen the conformance checking adding a *robustness* analysis consisting in the evaluation of the prediction capability of the discovered HAFs in Sect. 4.2. In the following analysis, we test the HAFs mined from the 2013–2015 dataset over the 2016 dataset, and labelling the patients from the 2016 dataset in accordance with the clusters Γ and Γ' obtained by the 2013–2015 dataset.

4.1 Generality and precision

We implement a conformance checking algorithm that, given as input a new event log E and a HAF model $\mathcal{F}$ returns a generality index g and a precision index p. The former checks how many traces of E are replicable using $\mathcal{F}$ and is defined as follows:

$$g = \frac{\text{number of traces in } E \text{ totally replicable in } \mathcal{F}}{|E|}.$$

The latter is a measure of how many traces generable from the model $\mathcal{F}$ represent behaviour that can occur in reality. We remark that, to the best of our knowledge, all wrong behaviour has been avoided a priori through the pre-processing phase and by the model implementation, that is the compliance with the framework in Fig. 2. However, some sequences of Tests & Care activities could be not possible in the real ED process, then we estimate a lower bound of traces allowed in the reality generating a set T of 10,000 traces from $\mathcal{F}$ in accordance with the clusters and edges probabilities, and then we check if they are contained in the event log D used for the process discovery. Then, the precision index p is computed as follows:

$$p = \frac{|T \cap D|}{|D|}.$$

As new event log E, we used the event log obtained applying the pre-processing algorithm (discussed in Sect. 3.1) to the ED dataset of the 29,155 patients arrived at the ED during the year 2016. Traces of the 2016 dataset is partitioned using the same decision tree used for Γ or Γ', depending on the considered model. Table 8 reports the conformance indices g and p for the 6 discovered process models.

As expected, models $\mathcal{F}_1$ and $\mathcal{F}'_1$ have the worst generality because of the overfitting of the event log used for the process discovery without adding any generalisation for other behaviour. Increasing the value of ℓ, we obtain higher values of g, close to the 100% when $\ell = 100$, while using the two clustering Γ and Γ' there is not a significant difference. Models with $\ell = 1$ have obviously the 100% of precision, because they allow us to replicate all but only traces in the event log D given as input. Increasing the value of ℓ the gain of generality involves a decrease of precision. However more of the 85% is guaranteed for the defined models, with a slightly improvement using the clustering Γ instead of Γ'.

4.2 Robustness

Section 4.1 shows that paths generated using the HAFs represent a behaviour compliant with the actual ED process. In order to provide a prediction tool, we are required to guarantee that the probabilities of the occurrence of the predicted events

Table 8 Conformance checking: values of generality and precision indices

Model	Clustering	ℓ	Generality		Precision	
			Traces replicated by $\mathcal{F}$	g (%)	Traces in D	p (%)
$\mathcal{F}_1$	Γ	1	26,289	90.17	10,000	100.00
$\mathcal{F}'_1$	Γ'	1	25,895	88.82	10,000	100.00
$\mathcal{F}_{30}$	Γ	30	28,517	97.81	8903	89.03
$\mathcal{F}'_{30}$	Γ'	30	28,353	97.25	8735	87.35
$\mathcal{F}_{100}$	Γ	100	28,913	99.17	8871	88.71
$\mathcal{F}'_{100}$	Γ'	100	28,828	98.88	8582	85.82

Table 9 Comparison between the frequencies of several events in 2013–2015 using the HATs of $\mathcal{F}_1$ and $\mathcal{F}'_1$, and real data of 2016

Cluster	a^{13-15} (%)	a^{16} (%)	f_H^{13-15}	f_H^{16}	$f_{H<N}^{13-15}$	$f_{H<N}^{16}$	f_X^{13-15}	f_X^{16}
$C_1 = C'_1$	1.01	1.12	0.893	0.871	0.566	0.675	0.023	0.036
$C_2 = C'_2$	1.52	1.69	0.928	0.931	0.764	0.723	0.011	0.005
$C_3 = C'_3$	1.26	1.19	0.830	0.852	0.732	0.734	0.013	0.011
$C_4 = C'_4$	2.21	2.89	0.900	0.913	0.793	0.816	0.018	0.033
$C_5 = C'_5$	1.70	1.94	0.781	0.783	0.717	0.713	0.062	0.089
$C_6 = C'_6$	27.81	27.50	0.680	0.691	0.666	0.676	0.178	0.172
$C_7 = C'_7$	1.11	1.22	0.482	0.615	0.352	0.511	0.091	0.104
C_8	2.56	3.04	0.713	0.745	0.555	0.631	0.012	0.016
C_9	60.80	59.41	0.353	0.386	0.308	0.345	0.139	0.127
C'_8	1.48	1.68	0.742	0.791	0.580	0.662	0.013	0.013
C'_9	1.08	1.36	0.673	0.689	0.520	0.593	0.011	0.020
C'_{10}	2.81	2.32	0.045	0.037	0.039	0.035	0.164	0.135
C'_{11}	1.97	1.34	0.007	0.000	0.006	0.000	0.381	0.535
C'_{12}	1.47	1.48	0.063	0.061	0.048	0.049	0.098	0.141
C'_{13}	7.31	7.55	0.572	0.585	0.490	0.515	0.037	0.023
C'_{14}	1.33	1.46	0.712	0.743	0.609	0.635	0.059	0.050
C'_{15}	10.46	9.17	0.323	0.329	0.316	0.324	0.276	0.269
C'_{16}	1.02	0.84	0.050	0.048	0.040	0.037	0.075	0.080
C'_{17}	2.13	1.71	0.126	0.180	0.105	0.151	0.166	0.138
C'_{18}	32.30	33.55	0.384	0.413	0.326	0.366	0.107	0.098

are robust enough with respect to the relative frequencies of the same events in accordance with the 2016 dataset.

In Table 9 we report frequencies of different events related to the patient paths computed with the HATs of $\mathcal{F}_1$ and $\mathcal{F}'_1$ obtained from the event log of the period 2013–2015 and we compare such values with the same frequencies of 2016. We remark that results are the same for the HATs of the two models when the clusters are equal, as reported in the first 7 rows of the table.

Columns denoted with a_{13-15} and a_{16} report the percentage of patients belonging to the clusters over the total. These results do not indicate significant variation of the cluster dimensions over time. The frequencies of executing at least one time the X-ray exams within the path are indicated with f_H^{13-15} and f_H^{16}, showing important differences in different clusters: for instance, a patient in C'_2 has a probability greater of 90% to make such an activity, while a patient in C'_{11} has a probability close to 0%. The difference of such frequencies between the period 2013–2015 and 2016 are very low, always under the 5%, except for the cluster $C_7 = C'_7$, which is one of the smaller clusters, with a difference of 13.3%. Columns indicated with $f_{H<N}^{13-15}$ and $f_{H<N}^{16}$ report the frequencies of executing the X-ray exams before the first revaluation visit, that are slightly lower than f_H^{13-15} and f_H^{16}, as expected. Also in this case the frequencies of 2013–2015 and 2016 are very similar, with an average difference of 3.8% and maximum 15.9% for the cluster $C_7 = C'_7$. The last two columns f_X^{13-15} and f_X^{16}

indicate the frequencies of a Discharge class activity immediately after the first visit. Also in this case, values vary for the different clusters, from value next to 0 up to the 53.8%. The average difference between 2013–2015 and 2016 is around 2%.

Observe that the clustering Γ' provides more detailed information with respect to Γ that could be useful making predictions. For instance, C'_{11} and C'_{14} are both subsets of C_9, but they have very different frequencies for the events reported in Table 9. Finally, no relevant differences in robustness for the clustering Γ and Γ' have been emerged from this analysis.

5 Conclusions

Although a flowchart of the ED process can be easily designed interviewing the ED staff, the high complexity and variability of the patient paths do not allow us a modelling without making significant assumptions. Such simplifications significantly impact on the replicability of the simulation model used to identify bottlenecks and to analyse policies to alleviate the overcrowding.

In this paper we propose an ad hoc process mining approach to discover a model capable to replicate the patient paths and to predict their possible evolutions over time. This requirement is due to the future need of implementing a simulation model for the evaluation of the real time allocation of the resources in order to reduce overcrowding. To this end, we would discover a fine-grained patient flow model satisfying the four main quality criteria, that is fitness, precision, generality and simplicity, which is a challenging task. Models mined with the application of standard process discovery approaches to the dataset of our case study differ a lot from such requirements.

Therefore we present an ad hoc approach divided into two phase. The first consists in the application of the Decision Tree to identify a clustering of patients with respect to their sequence of test and treatment activities after the first medical visit. Such clusters are then used in the second phase to build process models called Hybrid Activity Trees, which use a tree-structure to describe main paths and graphs to represent infrequent behaviour. The minimum frequency to consider sufficiently frequent a certain path evolution is defined by the parameter ℓ of the proposed algorithm.

Results prove the adequacy of the proposed approach in accordance with the above requirements. Clustering gives important insights to identify different behaviour depending on the patient characteristics. Then the conformance of the model is guarantee under two perspectives. Firstly, setting ℓ equal to 30 or 100 and taking into account a different dataset, almost the 100% of its traces are replicable. Furthermore, fixing $\ell = 1$, the frequency of several analysed events in our models is consistent in accordance with the paths of the such a dataset.

The conformance analysis suggests that we are now capable to develop a simulation model based on the discovered process models. Fixing ℓ equal to a value sufficient to guarantee statistical relevance, the Hybrid Activity Trees allow us to know the possible main behaviour depending of the already performed activities. As long as the patient remains within the main paths, we can use the

corresponding Hybrid Activity Trees with $\ell = 1$ in order to estimate probability of some events in real time during the treatment of the patients in accordance with their paths.

The next step of our research collaboration with the ED of Cantù will be the application of the proposed ad hoc process mining approach. We are developing a simulation model capable to represent the path of each single patient, exploiting the knowledge of the HAF to support real time decision making. Our purpose is to evaluate the impact of online optimisation methods to reduce the overcrowding and, more generally, to improve the ED management.

Future research could consider the application of the overall approach described in this paper to an ED having different characteristics of the ED on Cantù (e.g., dimension, organisation, patient population,...) for a further validation. From a methodological point of view, it could be interesting to investigate the use of different clustering techniques in the phase 1 of our approach in order to evaluate the impact on the quality criteria and on the robustness of our models.

Acknowledgements The authors wish to thank Alessandra Farina, Elena Scola and Filippo Marconcini of the ED at *Ospedale Sant'Antonio Abate di Cantù* for the fruitful collaboration and for providing us the dataset and allowing their use in this paper. The authors wish to thank the anonymous reviewers for their accurate reports and valuable suggestions.

References

Abo-Hamad W (2017) Patient pathways discovery and analysis using process mining techniques: an emergency department case study. In: Cappanera P, Li J, Matta A, Sahin E, Vandaele NJ, Visintin F (eds) Health care systems engineering. Springer, Cham, pp 209–219

Abouelji nane L, Sahin E, Jemai Z (2013) A review on simulation models applied to emergency medical service operations. Comput Ind Eng 66:734–750

Alvarez C, Rojas E, Arias M, Munoz-Gama J, Seplveda M, Herskovic V, Capurro D (2018) Discovering role interaction models in the emergency room using process mining. J Biomed Inform 78:60–77

Aringhieri R, Bruni M, Khodaparasti S, van Essen J (2017) Emergency medical services and beyond: Addressing new challenges through a wide literature review. Comput Oper Res 78:349–368, advance online publication 22 September 2016

Aringhieri R, Dell'Anna D, Duma D, Sonnessa M (2018) Evaluating the dispatching policies for a regional network of emergency departments exploiting health care big data. In: Nicosia G, Pardalos P, Giuffrida G, Umeton R (eds) International conference on machine learning, optimization, and big data. Lecture notes in computer science, vol 10710. Springer, pp 549–561, advance online publication 21 December 2017

Basole R, Braunstein M, Kumar V, Park H, Kahng M, Chau D, Tamersoy A, Hirsh D, Serban N, Bost J, Lesnick B, Schissel B, Thompson M (2015) Understanding variations in pediatric asthma care processes in the emergency department using visual analytics. J Am Med Inform Assoc JAMIA 22(2):318–323

Buijs J, van Dongen B, van der Aalst WMP (2014) Quality dimensions in process discovery: the importance of fitness, precision, generalization and simplicity. Int J Cooperative Inf Syst 23(1):1440,001/1–39

Calvello EJB, Broccoli M, Risko N, Theodosis C, Totten VY, Radeos MS, Seidenberg P, Wallis L (2013) Emergency care and health systems: consensus-based recommendations and future research priorities. Acad Emerg Med 20(12):1278–1288. https://doi.org/10.1111/acem.12266

Cildoz M, Mallor F, Ibarra A, Azcarate C (2017) Dealing with stress and workload in emergency departments. In: Cappanera P, Li J, Matta A, Sahin E, Vandaele NJ, Visintin F (eds) Health care systems engineering. Springer, Cham, pp 297–298

Derlet R (2002) Overcrowding in emergency departments: increased demand and decreased capacity. Ann Emerg Med 39(4):430–432

Derlet R, Richards J (2000) Overcrowding in the nation's emergency departments: complex causes and disturbing effects. Ann Emerg Med 35(1):63–68

Duma D, Aringhieri R (2017) Mining the patient flow through an Emergency Department to deal with overcrowding. In: 3rd international conference on health care systems engineering. Springer International Publishing AG, Springer Proceedings in Mathematics and Statistics, vol 210 **(to appear)**

Feng YY, Wu IC, Chen TL (2017) Stochastic resource allocation in emergency departments with a multi-objective simulation optimization algorithm. Health Care Manag Sci 20(1):55–75

Fitzgerald J, Dadich A (2009) Using visual analytics to improve hospital scheduling and patient flow. J Theor Appl Electron Commer Res 4(2):20–30

George F, Evridiki K (2015) The effect of emergency department crowding on patient outcomes. Health Sci J 9(1):1–6

Hoot N, Zhou C, Jones I, Aronsky D (2007) Measuring and forecasting emergency department crowding in real time. Ann Emerg Med 49(6):747–755

Hwang U, Concato J (2004) Care in the emergency department: how crowded is overcrowded? Acad Emerg Med 11(10):1097–1101

Koyuncu M, Araz OM, Zeger W, Damien P (2017) A simulation model for optimizing staffing in the emergency department. In: Cappanera P, Li J, Matta A, Sahin E, Vandaele NJ, Visintin F (eds) Health care systems engineering. Springer, Cham, pp 201–208

Kuo YH, Leung J, Graham C (2012) Simulation with data scarcity: developing a simulation model of a hospital emergency department. In: Proceedings—Winter Simulation Conference

Leemans S, Fahland D, van der Aalst W (2014) Discovering block-structured process models from event logs containing infrequent behaviour. Lecture notes in business information processing (LNBIP), vol 171, pp 66–78

Luscombe R, Kozan E (2016) Dynamic resource allocation to improve emergency department efficiency in real time. Eur J Oper Res 255(2):593–603

Mans R, Van Der Aalst W, Vanwersch R, Moleman A (2013a) Process mining in healthcare: data challenges when answering frequently posed questions. Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics) (LNAI), vol 7738, pp 140–153

Mans R, Van Der Aalst W, Vanwersch R, Moleman A (2013b) Process mining in healthcare: data challenges when answering frequently posed questions. Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics) (LNAI), vol 7738, pp 140–153

Partington A, Wynn M, Suriadi S, Ouyang C, Karnon J (2015) Process mining for clinical processes: a comparative analysis of four australian hospitals. ACM Trans Manag Inf Syst 5(4):19.1–19.18

Paul S, Reddy M, Deflitch C (2010) A systematic review of simulation studies investigating emergency department overcrowding. Simulation 86(8–9):559–571

Rebuge A, Ferreira D (2012) Business process analysis in healthcare environments: a methodology based on process mining. Inf Syst 37(2):99–116

Rojas E, Munoz-Gama J, Sepúlveda M, Capurro D (2016) Process mining in healthcare: a literature review. J Biomed Inform 61:224–236

Rojas E, Fernandez-Llatas C, Traver V, Munoz-Gama J, Sepúlveda M, Herskovic V, Capurro D (2017a) Palia-er: bringing question-driven process mining closer to the emergency room. In: CEUR workshop proceedings, vol 1920

Rojas E, Sepúlveda M, Munoz-Gama J, Capurro D, Traver V, Fernandez-Llatas C (2017b) Question-driven methodology for analyzing emergency room processes using process mining. Appl Sci (Switzerland) 7(3):302

Sinreich D, Jabali O, Dellaert N (2012) Reducing emergency department waiting times by adjusting work shifts considering patient visits to multiple care providers. IIE Trans (Inst Ind Eng) 44(3):163–180

Suriadi S, Andrews R, ter Hofstede A, Wynn M (2017) Event log imperfection patterns for process mining: towards a systematic approach to cleaning event logs. Inf Syst 64:132–150

Van Der Aalst W, Van Dongen B, Gnther C, Rozinat A, Verbeek H, Weijters A (2009) Prom: the process mining toolkit. In: CEUR workshop proceedings, vol 489

Weijters A, Ribeiro J (2011) Flexible heuristics miner (FHM). In: IEEE SSCI 2011: symposium series on computational intelligence–CIDM 2011: 2011 IEEE symposium on computational intelligence and data mining, pp 310–317

Yeh JY, Lin WS (2007) Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department. Expert Syst Appl 32(4):1073–1083

Davide Duma is a post-doctoral researcher at the Computer Science Department of University of Turin, Italy. He holds a B.S. in Mathematics and Computer Science from University of Salento, a M.S. in Mathematics from University of Turin, and a Ph.D. in Computer Science from the University of Turin. His main research interest focuses on the quantitative analysis and decision support systems for health care management.

Roberto Aringhieri is Associate Professor of Operations Research at the Computer Science Department of the University of Turin, Italy. He holds a M.S. in Computer Science and a Ph.D. in Mathematics for Economic Decisions and Operations Research from the University of Pisa. He is co-chair of the EURO Working Group on Operational Research Applied to Health Services. He is Associate Editor of Operations Research for Health Care. His main research interest focuses on simulation and optimization algorithms applied to the management of health services.