Online optimization methods applied to the management of health services

Davide Duma

October 12, 2018

Università degli Studi di Torino

Doctoral School of Sciences and Innovative Technologies
Doctoral in Computer Science

Doctoral thesis

Online optimization methods applied to the management of health services

Davide Duma

Supervisor Roberto Aringhieri

Reviewers Paola Cappanera
Università degli Studi di Firenze

Stefan Nickel
Karlsruhe Institute of Technology

October 12, 2018

Davide Duma
Online optimization methods applied to the management of health services
Doctoral thesis, October 12, 2018
Reviewers: Paola Cappanera and Stefan Nickel
Commettee: Paola Cappanera, Stefan Nickel and Elena Tànfani
Supervisor: Roberto Aringhieri

Università degli Studi di Torino
Doctoral in Computer Science
Doctoral School of Sciences and Innovative Technologies
Department of Computer Science
Corso Svizzera, 185
10149 – Torino (Italy)

Acknowledgement

My biggest and most heartfelt thanks goes to Roberto Aringhieri, who guided me and supported me during my first steps in the stimulating world of research. I could not have asked for a better PhD supervisor, for many reasons including his scientific value, the trust that he has always placed in my ideas and especially for his human value. I really appreciate his efforts in searching for the right directions to be taken in the research, looking forward also to what will be my career after my PhD.

I would thank the reviewers Stefan Nickel and Paola Cappanera for the precious time they spent on my doctoral thesis and for their remarks that allowed me to improve my work and suggested new developments. Together with Elena Tànfani, I thank them for coming in Torino and taking part in the doctoral committee.

Not much of this dissertation would have been possible without the fruitful collaboration with Alessandra Farina, Elena Scola and Filippo Marconcini of the Emergency Department Sant'Antonio Abate of Cantù. They provided data useful for the development of an important part of this thesis, but I would express my gratitude especially for introducing me with great willingness and professionalism to the complex world of the Emergency Department management, patiently providing the details of their work context and welcoming my proposals with enthusiasm and openness. My sincere thanks also go to Michele Sonnessa, Davide Dell'Anna, Luigi Casciaro and Simone Bocca for their contibution that enriched this thesis.

I would thank all my family for their fundamental support throughout my studies, which involved many sacrifices and an unpleasant distance. Last but not less important are the thanks to Samuele, a special presence in these years.

Davide Duma

Torino, October 2018

Contents

Research motivation and outline 1

> *The goal of real health care reform must be high-quality, universal coverage in a cost-effective way.*
>
> — **Bernie Sanders**

Global health broadly refers to "*an area for study, research, and practice that places a priority on improving health and achieving equity in health for all citizens*". The goal of achieving equity in health, namely the absence of systematic disparities in health or in the major social determinants of health between groups with different levels of underlying social advantage or disadvantage, has become more and more important in the years [17].

The rising costs of health care due to new technologies and demographic trends (in particular, the aging population), is a vitally important issue for health care policy makers. At the same time there is a paradigm shift in the service concept of health care. Patients are no longer prepared to accept poor quality service, either in terms of long waiting times or inconvenient appointment systems, and expect that services are well organized from a "customer" perspective. The service concept has shifted from optimizing the use of resources to finding a balance between service for patients and efficiency for providers [18].

1.1 Health care delivery

The Health Care Delivery is the process in charge of providing a health service while a Health Care Delivery System is the organization in charge of the management of several delivery processes. In order to deal with the above issues (equity, rising cost, more informed people, ...), the new way of thinking and organizing the health care delivery is to focus on the patient instead of only on the facilities.

A patient-centered approach to health care means to deliver a service which is "*closely congruent with and responsive to patients' wants, needs, and preferences*" [38]. In May 2004, the International Alliance of Patients' Organizations (IAPO) conducted a consultation with its member patients' organizations in order to investigate which health care policy issues were of most importance to them. The final report shown that the 74% of respondents indicated that "*defining patient-centered health care was very relevant to their organization*" [43].

Among the many fields where Operations Research and computers meet, health care delivery is surely one of the more vital nowadays. Health care delivery is a very relevant topic not only from the point of view of researchers, scholars and practitioners but also for the impact on public opinion and for fueling large discussions and debates [4]. The use of Operations Research in health care delivery has developed considerably over the years as documented, for instance, by the number of the special issues appeared in the last fifteen years (see, e.g., [4, 5, 6, 20, 21, 24, 34, 41, 47, 57]) and by the number of general or specific literature reviews appeared in the last years (and reported in this introduction and in the remaining of the thesis).

This increasingly interest of Operations Research in health care delivery is due to a number of reasons: health care has become a major industry, with many people involved either as employees in health care delivery organizations or as consumers of health care services. The characteristics of Operations Research in health care delivery – which make it different from Operations Research in industry or in commercial services – stem from the way health care organizations operate and from the type of health care system in use in a particular country [18].

The current development of the health care delivery is aimed to recognize the central role of the patient as opposed to the one of the health care providers. In this context, the attention from a single health benefit can be shifted to the whole health care chain thanks to Clinical Pathways (CPs), which are defined as "health-care structured multidisciplinary plans that describe spatial and temporal sequences of activities to be performed, based on the scientific and technical knowledge and the organizational, professional and technological available resources" [19]. A CP can be conceived as an algorithm based on a flowchart that details all decisions, treatments, and reports related to a patient with a given pathology, with a logic based on sequential stages [23]. A CP is therefore "the path" that a patient suffering from a disease walks in the National Health System. This path can be analyzed at a single local level of care (e.g. a single hospital, or a geographic area) or globally, taking into account every level of health-care from the diagnosis of diseases, treatment and recovery. CPs are specifically tailored to stimulate continuity and coordination among the treatments given to the patient through different disciplines and clinical environments, focusing on the patient perspective [44]. The aim of a CP is to enhance the quality of care by improving patient outcomes, promoting patient safety, increasing patient satisfaction, and optimizing the use of resources as stated by the European Pathway Association. Moreover, while many studies show that, appropriately implemented, CPs have the potential to increase patient outcome, reduce patient length of stay and limit variability in care, thereby yielding cost savings [48], little attention has been dedicated to study how CP can optimize the use of resources as reported by Addis et al. [1].

1.2 Online optimization

Traditional optimization research assumes complete knowledge of all data of a problem instance. However, the assumption that all information necessary to define a problem instance is available beforehand is unlikely in reality, in fact decisions may have to be made before complete information is available. Online optimization is characterized by the development of algorithms whose decisions are based only on past events without any solid information about future data. More generally, online (semi-online) optimization differs from offline optimization by the fact that the instance is completely (partially) unknown at the beginning of the computation [29].

Online optimization problems consist of a finite sequence of requests $r_1, \ldots, r_n$ to serve that are revelead step to step. For this reason, such an approach is suitable for real-time problems having robust factors of uncertainty. Unlike stochastic optimization, in which decisions are taken a priori on the basis of probability distributions, online optimization involves the use of algorithms that enrich their knowledge over time with the arrival of new information and make decisions accordingly. Recent developments formalize the concept of lookahead information, that is a limited overseen amount of future input data that can be exploited during the online computation [27]. Since only partial information can be exploited in its computation, an online algorithm can not provide an optimal solution, which can be computed only when all the necessary information are known, then it is usually based on a simple approximation algorithm technique (e.g. greedy algorithm or dinamic programming).

Let us consider the bin packing problem, one of the most classic in operations research. In the traditional version, a set I of n items $r_1, \ldots, r_n$ with a fixed size $s_1, \ldots, s_n \in (0, 1]$ has to be assigned to an infinite set B of bins $b_1, b_2, \ldots$ having capacity 1. Each item must be assign to a bin and the sum of the item sizes assigned to a bin must not exceed its capacity. The objective is to minimize the number of bins used, that is how many bins have at least one item assigned. The online version of this problem differs in the knowledge of the amount and size of the items to be assigned [50]. At each step i, the size s_i of the i-th item is made available, then it must be assigned to a bin before knowing if there will be another item at the next step and its size. The availabily of information over time makes the problem more challenging, because there is not an algorithm able to give the optimal soluzion for every intance. For this reason, simple approximation algorithm are used. Let us consider two well-known greedy approaches for the bin packing problem, the First-Fit and the Best-Fit algorithms. The former assigns each item to the first bin with sufficient free capacity with respect to its size, while the latter assigns each item in such a way to minimize the free capacity of the selected bin. In Figures 1.1

and 1.2 the two algorithms are compared for two different instances: the Best-Fist perform better than the First-Fit for the instance I and worst for the instance I'.

Fig. 1.1: First-Fit vs. Best-Fit: solutions for the instance $I = \{i_1, \dots, i_4\}$, with sizes $s_1 = \frac{1}{2}, s_2 = \frac{2}{3}, s_3 = \frac{1}{3}, s_4 = \frac{1}{2}$.

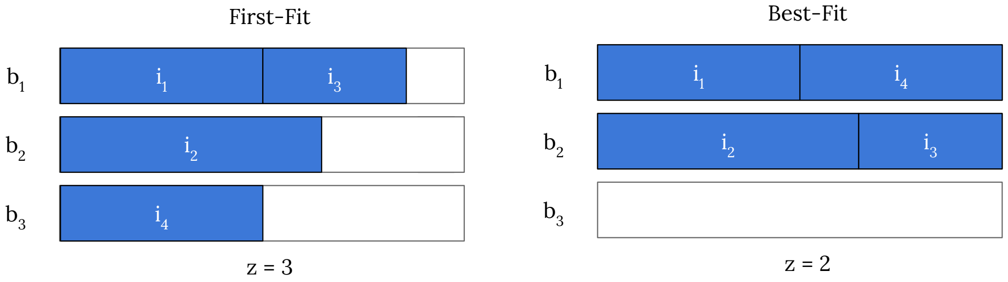
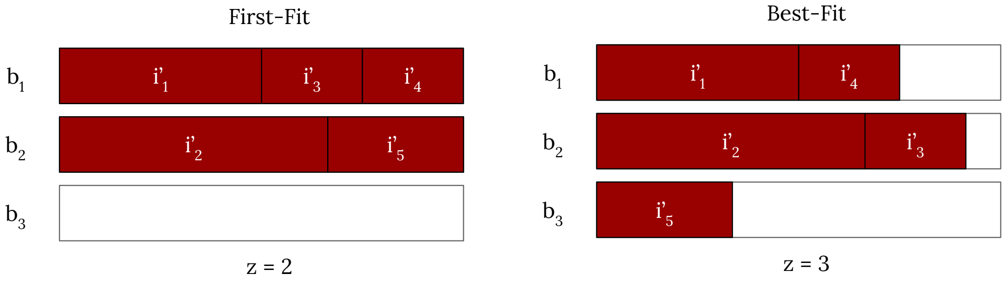


Fig. 1.2: First-Fit vs. Best-Fit: solutions for the instance $I' = \{i'_1, \dots, i'_5\}$, with sizes $s'_1 = \frac{1}{2}, s'_2 = \frac{2}{3}, s'_3 = \frac{1}{4}, s'_4 = \frac{1}{4}, s'_5 = \frac{1}{3}$.



In order to have a criterion for evaluating the quality of online algorithms, their solutions should be compared with those obtained by optimal offline algorithm(s). The standard approach for this comparison is the competitive analysis. Let $z^\star(I)$ and $z'(I)$ be respectively the value of the optimal offline and the value of the online solution for a given instance I of the optimization problem. In the case of a minimization problem, the online optimization algorithm is c-competitive if exist a real number $c \geq 1$ (called competitive ratio) such that $z'(I) \leq cz^\star(I)$ for each instance I. In other words, the competitive analysis is a sort of extension of the classical worst case analysis [51].

However, the competitive analysis has several drawbacks typical of the worst case analysis. Firstly, real world applications could have a high degree of complexity due to large number of dependent stochastic processes, which makes really complex the identification of the worst case. Then, a proper objective function in real world systems could have with different goals that are difficult to gather in a single multi-objective function for computing a meaningful competitive ratio, besides the fact that sometimes it could seem meaningless comparing an online algorithm with an omniscient offline algorithm. Furthermore, many applications need to have a greater effectiveness on the average case and, addressing real-time decisions, to satisfy further requirements such as computational efficiency. Starting from this limitations, alternative measures of efficiency for online algorithms are proposed and discussed in literature (see, e.g., [10, 11, 14, 15, 16, 25, 36, 37]). Starting from an analysis of such measures, Dunke and Nickel [26] propose a framework based on discrete event simulation to address the issue of modeling a complex real world system in which several stochastic processes are involved. In such an environment, online optimization algorithms can be tested on a large number of different instances of the problem in order to provide a quantitative analyisis based on the average case, which can be used to estimate a set of performance indices for the decision support.

From the point of view of the applications, real world systems are inherently complex and often contain optimization problems exhibiting online characteristics. The book of Grotschel et al. [32] illustrates many applications of the online optimization in large scale systems such as chemical engineering, robot control, and transportation. Online optimization has been applied also in health care. The majority of the contributions are related to the appointment scheduling problem to deal with unattended arrivals, overbooking and no-show patients [7, 12, 39, 40, 42, 49, 52, 55, 56, 58, 59]. Other specific applications are in online scheduling of chemotherapy and nuclear medicine [33, 45], dynamic transportation [9], and scheduling pick-up and delivery tasks in hospitals [30]. Finally, applications of online optimization in operating room, emergency medical service and emergency department are reported in detail in the next chapters.

1.3 Online optimization methods applied to the management of health services

The most challenging aspect in health care delivery stems from the high complexity of the system itself, its intrinsic uncertainty and its dynamic nature. Their management requires not only the expertise to analyze and to understand a large amount of information but also to organize that information on a cognitive base for adequate decision making and to promote a collaboration with researchers in other areas, such as doctors and economists.

Dealing at the same time with the complexity, the uncertainty and the dynamics of health care delivery problems requires the adoption of unconventional solution methodologies [4]. As a matter of fact, many deterministic and static optimization problems in health care delivery are $\mathcal{NP}$-hard and, by consequence, adding the uncertainty and the dynamic aspects makes them more challenging. The health care literature (see, e.g., [2, 3, 8, 13, 17, 22, 28, 31, 46, 53, 54]) shows off several and different approaches to deal with the uncertainty and the dynamics, such as chance-constrained programming, two stage stochastic programming with recourse, robust optimization, discrete event simulation, agent-based simulation, system dynamics, Monte Carlo simulation.

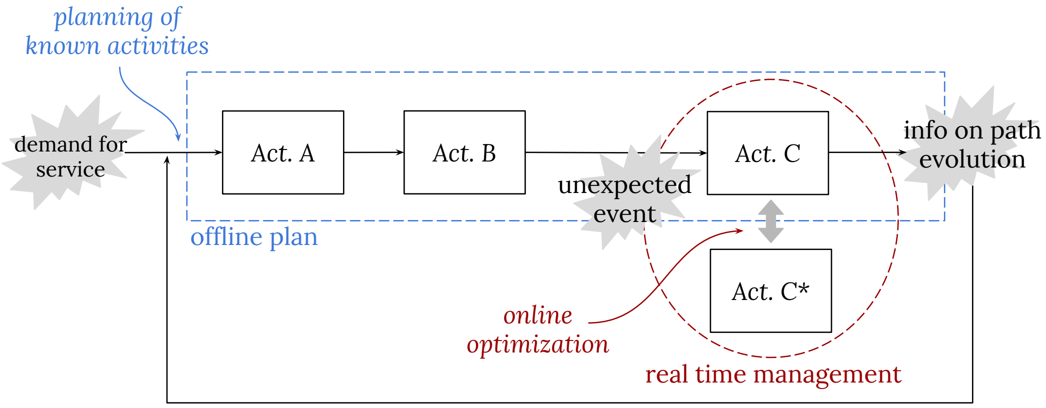
In this thesis, we would propose a different approach based on the online optimization methodology to manage the uncertainty and the dynamics of the health care delivery problem under consideration. Starting from a given offline planning solution (when it exists), the basic idea is to fix such a solution in real time (dynamics) as soon as an unattended event will occur (uncertainty) exploiting the available knowledge of the underlying clinical pathway (which could be static or dynamic).

We identified three health care delivery problems to illustrate and develop our approach. The three problems belong to two different clinical pathways, which are the surgical pathway and the emergency care pathway. The first problem arises in the context of Operating Room Planning (ORP), which is characterized by (i) a well-structured but complex pathway, and (ii) by several sources of uncertainty such as the arrival of unattended patients to be operated on, and the duration of a surgery and the length of stay. The second problem arises in the context of Emergency Medical Service (EMS) management, which is characterized by (i) a well-structured but simple pathway, and (ii) by several sources of uncertainty such as the arrival of unattended phone calls asking for an emergency requests to be served as soon as possible (depending on the level or urgency) by a not always available ambulance. Finally, the third problem arises in the context of Emergency Department (ED) management, which is characterized by (i) a non-structured but complex pathway, and (ii) by several sources of uncertainty such as the arrival of unattended patients to be served as soon as possible (depending on the level or urgency) by a possible overcrowded system, and the dynamic path evolution.

The ORP and the EMS management are *lasagna processes*, which is typical of well-structured pathways: the sequence of activities to be performed is known at the beginning of the CP and the possible path evolutions are limited. On the contrary, the ED management is a *spaghetti process*, which is typical of non-structured pathways: a large variety of path evolutions are possible and the sequence of activities to be performed is part itself of the lack of information.

A general CP is illustrated in Figure 1.3, in which white boxes with continuous border (e.g. Act. A) indicates medical activities and gray irregular shapes indicates unpredictable events that occur in a certain moment. The first of such events is the demand for a service, which is an important factor of uncertainty if that service should be provided in a very short time, such as a non-elective surgery, the request of an ambulance or the arrival of a patient at the ED. Then, one or a sequence of activities are planned on the basis of the available information, whose planning horizon changes in accordance with the knowledge of the CP and the time available to execute the activities. For instance, the planning could be the assignment of a time slot for a surgery or a sequence of scheduled activities at the ED. To deal with uncertainty, the resulting offline plan is then supervised by the real time management during its execution: if an unexpected event prevents its exact fulfillment, then online optimization is used to take decisions based on new information acquired and change the offline plan. For example, a surgery (Act. C) can be re-assigned to another time slot (becoming Act. C*). Further, the evolution of some CPs (e.g. the treatment of a patient in the ED) is discovered over time, then new activities need to be performed and the framework is reiterated.

Fig. 1.3: Example of CP: information available over time and online optimization.



Since the competitive analysis cannot be easily applied to the evaluation of our algorithms due to the complicated nature of the considered problems, an alternative evaluation framework is required. Exploiting the discrete structure of the problems under considerations, we evaluate the quality of our online solutions within a Discrete Event Simulation (DES) framework in accordance with Dunke and Nickel [26]. Basically, we adopt the DES to replicate the operative context in which the candidate algorithms operate using real world or realistic data. Further, it allows us to evaluate their impact over time, that is how the previous decisions can impact on the current decisions. In our analysis, we always consider a *baseline configuration* representing a basic organization equipped with some elementary decision making tools: it should provide a simple set of rules similar to those involved in delivering the health services in real case contexts, in such a way to have a comparison with the proposed methods. A sufficient number of independent simulation runs are performed to derive a conclusion about the solution quality of the proposed online algorithms. Further, in order to analyze the impact of such algorithms when they are used over time, a sufficiently large time horizon with respect to that of the single decision horizon is fixed.

1.4 Thesis outline

The thesis is organized in two parts composed of five chapters each.

Part I is concerned with the Surgical Pathway (SP) and the problem of the ORP. After an introduction and a literature review, the Real Time Management (RTM) is introduced in Chapter 2. Chapter 3 reports the RTM in the case of only elective patients describing also the DES framework of analysis. Such an approach is extended in Chapter 4 to consider also a flow of non-elective patients sharing the Operating Rooms (ORs) with the electives. A comprehensive comparison among dedicated, shared and hybrid policies for the management of non-elective patient is then presented in Chapter 5. Finally, the RTM is extended in Chapter 6 to consider several specialties that have OR sessions assigned by the Master Surgical Schedule (MSS) but shared overtime.

Part II is concerned with the Emergency Care Pathway (ECP) and the problem of overcrowding, which is mainly manifested with an excessive number of patients in the ED and long patient waiting times. A first attempt is to deal with overcrowding from the EMS side. In Chapter 8 the problem of reducing overcrowding is analyzed from a regional perspective exploiting Big Data to compare simple real time dispatching policies. Such an analysis is extended in Chapter 9 to consider a broad variety of dispatching, routing and redeployment real time policies. Then, we shift our attention to the analysis of an ED. In Chapter 10, we will propose a new framework to mine an ED process model capable to predict the use of the ED resources, and based on ad hoc process discovery tools. Such a process model is then exploited in Chapter 11 to enhance the quality of the real time management of the ED patient flow.

After summarizing the concluding remarks reported at end of each chapters, conclusions are discussed in Chapter 12.

Part I

The Surgical Pathway

2 Introduction and Literature Review

In this part of the thesis, we focus our attention on the analysis of a Surgical Pathway (SP), that is a generic Clinical Pathway (CP) comprising a surgical procedure, and we propose online optimization approaches for the problem of the Operating Room (OR) planning and scheduling. Such a problem is central with respect to the management of the SP, because ORs are the most critical and expensive resources of the whole pathway. Issues arising when dealing with this problem are usually classified into three phases corresponding to three decision levels, namely strategic (long term), tactical (medium term) and operational (short term) [127]. At the operational decision level, the problem arising in the OR management is also called “surgery process scheduling” and concern all the decisions regarding the scheduling of the allocation of resources for elective and non-elective surgeries.

An elective surgery is a planned and non-emergency surgical procedure. It may be either medically required (e.g., cataract surgery), or optional (e.g., breast augmentation or implant) surgery. Therefore, elective patients are inserted in a (usually long) waiting list and are scheduled through an ex-ante planning in accordance with several priority rules.

Dealing with elective patients, the surgery process scheduling is generally divided into two offline sub-problems referred to as “advanced scheduling” and “allocation scheduling” [107]. The first sub-problem consists in selecting patients from the waiting list and assigning to their surgery an OR session, that is an OR in a certain day over a planning horizon [65, 66, 72, 74, 75, 79, 92, 97, 109], trying also to take into account different stakeholder perspectives [76, 105, 106, 108, 112]. Given this advanced schedule, the second sub-problem consists in determining the precise sequence of elective surgical procedures and the allocation of resources for each OR session and day combination in order to implement it as efficiently as possible [77, 101, 104, 112, 115, 116, 117]. Usually, the two sub-problems have different objectives, which are the maximization of the operating room utilization, that is a facility-centered index, and the minimization of the number of surgeries delayed or canceled, that is a patient-centered index. Furthermore, especially when considering the inherent stochasticity of the problem, such objectives are conflicting as discussed in Beaulieu et al. [69]. Moreover, other aspects regarding the different stakeholder perspectives should be taken into account, such as the patient waiting times (patient-centered), the bed utilization and the use of overtime (facility-centered), and the balance of the workload along the planning period (medical staff perspective) [64, 67, 75, 76, 108]. For a complete overview of the problems arising in the OR planning and scheduling, the reader can refer to the exhaustive reviews [78, 96, 118].

A non-elective surgery is an unpredictable surgery that should be performed within a time limit, which is shorter than that of an elective surgery, due to the patient medical conditions. For this reason, non-elective patients cannot be inserted in the waiting list and scheduled through an ex-ante planning. Because of their unpredictability, the non-elective patient arrivals are therefore a further element of uncertainty, in addition to the stochasticity involving an elective surgery, whose most impactful component is its duration [60, 102, 119]. Non-elective surgeries deal with different time limits involving different goals: non-elective patients with a time limit of 30 minutes must be operated on as soon as possible while, when the time limit is equal to several hours, one can evaluate what is more beneficial between an immediate surgery or to postpone it waiting for the release of further ORs.

An immediate non-elective surgery can determine a negative impact on the elective patient scheduling. To limit or to avoid such a negative impact, the surgery can be postponed increasing the risk of exceeding the time limit for the non-elective. Such a trade-off should be taken into account when scheduling a non-elective surgery. In accordance with the analysis of the 31 papers considered in the literature review of Van Riet and Demeulemeester [130], the policies for handling elective and non-elective patients are classified into *dedicated*, *hybrid*, and *shared* (or *flexible*). The Dedicated Operating Room (DOR) policy consists in reserving, each day, one or more ORs to perform only non-elective surgeries. Conversely, the Shared Operating Room (SOR) policy allows to perform elective and non-elective surgeries in the same ORs. Furthermore, a hybrid policy is a mix of the two previous policies providing both dedicated and shared ORs. The issue of adopting one of these policies is debated in the literature. In Heng and Wright [99] and Wullink et al. [133], the DOR and the SOR policies are respectively promoted and the improvement of the non-elective waiting times is proved in both papers. Same remarks are reported by Ferrand et al. [94] in which different hybrid policies are evaluated. Since the conflicting conclusions reported in these papers could depend on the scenario and the operative conditions, a detailed comparison among the different policies is required. However, only two papers [93, 133] out of the 31 discussed in [130] provide a (partial) comparison between different policies.

A third online decision problem, called Real Time Management (RTM) of ORs, arises during the fulfillment of the surgery process scheduling. The contribution of the first part of this thesis is to describe the issues of such a new problem and to analyze online optimization for the OR planning and scheduling, which is also the aim of the papers [62, 63, 86, 87, 88].

The RTM addresses the problem of supervising the execution of the schedule when uncertainty factors occur, which are mainly two: (i) actual surgery duration exceeds the expected time, and (ii) non-elective patients need to be inserted in the OR sessions within a short time. In the former, the more rational decision regarding the surgery cancellation or the overtime assignment should be taken. In the latter, the decision concerns the OR and the moment in which non-elective patient are inserted, taking into account the impact on the elective patients previously scheduled.

The literature reports few attempts to address the problem as reported in Hans and Vanberkel [98]. Dexter et al. [83] showed how a computer assisted system could help mitigating the increase of over-utilization of the operating room resources such as overtime. The problem of tardiness from scheduled start times is addressed by Wachtel and Dexter [131] comparing the effectiveness of several procedures to reduce tardiness. The authors showed that the generation of a modified or auxiliary schedule that compensates for known causes of tardiness can be a good solution to reduce tardiness even if its impact proportionally increases as the number of cases involved. The problem of rescheduling the elective patients upon the arrival of emergency patients is addressed in Erdem et al. [89, 90]. The authors proposed a mixed integer linear programming model which considers the overtime cost of the operating rooms and/or the post-anesthesia care units, the cost of postponing or preponing elective surgeries, and the cost of turning down the emergency patients. They proposed a genetic algorithm for its approximate and faster solution. The results of the case study suggest that, instead of shuffling the elective surgeries, it would be worthwhile to consider performing the elective surgeries using the overtime of the operating rooms. Note that the problem of rescheduling patients can be addressed as a particular job shop scheduling problem [115, 124] but these experiences can not directly applied to the operating room context due to its peculiarity in the evaluation of a solution, as we will show in the following chapters. Strategies to move a patient from an operating room to another and based on statistical remarks are also proposed [80, 82, 113, 128]. Further, the impact of an online allocation scheduling has been analyzed in M'Hallah and Al-Roomi [113], in which ORs are assigned to a subset of patients of the waiting list only during their execution.

As discussed in Aringhieri et al. [4], health care optimization problems are challenging, often requiring the adoption of unconventional solution methodologies. The solution approach proposed herein belongs to this family. Our methodological approach is a hybrid simulation and optimization model. Simulation is used in order to generate a real situation with respect to the inherent stochasticity of the problem while optimization is used to take the best decisions in different points of the SP. Accordingly to Magerlein and Martin [107], we consider the operative decisions concerning the advanced scheduling and allocation scheduling of patients. Furthermore we consider the RTM of the operating room planning. The aim is to provide a tool for supporting decisions for the OR planning and scheduling, showing the impact of using offline and online optimization methods for the sub-problems of the surgery process scheduling. Different performance indices are defined and observed over time in order to take into account the different stakeholder perspectives. The generality of the proposed model allows us to replicate and to compare a wide range of possible scenarios and policies, in which most of the case studies of the literature can be included.

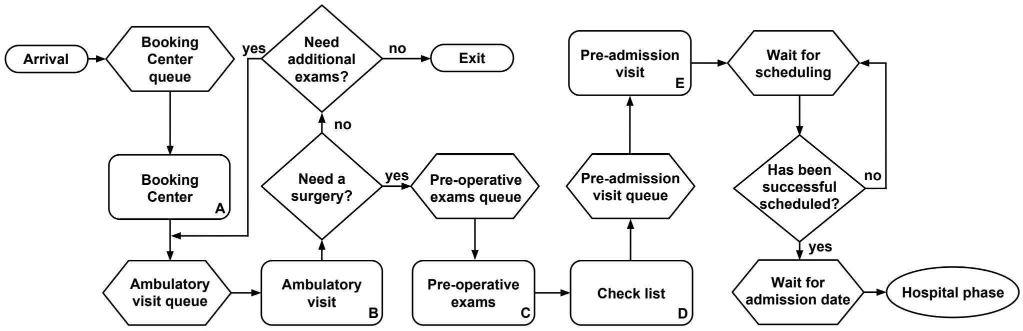
This part of the thesis is organized as follows. In Chapter 3 a hybrid model for the simulation of the elective patient flow is presented, embedding offline and online optimization methods for the surgery process scheduling and analyzing their impact over time. The non-elective patient flow is introduced in Chapter 4, showing its impact on the elective patients when the ORs are shared, proposing an online algorithm for the insertion of non-elective patients. A comparison among dedicated, shared and hybrid policies for the management of non-elective patient is then presented in Chapter 5, analyzing the common offline approaches by the literature for managing the OR sharing and combining them with our proposed online approaches. The RTM is extended in Chapter 6 for several specialties that have OR sessions assigned by the Master Surgical Schedule (MSS) but shared overtime.

3 The Real Time Management of elective patients

3.1 The Surgical Pathway of elective patients

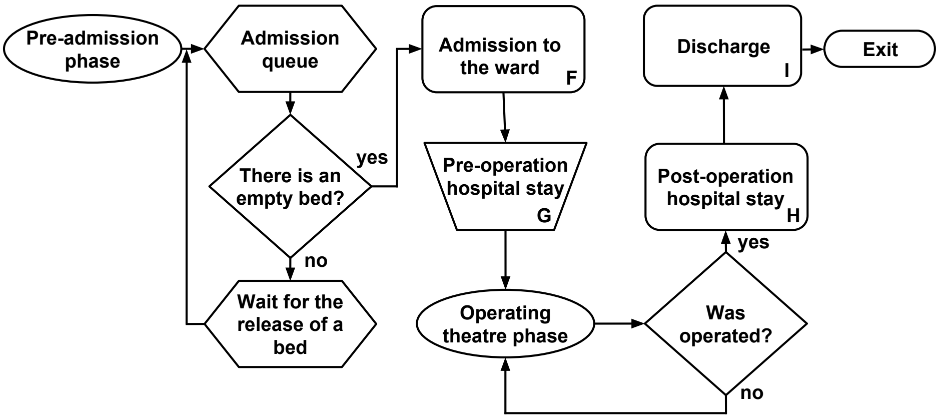
From a management point of view, a SP can be seen as made up of three phases. For each phase we present the problem that have to be addressed at the operational level by the surgery process scheduling, dealing with a single specialty and assuming to use a block scheduling approach. This means that ORs are assigned to specialties at the tactical level by a cyclic OR schedule, called MSS. For this reason, we can assume to have a fixed number of ORs available over a planning horizon and that the OR availability implies also the surgical team (e.g., surgeons and anesthetists) availability. In order to have an approach as general as possible, we consider the total duration of the surgical procedures as a unique service time. Along all this and the following chapters, we always consider a single specialty and its subset of OR sessions assigned by the MSS, except when indicated otherwise. However, such an assumption allows us to do not lose generality, since at the operative level in the close block scheduling the specialties work independently on the set of assigned OR sessions.

Fig. 3.1: Pre-admission phase flowchart



The first phase concerns the *pre-admission phase* and it is related to all the activities regarding patients before their admission (see Figure 3.1). In this phase, a relevant information is the Diagnosis Related Group (DRG), which defines a general time limit (i.e., days to surgery) before which the patient should be operated on. In our context, a Urgency Related Group (URG) is assigned to each patient belonging to the same DRG: the URG states a more accurate time limit called Maximum Time Before Treatment (MTBT). In other words, URG allows us to define a partition of the patients in the same DRG in order to prioritize their activity through the SP. The optimization problem arising in this phase is the *advanced scheduling*. It consists in the selection of patients from the waiting list and in their assignment to an *OR session* $(j, k) \in S$, which identifies a specific OR j of the set J of all the available ORs that have been assigned by the MSS to one specialty in the k-th day of the planning time horizon, whose days are included into the set K. Such a selection consist in a set of elective patients $L \subseteq I$ that is partitioned into n subsets L_{jk} corresponding to the patients that should be operated on within $(j, k) \in S$. Solving the advanced scheduling, several operative constraints need to be satisfied (bed capacity during the patient stay, total time available for the OR session, etc.) and one or more objectives can be fixed (maximization of the utilization, minimization of the waiting times, maximization of the fraction of patients scheduled on before the MTBT, etc.). This problem is well known in the literature as Surgical Case Assignment Problem (SCAP) [125].

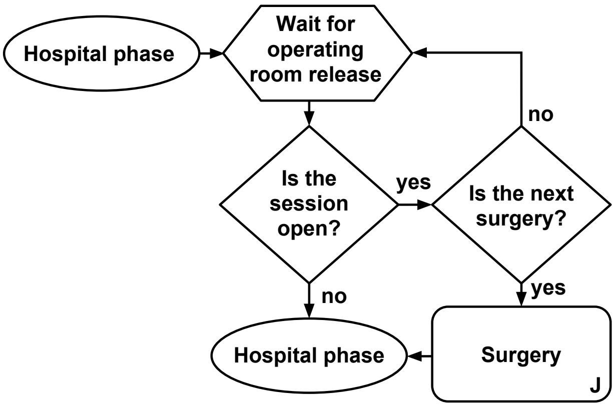
Fig. 3.2: Flowcharts of the hospital phases



The *hospital phase* concerns all the activities involving the admitted patient stay except for those related to the operating theater as shown in Figure 3.2. The relevant information in this phase is the Length Of Stay (LOS) of each patient, that is the number of days required before the discharge. The optimization problem arising in this phase is the allocation scheduling, which consists in finding a sequence of patients to determine the order in which they are operated on. For each OR session $(j, k) \in S$, all the m_{jk} patients in the set L_{jk} are listed in an ordered sequence $\lambda_{jk} = (i_1, \ldots, i_{m_{jk}})$. The objective is to minimize the risk of cancellation, while keeping an acceptable utilization rate with respect to the available operating time taking into account also the patient safety and satisfaction considering waiting times class of urgency and possible previous cancellations.

Figure 3.3 depicts the *operating theater phase*, which is a component of the hospital phase, as highlighted in Figure 3.2. Due to its importance in a SP, it requires to be treated separately. Patients assigned to a given OR session will be operated on following the sequence previously defined unless delays imposes to define a new sequence. Patients not operated on will be rescheduled. We could have a delay

Fig. 3.3: Flowcharts of the operating theater phases



as soon as the Estimated Operating Time (EOT) differs from the Real Operating Time (ROT). The RTM operates when such a delay become significant, that is exceeding the total operating time allowed. The following possible decisions should be considered:

- to use overtime reducing the total amount available for the current planning horizon;
- to cancel one or more surgeries and to re-schedule them;
- to change the sequence of the remaining surgeries in order to minimize the cancellation of patients that are close to their MTBT while keeping an acceptable level of OR utilization.

The first two choices are generally non-trivial and alternatives requiring to consider several aspects. For instance, the decision of postponing a patient could violate MTBT. Further, it determines an increased patient stay lowering the patient satisfaction and, by consequence, the quality of the service. On the other side, overtime is a scarce resource. So, it seems crucial to establish some criteria driving the decisions of using it to avoid cancellations.

The above definition of the SP presented here is a generalization of that describe and analyzed by Ozcan et al. [114] for the thyroid surgical treatment of elective patient. The reader can refer to this paper for further details.

3.2 The Hybrid Model

This section discusses the hybrid simulation optimization model proposed in this chapter. Simulation is exploited to model the inherent stochasticity that characterizes the problems arising in the OR management, that is the arrival of patients, the variability of patient length of stays and the variability of patient operating times (see, e.g., [100, 103, 126]). Furthermore, it allows us to easily replicate the three phases of the patient flow presented in Section 3.1. On this simulated SP, it is possible to embed the optimization modules to deal with the decision problems described in Section 3.1.

Fig. 3.4: Description of the hybrid simulation and optimization model

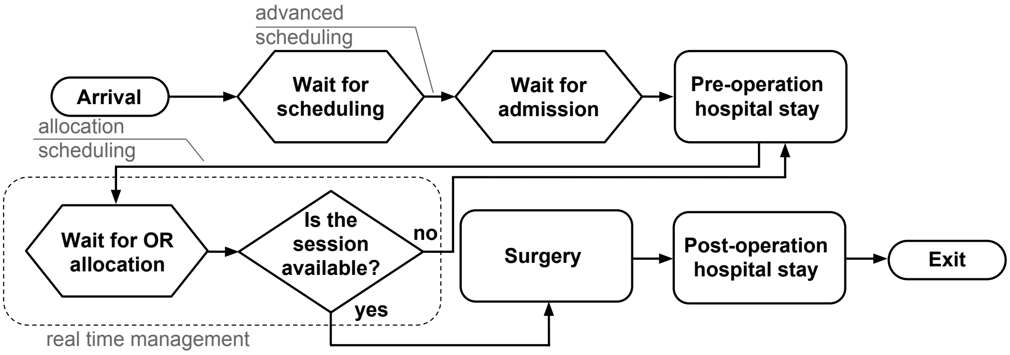


Figure 3.4 summarizes how the patient passes through the SP highlighting when the optimization operates: the advance scheduling manages its admission, the allocation scheduling manages its position in the surgery sequence and, finally, the RTM manages the ongoing operations before the surgery. Summing up, simulation allows to model the operative context required by the optimization modules to operate correctly over the time horizon needed to evaluate the impact of such optimization modules.

In the following, we will briefly describe the hybrid model through the description of its main components, that is the simulation framework and the three optimization modules.

3.2.1 The Discrete Event Simulation Framework

The simulation framework is based on a Discrete Event Simulation (DES) since it is the most suitable methodology to analyze a discrete and stochastic workflow. In general, simulation is a suitable methodology to analyze the impact of decisions on a set key performance indicators, which allow to take into account multiple goals, as reported by Rais and Viana [46] about numerous experiences from literature. Further, DES is the only approach capable to represent the single entities within a SP, which is a necessary condition to apply the proposed optimization planning modules in accordance with Dunke and Nickel [26] as discussed in Section 1.3. The proposed simulation model is a straightforward implementation of the SP represent in Figures 3.1, 3.2 and 3.3. The main parameters of the simulation model, and their distribution, are reported in 3.5. Optimization is embedded within the simulation model in correspondence of the decision points. Since our aim is to analyze the impact of online optimization methods, we use some of the most common approaches in the literature for the SCAP without the claim that it is the best solution. Then, we introduce new online algorithms in order to manage in real time the effects caused by uncertainty.

The hybrid model is implemented using AnyLogic 6.9 [73]: its Enterprise Library is exploited for the implementation of the DES simulation framework while the optimization modules are implemented from scratch in Java, which is the native programming language of AnyLogic.

3.2.2 Solving the Advanced Scheduling Problem

We propose a metaheuristic based on a greedy construction of an initial solution and then a local search to improve that solution. The proposed algorithm is a simplified version of that discussed in [65]. The operative context is represented by a long queue of patients from which we would like to select a subset of patients to be admitted taking into account the fact that the resources available can be reduced since patients admitted the previous planning horizon are already in the hospital phase, usually waiting for the discharge but also for their surgery.

Constructive greedy algorithm

The algorithm associates to each patient $i \in I$ the following values

$$w_i = \frac{t_i}{t_i^{\max}}, \tag{3.1}$$

$$\tilde{w}_i = \frac{t_i + \phi}{t_i^{\max}}, \tag{3.2}$$

where ϕ measures the days between the current day to the beginning of the next planning horizon. At the moment of determining a solution for the advance scheduling problem, ϕ is equal to the day before the beginning of the planning horizon plus its duration. The value w_i measures the ratio of the time elapsed before the surgery and the MTBT associated to the URG of the patient $i \in I$ while $\tilde{w}_i$ is a projection of w_i referred to the next planning horizon. In other words w_i and $\tilde{w}_i$ is the waiting time of the patient normalized with respect to the MTBT in correspondence of the current day and the first day of the next planning horizon, respectively.

Starting from the schedule containing the patients planned the previous planning horizon (i.e., patients rescheduled after their surgery cancellation), patients to be admitted and belonging to the admission queue are ordered by decreasing value of w_i in such a way to promote the scheduling of those patients which are close to their MTBT. Then, each patient is considered for the scheduling. A patient i will be inserted in the current schedule if there exists an OR session (j^*, k^*) available without exceeding its duration d_{jk} and the bed capacity b_k in the days $k = k^*, \dots, k^* + \ell_i - 1$, where ℓ_i is the LOS of the patient i.

Among different possible OR sessions $(j,k) \in S$, the algorithm tries to schedule the patient i first in the day k such that $k + \ell_i \leq \zeta$, where $\zeta = \max\{k|k \in K\}$ is the last day of the planning horizon. If it is not possible, the algorithm tries the insertion in the days k that minimize the bed utilization during the day with lower capacity (e.g. weekend stay beds could be a limited resource). This rule can be overridden when $\tilde{w}_i \geq 1$ assigning the patient to the first day $k = 1$, if possible, or to the second day $k = 2$, and so on. In this case, we would like to reduce the probability of not satisfying the URG requirements in case of cancellation. Finally, if a patient cannot be scheduled, the algorithm will consider the next patient. The algorithm terminates when all patients in the queue have been considered for the insertion in the current schedule.

Improvement local search algorithm

The Local Search tries to improve the greedy solution by exchanging pairs of patients already scheduled in such a way to cluster them in a reduced number of OR sessions and, by consequence, to allow the insertion of new patients previously not scheduled. Let (j^*, k^*) be the OR session having the maximum operating time yet available:

$$(j^*, k^*) = \arg \max_{(j,k) \in S} \left\{ d_{jk} - \sum_{i \in L_{jk}} e_i \right\}. \tag{3.3}$$

The Local Search algorithm follows these criteria to select the new incumbent solution:

- the new solution will be that providing the maximal increase of the time yet available of (j^*, k^*);
- otherwise, if the two schedules are equivalent in (j^*, k^*), the algorithm will consider the second least utilized OR session, and so on;
- otherwise, if the two schedules are equivalent in all OR sessions, the algorithm selects those solutions having OR sessions less utilized at the end of the planning horizon in such a way to favor the rescheduling of surgery canceled in the previous days in that OR sessions.

3.2.3 Solving the Allocation Scheduling Problem

Dealing with only elective patients, the allocation scheduling problem consists in establishing the sequence λ_{jk} in which patients in $L_{j,k}$ will be operated on in such a way to minimize the inefficiency due to possible cancellations. Since last positions of λ_{jk} have the higher probability to be cancelled because of the accumulation of delays, it is preferable to schedule at the beginning of the sequence: (i) the patients close to their MTBT that can not be postponed to the next planning horizon avoiding to exceed such a time limit, and (ii) the patients whose surgery has been already postponed. To deal with these special cases, we propose an algorithm in such a way to give different priorities to the patients in λ_{jk} as follows:

1. patients $i \in L_{jk}$ such that $\tilde{w}_i \geq 1$ are scheduled in decreasing order of $\tilde{w}_i$ at the beginning of the OR session;
2. patients previously postponed with $\tilde{w}_i \leq 1$ are scheduled in decreasing order of the days elapsed since the first cancellation;
3. finally, all the remaining patients in L_{jk} are scheduled using the LPT or the SPT rule with respect to e_i at the end of the OR session.

We refer to the two versions of algorithm with the terms *LPT-modified* (LPT$^+$) and *SPT-modified* (SPT$^+$) depending on the rule used at the last step.

3.2.4 An online approach for the Real Time Management

The solutions discussed in the previous sections provide a schedule based on the EOT, which is usually an estimate of the surgeons. Unfortunately, it is possible that the ROT differs from the EOT. Given L_{jk} and a patient $i \in L_{jk}$, the whole schedule could be delayed if $r_i > e_i$. When the overall delay could determine the exceeding of the jth OR session duration d_{jk}, the RTM should deal with the problem of postponing a surgery or using a part of the total overtime ν available for the whole planning horizon. Such a decision poses the problem of evaluating the impact of consuming overtime or to have a cancellation.

Let us consider the jth OR session on day $k =$ having duration d_{jk} and a list L_{jk} of scheduled and sequenced patients. Suppose that $m < |L_{jk}|$ patients are already operated on. Let ρ^{τ}_{jk} be the time elapsed in the OR session (j,k) from the beginning of the session at the time τ. If the surgery of the m-th patient belonging to the schedule of L_{jk} ends at time τ, the effective time elapsed to operate on the first m patients is

$$\rho^{\tau}_{jk} = \sum_{i=i_1,\ldots,i_m} r_i. \tag{3.4}$$

Let us introduce the following parameter:

$$\beta^{\tau}_{k} = 1 + \frac{\sum_{h>k} n_h}{n} - \frac{\nu^{\tau}_{k}}{\nu} \tag{3.5}$$

where ν^{τ}_{k} is the residual overtime after the surgery of patient i_m and n_h is the number of the OR sessions of the day h. Note that the second term of the right member in 3.5 is equal to 0 during the last day of the planning horizon (i.e., when $k = \zeta$). The value β^{τ}_{k} would measure the overtime still available with respect to the number of OR sessions to be still performed. Actually, β^{τ}_{k} is closed to 1 when the overtime was used proportionally; it is between 0 and 1 or it is greater than 1 when it was underused or overused, respectively. We remark that in the last day of the planning horizon β^{τ}_{k}

it is always less than or equal to 1 promoting the use of the residual overtime. The online algorithm starts every time a surgery ends and $\rho^{\tau}_{jk} > \sum_{i=i_1,\ldots,i_m} e_i$. It consists of three procedures.

Sequencing check. The sequencing of the remaining patients is checked in such a way to ensure that (i) all the remaining patients having $\tilde{w}_i > 1$ are scheduled prior to the other patients and (ii) those having $\tilde{w}_i > 1$ are ordered by decreasing value of $\tilde{w}_i$; if those patients run out the available operating time d_{jk}, the patients having $\tilde{w}_i \geq 1$ keep the same original order; otherwise, the free operating time is filled selecting a subset of the patients having $\tilde{w}_i < 1$ in such a way to fill the available operating time following a rule similar to the Best Fit rule for the Bin Packing problem.

Overtime allocation. Suppose that the overtime available ν^{τ}_{k} is sufficient to avoid the cancellation of a patient $i \in L_{jk}$. If the patient i is close to the MTBT (i.e., if $\tilde{w} \geq 1$), then the necessary amount of overtime is always allocated. Otherwise, the overtime is allocated if and only if the following criterion is satisfied:

$$\beta^{\tau}_{k} \left(\frac{\rho^{\tau}_{jk} + e_i}{d_{jk}} \right) \leq 1. \tag{3.6}$$

Rescheduling. At the end of the day, all the postponed surgeries must be rescheduled on OR sessions having enough free operating time. First the algorithm considers all the patients having $\tilde{w}_i > 1$ trying to insert each patient in the first OR session available. Then, the algorithm tries to insert iteratively subsets of patients having $\tilde{w}_i \leq 1$ according to the Bin Packing Best Fit rule. If an insertion is not possible, the patient will be scheduled on the first day available in the next planning horizon.

Finally, we remark that the algorithm for the insertion of a subset of patients, used both in the sequencing check and in the rescheduling procedures, is and adaptation of the dynamic programming discussed in Section 3.4.1 of [110]. For a description of the Best Fit rule for the Bin Packing, the reader can refer to Section 8.2 of [110].

Notation introduced in the problem statement (Section 3.1) and in the proposed solutions is reported in Table 3.1.

3.3 Quantitative analysis

This section reports the quantitative analysis performed in order to evaluate the impact of the online approach to the RTM and the additional optimization modules on the management of a SP. The main idea behind the proposed quantitative analysis is to evaluate their impact over time, that is how the previous decisions (e.g., determining less or more cancellations) can impact on the current decisions.

Tab. 3.1: Summary of the notation of problem statement and solutions.

Sets	
J: set of operating rooms	K: set of the days of the planning horizon
S: set of all OR sessions	I: set of patients in the waiting list
L: set of scheduled patients	L_{jk}: set of patients scheduled into (j,k)
λ_{jk}: sequence of patients scheduled into (j,k)	
Indices and cardinalities	
i: elective patient	j: index of the operating room
k: index of the day	ζ: index of last day of the planning horizon
n: number of OR sessions	n_k: number of OR sessions of the day k
m_{jk}: number of patient scheduled into (j,k)	b_k: stay bed units avalaible in the day k
Times and durations	
t_i: waiting time of patient i	$t_i^{\max}$: MTBT of patient i
w_i: normalized waiting time of patient i	$\tilde{w}_i$: value of w_i in the next planning horizon
e_i: EOT of patient i	r_i: ROT of patient i
d_{jk}: duration of (j,k)	τ: general instant during the OR session
ν: overtime available for one planning horizon	ν_k^τ: overtime available at instant τ of day k
ρ_{jk}^τ: time elapsed since the beginning of (j,k)	β_k^τ: parameter for the overtime criterion
ℓ_i: LOS of patient i (days)	

In our work, we are considering the surgery process scheduling problems arising at the operational level, which has usually a planning horizon of a week. The idea behind our quantitative analysis is therefore to evaluate the impact of such plannings over time, that is how the decisions made for the previous planning horizons (e.g., determining less or more cancellations) can impact on the current decisions.

Section 3.3.1 describes how the computational experiments are carried out reporting the possible configurations of the optimization modules, the performance indices and the different evaluation scenarios. Section 3.3.2 reports about the logical validation of the simulation model discussed in Section 3.2.1. Section 3.3.3 and Section 3.3.4 report the results of the computational tests made on two different evaluation scenarios. Section 3.3.5 extends the original hybrid model to deal with different trained surgery teams in order to prove the capability and the flexibility of our approach. Finally, Section 3.3.6 provides a brief analysis of the bed occupation over the week in order to evaluate the impact on the ward workload.

The results reported in the following sections are the average value among those obtained by running the hybrid model 30 times on a given configuration and, each time, starting from a different initial conditions. On average, one single run requires from 1.3 to 4.3 seconds when running with all the optimization approaches turned off or turned on, respectively. This means that no more than $4.3 \times 30 = 129$ seconds are needed to simulate two years of operating room management. Finally, we remark that the algorithms for the advanced scheduling are the most time consuming components while the running time required by the other optimization algorithms are negligible. Finally, we remark that all the simulation parameters are reported in 3.5. The Appendix describes and reports the parameters regarding the patient flow characteristics, the duration of the activities and their distributions, and all the other parameters characterizing our simulation such as the values for each class of URG and the number of beds available.

3.3.1 Test configurations, performance indices and scenarios

The optimization algorithms described in Section 3.2.2, 3.2.3 and 3.2.4 can be combined in different ways In order to evaluate their actual impact, we define a *baseline configuration* with respect to the three phases as follows:

Phase 1: advanced scheduling performed by a first-fit algorithm, that is (i) it considers patients by decreasing order of w_i, (ii) it scans the OR session from Monday to Friday and assigns the selected patient to the first one having enough operating time available (if possible);

Phase 2: the patient sequencing is that resulting from the patient assignment, that is, the first assigned to an OR session will be the first in the sequence, and so on;

Phase 3: overtime is assigned *a priori* uniformly to all OR sessions in an amount equal to $\frac{\nu}{n}$;

Phase 3: all the surgeries are rescheduled only at the end of the day using the first-fit algorithm, that is the first phase of the RTM rescheduling algorithm.

Besides the baseline configuration, we define further configurations to evaluate the impact of the optimization modules. Each configuration is defined with respect to the baseline configuration.

- **Phase 1**:
  - **option 1:** computing w_i w.r.t Monday instead of the previous Friday (in the simulation model, Friday is the day in which the advance scheduling is performed);
  - **option 2:** adopting the greedy explained in Section 3.2.2 (instead of the First-Fit algorithm);
  - **option 3:** adopting the Local Search described in Section 3.2.2;
- **Phase 2**:
  - **LPT/SPT:** use LPT$^+$ or SPT$^+$ rules in sequencing introduced in Section 3.2.3, respectively;
- **Phase 3**:
  - **option A:** adopting the RTM online algorithm after each surgery;
  - **option B:** adopting the rescheduling algorithm proposed in Section 3.2.4 at the end of the day (instead of the first-fit algorithm).

Table 3.2 reports the two types of indices adopted to evaluate the impact of the optimization modules. We define a set of patient-centered indices in such a way to evaluate the performance from a patient point of view. We also define a set of facility-centered indices in such a way to evaluate them against to the patient-centered ones. The indices w_{avg} and the f are a reformulation of the *need adjusted waiting days* proposed by Tànfani and Testi [125] while the remaining ones are reported in Cardoen et al. [78].

Tab. 3.2: Patient-centered and facility-centered indices

Index	Definition
	Patient-centered
o	number of patients operated on
c	fraction of cancellations
t_{avg}	average waiting time spent in the waiting list (days)
w_{avg}	average value of patient's w_i at the time of their surgery
f	fraction of patients operated within the MTBT
	Facility-centered
u_{bed}	OR session utilization
u_{OR}	bed utilization

It is quite evident that different indices can affect each other. For instance, the increase of the number of cancellations can affect the bed utilization and, in its turn, could reduce the fraction of patients operated within the MTBT.

Tab. 3.3: Scenarios: duration of the OR sessions (min)

(a) Scenario 1

	OR 1	OR 2	OR 3	OR 4	OR 5
Mon	300	360	420	420	420
Tue	300	360	420	420	
Wed	300	360	420		
Thu	300	360	420	420	420
Fri	300	360	420	420	

(b) Scenario 2

	OR 1	OR 2	OR 3
Mon	540	540	540
Tue	540	540	540
Wed	540	540	540
Thu	540	540	480
Fri	540	540	480

Table 3.3 describes the two different scenarios that differs for the number and the durations of the OR session, in which we evaluate the optimization solutions on different operating contexts. The two scenarios are characterized by about the same overall amount of operating time available ($7\,920$ vs. $7\,980$ min) distributed in a different way with respect to the number n of available OR sessions (21 vs. 15) and their duration d_{jk}, $(j,k) \in S$. Furthermore, the two scenarios have the same URG classification, MTBTs and probability distribution (i.e. fraction of patients of a URG class over the total) are reported in Table 3.4.

Tab. 3.4: Scenarios: URG classes, distribution and MTBT as in [114].

URG class	A	B	C	D	E	F	G
MTBT $t_i^{\max}$ (days)	8	15	30	60	90	120	180
distribution	2.45%	14.01%	41.36%	17.85%	11.40%	7.49%	5.44%

3.3.2 Simulation Model Validation

The validation of a simulation model requires a quite complex analysis. In our case, we are only interested in the logical correctness of the simulation model representing the SP. On the other side, we are not interested in the replication of a real system.

To this end, we adapted our simulation model to represent the inspiring case, that is the case study reported in Ozcan et al. [114]. We started from real data available in that paper, then (i) we replicated the patient distribution among the different URGs and the configuration of the available OR sessions, and (ii) we used statistics about the activity durations (average, minimum, maximum and modal values) to define distributions for modeling the service times in accordance with the suggestions in the literature (more details are repored in the Appendix 3.5). In that paper, the proposed model dealt with two patient flows having similar EOT but different LOS. Note that the LOS of the second flow is roughly the double of the first one while the number of patients in the first flow is roughly the double of the second flow. Since our model can generate only one type of patient flow, we adapted our patient flow generator in such a way to have, on average, the same number of patients having the LOS of the first flow which is the most numerous. In this validation scenario, we have $n = 7$ OR sessions having the same duration equal to 360 min. Two OR sessions are scheduled on from Tuesday to Thursday and one on Friday. The other parameters are set to the same value reported in the Appendix 3.5. Furthermore, we turn off all the optimization during the three phases. In Table 3.5 we compare the results of our adapted simulation model with those reported in [114].

Tab. 3.5: Model validation: comparison with real measures

	u_{bed}	u_{OR}
Real measures	51.1%	77.3%
Simulation model	49.1%	80.8%
Difference	2.0%	3.5%

The differences in the two performance indices can be accounted to the different composition of the patient flow as discussed above. For instances, the gap of 3.5% for u_{OR} expressed in minutes corresponds to the execution of one surgery having average duration. On the basis of these considerations, the comparison is satisfactory with respect to our objective, which is the validation of the logical correctness of our simulation model.

3.3.3 Scenario 1: analysis

We tested all the possible configurations that can be obtained combining the options defined in Section 3.3.1. Our aim is to identify the best configuration which increases the patient-centered indices without deteriorating the facility-centered ones. First, the impact of each optimization modules is evaluated through the quantitative analysis. Based on these results, two further configurations have been studied. The results are summarized in Table 3.6, which reports the value of the performance indices for each test configuration denoted by the value in the first column "id". All the results are compared with those obtained for the baseline configuration.

Tab. 3.6: Performance indices for each test configuration

	Option(s)						Performance indices						
id	1	2	3	seq.	A	B	o	c	t_{avg}	w_{avg}	f	u_{OR}	u_{bed}
(0)	baseline configuration						2 348	10.0%	55	1.17	32.6%	89.9%	63.6%
(1)	$\surd$						2 347	10.0%	56	1.11	31.9%	89.8%	60.2%
(2)	$\surd$	$\surd$					2 340	9.7%	58	1.16	26.0%	89.3%	60.6%
(3)			$\surd$				2 346	10.7%	52	1.12	36.0%	89.6%	60.4%
(4)	$\surd$		$\surd$				2 349	10.5%	53	1.06	35.3%	89.8%	60.3%
(5)	$\surd$	$\surd$	$\surd$				2 338	9.8%	58	1.17	27.2%	90.0%	60.8%
(6)				LPT$^+$			2 367	10.0%	48	1.03	47.9%	90.8%	60.5%
(7)				SPT$^+$			2 261	10.6%	72	1.51	12.1%	86.4%	58.6%
(8)					$\surd$		2 384	8.3%	35	0.80	74.6%	91.3%	59.3%
(9)						$\surd$	2 315	10.2%	55	1.18	30.7%	88.8%	72.6%
(10)					$\surd$	$\surd$	2 372	9.4%	37	0.83	73.0%	90.7%	64.0%
(11)			$\surd$	LPT$^+$	$\surd$		2 389	10.0%	32	0.73	79.9%	91.8%	60.3%
(12)	$\surd$		$\surd$	LPT$^+$	$\surd$		2 390	10.4%	34	0.71	85.5%	91.8%	60.6%

Regarding the impact of the advanced scheduling optimization module, we can observe a lower waiting time in the waiting list and an improvement of the performance indices related to MTBT in test configurations (3) and (4). On the other side, the minimal fraction of cancellations is obtained with configuration (2) but, at the same time, the fraction of patients operated on before their MTBT decreases consistently. Note that the use of Local Search allows to insert more patients determining the improvement measured in (3) and (4).

Regarding the impact of the allocation schedule optimization module, we can observe significantly better performances when LPT$^+$ policy is adopted. Figure 3.5 shows the trend of the waiting list length under the baseline, (6) and (7) configurations.

Regarding the impact of the online approach for the RTM, we observe a remarkable improvement of all the performance indices (see configurations (8) and in particular f). On the other side, we observe the negligible impact of the algorithm for the rescheduling postponed patients at the end of the day (see configurations (9) and (10)).

Fig. 3.5: Length (number of patients) of the waiting list over the 2nd year

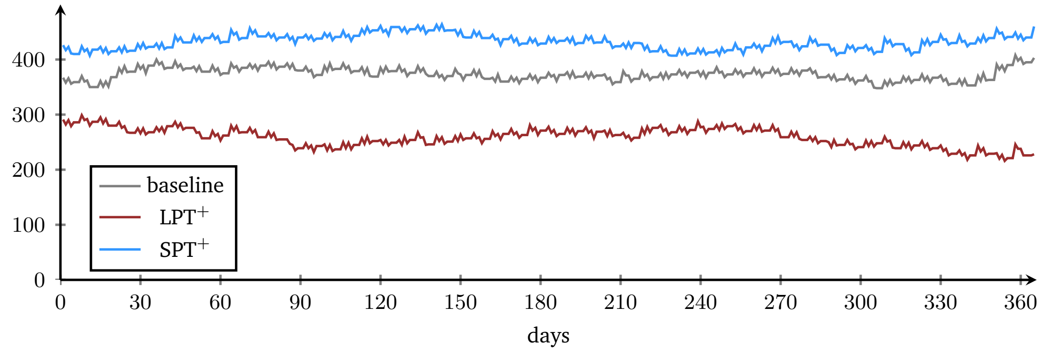


Figure 3.6 and 3.7 show respectively the trend of the waiting list length and the value of w_{avg} under the baseline and (8) configurations. Note that it is positive when $w_{\text{avg}} < 1$ which means that all the patients are operated on before their MTBT, on average.

Fig. 3.6: Length (number of patients) of the waiting list over the 2nd year

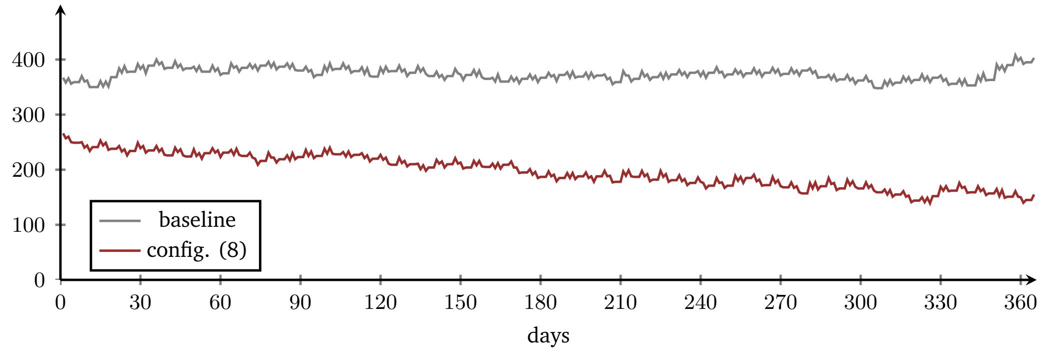
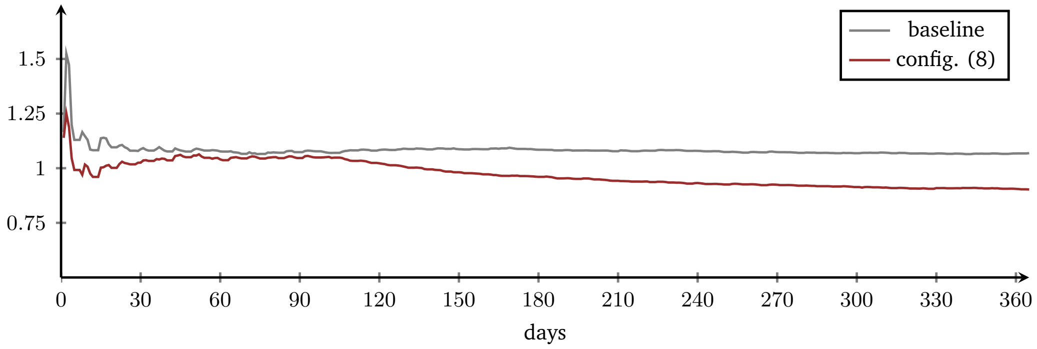
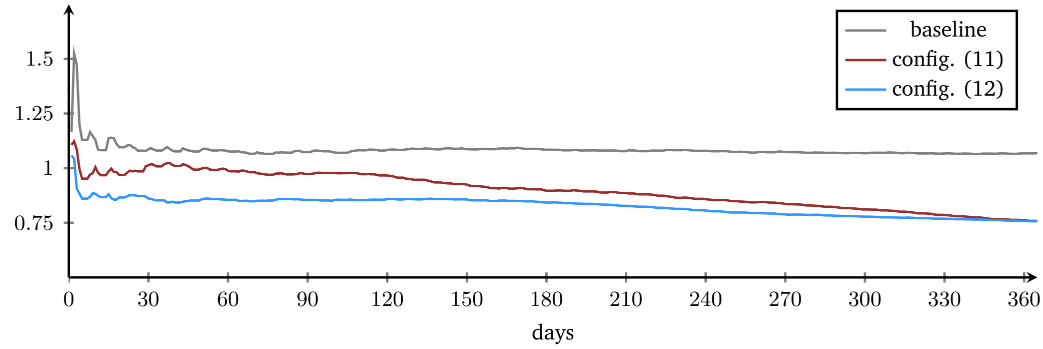


Fig. 3.7: Trend of w_{avg} over the 2nd year



Finally, configurations (11) and (12) report about the combination of the different best options. We note a further improvement of the performance indices except for that related to the number of cancellations if compared with configuration (8). This is due to the fact that Local Search allows to insert more patients in the advanced scheduling thus reducing the waiting time in the waiting list but increasing the probability of incurring in a cancellation. Figure 3.8 shows the trend of w_{avg} under the baseline, (11) and (12) configurations. While baseline configuration shows a value of w_{avg} always greater than 1, we remark that both configurations (11) and (12) tend to be less than 1. Further, configuration (12) seems more stable and powerful in reducing this index.

Fig. 3.8: Trend of w_{avg} over the 2nd year



3.3.4 Scenario 2: analysis

The second scenario differs from the first one in terms of the schedule of the OR sessions. As for scenario 1, the impact of each possible configuration is evaluated and then, based on these results, four further configurations have been studied. The results are summarized in Table 3.7. All the results are compared with those obtained for the baseline configuration.

Tab. 3.7: Performance indices for each test configuration

	Option(s)						**Performance indices**						
id	1	2	3	seq.	A	B	o	c	t_{avg}	w_{avg}	f	u_{OR}	u_{bed}
(0)	baseline configuration						2 411	8.7%	44	0.97	57.7%	92.0%	61.5%
(1)	√						2 408	8.3%	43	0.95	60.0%	92.1%	62.0%
(2)	√	√					2 403	7.5%	41	0.85	70.5%	91.8%	61.0%
(3)			√				2 404	8.3%	41	0.92	61.1%	92.0%	61.9%
(4)	√		√				2 409	8.4%	42	0.87	67.7%	92.0%	62.3%
(5)	√	√	√				2 398	8.3%	42	0.86	69.2%	91.9%	61.7%
(6)				LPT$^+$			2 462	7.1%	25	0.62	85.1%	94.3%	61.7%
(7)				SPT$^+$			2 346	8.6%	58	1.24	27.0%	89.5%	60.8%
(8)					√		2 422	7.3%	27	0.67	84.7%	92.6%	61.0%
(9)						√	2 405	7.9%	41	0.92	62.1%	92.0%	62.6%
(10)					√	√	2 411	7.3%	27	0.66	84.4%	92.5%	64.0%
(11)	√	√		LPT$^+$	√		2 430	6.4%	21	0.49	96.0%	92.8%	59.8%
(12)	√	√	√	LPT$^+$	√		2 434	8.4%	21	0.50	95.3%	93.1%	62.4%
(13)	√	√		LPT$^+$	√	√	2 426	6.6%	21	0.49	96.5%	92.8%	60.2%
(14)	√	√	√	LPT$^+$	√	√	2 419	8.3%	20	0.48	96.8%	92.6%	61.6%

Comparing the results for the two scenarios, we can observe that the number of cancellations with respect to the number of operated patients is almost the same. Further, the utilization indices (u_{OR} and u_{bed}) ranges around the same values, that is 60% and 90% for beds and OR sessions, respectively. The comparison of the results reported for configurations (6) and (7) confirms the fact that LPT^{+} can provide better results than SPT^{+}. The significant decrease of the waiting times and the value of w_{avg} is confirmed also in the analysis of the second scenario.

In both scenarios, the use of Local Search provides a higher number of patients scheduled in the OR sessions, increasing the OR utilization but also the fraction of cancellation. LPT^{+} give a significant improvement of all indices. Finally, the greater contribute for the lowering of the waiting times is obtained enabling the optimization module of the RTM, regarding the overtime allocation, which also provides a better OR utilization.

3.3.5 Dealing with differently trained surgical teams

A surgical team is a set of experts who perform surgery activities and related tasks together usually including surgeons, assistants, nurses, anesthetists and surgical technologists. Such roles require a long period of training to be specialized (especially for surgeons and anesthetists), with a significant impact on the variability of surgery duration.

Even if our focus is at the operational level to deal with the resource management, we would provide an evaluation of having surgical teams with different level of training. We suppose that a surgical team having less trained components could require additional time to accomplish their tasks.

The additional time added to the ROT is generated through an exponential distribution of parameter $a_j > 0$ set to have average delay $\frac{1}{a_j}$ (min). The exponential distribution has been chosen because such a probability function is positive and quickly decreasing.

We considered the new scenarios 1b and 2b obtained from the original one simply adding the value for parameter $\frac{1}{a_j}$, as reported in Table 3.8a and 3.8b.

Tab. 3.8: Scenarios with additional delay

(a) Scenario 1b

	OR 1	OR 2	OR 3	OR 4	OR 5
$\frac{1}{a_j}$	best	5	10	15	20

(b) Scenario 2b

	OR 1	OR 2	OR 3
$\frac{1}{a_j}$	best	10	20

Tables 3.9 and 3.10 show the results for the new scenarios corresponding to the most representative configurations. Comparing them with those obtained for the baseline configurations, adding further delay causes a substantial deterioration of

the performance indices, specially in correspondence of cancellations and number of patients operated on according to their MTBT.

Tab. 3.9: Performance indices of scenario 1b for best configurations in Table 3.6

id	o	c	t_{avg}	w_{avg}	f	u_{OR}	u_{bed}
(0)	2 162	17.5%	93	1.9	10.9%	87.5%	62.7%
(8)	2 298	12.2%	64	1.4	18.0%	93.5%	60.5%
(11)	2 345	15.2%	52	1.1	38.1%	95.4%	63.5%
(12)	2 342	15.7%	51	1.0	43.7%	95.1%	63.7%

Tab. 3.10: Performance indices of scenario 2b for best configurations in Table 3.7

id	o	c	t_{avg}	w_{avg}	f	u_{OR}	u_{bed}
(0)	2 248	15.3%	78	1.65	9.4%	90.6%	63.5%
(8)	2 354	11.3%	50	1.09	43.1%	95.0%	62.6%
(11)	2 400	13.8%	39	0.83	75.6%	96.5%	64.1%
(12)	2 396	14.6%	40	0.84	75.7%	96.4%	64.6%
(13)	2 375	14.4%	38	0.81	78.2%	95.6%	68.3%
(14)	2 386	14.9%	40	0.84	73.4%	96.3%	67.7%

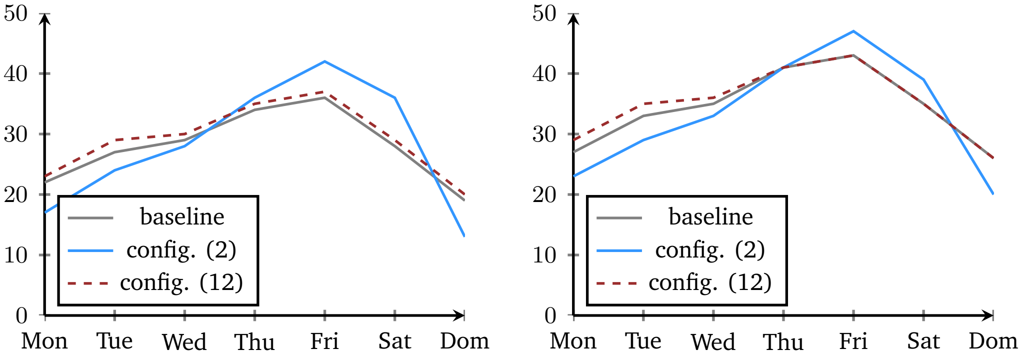
The positive impact of the optimization persists on the observed indices. Actually, the fraction of patients operated within the time limit presents a more significant improvement: it ranges from 10.9% to 43.7% in the scenario 1b, and from 9.4% to 73.4% in the scenario 2b.

3.3.6 Bed Levelling

Among many different performance criteria, the evaluation of the ward stay bed levelling seems to be one of the more challenging [71, 78]. A planning leading to a smooth – without peaks – stay bed occupancy, will determine a smooth workload in the ward and, at the end, an improved quality of care provided to patients. In this section we provide a brief analysis of the bed occupation over the week in scenario 1, that is the most challenging due to the different number and duration of the OR sessions in the day of the week.

Figure 3.9 reports the bed occupancy during the week reporting both the average (Figure 3.10a) and 95th percentile (Figure 3.10b) values. The results for the baseline and configuration (12) show a peak on Friday determined by an increased bed occupancy of about the 50% with respect to Monday. This behaviour seems not affected by the optimization of the SP since the baseline configuration and configuration (12) are really similar. Indeed, the online optimization modules involved in configuration (12) slighty increase the bed occupancy over all the week, because of the higher number of patients operated on, but the difference between the maximum and the minimum occupancy is unchanged.

Fig. 3.9: Bed occupancy during the week (average and 95th percentile values).



(a) average

(b) 95th percentile

On the other side, the behaviour of the configuration (2) shows how planning decisions can affect the bed levelling during the week. In the case of configuration (2), the advance scheduling would compute a solution limiting the use of the weekend stay beds since they are limited in number. This decision largely affects the bedlevelling as shown both in Figure 3.10a and 3.10b where beds occupancy is doubled, approximately. These results confirms those available in literature leading to the need of ad hoc optimization methods for bed levelling as in [64, 66, 67].

3.4 Concluding remarks

In this chapter we proposed a model for the Real Time Management of operating rooms. Given an OR schedule, it consists in a sort of centralized surveillance system whose main task is to supervise the execution of such a schedule and, in the case of delays, to take the more rational decision regarding the surgery cancellation or the overtime assignment. We evaluated its impact on the performance of a generic SP for elective patients. To this end, we developed a hybrid simulation and optimization model.

The extensive quantitative analysis discussed in Section 3.3 showed the positive impact of the optimization in the management of a SP through the evaluation of a set of patient-centered and facility-centered indices.

The online algorithm developed for the RTM is capable to determine a general improvement of all the performance indices. Comparing the baseline configuration with the best configuration in the two scenarios considered, we observed a significant improvement of the performance indices related to the waiting times. This allow to almost double the fraction of the patients operated on before their MTBT time limit. These improvements can determine a general increasing of the quality of service from a patient-centered point of view without deteriorating the facility-centered performance indices (i.e. the resource utilization). The quantitative analysis confirms the trade-off between the number of cancellations and the number of operated patients (or, equivalently, the OR session utilization) as discussed in [69]. Further, results obtained in the baseline configurations (e.g. OR utilization) are consistent with those presented in [69], which has been used for the design of the CP and the parametrization. The analysis provided in Section 3.3.5 demonstrates the capability and the flexibility of our hybrid model to deal with different OR settings. This analysis also showed how the overtime could be interpreted as a really flexible resources that can be used to bring under control challenging situations.

From an OR management point of view, the quality of the provided results and the low computation time suggest the development of a decision support system based on the online algorithm for the RTM powered by an ICT infrastructure to track the surgeries within the operating rooms. Such a system could support the OR supervisor(s) in the management of the current schedule optimizing the use of the overtime.

Further works could extend the analysis to deal with the no-show phenomena, whose effect is a lowering of both the OR utilization and the patient satisfaction. No-show consists in a particular type of cancellation and, in some cases, the surgery of the involved patient can be rescheduled. Althoug we considered only cancellations caused by management issues, our hybrid model can be used to study the impact and the robustness of our optimization modules dealing with different rates of no-show and rescheduling policies.

Although very promising results have been provided in the quantitative analysis, in order to further demonstrate the effectiveness of online optimization methods, future research avenues could provide a comparison of such approaches with the other alternative well-known methodologies taking into account uncertainty, such as stochastic programming and robust optimization.

3.5 Appendix: Parameters

In this appendix, we report the parameters of the simulation model and its setting both for the model validation (Section 3.3.2) and for the quantitative analysis (Sections 3.3.3–3.3.5).

Table 3.11 lists the parameters used by the simulation model regarding. Table 3.12 shows the distributions used to generate the required time for the execution of the activities A–J. Table 3.13 reports the values assigned to the parameters for the model validation and for the quantitative analysis.

Starting from the values reported in [114], that is minimum, maximum, average and modal values, we use a Gamma distribution because, empirically, those values suggested a distribution whose shape recalls the Gamma. The parameters k and ϑ

Tab. 3.11: Definition of the parameters

Index	Definition
	Flow and patient characteristics
r_0	patient interarrival rate
R_0	initial length of the pre-admission waiting list
p_1	patient probability to require a surgical treatment during the ambulatory visit (see Fig. 3.1)
p_2	patient probability to do not require a surgical treatment but requiring further exams during the ambulatory visit (see Fig. 3.1)
	Activity durations
$T^{\text{min, max, mod}}_{\text{A},\ldots,\text{F,I}}$	minimum, average and modal time for the execution of A–F and I (see Figures 3.1–3.3)
$\ell^{\text{min, max, mod}}_{\text{A},\ldots,\text{G}}$	minimum, maximum and modal LOS for patients of urgency class A–G
$\bar{\epsilon}_{\text{A},\ldots,\text{G}}$	average EOT for the surgery of a patient of urgency class A–G
$e_{\max}$	maximum duration of a surgery
$\sigma_{\text{A},\ldots,\text{G}}$	EOT standard deviation for the surgery of a patient of urgency class A–G
Δ	discretization constant for the EOT
σ	ROT standard deviation for each patient
d_{tol}	tolerance time within which the surgical team operates a patient at the end of OR session without resorting to the overtime

Tab. 3.12: Distribution of the activity durations

Activities	Durations	Parameters
A – F, I	$T^{\text{A},\ldots,\text{F,I}}_{\min} + T,$	$k = T^{\text{A},\ldots,\text{F,I}}_{\text{avg}} - T^{\text{A},\ldots,\text{F,I}}_{\text{mod}},$
	$T \sim \text{Gamma}(k, \vartheta)$	$\vartheta = \frac{T^{\text{A},\ldots,\text{F,I}}_{\text{avg}} - T^{\text{A},\ldots,\text{F,I}}_{\min}}{T^{\text{A},\ldots,\text{F,I}}_{\text{avg}} - T^{\text{A},\ldots,\text{F,I}}_{\text{mod}}}$
H (LOS)	$\lfloor \text{Triangular}(\ell^{\min}_{\text{A},\ldots,\text{G}}, \ell^{\max}_{\text{A},\ldots,\text{G}}, \ell^{\text{mod}}_{\text{A},\ldots,\text{G}}) + \frac{1}{2} \rfloor$	
J (EOT)	$\min\left\{\max\left\{\lfloor \frac{T}{\Delta} + \frac{1}{2} \rfloor \Delta, 0\right\}, e_{\max}\right\},$	$\mu = \log \epsilon_{\text{A},\ldots,\text{G}} - \frac{1}{2} \log\left(\frac{\sigma^2_{\text{A},\ldots,\text{G}}}{\epsilon^2_{\text{A},\ldots,\text{G}}} + 1\right),$
	$T \sim \text{Lognormal}(\mu, s^2)$	$s = \sqrt{\log\left(\frac{\sigma^2_{\text{A},\ldots,\text{G}}}{\epsilon^2_{\text{A},\ldots,\text{G}}} + 1\right)}$
J (ROT)	$\min\{\max\{0, T\}, e_{\max}\},$ $T \sim \text{Gaussian}(\text{EOT}, \sigma^2)$	

were obtained in such a way to equal the expected and the modal values reported in [114]. Further, we compute the value of the survival function on the maximum time for the execution of activities, obtaining a value less than 10% that guarantee a reasonable truncation.

The EOT of the patient i represents a prediction of the surgery duration performed by the surgeons at the moment of the ambulatory visit who indicates the mean duration of similar surgeries on the basis of the own personal experience (in absence of historical data). In the literature, *a priori* surgery duration generally follows a Lognormal distributions (see, e.g.,[111, 120, 123]). Then, the ROT of the patient i has been generated in such a way to replicate the uncertainty pertaining the prediction

Tab. 3.13: Parameters used in the simulation framework

Parameters	unit of measure	Validation	Quantitative analysis
r_0	patients/min	$5.8 \cdot 10^{-3}$	$2.0 \cdot 10^{-2}$
R_0	patients	140	420
p_1, p_2		0.2, 0.1	0.2, 0.1
$T_{\text{min}}^{\text{A},\ldots,\text{F,I}}$	min	$5, 25, 25, 25, 40, 25, 35$	$5, 25, 25, 25, 40, 25, 35$
$T_{\text{avg}}^{\text{A},\ldots,\text{F,I}}$	min	$7.5, 31.5, 31, 28, 62.5, 32, 41$	$7.5, 31.5, 31, 28, 62.5, 32, 41$
$T_{\text{mod}}^{\text{A},\ldots,\text{F,I}}$	min	$6, 30, 26, 25, 50, 30, 40$	$6, 30, 26, 25, 50, 30, 40$
$\ell_{\text{A},\ldots,\text{G}}^{\text{min}}$	days	$2, 1, 1, 1, 1, 1, 1$	$2, 1, 1, 1, 1, 1, 1$
$\ell_{\text{A},\ldots,\text{G}}^{\text{max}}$	days	$29, 16, 7, 9, 5, 5, 5$	$29, 16, 7, 9, 5, 5, 5$
$\ell_{\text{A},\ldots,\text{G}}^{\text{avg}}$	days	$3, 2, 2, 2, 2, 2, 2$	$3, 2, 2, 2, 2, 2, 2$
e_{max}	min	360	420
$\bar{\epsilon}_{\text{A},\ldots,\text{G}}$	min	$145, 171, 149, 153, 171, 164, 166$	$145, 171, 149, 153, 171, 164, 166$
$\sigma_{\text{A},\ldots,\text{G}}$	min	$85, 85, 66, 60, 61, 51, 60$	$85, 85, 66, 60, 61, 51, 60$
σ	min	0	30
d_{tol}	min	30	10
ν	min	0	300
Δ	min	30	30
$b_1, \ldots, b_7$	beds	$18, 18, 18, 18, 18, 18, 18$	$50, 50, 50, 50, 50, 35, 35$

made by the surgeons: we generate a value X using a Gaussian distributions with average 0 and standard deviation σ; then, the ROT value r_i is computed as $e_i + X$.

We observe that the simulation model generates activity durations on the basis of few information reported in [114]. In presence of historical data about surgery durations, it should be used for replicating and predicting surgery durations depending on the characteristics of the patient, learning from the previous experiences. For this purpose several Bayesian methods can be used as reported in the literature (see, e.g., [81, 84, 85]).

4 The Real Time Management of non-elective patients

Non-elective patients should require to be operated on within different but usually tight time limits depending on their urgency. Such time limits can range from "*as soon as possible*" to "*within 24 hours*" [130]. Since the insertion of non-elective patients could have a negative impact on the elective patient scheduling, an appropriate handling of non-elective patients could significantly improve the performance. As discussed in Chapter 2, two main policies can be adopted to deal with both elective and non-elective patient flow, which can be operated on within dedicated (DOR policy) or shared (SOR policy) ORs.

In Chapter 3 we addressed the optimization problems dealing with elective patients, which we call Elective-Oriented Optimization (EOO) problems hereafter. The analysis performed in Chapter 3 tested the effectiveness of the EOO modules dealing with the uncertainty given by the surgery duration. The unpredictable arrival of non-elective patients produce a further uncertainty factor, which introduce a set of Non-elective-Oriented Optimization (NOO) problems.

In this chapter we deal with the real time insertion of non-elective patients, which share the ORs used for the elective surgery (SOR policy), that is the more challenging setting for the optimization of the RTM of ORs. Firstly we present a preliminary analisys performed to test the effectiveness of the approach proposed in Chapter 3 when dealing with an additional non-elective patient flow and using a simple policy consisting in the as-soon-as-possible insertion of such patients. When the whole OR session capacity is allocated to plan elective patients, such insertions will cause an overload that involves an higher demand of overtime, which generally is a scarce resource. Furthermore, from the non-elective patients perspective, the responsiveness of the SP (i.e. the speed at which an OR is available for that surgery) is crucial to guarantee a positive final outcome. Therefore, we would provide an online algorithm that addresses the crucial decision of the insertion of a non-elective patient in one of the available OR sessions.

To this purpose, we extend our simulation model presented in Figure 3.4 of Chapter 3 adding a non-elective patient flow that shares ORs with elective patients, as shown in Figure 4.1. Accordingly, we suppose that non-elective patients have dedicated stay bed units. In other words, the SP of a non-elective patient consists of only two phases, which are the hospital phase and the operating theater phase, but only in the first phase resources are shared with elective patients.

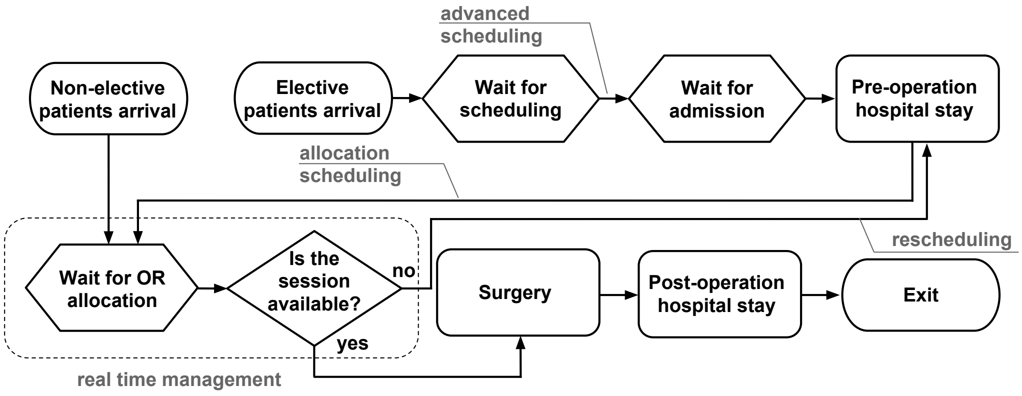


Fig. 4.1: The hybrid model extended to non-elective patients.

This chapter is organized as follows. In Section 4.1 the impact of the insertion of a non-elective patient flow on an optimized SP through EOO modules is analyzed. In Section 4.2 we propose an online approach for the ex-ante solution of the real time insertion of non-elective patients within shared ORs. An offline model is then provided in Section 4.3 for the ex-post solution of such a problem. The effectiveness of the online approach is tested in Section 4.4 with a quantitative analysis that is divided in two phases: in the former an extension to the one provided for the EOO in Section 3.3 of Chapter 3 is provided, while in the latter we make a competitive analysis comparing the ex-ante and the ex-post solutions. Section 4.5 closes the chapter.

4.1 A preliminary analysis

In this section, we would like to evaluate the impact of introducing a patient flow of non-elective emergency surgeries within an optimized SP. Basically, we would evaluate the capability of the offline and online EOO approaches proposed in Chapter 3 of dealing with also elective patients.

In our setting, a patient requiring an emergency surgery is operated as soon as an OR becomes available. This means that no changes are considered in the algorithms for determining a solution for the advanced and the allocation scheduling. More sophisticated algorithm for the insertion of non-elective patients are presented in the next sections of this chapter and in Chapter 5.

The non-elective emergency patient flow is generated in such a way to have, on average, one emergency patient each day having the same EOT and ROT of an elective patient with the highest level of URG (class A) but a short time limit that varies between 30 and 240 min, with step 30 min. We test the SP in Figure 4.1 on scenario 2 (Table 3.3b) of Chapter 3 taking into account the baseline, (13) and (14) configurations, which are those that provide the best results in the analysis for the elective patient flow. The choice to analyze scenario 2 instead of scenario 1 is because a lower number of parallel OR sessions makes more challenging the problem of inserting non-elective emergency patients without worsening the solution from the elective patients point of view. Table 4.1 reports the performance index f_{NE} defined as the fraction of non-elective surgeries started before a fixed time limit. Observe that each column of Table 4.1 consists in a different scenario.

Tab. 4.1: f_{NE} for emergency patients w.r.t. different MTBT

	Non-elective time limit (min)							
id	30	60	90	120	150	180	210	240
(0)	55.7%	75.4%	88.6%	95.1%	98.3%	99.7%	100.0%	100.0%
(13)	52.2%	72.9%	84.1%	88.3%	95.5%	98.1%	99.4%	100.0%
(14)	53.5%	73.1%	83.3%	89.7%	95.0%	98.1%	99.6%	99.9%

Table 4.2 reports the performance of the whole SP after introducing the non-elective patient flow. Note that the last two columns report the value for the indices w_{NE} and f_{NE} referred to the non-elective patients with time limit set to 60 min, where w_{NE} is defined as well as w_{avg}.

Tab. 4.2: Evaluating performance of best configurations in Table 3.7 of Chapter 3

id	o	c	t_{avg}	w_{avg}	f	u_{OR}	u_{bed}	w_{NE}	f_{NE}
(0)	2 302	12.8%	66	1.4	16.4%	95.1%	62.9%	0.60	75.4%
(13)	2 271	16.3%	50	1.0	48.2%	94.4%	73.9%	0.68	72.9%
(14)	2 300	16.0%	52	1.1	41.1%	95.6%	71.0%	0.68	73.1%

As one might expect, it may be noted a general worsening of the patient-centered indices for the elective patients. On the other side, the indices referred to the non-elective patients show quite satisfactory results considering the really tight MTBT and the absence of any NOO approach.

Recalling the model introduced in Section 3.2.4 of Chapter 3, RTM decisions largely depend on the ratio $\frac{\nu_k^\tau}{\nu}$, that is from the total amount ν of overtime available for each planning horizon. Therefore, we would evaluate the overtime available (and the overtime really used) to guarantee the same performance before the introduction of the non-elective patient flow as suggested by Erdem et al. [90]. The amount of overtime available can be interpreted as the hours available of a DOR for non-elective surgeries. Table 4.3 reports about such tests. The first column reports the extra overtime ν_+ available, that is the number of overtime hours added to the initial overtime of 5 h (see Table 3.13 of Chapter 3), that is 1 h/day. Column headed with the index f_{over} reports the average fraction of overtime actually used to operate patients in that scenario.

The first remark is concerned with the overtime percentage effectively used, which decreases as soon as the number of hours weekly available increases. On the other

Tab. 4.3: Overtime estimation

ν_+	o	c	t_{avg}	w_{avg}	f	u_{OR}	u_{over}	u_{bed}	w_{NE}	f_{NE}
0	2 271	16.3%	50	1.03	48.2%	94.4%	98.5%	73.9%	0.68	72.9%
5	2 290	14.5%	41	0.90	71.9%	95.1%	90.5%	71.9%	0.69	72.1%
10	2 351	13.5%	35	0.75	84.6%	96.8%	76.5%	72.6%	0.70	71.6%
15	2 231	13.0%	35	0.80	84.8%	92.7%	66.0%	74.2%	0.67	73.2%
20	2 250	12.1%	33	0.79	86.9%	93.6%	54.8%	72.6%	0.69	72.1%
25	2 316	11.2%	29	0.65	90.6%	97.5%	44.3%	68.9%	0.71	72.4%
30	2 397	10.2%	26	0.59	93.7%	98.6%	37.4%	66.6%	0.73	70.8%
35	2 403	9.7%	28	0.61	93.6%	98.9%	33.8%	66.8%	0.72	71.7%
40	2 424	8.6%	27	0.61	92.8%	99.5%	29.3%	64.5%	0.74	70.7%
45	2 422	8.5%	23	0.54	95.2%	99.5%	26.9%	64.3%	0.73	70.9%
50	2 413	8.0%	23	0.53	96.0%	99.5%	24.8%	62.9%	0.72	71.3%
55	2 406	7.6%	21	0.50	96.8%	99.1%	22.7%	63.5%	0.74	70.1%

side, it seems that about 800 min of overtime are those really used to deal with the emergency surgery flow under the second scenario.

The available overtime seems the more influencing factor. Actually, we can reach about the 90% of elective patients operated within their MTBT by making available 25 h of overtime but using only the 44.3%. On a schedule of five days, the 25 h of overtime could correspond to the availability for 5 h/day of 1 dedicated OR for non-elective surgery. Moreover, a detailed analysis of the DOR policy is presented in Chapter 5.

4.2 Ex-ante approach: an online algorithm

Online NOO approaches have to decide in which OR session the non-elective patient has to be scheduled. Such a decision can determine a different need of overtime or the cancellation of the elective patients previously scheduled. To deal with this online problem, we introduce an algorithm called Non-Elective Worst-Fit (NEW-Fit).

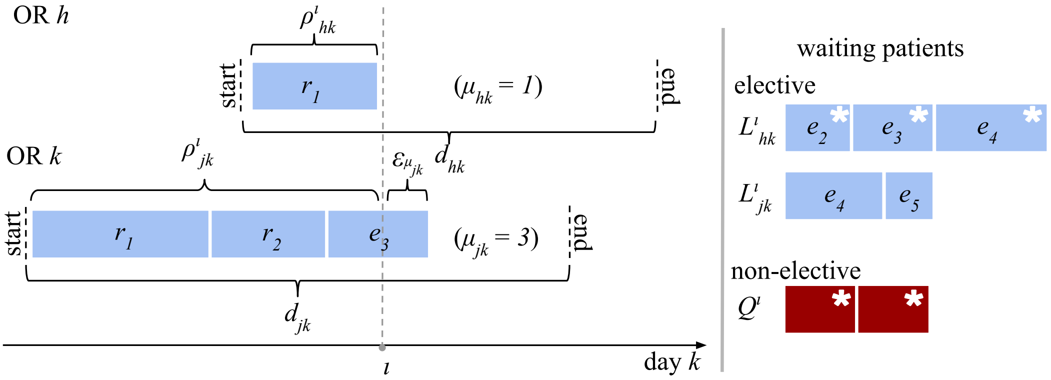
4.2.1 Parameters

We take into account the arrivals of non-elective patients which must be treated within the end of the current day k taking into account time limits $t_i^{\max} < 24$ h. Let S_k be the set of the OR sessions planned on the day k, in which $m_{jk} = |L_{jk}|$ patients are scheduled. At the instant ι of the day k, let h be an operating room available after having operated on μ_{hk} patients $i_1, \dots, i_{\mu_{hk}}$. Let $i_{\mu_{jk}}$ be the patient that is still within the OR, with respect to the other OR sessions (j, k), $j \neq h$. Let L^ι_{hk} be the set of the waiting elective patient scheduled in $(h, k) \in S_k$, that are ordered in the last $m_{hk} - \mu_{hk}$ positions of the sequence λ_{hk} (i.e., $i_{\mu_{hk}+1}, \dots, i_{m_{hk}}$). Let Q^ι be the set of all the waiting non-elective patients at the instant ι. If at that moment the operating room j is available, then the next patient should be selected from $L^\iota_{hk} \cup Q^\iota$. Note that the problem arises only if $Q^\iota \neq \emptyset$. Let us introduce the parameter

$$\epsilon_{\mu_{jk}} = \begin{cases} \max\left(\sum_{i_1,\dots,i_{\mu_{jk}-1}} r_i + e_{i_{\mu_{jk}}} - \rho^\iota_{jk}, 0\right) & \text{if } j \neq h \\ 0 & \text{otherwise} \end{cases}$$

that is the estimated time for the next release of the OR session (j,k).

Fig. 4.2: Parameters defined at the releasing of an operating room.



In Figure 4.2 an example of OR release at the instant ι is reported. In the OR session (h,k) the first surgery is concluded after r_1 minutes, that is the ROT of the first patient, then $\rho^\iota_{hk} = r_1$ and $\epsilon_{\mu_{hk}} = 0$. The time elapsed in the OR session (h,k) is equal to the sum of the ROTs of the operated patient plus the time elapsed from the entry of the current patient. The time required for the end of the surgery of such a patient is estimated by $\epsilon_{\mu_{jk}}$ computed using its EOT. The waiting patients that are candidates for the allocation of (h,k) are represented by boxes marked with an asterisk.

4.2.2 The NEW-Fit algorithm

The algorithm provides an online greedy construction of an alternative schedule of the patients in which we try to insert the non-elective patients in Q^ι. On the basis of this auxiliary schedule, the NEW-Fit re-determines the sequence of surgeries λ_{hk} establishing if to continue with the planned schedule or to insert a non-elective patient as next surgery in (h,k) in such a way to reduce the maximum exceeding time with respect to the duration of the sessions.

The pseudo-code reported in Algorithm 1 describes the algorithm NEW-Fit, having the parameter $\delta \in (0,1]$ which is used to define, for each non-elective patient i, an *early deadline* $\delta t^{\max}_i$ until which the insertion can be planned. The early deadline is introduced in such a way to deal with the uncertainty of the surgery duration. When δ is close to 0, NEW-Fit reduces the risk of exceeding the non-elective time limit. On

Algorithm 1: Non-Elective Worst Fit

```
Input  : $\delta$;
begin
    $p^s \leftarrow i_{\mu_{hk}+1}$;                    /* next elective patient $p^e$ in $L^\tau_{hk}$ */
    $p^u \leftarrow \arg\min_{i \in Q^\iota} (t_i^{\max} - t_i)$;
    $Q' \leftarrow Q^\iota$;
    foreach OR session $(j,k)$ do $L'_j \leftarrow L^\iota_{jk}$; $\epsilon'_{\mu_{jk}} \leftarrow \epsilon_{\mu_{jk}}$;
    $S = (i_1, \dots, i_{\mu_{hk}}, i_{\mu_{hk}+1}, \dots, i_{m_{hk}})$;
    flag $\leftarrow$ false;
    stop $\leftarrow$ false;
    while $Q' \neq \emptyset$ and not stop do
        $p^{ne} \leftarrow \arg\min_{i \in Q'} (t_i^{\max} - t_i)$;
        $x^\star \leftarrow +\infty$;
        $j^\star \leftarrow -1$;
        foreach $(j,k) \in S_k$ do
            $x = \rho^\iota_{jk} + \epsilon_{\mu_{jk}} + \sum_{i \in L'_j} e_i - d_{jk}$;
            if $x < x^\star$ and $\epsilon'_{\mu_{jk}} \leq \delta(t^{\max}_{p^{ne}} - t_{p^{ne}})$ then
                $x^\star \leftarrow x$;
                $j^\star \leftarrow j$
        if $j^\star = -1$ then
            $S = (i_1, \dots, i_{\mu_{hk}}, p^u, i_{\mu_{hk}+1}, \dots, i_{m_{hk}})$;
            stop $\leftarrow$ true
        if $j^\star = h$ and flag = false then
            $p^s \leftarrow p^{ne}$;
            flag $\leftarrow$ true
        $L'_{j^\star} \leftarrow L'_{j^\star} \cup \{p^{ne}\}$;
        $Q' \leftarrow Q' \setminus \{p^{ne}\}$;
        $\epsilon'_{\mu_{j^\star k}} \leftarrow \epsilon'_{\mu_{j^\star k}} + e_{p^{ne}}$;
    if flag = true then $S = (i_1, \dots, i_{\mu_{hk}}, p^s, i_{\mu_{hk}+1}, \dots, i_{m_{hk}})$;
Output: $S$;
```

the contrary, when δ is close to 1, the number of feasible insertions increases and better global solutions can be computed.

After the initialization of the auxiliary data structures, the algorithm starts a loop to determine the auxiliary schedule. At each iteration, the current non-elective patient p^{ne} is scheduled on one of the OR sessions (j,k) such that the condition of the early deadline in correspondence of the instant of insertion

$$\epsilon'_{\mu_{jk}} \leq \delta(t^{\max}_{p^{ne}} - t_{p^{ne}}) \tag{4.1}$$

is satisfied, where $\epsilon'_{\mu_{jk}}$ is equal to $\epsilon_{\mu_{jk}}$ plus the sum of the EOTs of the non-elective patients planned in (j,k) in the previous iterations. The algorithm selects the OR session that minimizes the difference between the estimated total duration of the operated and non-operated patients in L_{jk}

$$\rho^\iota_{jk} + \epsilon_{\mu_{jk}} + \sum_{i \in L'_j} e_i \tag{4.2}$$

and its duration d_{jk}. Such a rule corresponds to insert the patient p^{ne} in the OR session with the maximum unused OR time in such a way to minimize the overtime demand, when d_{jk} is greater than (4.2). The aim is to balance the workload among the OR sessions. At a certain iteration, if the condition (4.1) is not satisfied for any OR session of the day, it means that we are not able to plan all the non-elective patients before their early deadlines, then the NEW-Fit terminates inserting the most urgent non-elective patient p^u as next operation within the sequence λ_{hk}, that is at the $(\mu_{hk} + 1)$-th position. When all the insertions are feasible within the time limits and at least one non-elective patient has been inserted in the OR session (h, k), the NEW-Fit returns adding at the $(\mu_{hk} + 1)$-th position of the sequence λ_{hk} the one with the shortest deadline. Otherwise, the sequence λ_{hk} remains unchanged and the elective patient $i_{\mu_{hk}+1}$ will be the next to be operated on in (h, k).

Finally, we would like to remark that the amount of effective used overtime could slightly exceed the maximum overtime available ν. It depends on whether the overtime is assigned basing the decision on the EOT since ROT is not available. Under special circumstances, extra overtime can be required for the surgery completion but all the available overtime has been previously assigned. In this case, we assume to allow the surgery completion setting the parameter ν equal to the effectively used overtime.

In Table 4.4 we report a summary of the main notation introduced heretofore.

Tab. 4.4: Summary of the main notation used in this Chapter.

Sets	
S: set of all OR sessions	S_k: set of all OR sessions of day k
L_{jk}: set of patients scheduled into (j, k)	λ_{jk}: sequence of patients scheduled into (j, k)
Q^ι: set of waiting non-elective patients	
Indices and cardinalities	
i: elective patient	j: index of the operating room
k: index of the day	h: index of the released operating room
n: number of OR sessions	n_k: number of OR sessions of the day k
m_{jk}: number of patient scheduled into (j, k)	m_{jk}: number of patient operated on into (j, k)
Times and durations	
t_i: waiting time of patient i	$t_i^{\max}$: MTBT of patient i
w_i: normalized waiting time of patient i	$\tilde{w}_i$: value of w_i in the next planning horizon
e_i: EOT of patient i	r_i: ROT of patient i
d_{jk}: duration of (j, k)	ι: general instant during the OR session
ν: overtime available for one planning horizon	ρ^ι_{jk}: time elapsed since the beginning of (j, k)

4.3 Ex-post approach: the offline solution

The online solution is characterized by the lack of knowledge about what might happen in the remaining of the planning horizon. This is due to the difference between estimated and real duration of a surgery, and to the unforeseeable arrivals of non-elective patients.

On the contrary, at the end of the planning horizon we have a complete information about what is happened. Thus is possible to evaluate what would be the optimal decisions to be taken assuming to know in advance all the information that are acquired over time by the online solution. In our case, such information includes the ROTs of the elective patients and the surgery demand of the non-elective patients, that is their amount, the ROTs and the day in which they must be operated on.

We denote this set of decisions as *offline solution*. Such a solution provides a significant contribution to evaluate the effectiveness of the online approach. In this section, first we provide a linear programming model to compute the optimal offline solution in the case of only elective patients. Then, we extend this model to take into account also the non-elective patients.

Fig. 4.3: Surgery process scheduling vs. online scheduling: elective patients.

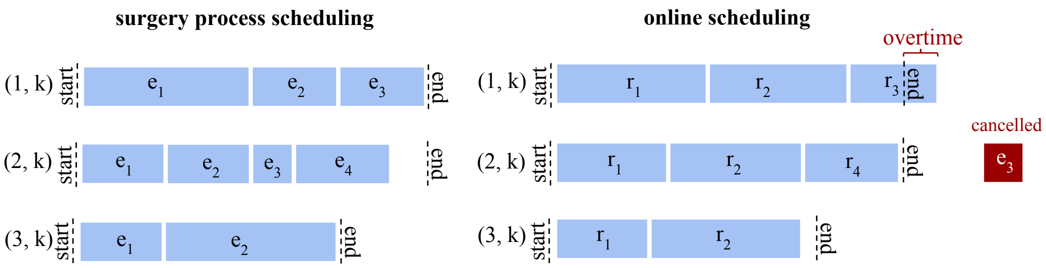


Figure 4.3 reports an example in which the difference between EOTs and ROTs caused the request of an amount of overtime. In the OR session $(1, k)$ the overtime has been allocated in order to operate on the last patient, while in the OR session $(2, k)$ the surgery of the patient with index 3 has been postponed.

To determine an offline solution in the case of dealing with only the elective patients, the only relevant decision is that of postponing the surgery interventions. Since the ROTs are known, we remark that any sequencing of the surgery planned into an OR session determines the same outcome. Thus the sequencing is not relevant for the offline solution. Let us introduce the following decision variables

$$x_i = \begin{cases} 1 & \text{if the surgery intervention of the patient } i \in L \text{ is postponed} \\ 0 & \text{otherwise} \end{cases}$$

and the non-negative integer $\nu_{jk} \in \mathbb{Z}_+$ measuring the overtime assigned to the OR session (j, k).

To model the offline solution we introduce the following constraints:

$$\sum_{i \in L_{jk}} (1 - x_i) r_i \leq d_{jk} + \nu_{jk} \qquad \forall (j,k) \in S \tag{4.3}$$

$$\sum_{(j,k) \in S} \nu_{jk} \leq \nu \tag{4.4}$$

$$x_i = 0 \qquad \forall i \in L_{\text{first}} \tag{4.5}$$

Constraints (4.3) ensure that the overall duration of the surgery performed during the OR session (j,k) can not exceed the duration of the OR session plus the additional overtime assigned. Constraint (4.4) limits the use of the overtime to the maximum overtime available. Finally, we remark that the first patient scheduled in each OR session (j,k) is not the subject of an online decision, that is he/she will be always operated on. Therefore, we are required to model this fact in our offline solution introducing the constraints (4.5) where $L_{\text{first}} \subset L$ is the set of all the patients sequenced as the first of an OR session.

We recall that our online solution would maximize the utilization of the OR sessions and to minimize the number of postponed patients whose $\widetilde{w} > 1$, that is those patients for which the MTBT will be exceeded. Thus our objective function should take into account these requirements.

We define the overall utilization of the OR sessions as the ratio between the total duration of the operated patients and the sum of the duration of all the OR sessions, limited to 1 to avoid greater values, that is when using the overtime

$$u = \min \left\{ \frac{\sum_{i \in L} (1 - x_i) r_i}{\sum_{(j,k) \in S} d_{jk}}, 1 \right\}.$$

To promote a solution with higher utilization, we introduce an auxiliary continuous variable $u \in [0,1]$ and the constraint

$$u \sum_{(j,k) \in S} d_{jk} \leq \sum_{i \in L} r_i (1 - x_i). \tag{4.6}$$

Our aim is to maximize the objective function defined as follows

$$z \equiv (1 - \alpha) u + \alpha \frac{\sum_{i \in L} (1 - x_i) - \sum_{i \in L_{\widetilde{w} > 1}} x_i \psi_i}{|L|}, \tag{4.7}$$

which is the convex combination of two terms in $\alpha \in [0,1]$. The former is the utilization defined by the constraint (4.6). The latter is the number of the patients operated on minus a sum of the penalties associated to those patient whose surgery is postponed and their $\widetilde{w} > 1$. Since the utilization ranges in $[0,1]$, the latter term is normalized on the overall number of scheduled patient $|L|$. The penalties are defined as

$$\psi_i = \widetilde{w}_i^2 \,. \tag{4.8}$$

in order to limit the impact of the symmetries as reported in [95].

Finally, the offline solution in the case of only elective patients can be computed by finding the optimal solution of the following mixed-integer linear program

$$\begin{aligned} M^e: \qquad \max z \quad & \text{s.t. (4.3)–(4.6)} \\ & x_i \in \{0,1\} \quad \forall i \in L \\ & \nu_{jk} \in \mathbb{Z}_+ \quad \forall (j,k) \in S \\ & u \in [0,1] \end{aligned}$$

Fig. 4.4: Surgery process scheduling vs. online scheduling: elective and non-elective patients.

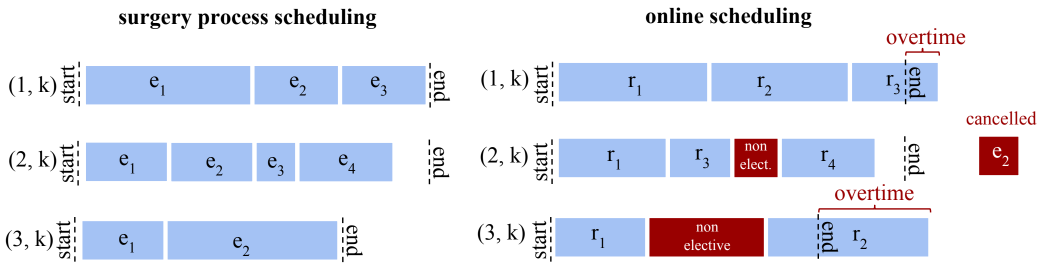


Figure 4.4 reports an example of solution of the problem in which two non-elective patients have been scheduled, determining the request of an amount of overtime for some of the OR sessions.

Model M^e can be modified to address also the management of the non-elective patients. Let Q_k be the set of the non-elective patient arrived the day k. We introduce the following decision variable

$$y_{ijk} = \begin{cases} 1 & \text{if the patient } i \in Q_k \text{ is inserted in the session } (j,k) \\ 0 & \text{otherwise} \end{cases} .$$

The constraints

$$\sum_{(j,k')\in S: k'=k} y_{ijk'} = 1 \quad \forall i \in Q_k,\, \forall k \in K \tag{4.9}$$

ensure that each non-elective patients in Q_k is operated on during only one OR session (j,k). Although the NEW-Fit algorithm presented in Section 4.2 deal with non-elective patients having different time limits $t_i^{\max} \leq 24$ h, we assume that all non-elective patients have a time limit $t_i^{\max} = 24$ h. This allows us to have a simple model for the offline solution. Moreover, this assumption is relaxed in Chapter 5, where a more general analysis is provided. We remark that we have to modify the constraints (4.3) and (4.6) to take into account the insertion of the non-elective patients. By consequence, the new constraints are

$$\sum_{i \in L_{jk}} (1 - x_i) r_i + \sum_{i \in Q_k} y_{ijk} r_i \leq d_{jk} + \nu_{jk} \qquad \forall (j,k) \in S \tag{4.10}$$

$$u \sum_{(j,k) \in S} d_{jk} \leq \sum_{i \in L} (1 - x_i) r_i + \sum_{i \in Q} r_i \tag{4.11}$$

where $Q = \bigcup_{k \in K} Q_k$.

Finally, the offline solution in the case of elective and non-elective patients can be computed by finding the optimal solution of the following mixed-integer linear program

$$\begin{aligned} M^{ne}: \qquad \max z \quad & \text{s.t.} \;\; (4.4)\text{–}(4.5),\ (4.9)\text{–}(4.11) \\ & x_i \in \{0,1\} \quad \forall i \in L \\ & \nu_{jk} \in \mathbb{Z}_+ \quad \forall (j,k) \in S \\ & y_{ijk} \in \{0,1\} \quad \forall i \in Q,\ \forall (j,k) \in S \\ & u \in [0,1] \end{aligned}$$

4.4 Quantitative analysis

This section reports the quantitative analysis performed under several scenarios to evaluate the effectiveness of the proposed online methods providing two different but complementary analysis. The first one is to embed our online approaches on the simulated SP in Figure 4.1 reported at the beginning of this chapter, in such a way to evaluate their impact on the RTM week by week, that is how the previous decisions (e.g., determining less or more cancellations) impact on the current decisions. The second one exploits the computation of the corresponding offline solutions in such a way to assess the competitive of the proposed online solutions.

The optimization modules embedded in the hybrid model are the RTM algorithms presented in Section 4.2 and the following:

Advanced scheduling: as well as in Chapter 3 when all EOO modules (options A–C) are enabled;

Allocation scheduling: patients are sequenced in decreasing order of $\tilde{w}_i$.

Rescheduling: canceled surgeries are rescheduled in one of the OR sessions of the first day of the next week.

We recall that the advanced scheduling is aimed at maximizing the OR utilization. This fact directly influences the number of possible cancellations during the scheduling posing a challenge for the RTM. Furthermore, we recall that it makes really difficult to insert a patient whose surgery has been postponed by the RTM, as reported in Section 3.3 of Chapter 3. This justify our choice to schedule on the next week all the postponed patients. Observe that the policies provided for the allocation scheduling and the rescheduling are simplifications of the ones proposed in Section 3.2 of Chapter 3 that allows us to use simple models for the offline solution.

4.4.1 Scenarios and indices

To avoid to get trapped on a single case study, which could be a limitation from our point of view, we introduce four scenarios in such a way to provide more accurate insights from our quantitative analysis. To this end, we will consider four scenarios (E, NE1, NE2, NE2b) obtained by varying the non-elective arrival ratio and available overtime while the other parameters characterizing them are fixed. All the parameters are reported in Table 4.5.

Tab. 4.5: Parameters characterizing the four scenarios.

		Varying parameters			
scenario	non-elective	Ω	scenario	non-elective	Ω
(E)	—	10 hours	(NE2)	30 per week	50 hours
(NE1)	15 per week	15 hours	(NE2b)	30 per week	40 hours
		Common parameters to all scenarios			
parameter	unit	value	parameter	unit	value
elect. arrival rate	patients/day	25	initial $\|I\|$	patients	500
avg. EOT	minutes	140	s. dev. EOTs	minutes	75
s. dev. $r_i - e_i$	minutes	30	max ROT	minutes	480
n	sessions/week	45	d_{jk}	minutes	480

The flow of elective patients is described in terms of urgency class, frequency and MTBT in Table 4.6. Finally, all the simulation model parameters are the same of those reported in Section 3.5 of Chapter 3.

Tab. 4.6: Urgency classes and MTBTs of the elective patients.

URG class	frequency	MTBT (days)	URG class	frequency	MTBT (days)
A	3%	8	B	5%	15
C	7%	30	D	10%	60
E	15%	90	F	25%	120
G	35%	180			

4.4.2 Results

In this section we report the results of our quantitative analysis obtained by running our methods on the four different scenarios.

In order to provide a term of comparison, we introduce a baseline configuration in which the solution for the RTM is simpler than those proposed in Section 4.2. In the baseline configuration, the resequencing is not performed, the overtime is a priori uniformly distributed among the OR sessions and the non-elective patients are assigned to the first free OR session. The baseline configuration does not claim to fit perfectly the clinical practice (since we are not dealing with a specific case study) but it would represent a more general operative context in which some optimization approaches are performed on the planning side but without taking into account the inherent uncertainty arising in the management of a SP. The introduction of the baseline configuration allows us to evaluate the actual impact of the RTM on the management of the surgical pathway.

Two further configurations are introduced to properly evaluate the online approach in the case of non-elective patients, that is one configuration with the NEW-Fit algorithm (conf. 2) and one without (conf. 1). When NEW-Fit is not considered, the non-elective patients are scheduled as soon as possible.

Tab. 4.7: Performance indices for each scenario and RTM configuration.

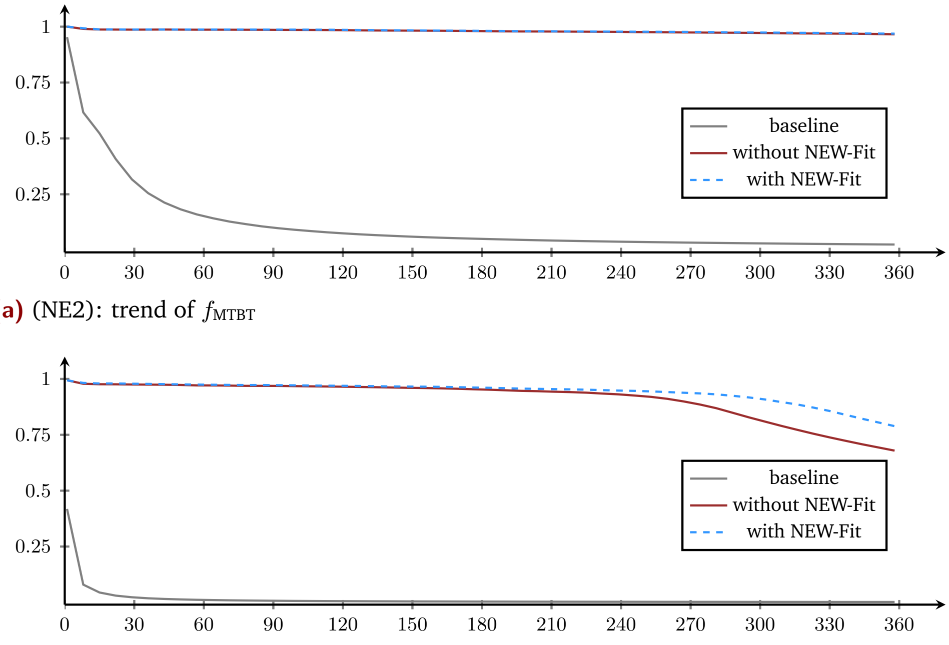
Scenario id	RTM config.	Performance indices o	 c	 t_{avg}	 w_{avg}	 f	 u_{OR}	 u_{over}
(E)	baseline	7 532	6.7%	111	1.05	44.0%	83.4%	13.3%
	1	7 991	5.0%	86	0.81	91.9%	88.5%	56.9%
(NE1)	baseline	7 050	14.2%	130	1.25	17.3%	86.5%	17.5%
	1	7 657	9.1%	98	0.92	65.6%	93.1%	100.0%
	2	7 796	7.6%	93	0.87	76.2%	94.6%	100.0%
(NE2)	baseline	6 866	16.3%	147	1.42	2.5%	93.9%	21.2%
	1	8 131	4.5%	81	0.75	96.6%	100.0%	95.8%
	2	8 147	4.4%	80	0.74	96.8%	100.0%	96.2%
(NE2b)	baseline	6 717	18.5%	154	1.50	0.2%	92.3%	20.2%
	1	7 780	7.4%	99	0.92	66.7%	100.0%	98.0%
	2	7 878	6.9%	96	0.89	77.4%	100.0%	98.4%

Table 4.7 reports the value of the performance indices, which are obtained by taking the corresponding average value running the simulation model 30 times on a given configuration and, each time, starting from a different initial condition. Each run replicates two years of operating room management. Data are collected only on the second year. Remarks on running time are reported in Section 4.4.3.

With respect to the baseline configuration, the reported results showed that the adoption of the online approach for the RTM – both in the case of only elective or non-elective patients – can largely improves the patient-centered performance indices while maintaining the facility-centered ones.

The most relevant improvement is that regarding the fraction f of patients operated within the MTBT, which measure the capability of the hospital to respect their deadlines ensuring to deliver a surgery in a proper way. The results prove the positive impact of the NEW-Fit algorithm. For instance, an improvement of more than the 10% can be observed for the index f on the scenarios (NE1) and (NE2b) while this improvement is limited in the scenario (NE2).

Fig. 4.5: Trend of f (data referred to the 2nd year, days on x-axis, percentage of patients on y-axis).



(a) (NE2): trend of f_{MTBT}

(b) (NE2b): trend of f

Figure 4.5 reports the trend of the f_{MTBT} value during the simulation in order to compare the behavior in the scenario (NE2) (Figure 4.6a) and (NE2b) (Figure 4.6b). As reported in Table 4.5. the difference between the two scenarios is the amount of available overtime, that is 50 hours for (NE2) and 40 for (NE2b). While in Figure 4.6a the f is quite stable along the time, we observe that in Figure 4.6b a drop is reported after about 240 days. This highlights the fundamental role of the overtime as a flexible resources, when the required amount is correctly evaluated. We also remark that the NEW-Fit is able to limit the negative impact of an overtime underestimation (Figure 4.6b).

Table 4.8 reports the competitive analysis, that is the comparison between online and offline solutions. The results are obtained as follows. Among the 30 runs reported before, we selected the run whose performance indices are closest to the average values of the performance indices. From this run, we extracted the information required to generate the 52 instances corresponding to the second year of simulated time. Finally, we computed the optimal solution for each of the 52 instances by solving the corresponding mixed-integer linear problem.

Therefore, Table 4.8 reports the average values over 52 instances of the following quantities: π and π' are respectively the number of elective patients scheduled and the number of elective patients scheduled whose $\tilde{w}_i > 1$; γ and γ' are respectively the number of cancellation and the number of cancellation whose corresponding patients have $\tilde{w}_i > 1$; z_{avg} is the value of the objective function (4.7). Finally, the columns regarding the competitive ratio report both the experimental average and worst ratio.

The competitive analysis confirms the quality of the online solutions. In particular, the analysis of the z_{avg} values and the average and the worst competitive ratio values validates the remark about the positive impact of the NEW-Fit algorithm.

Tab. 4.8: Comparison between online and offline solutions.

Scenario	RTM	input	online sol		offline sol		comp. ratio		time
id	config.	π (π')	γ (γ')	z_{avg}	γ (γ')	z_{avg}	avg.	worst	secs.
(E)	baseline	157 (78)	12 (5)	0.8607	3 (1)	0.9347	1.09	1.15	0.12
	1	161 (8)	7 (0)	0.9243	4 (0)	0.9467	1.02	1.09	0.11
(NE1)	baseline	156 (109)	20 (13)	0.8022	2 (1)	0.9830	1.23	1.45	0.96
	1	161 (42)	13 (2)	0.9188	2 (0)	0.9883	1.08	1.26	35.31
	2	161 (24)	12 (0)	0.9353	2 (0)	0.9894	1.06	1.13	3.98
(NE2)	baseline	154 (138)	22 (20)	0.7708	0 (0)	0.9978	1.30	1.63	0.20
	1	163 (6)	8 (0)	0.9768	1 (0)	0.9971	1.02	1.06	68.75
	2	163 (5)	7 (0)	0.9786	1 (0)	0.9972	1.02	1.06	0.53
(NE2b)	baseline	152 (147)	23 (22)	0.7317	1 (1)	0.9929	1.37	1.78	0.49
	1	161 (37)	12 (1)	0.9616	2 (0)	0.9932	1.03	1.12	117.31
	2	161 (31)	10 (0)	0.9681	2 (0)	0.9939	1.03	1.07	110.16

The analysis of the average competitive ratio proves the challenging of the problem of dealing with the management of a flow of elective and non-elective patients. Actually, the competitive ratio of the baseline solution largely increases as soon as the arrival rate of the non-elective increases or the available overtime is tight. On the contrary, the competitive ratio of the configurations 1 and 2 is quite stable. Furthermore, the gap between the two competitive ratios (baseline vs. configuration 1 or 2) is quite acceptable for the scenario (E) while increases for the other non-elective scenarios demonstrating the need of an online solution to cope in a effective way the management of non-elective patients.

4.4.3 Computational remarks

The results reported in Section 4.4.2 are obtained running our computational tests on a 64 bit Intel Core i5 CPU with 3.33GHz and 3.7GB of main memory.

On average, one single run of the simulation model requires from 7.0 to 20.5 seconds when running with scenario (E) and baseline configuration or with scenario (NE2b) and configuration 2, respectively. This means that no more than 615 seconds are needed to simulate two years of operating room management. Finally, we remark that the algorithm for the advanced scheduling is the most time consuming component while the running time required by the other components is negligible.

The mixed-integer linear programs are solved using IBM ILOG CPLEX Optimization Studio 12.3. The CPLEX running time are reported in the last column of Table 4.8. Note that usually few seconds are enough to solve an instance of the offline problem. The high average values are determined by few instances requiring a lot of time to close the optimality gap. For example, the number of instances requiring more than 5 seconds in the scenario (NE2b) are 10 and 4 for configuration 1 and 2, respectively. This is probably due to the large number of symmetries determined by the decision variables y_{ijk} for a given day k.

4.5 Concluding remarks

In this chapter, we considered a challenging extension of the RTM of ORs, considering a joint flow of elective and non-elective patients. We evaluated the effectiveness of the RTM on a simulated surgical clinical pathway under several scenarios and also reporting a competitive analysis with respect to an offline solution obtained by solving a mixed-integer programming model. The quantitative analysis showed the capability of the online solutions to address the inherent uncertainty of the RTM determining a general improvement of the patient-centered indices without deteriorating the facility-centered ones. Further, the analysis of the competitive ratios confirmed the challenging of the problem of dealing with a flow of non-elective patients sharing the ORs with a flow of elective patients.

From a methodological point of view, our analysis suggested that online optimization can be a suitable methodology to deal with the inherent stochastic aspects arising in the majority of the health care problems. Although online optimization does not exploit sophisticated mathematical approaches, the competitive analysis reported in Table 4.8 suggested its capability to deal with the stochastic aspects of a problem whenever such aspects are embedded into a well-structured optimization problem, such as those arising in the health care management.

5 Dedicated vs. Shared Operating Room Policies

The aim of this chapter is to exploit the hybrid model presented in Chapters 3 and 4 to enable the analysis and the comparison of the DOR and the SOR policies for the management of both elective and non-elective surgery. In this way we provide a tool capable to support exhaustively the decision problems in the surgery process scheduling. Indeed, the generality of the proposed model allows us to replicate and to compare a wide range of possible scenarios and policies, in which most of the case studies of the literature can be included.

The chapter is organized as follows. Common approaches in literature for the DOR and SOR policies are reported in Section 5.1. Online and offline algorithms for the optimization of the DOR and SOR policies are described in Section 5.2. The computational environment is defined in Section 5.3 describing scenarios, configurations and performance indices. Based on this environment, a comprehensive quantitative analysis is reported in 5.4: we evaluate the DOR, the SOR and the hybrid policies determining, for each policy, the best configuration with respect to the considered scenario; then, we use such configurations in order to compare the three policies in such a way to derive some insights in terms of supporting decision making; finally, the analysis also proves the effectiveness of the proposed approaches. Section 5.5 closes the chapter.

5.1 Common problems and approaches from literature

Since the DOR policy allows us to consider separately the two flows of patients, the elective patient flow is managed considering $S_E \subset S$, that is the set of OR sessions dedicated to the elective surgery, as we proposed in Chapter 3, while the non-elective surgery flow is simply managed in the remaining and dedicated OR sessions. In addition to the decisions regarding the management of the elective patients, the DOR policy imposes a further decision, that is how many OR sessions should be allocated for elective and non-elective surgery.

Under the SOR policy, Van Riet and Demeulemeester [130] identify two classes of methods to deal with the *non-elective insertion*, that is the *slack management* and the *Break-In-Moment optimization*.

When elective and non-elective surgeries are performed in the same ORs, the schedule of the elective patients should take into account the possible insertion of non-elective patients during the execution of the OR sessions. If the whole session capacity is allocated to plan elective patients, such insertions will cause an overload that involves an higher demand of overtime, which generally is a scarce resource. Therefore slack management policies are introduced to avoid the increase of the cancellations [60, 132, 134, 135]. Different policies are obtained on the basis of two decisions, that is (i) the total amount b_{jk} of time reserved during the elective advanced scheduling and (ii) the distribution of the slacks within the schedule. Note that such decisions deal with the trade-off between cancellations and OR utilization, as well as having a different impact on the two flows of patients.

Tab. 5.1: Categorization of non-elective patients with respect to their time limit (extracted from [130]).

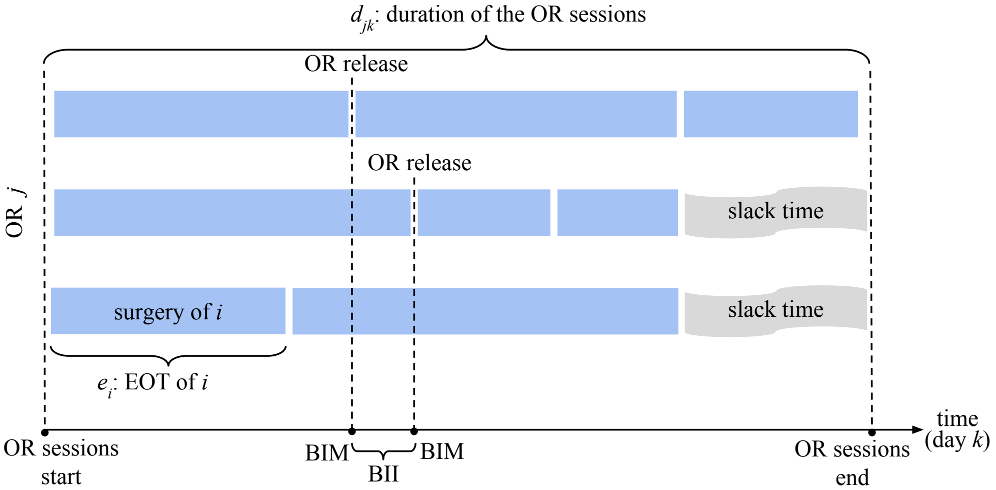
Category	Time Limit	Category	Time Limit
Trauma	30 minutes	*Emergent*	from 30 minutes to 24 hours
Urgent	from 4 to 24 hours	*Semi-urgent*	from 8 hours to 3 days
Add-on	24 hours	*Work-in*	from 24 hours to 3 days

Non-elective surgeries should be performed within a time limit that varies in accordance with their urgency, as reported in Table 5.1, taken from the survey by Van Riet and Demeulemeester [130]. Such surgeries can be inserted in the schedule of the elective patients once an going elective surgery has finished. We denote these completion times of the elective surgeries by Break-In-Moments (BIMs). Then, the waiting time for emergency surgeries depends on these BIMs, whose optimization consists in sequencing the elective surgeries in their assigned OR, such that the intervals between consecutive BIMs is minimized. To the best of our knowledge, only Essen et al. [91] deal with the BIM optimization problem, proposing an offline approach.

Let ι be the BIM in which the m-th patient leaves the OR session (j,k) and let Q^{ι} be the set of the non-elective patients waiting for an insertion at that instant. If $Q^{\iota} \neq \emptyset$, then the sequence λ_{jk} could be modified inserting one of such non-elective patients $i^{ne} \in Q^{\iota}$ at the position $m+1$ and shifting of one position the last $m_{jk} - m$ patients. Otherwise the sequence λ_{jk} could remain unchanged and the non-elective patients will wait the next BIMs for the insertion. The BIM optimization consists in determining for each day k of the planning horizon the set Λ_k of all the sequences of surgeries λ_{jk} that minimizes the time between two consecutive BIMs, called Break-In-Interval (BII), in such a way to lower waiting time for the non-elective patient. The information available for the computation of the BIMs is the EOT of the patients $i \in L_{jk}$.

This problem requires to be addressed during the allocation scheduling since the only way to change the BIMs is to determine an alternative surgery sequencing. Figure 5.1 reports an example of scheduling with three ORs planned for the day k, reserving slacks in two of them. We supposed to have OR sessions with the same duration and starting at the same time. Each gray box represents the surgery of an elective patient that has been placed according to λ_{jk}. The length of the box expresses the EOT of the corresponding patients, causing different BIMs corresponding to all the OR releases during the day k. Two consecutive BIMs have been indicated with a dashed vertical line: their distance of time determines one of the BIIs. From a real time management perspective, the uncertainty can change the BIMs determining the need of an online re-sequencing.

Fig. 5.1: Slacks, BIMs and BIIs – example of configuration with three OR sessions.



In addition, the non-elective insertion problem should consider also the real time decision of the OR sessions in which the surgery of the non-elective patient should be inserted: the decision should reach a good trade-off between the waiting time of the non-elective patients and the cancellation of the elective surgeries. We call this problem Non-Elective Real Time Insertion (NERTI). The literature analysis reveals that such a decision is not considered, and for this reason we propose an online approach in Section 5.2.3.

5.2 Optimization of the non-elective patient flow

As discussed in Section 5.1, the problem of inserting non-elective patients can be tackled with three methods, that is the slack management, the BIM optimization, and the NERTI.

5.2.1 Slack management

Before scheduling the elective patients, there are two different choices regarding the slack management that should be taken. The first decision is in which OR session to provide a slack, which means to decide the number $n_s < n_k$ of ORs that will contain a slack during the day k. The second decision is about the fraction π of time to reserve in each of those OR sessions with respect to their total duration. The couple of parameters (n_s, π) will indicate that a slack of duration πd_{jk} has been reserved in each of the OR sessions (j, k), with $j = n_k - n_s + 1, \ldots, n_k$ and $k \in K$, while the OR session (j, k), with $j \leq n_k - n_s$ are entirely available during the advanced scheduling.

5.2.2 BIM optimization

With BIM (or BII) optimization we refer to the problem of sequencing the surgery of scheduled patients within the OR sessions of the day k in such a way to minimize the waiting time of possible non-elective patient arrivals. The problem has a strong stochastic component because of the unpredictability of the non-elective and their characteristics, that is the time of arrival, the surgery duration and the urgency (with the corresponding time limit). Although in literature such a problem is addressed before the beginning of the OR sessions, that is during the allocation scheduling, we also take into account the possibility of optimizing the BIIs configuration during the execution of the OR session, in such a way to exploit the updated information, that is the ROT r_i of the patient i operated on (instead of the estimation e_i) and the insertion of non-elective surgery already performed.

We propose the Break-In Layout Local Search (BILLS), an algorithm inspired to that proposed in [91], but capable to deal with the elective patients close to their MTBT. The algorithm tries with a local search to improve an initial solution $\Lambda_k = \{\lambda_{jk}\}_j$ exchanging pairs of patients in the same sequence λ_{jk} in such a way to minimize an objective function accounting for the waiting time of the elective patients. The algorithm ends when there is no improvement of this function in the neighborhood. We propose two alternative objective functions to

$$z_1 = \max_{m \geq 1} (\iota_m - \iota_{m-1}) \tag{5.1}$$

$$z_2 = \frac{1}{d} \int_0^d \beta(t) dt \tag{5.2}$$

where ι_0 is the instant in which the OR sessions begin, ι_m are all the instants corresponding to all the other ordered BIMs $(m = 1, 2, \ldots)$, d is the duration of the OR sessions of the day and $\beta : [0, d] \to [0, d]$ is the function which associates to each instant the time remaining to the release of the next OR. The former objective function represents the longest time interval between two consecutive BIMs. The latter is the average value of the estimated waiting time with respect to the overall duration of the OR sessions. Note that using deterministic surgery durations (EOTs), equation (5.2) is equal to

$$\frac{1}{d} \sum_{m \geq 1} \frac{(\iota_m - \iota_{m-1})^2}{2}$$

therefore we can define the equivalent, but simpler, objective function

$$z_2' = \sum_{m \geq 1} \left(\iota_m - \iota_{m-1}\right)^2 .$$

Fig. 5.2: BIM optimization: computation of the objective functions z_1 and z_2 (we supposed that all the OR sessions begin at the same instant and that $d_{jk} = d$ for all (j,k)). In the figure, the decreasing function $b(t)$ from ι_i to ι_{i+1} $(i = 0, \dots, 4)$ measures the remaining time to ι_{i+1}, that is the waiting time of a non-elective patients arriving at $t \in (\iota_i, \iota_{i+1})$.

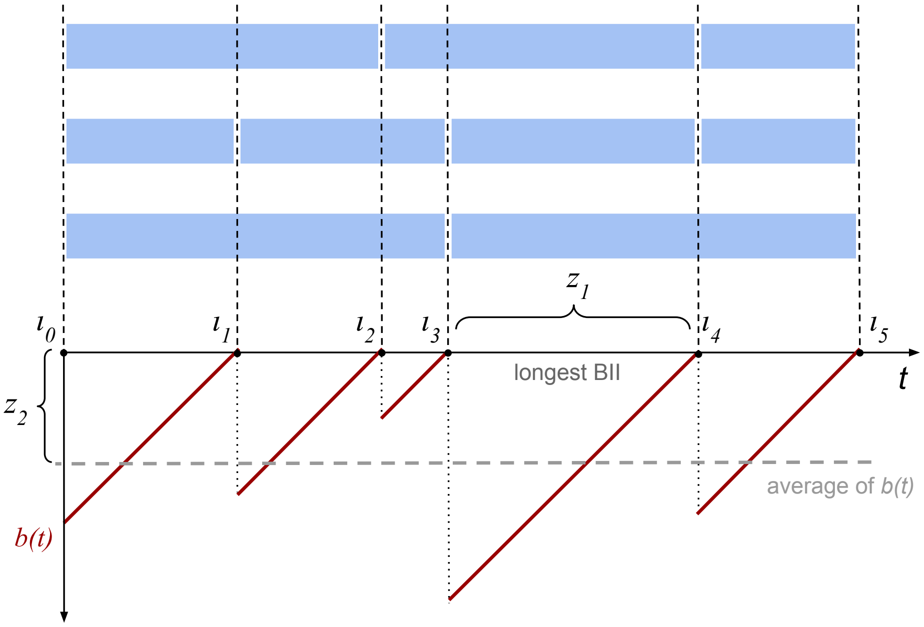


Figure 5.2 shows an example for the elective surgeries schedule and the corresponding values of the objective functions z_1 and z_2. In the lower part of the figure, the piece-wise linear function $b(t)$ has been obtained starting from the BIMs $\iota_0, \dots, \iota_5$. Note that $b(t)$ is equivalent to $\beta(t)$ when deterministic times are considered.

Since patients close to their MTBT are scheduled by the LPT$^+$ algorithm (reported in Section 3.2 of Chapter 3) to avoid a cancellation, we impose that such patients can not be swapped during the local search.

We use two versions of this algorithm: an offline version will be used for the allocation scheduling at the beginning of each day while an online version will be used every time an operating room is released.

5.2.3 NERTI

The insertion of a non-elective patient within a certain OR session determines (i) the shift of the remaining elective surgeries and (ii) the variation of the BIIs configuration determining an effect to the other non-elective patients. Such modification can have an impact that should be considered. With NERTI we refer to the problem of deciding when and in which OR session a non-elective patient could be inserted, which requires an online approach because such a patient could arrive in any instant during the day asking for a surgery within a short time limit. Such a decision could determine (i) a different need of overtime or the cancellation of the elective patients previously scheduled and (ii) longer waiting times of further more urgent non-elective patients (still not arrived). To deal with the two different impacts discussed above, we propose two algorithms: the NEW-Fit algorithm and the Non-Elective Insertion Criterion (NEIC). The latter is presented in Section 4.2 of Chapter 4, the former is explained below.

The NEIC algorithm establishes the best BIM for inserting a non-elective patient on the basis of the number of BIMs available up to the end of the OR session. The idea is to schedule a non-elective patient only when a sufficient number of BIMs is available in the next minutes, in such a way to guarantee the insertion of further and more urgent arriving non-elective patients.

Let $\delta \in (0, 1]$ be the parameter for defining the early deadline of the non-elective patients' time limit, that is $\delta t_i^{\max}$. Let i be the non-elective patient with the minimum value of $t_i^{\max} - t_i$. On the basis of the EOTs of the elective schedule, let $\bar{\iota}$ be the time estimated for the next OR release, which is the first BIM after the time t. Finally, let $\eta(t_0, T)$ be the number of BIMs within a certain interval of time (t_0, T). Then the patient i is inserted in the released OR (h, k) if and only if at least one of the following conditions is satisfied

$$\begin{aligned} t_i^{\max} - t_i &\leq T && (5.3)\\ \delta(t_i^{\max} - t_i) &< \bar{\iota} && (5.4)\\ \frac{\eta(t, t + \delta(t_i^{\max} - t_i))}{\delta(t_i^{\max} - t_i)} &\leq \frac{\eta(t, t + \delta T)}{\delta T} && (5.5) \end{aligned}$$

otherwise the schedule remains unchanged.

Let i^+ be a possible further patient that arrives right after the entry of a patient in the OR session (h, k), which we have to allocate within the shortest time limit T, that is the worst case for our online problem. The condition (5.3) is satisfied if the patient i is closer to the time limit than i^+, while the condition (5.4) is satisfied if patient i can not wait until the next OR release without exceeding the early deadline. In both cases, it is not convenient to optimize the waiting time of further non-elective patients, because of the short time limit of an already waiting patient. The condition (5.5) is satisfied when the frequency of BIMs in the next $\delta(t_i^{max} - t_i)$ minutes, that is the early deadline of i, is lower than the frequency of BIMs in the next δT minutes, that is the early deadline of i^+. In this case it is better to insert i even if $\delta(t_i^{\max} - t_i) > \delta T$ because there are more frequent BIMs in the next minutes than hereafter.

5.3 Setting up the computational environment

We performed a quantitative analysis in order to assess the impact of the different policies (the DOR, the SOR and hybrid policies) and the optimization approaches when they are used separately or jointly.

5.3.1 Scenarios

We introduce four scenarios S_1–S_4 in such a way to provide more accurate insights from our quantitative analysis. Such scenarios are obtained by varying the number of OR sessions per day and the distribution of the surgery durations to represent different settings of the operating theater and different characteristics of the patient population, respectively. We fixed the arrival rate of the patients in such a way to match the the surgery time per week with the total number of arriving patients per week multiplied for the average surgery duration. Note that the total duration of the OR sessions is just sufficient to operate all the patients if their durations were deterministic and if it were possible to predict the non-elective arrivals and to perform the advanced scheduling with the 100% of the OR utilization. This choice allows us to have four scenarios in which the capacity is adequate to the need of interventions, but extra time is necessary to deal with uncertainty: we provided 30 minutes of overtime per OR session.

Tab. 5.2: Parameters characterizing the four scenarios.

Varying parameters

scenario	**capacity**		**EOT distribution**			**patients**	
id	OR sessions	ν	e_0	μ_{EOT}	σ_{EOT}	arrival rate	initial
S_1	10 per day	25 hours	60 min	150 min	60 min	160 per week	400
S_2	5 per day	12.5 hours	60 min	150 min	60 min	80 per week	200
S_3	10 per day	25 hours	30 min	90 min	15 min	266 per week	400
S_4	5 per day	12.5 hours	30 min	90 min	15 min	133 per week	200

Common parameters to all scenarios $S_1 - S_4$

		patient distribution			
		elective (85% of the total)		**non-elective** (15% of the total)	
parameter	**value**	**urgency** (freq.)	**MTBT**	**type** (freq.)	**time limit**
d_{jk}	480 min	A (3%)	8 days	trauma (20%)	30 min
$e_{\max}$	480 min	B (5%)	15 days	emergent (40%)	90 min
q	15 min	C (7%)	30 days	urgent (30%)	3 hours
σ_{ROT}	30 min	D (10%)	60 days	add-on (10%)	24 hours
$r_{\min}$	15 min	E (15%)	90 days		
$r_{\max}$	480 min	F (25%)	120 days		
		G (35%)	180 days		

The main parameters used in the four scenarios $S_1 - S_4$ are summarized in Table 5.2, in which we adopted the same terminology introduced in Table 5.1 for the definition of the time limit for non-elective patients, that is *trauma*, *emergent*, *urgent*, *add-on*. According to [121, 122], the EOT of each patient is obtained generating a value with a 3-parameters Lognormal distribution of minimum value e_0, average μ_{EOT}

and standard deviation σ_{EOT}, truncated to the maximum value $e_{\max}$. Such values are then approximated to the nearest multiple of a discretization constant q, which models the estimate made by the physician during the pre-operative visit. Once the EOT e_i has been determined, the ROT is generated with a Gaussian distribution with average e_i and standard deviation σ_{ROT}, truncated to the minimum and maximum values $r_{\min}$ and $r_{\max}$. We observe that if m_{jk} patients are scheduled in the OR session (j, k), then the total duration of their surgeries is a random variable with average equal to the sum of the EOTs and standard deviation $\sqrt{m_{jk}}\sigma_{\text{ROT}}$. For instance, if 4 patients with EOT of 75, 105, 120 and 180 min are scheduled in an OR session of duration 480 min and we fix $\sigma_{\text{ROT}} = 30$ min, then the total surgery duration will have average 480 min and standard deviation 60 min, which means that the probability of exceeding more than 30 and 60 min the closing time is 31% and 16%, respectively.

Finally, we remark that our model allows to modify the settings of all the parameters reported in Table 5.2 in order to represent a large variety of operative conditions.

5.3.2 Configurations

In order to provide a term of comparison in our quantitative analysis, we introduce a baseline configuration valid for the DOR, the SOR and the hybrid policies, in which:

- advanced scheduling is performed using the same greedy algorithm, that is executing only the first step of the metaheuristic introduced in Section 3.2 of Chapter 3;
- allocation scheduling is performed by ordering the elective patients in decreasing order of w_i;
- resequencing is never performed;
- the overtime is subdivided a priori among assigning the amount $\frac{\nu}{n}$ to each OR session, which is always allocated to patients in need until exhaustion;
- all non-elective patients are inserted as soon as possible in the first dedicated or shared (in accordance with the policy used) OR session, giving the priority to the patient closest to the time limit.

We remark that the baseline configuration for the DOR policy has an additional parameter representing the number of daily ORs dedicated to non-elective patients.

Starting from the baseline configuration, further configurations can be obtained enabling several optimization modules. Table 5.3 reports all the EOO and the NOO modules we will consider in the quantitative analysis reported in Section 5.4, specifying the problem in which they are included and the parameters required ($z = z_1$ or z_2 is the objective function used in the BILLS algorithm). Observe that module $\mathcal{A}$ is introduced in Section 3.2 of Chapter 3, while module $\mathcal{E}$ is presented in

Tab. 5.3: Optimization modules available for the three different phases of the surgery process scheduling: the first column denotes the optimization module and its parameter(s).

mod.(par.)	**description**	type		advanced	allocation	
		EOO	NOO	sched.	sched.	RTM
$\mathcal{A}$	Greedy + Local Search	√		√		
	LPT modified	√			√	
	Best Fit resequencing	√				√
	Overtime criterion	√				√
$\mathcal{B}(n_s, \pi)$	Slack		√	√		
$\mathcal{C}(z)$	offline BILLS		√		√	
$\mathcal{D}(z)$	online BILLS		√		√	√
$\mathcal{E}(\delta)$	NEW-Fit		√			√
$\mathcal{F}(\delta)$	NEIC		√			√

Section 4.2 of Chapter 4, and all the other modules are proposed inSection 5.2 of this chapter. Since the aim of this study focuses on the impact of the DOR, the SOR and the NOO optimization approaches, we will study only the overall impact of the EOO approaches (module $\mathcal{A}$). We refer to Chapter 3 for a complete analysis, on the basis of which we define an unique best EOO module (the configuration giving the best overall performance) that will be used in our quantitative analysis.

Finally, we remark that in Section 5.4, we report only the results of several representative configurations with the aim of giving a general idea of the analysis that our model allows us to do. This choice is determined by the high numbers of possible configurations: as a matter of fact, limiting both the parameters π and δ to 4 different values, there are 104 possible configurations for the DOR, $11\,160$ for the SOR and $145\,080$ for the hybrid policies.

We use the set indices in such a way to evaluate the performance of each representative configurations from both the patient and the facility point of view. Table 5.4 reports the indices used for the quantitative analysis. Observe that to the most representative indices already defined in Table 3.2 of Chapter 3 and the overtime utilization f_{over}, new indices are introduced to represent the non-elective patients perspective.

The strong trade-off among the facility- and the patient-centered indices does not allow us to state what configurations are better than the others, because it depends on the particular scenario and the individual objectives of hospital managers. In order to provide a concise analysis, we define an objective function Z that allow us to determine uniquely what is the more rational configuration, that is

$$\max \quad Z = 3f_{\text{E}} + (1 - c) + 4f_{\text{NE}} + 2u_{\text{OR}}. \tag{5.6}$$

We derived the equation (5.6) in such a way to balance the contribution of the performance indices related to different stakeholders. We included four performance

Tab. 5.4: Patient-centered and facility-centered indices.

Index	Definition
	Facility-centered
u_{OR}	OR utilization
u_{over}	overtime utilization
	Elective patient-centered
o	number of elective surgeries performed
c	fraction of cancellations
t_{avg}	average waiting time spent by elective patients in the waiting list
f	fraction of elective patients operated within the MTBT
w_{avg}	average value of elective patient's $w_i = t_i/t_i^{\max}$ at the time of their surgery
	Non-elective patient-centered
w_{NE}	average value of all non-elective patient's w_i at the time of their surgery
$w_{\text{tr,em,ur,ad}}$	average value of patient's w_i at the time of their surgery in the classes "trauma", "emergent", "urgent" and "add-on"
f_{NE}	fraction of all non-elective patients operated within the time limit
$f_{\text{tr,em,ur,ad}}$	fraction of patients operated within the time limit in the classes "trauma" "emergent", "urgent" and "add-on"

indices from Table 5.4: f and c to consider the the elective patients point of view, f_{NE} to take into account the non-elective patients point of view and u_{OR} for the facility-centered aspect. The coefficients have been fixed in order to assign the 40% of the weight to both elective and non-elective patients and the remaining 20% for the efficiency point of view. We observe that $Z \in [0, 10]$ is equal to 10 when the OR sessions are fully utilized, there are not cancellations and all (elective and non-elective) patients are operated within their time limits. The objective function Z can be redefined changing the weights and/or involving other indices in such a way to account for the different perspectives of the stakeholders.

5.4 Quantitative analysis

In this section we report the results obtained by the quantitative analysis described in Section 5.3. In Section 5.4.1 we discuss the analysis for the DOR policies, which are more straightforward than the SOR ones because of the reduced number of possible configurations. Then, the analysis of the SOR policies is reported and discussed in Section 5.4.2 with a particular attention to the evaluation of the several NOO algorithms introduced in Section 5.2. Starting from the best configuration for the DOR policies, we provided the analysis for the hybrid policies in Section 5.4.3. Finally, we compare the performance of all the best configurations of the different policies in Section 5.4.4.

All the results reported in this section are the average value of the performance indices over 30 different simulation runs for each scenario and configuration. Each run starts from a different seed in such a way to obtain an independent and identical distributed replication. A time horizon of two years has been fixed: after a warm up period of one year, the steady state results are collected over the second year. This allows us also to appreciate the impact of decisions over time and not only over the single planning horizon of one week. Such parameters are those already used in Chapter 3 in which the patient pathway has been validated.

For each policy, we focus on the results of the scenario S_1 showing the impact of the EOO and, after, the effect of each single NOO module on the performance. On the basis of the best values of Z, we also evaluate the impact of enabling all the best NOO modules at the same time. Because of the huge number of configurations, we will show only the best configuration for the scenarios S_2–S_4 to remark that different scenarios could require a different approach.

The average execution time for a single simulation running over the whole time horizon ranged between 7 and 348 seconds, depending on the fixed scenario and configuration. The scenario S_3 required the longest computational times, because of the higher amount of patients. For the same reason, S_2 had the best performance in terms of execution time: all configurations required on average less than 23 seconds for each complete run. In general, the greater impact is given by the optimization modules $\mathcal{A}$ and $\mathcal{D}$, because of the use of a local search algorithm in both of them. However, the times required are satisfactory for our aims.

5.4.1 Dedicated Operating Room

We simulate different configurations of the DOR, which are obtained varying the number of dedicated OR sessions over the total number of 10, using or not the EOO modules and adopting or not a policy $\mathcal{G}$ for the immediate insertion of trauma patients: while as default they can access only to the dedicated ORs, adopting this policy they are allowed in any OR that is released first. The reason of such a policy is the need of an immediate intervention for the patients of this type. The main results about the scenario S_1 are reported in Table 5.5, in which several baseline configurations are obtained varying the number of dedicated ORs.

As expected, increasing the number of ORs dedicated to non-elective patients, their waiting times decrease allowing the respect of the time limits. However this causes a worsening of the elective patient performance, which have less available resources, but also a lower OR utilization.

Figure 5.3(a) remarks the strong trade-off between the percentage of elective and non-elective patients operated on time, while Figure 5.3(b) focuses on the different

Tab. 5.5: DOR – Scenario S_1 with 10 ORs – main performance indices.

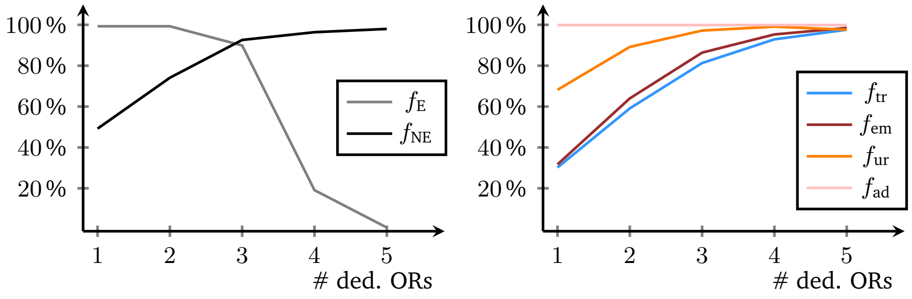
conf. id	# ded. ORs	enabled modules	Performance indices u_{OR}	u_{over}	o	c	t_{avg}	f	w_{avg}	f_{NE}	w_{NE}	Z
A1	1		78.0%	9.3%	6.5k	6.9%	58	99.0%	0.52	47.1%	2.01	7.344
B1	1	$\mathcal{G}$	78.6%	9.3%	6.5k	6.9%	56	99.0%	0.50	49.3%	1.89	7.446
C1	1	$\mathcal{A}, \mathcal{G}$	86.6%	57.8%	7.2k	0.0%	11	99.3%	0.13	49.2%	1.88	7.681
A2	2		76.4%	9.0%	6.0k	7.1%	92	80.1%	0.82	72.2%	1.01	7.747
B2	2	$\mathcal{G}$	76.8%	9.0%	6.0k	7.1%	91	81.2%	0.81	73.7%	0.92	7.849
C2	2	$\mathcal{A}, \mathcal{G}$	85.6%	57.0%	6.8k	0.0%	39	99.3%	0.36	74.1%	0.90	8.656
A3	3		71.6%	8.7%	5.4k	7.3%	132	22.5%	1.22	87.1%	0.43	6.520
B3	3	$\mathcal{G}$	71.8%	8.6%	5.4k	7.2%	131	23.7%	1.21	87.8%	0.42	6.586
C3	3	$\mathcal{A}, \mathcal{G}$	80.5%	53.2%	6.1k	0.0%	85	89.9%	0.76	92.6%	0.35	8.984
A4	4		64.3%	7.8%	4.6k	7.4%	171	1.6%	1.68	93.4%	0.22	5.993
B4	4	$\mathcal{G}$	64.4%	7.8%	4.7k	7.2%	170	1.6%	1.67	93.5%	0.21	6.007
C4	4	$\mathcal{A}, \mathcal{G}$	71.1%	38.8%	5.1k	0.0%	136	19.1%	1.27	96.4%	0.12	6.852
A5	5		56.0%	6.7%	3.9k	7.3%	201	0.1%	2.16	95.5%	0.15	5.873
B5	5	$\mathcal{G}$	56.0%	6.8%	3.9k	7.2%	204	0.1%	2.18	95.6%	0.14	5.875
C5	5	$\mathcal{A}, \mathcal{G}$	61.7%	31.8%	4.3k	0.0%	183	0.8%	1.86	98.8%	0.04	6.211

type of non-elective patients and shows that the most urgents have a higher risk of exceeding the time limits when the number of dedicated ORs is not sufficient.

Regardless of the number of dedicated ORs, the EOO is able to better exploit the overtime than the baseline configurations. Then the OR utilization is significantly improved and cancellations are almost totally annulled. This fact is also due to the lower uncertainty that the DOR policy has because of the insertion of non-elective patients does not affect on the risk of elective patients cancellation, as in the SOR.

When the module $\mathcal{G}$ is enabled, a slight improvement of the non-elective waiting times has been observed, but an even greater contribution is given by the EOO. In Table 5.6 can be seen that this fact is more evident for trauma and emergent patients. In particular, $w_{\text{tr}} = 1.24$ in the baseline configuration, that means that the average waiting time is 7 minutes over the time limit, but enabling the modules $\mathcal{A}$ and $\mathcal{G}$ such exceeding is less than one minute and there is 3.6% more trauma patients operated

Fig. 5.3: DOR – Scenario 1 – Percentage of patients treated in time for various numbers of dedicated ORs, with $\mathcal{A}$ and $\mathcal{G}$ enabled.



(a)elective vs. non-elective

(b)non-elective classes

on time. Therefore, the proposed EOO approaches have also a positive impact on the non-elective patients, although they are designed for an elective patient flow.

Tab. 5.6: DOR – Scenario S_1 – Focus on non-elective patient-centered indices, 3 dedicated ORs

config. id	# ded. ORs	enabled modules	Performance indices f_{NE}	f_{tr}	f_{em}	f_{ur}	f_{ad}	w_{NE}	w_{tr}	w_{em}	w_{ur}	w_{ad}
A3	3		87.1%	77.7%	82.4%	95.4%	100%	0.43	1.24	0.34	0.17	0.03
B3	3	$\mathcal{G}$	87.8%	78.5%	83.4%	95.7%	100%	0.42	1.20	0.31	0.16	0.03
C3	3	$\mathcal{A}$, $\mathcal{G}$	89.9%	81.3%	86.4%	97.2%	100%	0.35	1.01	0.27	0.13	0.02

Although the baseline configuration A2 is better than the baseline configuration A3, when the optimization modules are enabled configuration B3 has a greater value of Z than B2. Then the number of dedicated ORs that maximizes the performance strictly depends on the optimization approaches that are used.

Tab. 5.7: DOR – Scenarios $S_1 - S_4$ – Best configurations.

scen. id	config. id	# ded. ORs	enabled modules	Performance indices u_{OR}	u_{over}	p	c	t	f_{E}	w_{E}	f_{NE}	w_{NE}	Z
S_1	C3	3	$\mathcal{A}$, $\mathcal{G}$	80.5%	53.2%	6.1k	0.0%	85	92.6%	0.76	89.9%	0.35	8.984
S_2	C1	1	$\mathcal{A}$, $\mathcal{G}$	81.9%	52.7%	3.3k	0.0%	46	99.3%	0.42	64.6%	1.30	8.202
S_3	C2	2	$\mathcal{A}$, $\mathcal{G}$	83.6%	60.6%	10.8k	0.1%	55	99.3%	0.48	71.4%	1.01	8.505
S_4	C1	1	$\mathcal{A}$, $\mathcal{G}$	81.3%	69.1%	5.4k	0.3%	57	99.2%	0.50	60.5%	1.43	8.018

All the previous remarks for the scenario S_1 are confirmed also for the other scenarios, whose best configurations are listed in Table 5.7. We observe that, in scenario S_1, the best configuration C3 provides the 30% of the ORs to the non-elective patients, that are the 15% of the total, because the unpredictability of such patients requires a higher amount of resource to deal with the time limits. Differently, the scenario S_3 maximizes the objective function with the configuration C2, which provides the 20% of the ORs to the non-elective. This result indicates that the need of dedicated ORs depends also on the surgery duration distribution, that is the only difference between the two scenarios.

5.4.2 Shared Operating Room

Starting from the unique baseline configuration D1 defined for the SOR, Table 5.8 reports the results of different configurations obtained enabling and combining the optimization modules to maximize the objective function Z for the scenario S_1.

The configuration E1 corresponds to the configurations C1–C5 of the DOR, because of the use of the EOO and the implication of the module $\mathcal{G}$ in any SOR policy. All the performance indices are improved by the module $\mathcal{A}$: OR utilization and waiting times of both elective and non-elective patients are better than those of the DOR. The higher OR utilization is due to the advanced scheduling that plans elective patients in all the OR session and, as opposed to the DOR, never an OR slot is

Tab. 5.8: SOR – Scenario S_1 – Performance indices.

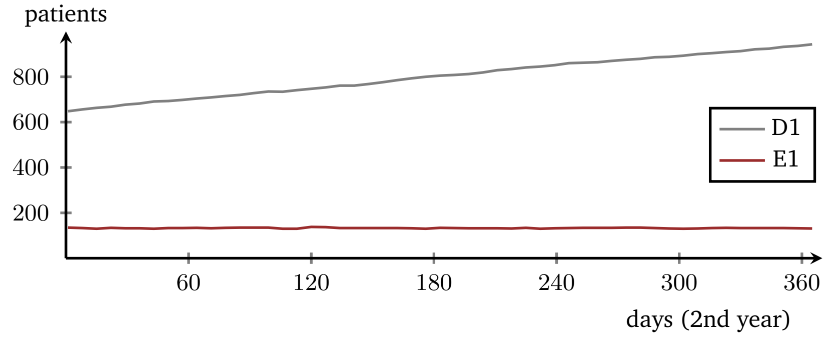
conf. id	enabled modules	Performance indices									
		u_{OR}	u_{over}	o	c	t_{avg}	f	w_{avg}	f_{NE}	w_{NE}	Z
D1		89.0%	12.2%	6.8k	21.4%	40	98.8%	0.37	91.0%	0.35	9.167
E1	$\mathcal{A}$	93.1%	99.9%	7.1k	11.5%	8	99.3%	0.11	92.4%	0.32	9.423
F1	$\mathcal{A}, \mathcal{B}(5, 0.3)$	89.6%	99.7%	6.8k	25.2%	48	99.1%	0.43	92.9%	0.35	9.208
F2	$\mathcal{A}, \mathcal{B}(5, 0.4)$	88.6%	93.1%	6.7k	21.0%	54	99.1%	0.48	92.8%	0.33	9.247
F3	$\mathcal{A}, \mathcal{B}(5, 0.5)$	85.6%	80.1%	6.4k	13.7%	73	99.2%	0.64	93.8%	0.29	9.300
F4	$\mathcal{A}, \mathcal{B}(5, 0.6)$	79.6%	63.6%	5.9k	10.2%	103	61.2%	0.92	94.9%	0.23	8.120
F5	$\mathcal{A}, \mathcal{B}(10, 0.15)$	93.3%	90.0%	7.1k	4.5%	8	99.3%	0.11	93.0%	0.31	9.520
F6	$\mathcal{A}, \mathcal{B}(10, 0.2)$	93.1%	84.5%	7.1k	2.7%	17	99.3%	0.17	92.8%	0.32	9.528
F7	$\mathcal{A}, \mathcal{B}(10, 0.25)$	87.9%	63.7%	6.7k	0.6%	52	99.3%	0.46	93.2%	0.30	9.459
F8	$\mathcal{A}, \mathcal{B}(10, 0.3)$	84.6%	51.6%	6.4k	0.2%	69	99.2%	0.61	93.6%	0.28	9.410
G1	$\mathcal{A}, \mathcal{C}(z_1)$	92.9%	99.6%	7.1k	11.0%	8	99.3%	0.11	93.2%	0.29	9.454
G2	$\mathcal{A}, \mathcal{C}(z_2)$	93.0%	99.2%	7.1k	10.5%	8	99.3%	0.11	94.2%	0.25	9.503
G3	$\mathcal{A}, \mathcal{D}(z_1)$	92.6%	99.6%	7.1k	10.7%	8	99.3%	0.11	93.6%	0.28	9.469
G4	$\mathcal{A}, \mathcal{D}(z_2)$	92.7%	99.3%	7.0k	10.3%	8	99.3%	0.11	94.2%	0.25	9.499
H1	$\mathcal{A}, \mathcal{E}(0.25)$	92.6%	99.9%	7.1k	10.1%	8	99.3%	0.11	87.7%	0.46	9.240
H2	$\mathcal{A}, \mathcal{E}(0.5)$	92.9%	100%	7.1k	10.0%	8	99.3%	0.11	86.2%	0.51	9.183
H3	$\mathcal{A}, \mathcal{E}(0.75)$	92.5%	99.9%	7.1k	9.5%	7	99.3%	0.11	85.6%	0.53	9.158
H4	$\mathcal{A}, \mathcal{E}(1)$	92.6%	99.8%	7.1k	9.6%	8	99.3%	0.11	84.4%	0.55	9.112
I1	$\mathcal{A}, \mathcal{F}(0.25)$	93.1%	100%	7.1k	11.8%	8	99.3%	0.11	92.5%	0.33	9.424
I2	$\mathcal{A}, \mathcal{F}(0.5)$	92.9%	99.9%	7.1k	11.6%	8	99.3%	0.11	91.8%	0.35	9.393
I3	$\mathcal{A}, \mathcal{F}(0.75)$	92.5%	99.8%	7.1k	11.4%	8	99.3%	0.11	91.6%	0.37	9.380
I4	$\mathcal{A}, \mathcal{F}(1)$	93.1%	100%	7.1k	11.3%	8	99.3%	0.11	90.9%	0.40	9.365
J1	$\mathcal{A}, \mathcal{B}(10, 0.2), \mathcal{E}(1)$	92.9%	82.8%	7.1k	2.5%	17	99.3%	0.18	92.9%	0.32	9.528
J2	$\mathcal{A}, \mathcal{C}(z_2), \mathcal{F}(0.25)$	92.7%	99.2%	7.1k	10.4%	8	99.3%	0.11	94.3%	0.25	9.501
K1	$\mathcal{A}, \mathcal{B}(10, 0.2), \mathcal{C}(z_2),$ $\mathcal{F}(0.25)$	92.8%	83.4%	7.1k	2.3%	18	99.3%	0.19	94.8%	0.24	9.605

unused because of the lack of non-elective patients to operate on, this allows us to operate on more elective patients per week and to have shorter waiting times. On the other hand, non-elective patients do not need the release of a specific dedicated OR to be inserted, then their waiting times are lower than those of the DOR up to 3 dedicated ORs. The high utilization of the overtime suggests that the online approach included in the EOO avoids a high number of cancellations, nevertheless the value of c is high because of the uncertainty determined by the insertion of non-elective surgery in almost filled OR sessions. Figure 5.4 shows that the used EOO approaches avoid the lengthening of the waiting list. Because of the general improvement given by the EOO optimization, the module $\mathcal{A}$ has been always enabled in the further configurations involving NOO approaches.

Configurations F1–F8 concern the slack management and have been obtained fixing the number n_s equal to 5 (half of daily ORs) and 10 (all daily ORs), and ranging the parameter π in such a way to reserve a percentage between 15% and 30% of the total time with a 5% step, that is π ranges between 0.3 and 0.6 when $n_s = 5$ and between 0.15 and 0.3 when $n_s = 10$.

As shown in Figure 5.5, at the increasing of π both the OR utilization and the number of cancellations decreases, because there is less probability to exceed the total duration of the OR sessions that involves a lower request of overtime, which is proved by a lower value of u_{over}. Conversely, the waiting times of the elective patients

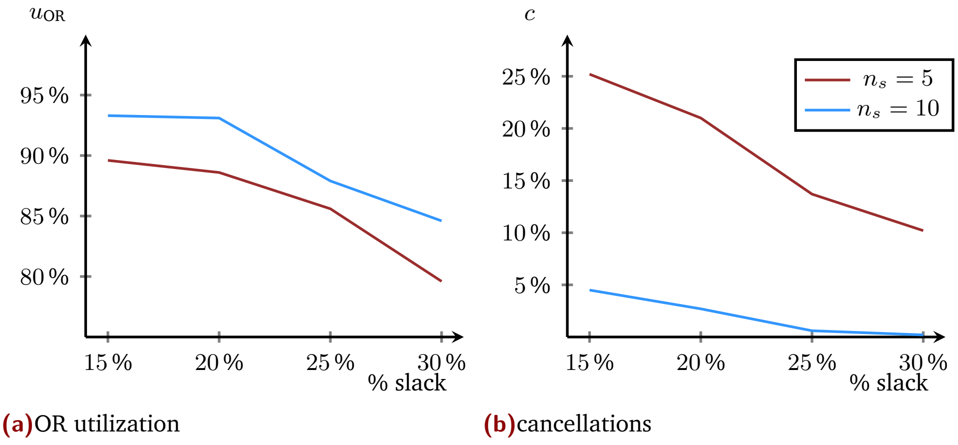
Fig. 5.4: SOR – Scenario S_1 – Length of the elective waiting list with and without EOO.



raise when π increases, causing a significantly lowering of f_E for the configuration F4. In all the other configurations involving slacks, more than 99% of elective patients are operated within the MTBT, but the growth of the waiting list can be unmanageable on a longer period. Furthermore, slacks lead to a slight improvement of non-elective patients performance.

The effectiveness of BIM optimization has been tested in configurations G1–G4, using the two different objective functions z_1 and z_2 in the BILLS algorithm, for both the offline and the online version. A first difference with the configuration E1 is the higher percentage of non-elective patients operated on within the time limits, that is more remarkable for the trauma patients using the objective function z_2, as can be seen in Table 5.9. However, it seems that the online version of the algorithm does not provide a further improvement respect to the offline version. Furthermore, the NOO modules $\mathcal{C}$ and $\mathcal{D}$ slightly impact also on the trade-off between OR utilization and cancellations, fostering the improvement of the latter at the expense of the former.

Fig. 5.5: SOR – Scenario S_1 – OR utilization and fraction of cancellation, varying the fraction of slack on the total surgery time.



(a) OR utilization **(b)** cancellations

Tab. 5.9: SOR – Scenario S_1 – Impact of BIM optimization on the non-elective patients.

config. id	enabled modules	Performance indices									
		f_{NE}	f_{tr}	f_{em}	f_{ur}	f_{ad}	w_{NE}	w_{tr}	w_{em}	w_{ur}	w_{ad}
E1	$\mathcal{A}$	92.4%	68.9%	96.8%	99.8%	100%	0.32	0.95	0.24	0.12	0.02
G1	$\mathcal{A}, \mathcal{C}(z_1)$	93.2%	71.2%	97.7%	99.7%	100%	0.29	0.84	0.21	0.10	0.02
G2	$\mathcal{A}, \mathcal{C}(z_2)$	94.2%	74.9%	98.3%	99.8%	100%	0.25	0.72	0.18	0.09	0.02
G3	$\mathcal{A}, \mathcal{D}(z_1)$	93.6%	72.9%	97.8%	99.8%	100%	0.28	0.81	0.21	0.10	0.02
G4	$\mathcal{A}, \mathcal{D}(z_2)$	94.2%	74.9%	98.3%	99.8%	100%	0.25	0.74	0.18	0.09	0.02

The NEW-Fit algorithm has been used in configurations H1–H4 varying the value of the parameter δ between 0.25 and 1.00 with step 0.25. We observe that when $\delta = 1$ the early deadline for the non-elective patients is fixed equal to the time limit, while decreasing δ to 0 the algorithm uses an even more restrictive early deadline. As expected, enabling $\mathcal{E}$ the number of cancellations decreases because the non-elective patients are inserted in such a way to balance the workload among the OR sessions. In proportion to the value of δ, this causes higher non-elective patients waiting times that induce to a higher number of patients exceeding the time limit due to the uncertainty of the surgery durations of the previous patients. The reason of the limited impact of the NEW-Fit algorithm is probably due to the workload of the chosen scenario, whose baseline configuration D1 shows high performance from the elective patients point of view ($f = 98.8\%$). As a matter of fact, the effectiveness of the the NEW-Fit algorithm has been proved in other more overloaded scenarios, as reported in Section 4.4 of Chapter 4. Furthermore we analyzed the impact of the NEW-Fit algorithm jointly to the insertion of slacks in the schedule. Configuration J1 gives the best value of Z varying δ in $(0, 1]$, that is the combination of the configurations F6 and H4. The results are very similar to that computed for the configuration F6, but it is interesting that the negative impact of the NEW-Fit algorithm on the non-elective patients can be canceled using slacks.

An analogous analysis has been provided for the NEIC algorithm in the configurations I1–I4. All the performance indices are very close to those of the configuration E1, except the waiting times of the non-elective patients that increases slightly. However the NEIC algorithm is implemented to preserve the BIM optimization when the non-elective patients are inserted, therefore in configuration J2 we tested the impact when the BIM are optimized, but without better results.

Finally, the configuration K1 is the configuration that maximizes Z and, compared to E1, improves all the performance indices except a very slight loss of the OR utilization, with the advantage of using 16.5% less overtime. Table 5.10 lists the best configurations of all scenarios $S_1 - S_4$. We observe that the EOO and the same configuration of slacks are always enabled, while the BILLS algorithm is used in the offline or online version using the objective function z_2. Both NEW-Fit and NEIC contribute in the best configuration with a slight improvement in only one of the four scenarios. The most interesting thing is that the four not very different scenarios provide four different best configurations, which remarks the usefulness of a decision support system that is specifically adapted to the operative environment.

Tab. 5.10: SOR – Scenario $S_1 - S_4$ – Best configurations.

scen.	conf.	enabled	Performance indices									
id	id	modules	u_{OR}	u_{over}	o	c	t_{avg}	f	w_{avg}	f_{NE}	w_{NE}	Z
S_1	K1	$\mathcal{A}, \mathcal{B}(10, 0.2), \mathcal{C}(z_2), \mathcal{E}(1)$	92.8%	83.0%	7.0k	2.3%	10	99.3%	0.13	94.6%	0.24	9.595
S_2	K2	$\mathcal{A}, \mathcal{B}(5, 0.2), \mathcal{C}(z_2)$	91.8%	81.2%	3.5k	2.8%	14	99.3%	0.16	90.1%	0.39	9.391
S_3	K3	$\mathcal{A}, \mathcal{B}(10, 0.2), \mathcal{D}(z_2), \mathcal{F}(1)$	94.7%	79.6%	11.8k	1.1%	14	99.3%	0.15	98.4%	0.13	9.793
S_4	K4	$\mathcal{A}, \mathcal{B}(5, 0.2), \mathcal{D}(z_2)$	94.3%	79.6%	5.9k	2.0%	17	99.2%	0.18	95.6%	0.21	9.668

5.4.3 Hybrid policies

A hybrid policy is a mix of dedicated and shared policies. In our settings, a number of ORs are reserved for the non-elective patients (as in the DOR) while the remaining ORs are used to operate on both elective and non-elective patients (as in SOR). The elective patients are scheduled into the shared ORs. When a non-elective patient arrives, his/her surgery is scheduled into a dedicated OR, if available at that time; on the contrary, the surgery is scheduled into the first released (dedicated or shared) OR. If two or more non-elective patients are waiting for the insertion, the priority is given to the patient that is closer to his/her time limit. We observe that the module $\mathcal{G}$ for the immediate insertion of trauma patients is already included in this hybrid policy.

Tab. 5.11: Hybrid policy – Scenario S_1 – Performance indices.

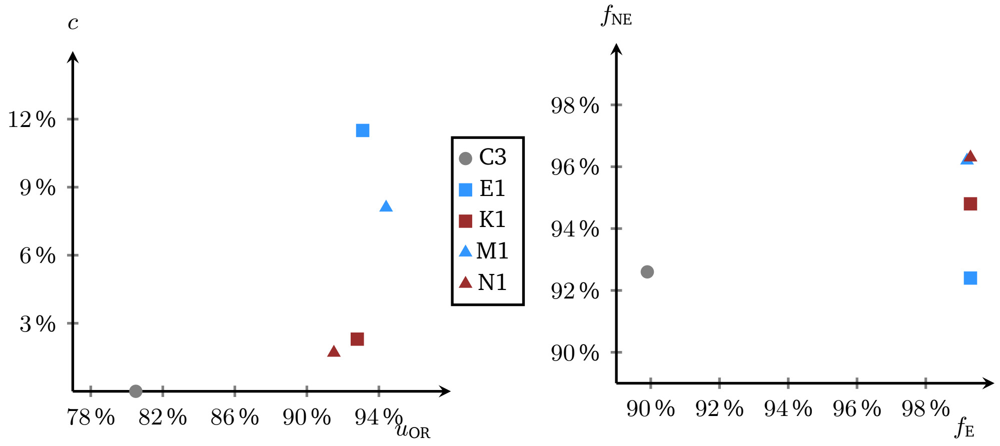
conf.	# ded.	enabled	Performance indices									
id	ORs	modules	u_{OR}	u_{over}	o	c	t_{avg}	f	w_{avg}	f_{NE}	w_{NE}	Z
L3	3		72.6%	8.9%	5.4k	9.3%	131	23.5%	1.22	95.7%	0.15	6.893
L2	2		79.7%	10.1%	6.0k	11.6%	90	82.6%	0.81	95.4%	0.17	8.769
L1	1		85.2%	11.0%	6.5k	15.8%	58	98.8%	0.52	94.0%	0.24	9.270
M1	1	$\mathcal{A}$	94.4%	97.8%	7.2k	8.1%	8	99.2%	0.12	96.2%	0.17	9.633
N1	1	$\mathcal{A}, \mathcal{B}(10, 0.2), \mathcal{C}(2)$	91.5%	78.1%	6.9k	1.7%	26	99.3%	0.25	96.3%	0.16	9.659

Table 5.11 reports the results for the scenario S_1. We started considering 3 dedicated ORs (configuration L3), which corresponds to the best configuration of the DOR, and we decreased this number (configuration L2 and L1) in order to improve the elective patients performance. In addition the OR utilization has been improved using less dedicated ORs. As expected, allowing the access of non-elective patients to resources that the DOR dedicated to elective patients, the distribution assignment of the operating rooms to the two patient flows needs to be changed to have a fair balance. Configuration M1 is obtained from configuration L1 enabling the EOO modules that, also in this case, provide a very significant and general improvement. Finally, configuration N1 is the best one using also the NOO approaches. Similar results are obtained for the other scenarios $S_2 - S_4$.

5.4.4 Comparing policies

We evaluate the impact of the best configurations for the DOR, the SOR and the hybrid policy comparing their results on the four indices involved in the objective function Z. Such a comparison is summarized in Figure 5.6: a facility-centered index is compared with an elective one in 5.6(a) while an elective patient-centered index is compared with a non-elective one in 5.6(b). We plotted the configurations using only the EOO approaches for the SOR and the hybrid policy, in order to appreciate the impact of the NOO. Finally, note that the results for the configurations M1 and N1 are overlapping in Figure 5.6(b).

Fig. 5.6: Scenario S_1 – Comparing DOR, SOR and hybrid on the main performance indices.



(a) u_{OR} vs. c

(b) f_{E} vs. f_{NE}

In Figure 5.6(a) the trade-off between OR utilization and cancellations is evident. We remark that all the configurations represent a different compromise between the two indices, except the configuration E1 that is dominated by K1 and M1. Further, the best compromise seems to be provided by K1 and N1 configurations. On the other side, considering the trade-off between the percentage of elective and non-elective patients operated within their MTBT, Figure 5.6(b) shows that hybrid configurations (M1 and N1) dominate the DOR and the SOR configurations. Globally, the configuration N1 seems the more rational one in accordance with the coefficients adopted in (5.6).

5.5 Concluding remarks

In the literature, the problem of sharing operating rooms between elective and non-elective patients counts a number of different approaches whose results are usually conflicting. Such approaches are applied and tested to different operative conditions making their comparison very difficult. In order to determine the best approach under certain operative conditions an ad hoc study is therefore necessary. In this chapter, we fill this research gap providing a hybrid model capable to represent a large range of operative and decision-making conditions to study and evaluate the impact of such approaches.

Our hybrid model uses discrete event simulation to represent the general surgical pathways of the two patient flows (elective and non-elective) under the dedicated and shared policies, and their hybrid versions. A set of optimization approaches are embedded within the model. We consider the approaches proposed in the literature to deal with the sharing of the operating rooms (the reservation of slacks and the break-in-moment optimization). Further, we introduce three new online algorithms for the real time management of the non-elective patients, that is the Break-In Layout Local Search, the Non-Elective Worst Fit and the Non-Elective Insertion Criterion. In particular, the last two algorithms deal with the Non-Elective Real Time Insertion problem, which suffers from a lack of studies in the literature.

Because of the capability of the model to represent a high number of scenarios and configurations, we have to restrict the quantitative analysis to four representative scenarios, choosing conditions such that the workload caused by the elective patients is proportional to the available resources. For each of the four scenarios, we show the performance of both the dedicated and shared operating room policies. Further, for the latter we observe the impact of the optimization modules when they are enabled separately and, on the basis of their results, we combine them to find the best configuration with respect to both the (elective and non-elective) patient and facility points of view. Furthermore, the impact of a hybrid policy is evaluated.

From the management policy point of view, the results confirm the strong trade-off between the OR utilization and number of cancellations, which is widely discussed in literature. While the dedicated operating room policy allows us to have a very low probability of elective-patients cancellations, the shared operating room policy is able to increase the use of the resources and, consequently, to reduce the length of the waiting list. However, a better trade-off between the performance of the elective and non-elective patients is given by the shared operating room policy. We also show that hybrid policies could provide a further performance improvement.

In summary, our analysis suggests the use of a hybrid policy to manage elective and non-elective patients even if shared and dedicated policies can be a good compromise in certain operative conditions. However, to account for the different perspectives of the stakeholders, it is recommended to provide an ad hoc analysis.

From an algorithmic point of view, we prove the effectiveness of the elective-oriented optimization approaches: they are able to manage the elective patient flow and, counter-intuitively, also some non-elective performance indices take advantage from them. This result suggests that an appropriate management of the elective patient flow is a necessary condition to have a positive impact when dealing also with the non-elective patient flow.

Regarding the non-elective-oriented optimization algorithms, our analysis suggests that an appropriate slack management can overcome the limitation of a dedicated policy, at parity of overall operating time, causing less cancellations but lowering the OR utilization, with a slight improvement of the non-elective patients performance. The Break-In Layout Local Search algorithm has a positive impact on non-elective patients without deteriorating the performance of the elective patients. Finally, the impact of the the Non-Elective Worst Fit and the Non-Elective Insertion Criterion seems limited in the operative conditions represented by the four scenarios. However, the effectiveness of the Non-Elective Worst-Fit is proved in Section 4.4 of Chapter 4 in more crowded scenarios.

6 Sharing operating rooms among surgical pathways

In the previous chapters we dealt with the surgery process scheduling over a fixed set of assigned ORs for each day of the planning horizon. Such a specific assignment of OR sessions is defined at the tactical level by the MSS, which provides to allocate ORs and, more generally, resources among different SPs in a cyclic manner [70, 129].

The MSS must be updated whenever the total amount of OR time or the requirements of some surgical CPs change. This can occur not only as a response to long term changes in the overall OR capacity or staffing fluctuations, but also in response to seasonal fluctuations in demand. Therefore, the MSS should dynamically adapts to the current state of the different waiting lists of the specialties that share the same ORs [61, 68].

In this chapter, we propose several simple approach to the MSS problem in such a way to study in different contexts the impact of sharing or not the overtime among SPs in the RTM of the ORs. As already discussed in the previous sections, the RTM deals with the overtime management in the case of a single surgical SP. However, when the overtime is a shared resources, the online decision of using the overtime, or to cancel a surgery, should take into account a fairness criterion. The objective of the management of the shared resources is therefore to have a fair assignment of the overtime and the OR sessions to surgical SPs.

To this purpose, we consider two or more SPs corresponding to different specialties. Each SP is essentially the same surgical pathway described in Figure 3.4 of Chapter 3 for the elective patient flow. We will consider only elective patients since we would like to have a more clear idea of the impact of the proposed sharing policies: actually, in the current context, a flow of non-elective patients would correspond to a higher workload, that we are able to manage as shown in Chapters 4 and 5. For the sake of simplicity, we refer hereafter to those surgical SPs as specialties.

After defining policies for the resource sharing among different specialties in Section 6.1, a quantitative analysis is reported in Section 6.2. Conclusions closes the chapter.

6.1 Policies for sharing resources

We propose several optimization approach for the MSS in Section 6.1.1, taking into account the assignment of the ORs to the specialties. In Section 6.1.2 two online approaches for the overtime allocation during the RTM are defined.

6.1.1 Policies for sharing ORs

We would define how to assign the available OR sessions among different specialties in such a way to ensure enough and balanced OR sessions to each specialty. In our operative context, the MSS is updated every time period (usually one month or one week). To this end, we define three alternative policies.

The first policy is Based on Lengths (BL), that is, every four weeks, the OR sessions are reassigned so that they are proportional to the number of patients in the waiting list of each specialty. On the contrary, the second policy is Based on EOTs (BE) in which every four weeks, the OR sessions are reassigned so that they are proportional to the sum of the EOTs of the patients in the waiting list of each specialty.

The last policy consists in a Flexible Scheduling (FS) in which MSS and Advanced Scheduling are solved at the same time every week. The algorithm implementing the FS policy is an adaptation of that proposed in [65]. It consists of a greedy construction of the initial solution and an improvement phase performed by a local search engine: (i) at the beginning of the greedy construction, the patients are ordered by decreasing value of the ratio between the waiting time and the MTBT $\widetilde{w}_i$, and each OR session is not assigned to any specialty, except for the OR sessions used to reschedule the patients postponed during the last week; (ii) during the greedy construction, patients can be inserted only into OR sessions assigned to their specialty, or into OR sessions not already assigned (that is empty OR sessions); in the latter case, such an OR session is assigned to the specialty of the patient; (iii) during the local search, only swaps between patients belonging to the same specialty are allowed.

6.1.2 Policies for sharing overtime

When sharing overtime, we are interested in guaranteeing a fair access to the available overtime from the different specialties. We propose two alternative policies. The first policy is called Dedicated Overtime Allocation (DOA). in which a dedicated amount ν_Σ of weekly overtime is allocated to the specialty Σ, so that it is proportional to the number of OR sessions assigned by the MSS. By consequence, the RTM will take into account ν_Σ as the overtime available when applying the criterion for the overtime allocation to elective surgery of a single SP, as defined by equation (3.6) in Section 3.2 of Chapter 3.

The second policy is called Flexible Overtime Allocation (FOA), in which all the specialties share the total available overtime ν; in order to foster a balanced use of the overtime, we adapt the criterion (3.6) as follow:

$$\beta_k^\tau \left(\frac{e_i+\rho_{jk}^\tau}{d_{jk}}\right)\left(1+\frac{\nu_\Sigma}{\nu}-\frac{n_k^\Sigma}{n}\right) \leq 1 \tag{6.1}$$

where ν_Σ is the amount of weekly overtime used by the specialty Σ until that time and n_k^Σ is the number of OR sessions of that specialty from the day after k.

The policy FOA introduces a new factor which measure the overtime still available with respect to the number of OR sessions to be still performed by the specialty. This value is closed to 1 when the overtime has been used proportionally with respect to the assigned and completed sessions. On the contrary, it is between 0 and 1 or it is greater than 1 when it is underused or overused, respectively.

6.2 Quantitative analysis

In the current analysis, we consider as "baseline" the configuration (1) introduced in Section 4.4.2, that is the best one when dealing with only elective patients. Such a configuration includes a greedy construction for the advanced scheduling, a sequencing in decreasing order of $\tilde{w}_i$ for the allocation scheduling, while overtime is always assigned until the amount dedicated that OR session does not exhaust. In the baseline configuration, the overtime is allocated using the DOA rule while the number of OR sessions are proportional to the arrival rate and it does not change over time.

Tab. 6.1: Parameters of the two scenarios.

Parameters	unit of measure	Scenario S_1	Scenario S_2
arrival rate pathways 1, 2, 3	patients/day	$12.5, 12.5, -$	$24.0, 12.0, 4.0$
initial waiting list	patients	1000	1500
MTBT URG $A, \cdots, G$	days	$8, 15, 30, 60, 90, 120, 180$	$8, 15, 30, 60, 90, 120, 180$
frequency URG $A, \cdots, G$			
pathway 1		$5\%, 15\%, 40\%, 15\%, 10\%, 10\%, 5\%$	$5\%, 15\%, 40\%, 15\%, 10\%, 10\%, 5\%$
pathway 2		$14\%, 14\%, 14\%, 14\%, 15\%, 15\%, 15\%$	$14\%, 14\%, 14\%, 14\%, 15\%, 15\%, 15\%$
pathway 3			$8\%, 9\%, 11\%, 12\%, 15\%, 15\%, 30\%$
EOT average pathways 1,2,3	min	$120, 180, -$	$150, 120, 180$
EOT deviation pathways 1,2,3	min	$75, 75, -$	$75, 75, 75$
n	ORs	50 (10 a day)	75 (15 a day)
d_{jk}	min	480	480
ν	min	600	900

Table 6.1 describes the two scenarios in which we evaluate our sharing policies. The two scenarios differ from (i) the number of specialties, (ii) the amount of available resources (number of operating rooms and weekly overtime), and (iii) the

patient features, namely the arrival rates, the EOT distributions and the urgency distributions.

Tab. 6.2: Scenario S_1: performance indices (B is the baseline configuration).

	Policies		Performance indices						
id	MSS	FOA	o	c	t_{avg}	w_{avg}	f	u_{OR}	u_{over}
(0)			8 380	4.2%	56	0.99	56%	88%	44%
(1)	BL		9 058	4.3%	43	0.75	91%	97%	77%
(2)	BE		9 040	4.2%	44	0.79	87%	97%	74%
(3)	FS		9 050	4.2%	42	0.74	93%	97%	79%
(4)	BL	$\surd$	8 903	6.1%	50	0.87	81%	95%	77%
(5)	BE	$\surd$	8 895	5.5%	53	0.94	61%	95%	59%
(6)	FS	$\surd$	8 850	7.1%	53	0.94	67%	95%	69%

Tables 6.2 and 6.3 report the analysis of the proposed policies for the resource sharing. We denote with the id (0) the baseline configuration while those obtained by combining the different policies are denoted with an integer from (1) to (6). The combination of the different policies is described in the second and the third columns: column “MSS” denotes the policy used for OR sharing while column “FOA” indicates when the FOA policy has been adopted in alternative to the DOA one. Finally, for each configuration, the performance indices are reported in accordance with the definitions in Table 5.4.

Tab. 6.3: Scenario S_2: performance indices (B is the baseline configuration).

	Policies		Performance indices						
id	MSS	FOA	o	c	t_{avg}	w_{avg}	f	u_{OR}	u_{over}
(0)			13 547	4.5%	56	1.09	33%	92%	26%
(1)	BL		14 189	4.4%	51	0.91	58%	97%	52%
(2)	BE		14 198	4.3%	51	0.92	60%	97%	54%
(3)	FS		14 238	4.0%	51	0.88	83%	97%	74%
(4)	BL	$\surd$	14 031	5.1%	57	1.03	40%	96%	46%
(5)	BE	$\surd$	14 009	5.1%	58	1.04	38%	96%	45%
(6)	FS	$\surd$	13 858	6.3%	60	1.05	39%	95%	46%

Considering the scenario S_1 in Table 6.2, we remark that all the configurations (1)–(6) indicates a general improvement of the performance indices with respect to the baseline configuration, except for the number of cancellations. This justify the need of *ad hoc* solutions to deal with the resource sharing. In particular, we observe a higher overtime utilization that, in accordance with the analysis in the previous chapters, is due to the online optimization modules for the RTM. Furthermore, the DOA rule seems to be more effective than the FOA, especially when considering the patient-centered indices. These considerations are confirmed also by the analysis reported in Table 6.3 for the Scenario S_2, in which the effectiveness of the DOA with respect to the FOA is more evident.

Considering both scenarios S_1 and S_2, the configuration 3 – that is, that adopting a flexible scheduling policy for the MSS and a dedicated allocation for the overtime

– leads to a general and robust improvement of all the indices resulting as the best configuration. Table 6.4 reports in more detailed way the results for that configuration reporting also the performance indices for the two pathways analyzed in the scenario S_1.

Tab. 6.4: Scenario S_1 – configuration (3) vs. baseline: detailed analysis.

id	pathways	**Performance indices**						
		o	c	t_{avg}	w_{avg}	f	u_{OR}	u_{over}
(0)	both	8 380	4.2%	56	0.99	56%	88%	44%
	1	4 587	5.5%	8	0.25	99%	80%	77%
	2	3 793	2.7%	114	1.89	5%	96%	11%
(3)	both	9 050	4.2%	42	0.74	93%	97%	79%
	1	4 544	6.4%	36	0.75	94%	98%	92%
	2	4 507	2.0%	47	0.73	91%	97%	70%

From the results for each pathway reported in Table 6.4, it is evident the effectiveness of the flexible scheduling policy: actually, the results for configuration 3 demonstrate a good balance of the performance indices relative to the two pathways, especially that regarding the percentage of surgeries performed within their MTBT threshold.

6.3 Concluding remarks

In this chapter, we extended the model presented in Chapter 3 in order to deal with several SPs that share ORs and overtime.

Our analysis demonstrated that different pathways can benefit from sharing the resources when adequate policies are adopted. In particular, a flexible approach for the MSS to share ORs among specialties could provide significant improvement of the patient-centered indices.

The simple extension of the RTM implemented to share overtime among specialties demonstrated to do not be more effective than that proposed for a single specialty. However, the online approach showed also in this more general context to be useful to optimize the available resources.

Part II

The Emergency Care Pathway

7 Introduction and Literature Review

The Emergency Care Delivery System (ECDS) is usually composed of an Emergency Medical Service (EMS) serving a network of Emergency Departments (EDs). ECDS plays a significant role as it constitutes an important access point to the national health system, saving people's lives and reducing the rate of mortality and morbidity [150].

The EMS receives a phone call from a citizen asking for an emergency care for himself or for a third person. The operators at the EMS's operation center are in charge of answering the calls and assigning a color code to each emergency request, based on the severity of injury, through a phase called *triage*. After the triage phase the operator dispatches an ambulance following a predefined dispatching policy. Ambulance crew rescues the patient and, if necessary, transports him/her to a hospital. Note that usually the ambulance crew is in charge of the patient until he/she is handed to the hospital staff.

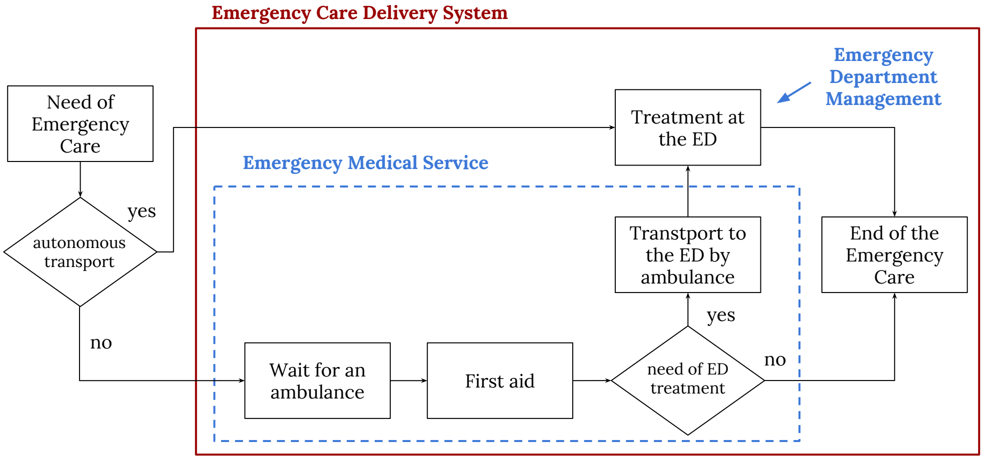
An ED is a medical treatment facility inside of a hospital or in other primary care center and is specialized in emergency medicine providing a treatment to unplanned patients, that is patients who present without scheduling. The ED operates 24 hours a day, providing initial treatment for a broad spectrum 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. The patient flow of an ED is composed of two distinct flows: the former is made of the patients transported by the EMS while the latter is that of the patients arrived at the ED by their own.

A phenomenon that affects EDs all over the world reaching crisis proportions is the overcrowding [187]. It is manifested through an excessive 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 [166]. 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 because of the decreased productivity [161]. Moreover, the ED overcrowding causes stress among the ED staff, patient dissatisfaction and episodes of violence [153, 156].

Since the perception of the crowding level from the staff is subjective [191] and because of the need to adequately prevent the phenomenon, several indices for the real time measurement of overcrowding have been introduced and studied. The most popular overcrowding measures are the National Emergency Department Overcrowding Scale (NEDOCS) [200], the Emergency Department Work Index (EDWIN) [146] and the Demand Value of the Real-time Emergency Analysis of Demand Indicators (READI) [190]. They are based on different indices about the current operating status: the amount of available resources, the number of patients in the ED involved in some activities or waiting for a resource, their waiting times, the patient outcome and the predicted arrivals. However, the analysis by Hoot et al. [165] shown that none of most popular overcrowding measures is capable of providing an adequate forewarning.

The Emergency Care Pathway (ECP) was introduced by Aringhieri et al. [141] formalizing, from an Operations Research perspective, the idea of the ECDS. The ED overcrowding can be addressed in different points of the ECP and, in particular, into the following two phases: (i) the ambulance rescue performed by the EMS and (ii) the management of the ED patient flow. The focus of the second part of this thesis is therefore on the management of the ECDS in order to reduce the overcrowding adopting an online optimization approach to deal with the inherent uncertainty of the ECDS system. While the Clinical Pathway (CP) of an EMS is structured and quite simple, the CP of the ED is (at high level) structured and complex. The high level structure means that the pathway should describe the broad variety of patients entering in the ED. This means that the ED pathway is less descriptive for our purpose. The relation between ECDS, EMS, and EDs with respect to the ECP is represented by the framework in Figure 7.1.

Fig. 7.1: General framework of ECDS.


Emergency Care Delivery System
Emergency Department Management
Need of Emergency Care
autonomous transport
yes
no
Treatment at the ED
Emergency Medical Service
Transtport to the ED by ambulance
End of the Emergency Care
yes
Wait for an ambulance
First aid
need of ED treatment
no

At the regional level, the ECDS can be seen as a network of EDs cooperating to maximize the outputs (number of patients served, average waiting time, ...) and outcomes in terms of the provided care quality. Many EDs, especially those serving a large amount of people, complain about the large number of non-urgent patients usually transported by the EMS ambulances. Further, EMSs usually do (or can) not take into account the ED workload level when assigning and transporting a patient to an ED. In Chapter 8, we discuss how quantitative analysis can provide a tool to evaluate the impact of a simple dispatching policy for an ECDS at the regional level. More generally, the real-time management of the ambulances of an EMS is an online optimization problem in which three main decisions should be addressed to serve an emergency request, that is i) the dispatching of the ambulance, ii) the selection of the ED facility to which the patient will be transported, and iii) the redeployment of the ambulance. In Chapter 9, we provide a comprehensive analysis of such an online optimization problem.

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, which are fundamental in online optimization. Actually, the challenge in modeling the ED behavior is to replicate such different paths. In Chapter 10, 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. In Chapter 11, we propose a new ED simulation model for the evaluation of several online resource allocation methods, which are based on the current state of the ED and on the prediction of the next activities provided by the above process model.

The remaining of the chapter is devoted to a literature review of the above problems restricted to online optimization approaches.

Emergency Medical Services and overcrowding

The importance and sensitivity of decision making in the EMS field have been recognized by researchers who studied many problems arising in the management of EMS systems since the 1960s, as recently reported by Aringhieri et al. [141], Reuter-Oppermann et al. [192] and Bélanger et al. [145]. Further, Aboueljinane et al. [137] present a review of the many simulation models that have been developed over the years to deal with the EMS analysis.

In the real practice, many EMSs apply simple dispatching policies, such as the closest-idle policy, which always sends the closest available ambulance to a call, which is selected based on the priority of incoming emergency calls. The closest-idle policy has gained much attention in EMS planning due to its simplicity although it results in sub-optimal solutions [139, 143, 151, 173, 182, 194], as well as increasing the waiting time for subsequent calls. Recently, Jagtenberg et al. [168] provide a bound existing online solutions in comparison with three different methods to compute the optimal offline dispatch policy for problems with a finite number of incidents. The performance of the offline optimal solution serves as a bound for the performance of an unknown optimal online dispatching policy. Then, they compare such an offline solution to the closest idle vehicle dispatching policy obtaining a bound of 2.7 times on the fraction of late arrivals.

To overcome the drawbacks of the closest-idle policy, many researchers suggested to take the priority of calls into consideration. McLay and Mayorga [182] present a Markov decision process to dispatch distinguishable ambulances to prioritized calls while considering the fact that errors in the classification of patient priorities might occur. The model determines the dispatching policy that maximizes the expected coverage of true high-risk calls. The results of their study show that over-responding (under-responding) is preferable only when there is a high (low) rate of classification errors. Then, McLay and Mayorga [183] extended their work to obtain an equity- and efficiency- based dispatching model in which distinguishable ambulances as well as emergency call priorities are addressed. The efficiency-based objective function of the model focuses on the fraction of covered calls. In addition, four equity constraints are considered to reflect the fairness from both the patient's and vehicle's points of view. The study also investigates how the incorporation of different equity measures can affect the dispatching policy. However, the authors also mention that incorporating equity might lead to a lower service or negative outcomes.

Bandara et al. [143] present a priority-based heuristic dispatching rule based on static dispatching rules and fixed deployment. It always sends the closest available ambulance to priority urgent calls and sends a less busy ambulance to non-urgent calls. The results of applying this dispatching policy indicate better performance for urgent calls in comparison with the closest-idle policy. On the contrary, a slightly worsening of the performance for non-urgent calls is reported.

In contrast to the papers highlighting the role of call priority, there are other dispatching policies that follow a different stream. These works mostly integrate the closest assignment policy with some notions borrowed from other fields.

Lee [174, 175] introduced the centrality-based dispatching policy for the EMS serving at the same time rural and urban area: it combines the centrality notion, used in complex networks, and the closeness notion. The main idea behind incorporating centrality into dispatching decisions is that by responding to the most central calls, especially when the probability of transferring patients to a hospital is low, the chance of in-time response to the next calls will increase. As the author mentions, the applicability of the centrality-based policy highly depends on the probability of transferring the patients to the hospital (hospital probability) which in turn depends on many factors such as crew expertise, the nature and severity of accidents, and resource scarcity. To overcome the shortcoming of a pure centrality dispatching policy, Lee [176] adopted the notion of parallelism to develop a centrality-based dispatching policy in which both idle and busy ambulances are considered simultaneously. The idea originates from the fact that a currently busy server might respond sooner to a call after completing the service, than idle vehicles located farther from the call.

Simulation is often exploited for the analysis of the overcrowding at the ED and its impact on ambulance diversion. Nafarrate et al. [184] found that the number of patients in the waiting room is a better trigger for ambulance diversion than inpatient bed availability, as it provides the best balance between accessibility and waiting times. Ramirez-Nafarrate et al. [189] use a simulation model to determine the effect of several ambulance diversion policies on the patient's waiting time. More generally, Buuren et al. [149] uses simulation to evaluate several dynamic dispatching strategies while Maxwell et al. [180] evaluates redeployment policies computed by approximate dynamic programming using simulation. Other analysis of dispatching policies are those reported in [164, 169].

Usually, the approaches reported in the literature for the redeployment of the idle ambulances do not explicitly consider real-time system changes due to ambulances becoming idle or new calls arriving [160, 195]. Jagtenberg et al. [167] develop a heuristic algorithm for real-time ambulances redeployment. The dispatching policy for idle ambulances is based on maximizing the marginal expected coverage of the corresponding ambulances. In [181], Maxwell et al. present an approximate dynamic programming model that redeploys idle ambulances such that the coverage is maximized. The dynamic programming formulation captures the real-time evolution of the system while solutions can be computed quickly. This also holds for realistically sized instances, since only a simple optimization problem has to be solved.

A first attempt to consider real-time redeployment policies is due to Ni et al. [186]. In their analysis, the authors uses simulation to devise and to evaluate redeployment policies. Recently, Barneveld et al. [144] evaluate the impact of typical factors influencing the performance of an EMS such as (i) the frequency of redeployment actions, (ii) time bounds on the ambulance relocation, and (iii) the inclusion of busy ambulances in the decision process. The main insights derived by their research are that adding more relocation action is highly beneficial for rural areas and considering ambulances involved in dropping off patients available for newly coming incidents reduces relocation times only slightly.

Finally, Nasrollahzadeh et al. [185] develop a flexible optimization framework for real-time ambulance dispatching and relocation. They formulate the problem as a stochastic dynamic program. Because of the unbounded state space, the authors propose an approximate dynamic programming framework to generate high-quality solutions. Their analysis is performed on an available benchmarks on an EMS system in Mecklenburg County, North Carolina.

Emergency Department and overcrowding

Simulation is widely used to test what-if scenarios to deal with overcrowding [187], analyzing 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 [155, 156]. Then human and equipment resources available should be used as efficiently as possible optimizing existing resources and processes. For this reason, research addressing short-term decision problems are increasing in the recent years [138]. Placing in the perspective to alleviate the ED overcrowding without changing the ED resources and settings, there are two way to act: (i) changing the human resources planning [159, 197, 203] or (ii) adopting different policies in the allocation of the human and equipment resources [158, 170, 171, 178].

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 [136, 179], ensuring their understanding in order to generate benefits associated with efficiency [193]. In literature there are several process mining approaches that use specialized data-mining algorithms to extract knowledge from data-set, creating a process model that takes into account dependency, order and frequency of events, but also decision criteria and durations. In [157] 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 to 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 [157].

8 The impact of dispatching policies at the regional level

At the regional level, the ECDS can be seen as a network of EDs cooperating to maximize the outputs (number of patients served, average waiting time, ...) and outcomes in terms of the provided care quality. The development of models for the analysis of a health system as a whole is one of the main challenges in the health care management field. The basic idea is to have a tool capable to validate management policies at health system level modeling the patient flow through the care pathway. As a matter of fact, the current trend in the analysis of health care systems is to shift the attention from single departments to the entire health care chain in such a way to increase patient's safety and satisfaction, and to optimize the use of the resources.

In order to apply such an approach to the analysis of a regional ED network, one of the main difficulties is the collection of all the information regarding the transportation of the patients from the emergency scene to the ED. Nevertheless this problem can be now overcome exploiting the immense amounts of data generated by health care systems. Health Care Big Data (HCBD) are a key enabling technology to support detailed health system analysis: exploiting the HCBD, one can replicate the behavior of the health system modeling how each single patient flows within her/his care pathway.

In this chapter we discuss how quantitative analysis based on the HCBD can provide a tool to evaluate dispatching policies for a regional network of emergency departments: the basic idea is to exploit clusters of EDs in such a way to fairly distribute the workload. We present a simulation model based on the case study of the Piedmont in Italy, and powered by the knowledge provided by the analysis of regional HCBD.

The chapter is organized as follows. In Section 8.1, we discuss how big data can enable a novel methodological approach to the health system analysis. In Section 8.2, we describe the case study under consideration. In Section 8.3, we report how we implemented the simulation model. In Section 8.4, we report a quantitative analysis of the results obtained running the simulation model. Conclusions and future works are discussed in Section 8.5.

8.1 Methodological Motivations

The development of models for the analysis of a health system as a whole is one of the main challenges in the health care management field [188]. The basic idea is to have a tool capable to validate management policies at health system level modeling the patient flow through the corresponding CP. Literature indicates that System Dynamics (SD) seems to be the most appropriate methodology. A first attempt has been made by Wolstenholme during his collaboration with the NHS. In [201], he applies SD to the development of national policy guidelines for the U.K. health service. The tested policies include the use of “intermediate care” facilities aimed at preventing patients needing hospital treatment. Intermediate care, and the consequent reductions in the overall length of stay of all patients in community care, is demonstrated here to have a much deeper effect on total patient wait times than more obvious solutions, such as increasing acute hospital bed capacity. More generally, as discussed in [202], the key message is that affordable and sustainable downstream capacity additions in patient pathways can be identified, which both alleviate upstream problems and reduce the effort for their management.

A SD model has been used as a central part of a whole-system review of emergency and on-demand health care in Nottingham, as reported in [147]: due to a growing emergency care demands, the hospital systems were unlikely to achieve some government performance and quality targets. Such a model discovered a range of undesirable outcomes associated with the growing demand and, at the same time, suggested policies capable to mitigate such impacts. Vanderby and Carter [198] were interested in determining whether SD can be an appropriate methodology to model the patient flow in a hospital, and to analyze it from a strategic planning perspective. The SD model were developed in collaboration with the General Campus at The Ottawa Hospital with particular attention to the delays experienced by patients in the ED. The authors reported about the modeling techniques, validation and scenarios tested, accompanied by their comments regarding the appropriateness of SD for such a strategic analysis.

From a modeling point of view, SD is a simulation methodology whose main elements are stocks and flows: a stock is any entity that accumulates or depletes over time; a flow is the rate of change of a stock. For instance, in health care a stock can represent the waiting list for a surgery, that is a number of people requiring a surgery, while a flow can be the rate of a new insertion in the list. One of the main limitation of using SD for health system analysis is that patients are indistinguishable from each other within stocks and flows. On the contrary, health care services are generally characterized by a large variety of different patients suffering from the same diseases and flowing in the same care pathway.

From the above remarks, Discrete Event Simulation (DES) seems the more appropriate methodology to model such a large variety of patients flowing in their corresponding pathway because of DES has the capability of representing each single patient (or entity) within one of more pathways. Further, DES can easily enable the application of optimization algorithms to take the best (or the most rational) decision regarding a single or a group of entities modeling a single or a group of patients.

It is worth noting that such a modeling approach requires a lot of detailed data, that is all the data needed to replicate the behavior of each single patient flowing in its corresponding pathway. Moreover, in terms of health system analysis, such a model requires the availability of all the data for all patients flowing in all pathways of the same type in the health system under consideration. A defining characteristic of today's data-rich society is the collection, storage, processing and analysis of immense amounts of data. This characteristic is cross-sectoral and applies also to health care.

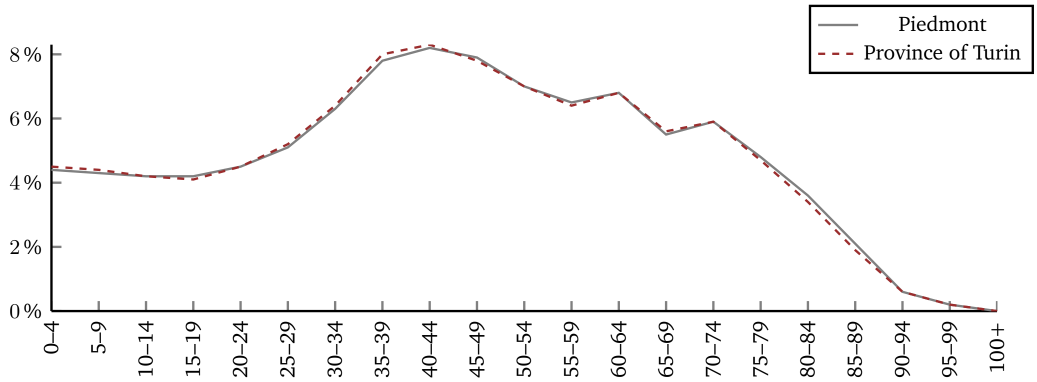
We argue that the HCBD can power a detailed health system analysis using DES methodology: exploiting the HCBD, one can replicate the behavior of the health system modeling how each single patient flows within her/his care pathway. The novelty of the proposed approach is therefore the use of the DES methodology for the health system analysis exploiting the Big Data in order to better represent the variety of the patients accessing the health system.

8.2 The Emergency Department network in Piedmont Region

Piedmont (Italian: Piemonte) is one of the 20 regions of Italy. It has an area of 25,402 square kilometers and a population of about 4.6 million. The capital of Piedmont is Turin. Piedmont is organized in 7 provinces. The province of Cuneo is the largest one while the province of Turin is the most populated one: actually, about 2.3 million of inhabitants are living in, and 1.4 million are living in the area of Turin. Figure 8.1 reports the number of inhabitants living in Piedmont and in the province of Turin, divided in different age classes.

According to the 2015's report of the "Programma Nazionale Esiti" by the Ministry of Health, the waiting time for a urgent and a non-urgent code could exceed respectively 60 minutes and 450 minutes, in the worst case. In other similar Italian regions, such waiting times are about 20% lower. We remark that in Italy, the Regions are in charge of providing the health services in accordance with the minimal level decided at the national level by the Ministry of Health. This comparative analysis demonstrates the need of investigating the reasons of such differences and, eventually, to individuate some possible improvements.

Fig. 8.1: Population of Piedmont and province of Turin: age distribution (ISTAT 2011).



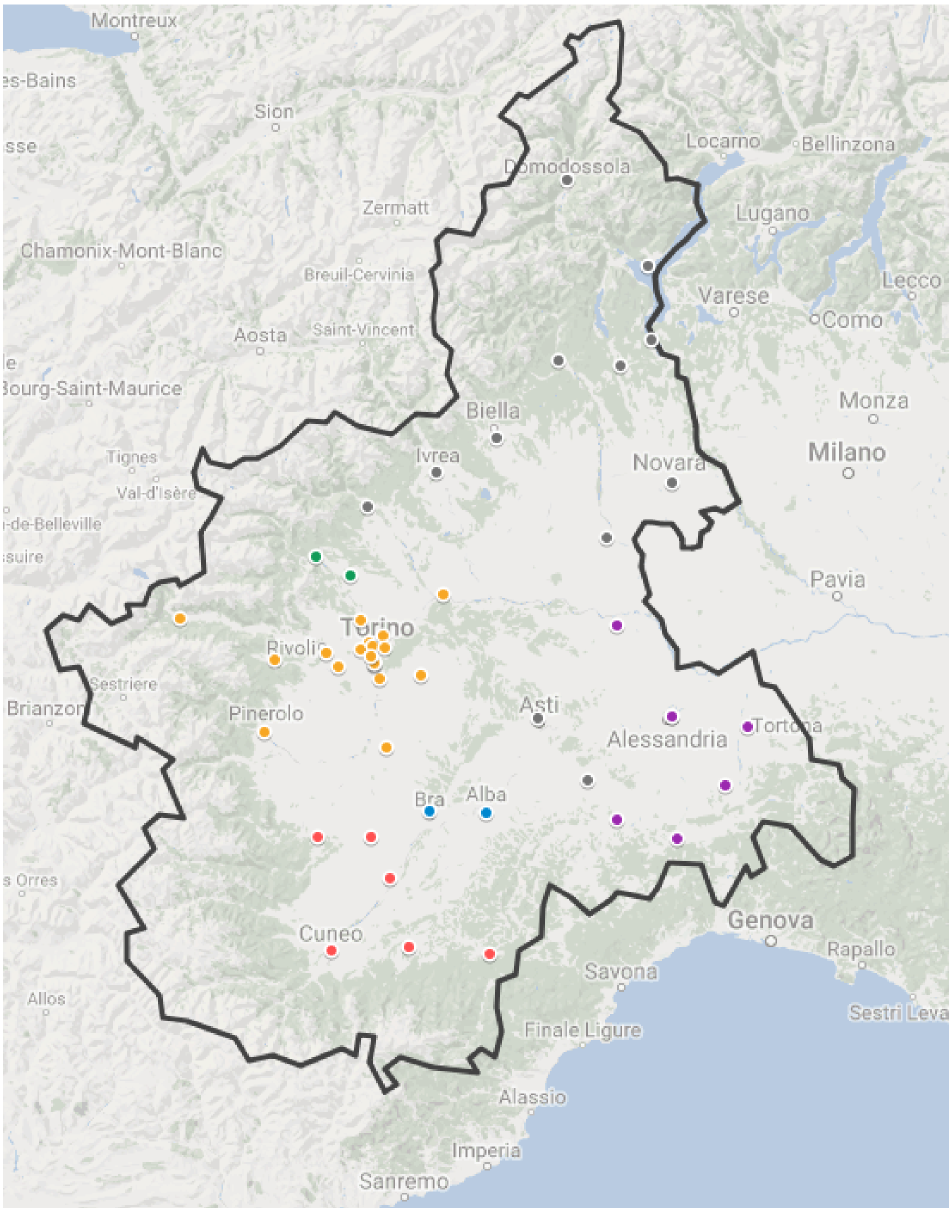
From more than 10 years, the Piedmont region is collecting data about the regional health system, and released a regional law to unify the flows of data gathered from all the health care providers operating in Piedmont, that is, local health agencies, hospitals, and all the private structures in agreement. Such a regional law guarantees the quality of the data collected in accordance with the national standards: all the information must respect a standard format and their consistency is checked for financial reasons since health providers are reimbursed w.r.t. the number and the type of treatments.

Concerning the access to the network of EDs, the HCBD contains all the information regarding the access: encrypted patient ID, patient registered residence, times (arrival, discharge, ...), urgency code, ED, treatment(s), etc.. Each year they collect all the information regarding about $1,800,000$ accesses to the regional network of EDs: for instance, in 2013, there were $1,768,800$ accesses; among them, the 90.53% were non-urgent. The network is composed of 49 EDs, mostly – about 20 – located in the province of Turin as reported in Figure 8.2. The EMS usually transports patients to the closest ED, apart some particular – limited in number – cases.

8.3 A two-phase Discrete Event Simulation model

We propose a quantitative model for the analysis of the network of EDs operating in Piedmont. The proposed model is organized in two phases, and it operates on a time horizon of one month. The first phase is devoted to data analysis concerning the time horizon taken into account in order to determine the appropriate value of the parameters of the DES model, which is the main part of the second phase. As a matter of fact, the emergency demand depends on the day of the week and the time of day [196]. Further, a not accurate forecasting can lead to managerial solutions that worsen the EMS performance, and by consequence the quality of the access to the ED, even if more resources are used, as discussed in [140].

Fig. 8.2: The ED network in Piedmont.



8.3.1 Dynamic estimation of the parameters.

In order to have a proper representation of the main parameters of the network of EDs, the first phase of our quantitative model concerns the analysis of the big data relative to the time horizon considered in the running experiments. Parameters and their corresponding distributions are empirically computed over adequate time intervals in such a way to fit the model on a given and fixed time horizon, and to replicate both the patients flow and their management by the EDs.

The main parameters dynamically evaluated are the emergency demands and their urgency code, the capacity and the service time of each ED, and how the patients are distributed to EDs with respect to their geographic origin. The general evaluation procedure consists in scrolling the data concerning each access in chronological order to keep track of the information needed to estimate the considered parameters and their corresponding empirical distribution, as we describe in the following.

Emergency demands. The emergency demand consists in the number of accesses to the whole regional ED network. Such a distribution is computed counting the average number of accesses in each time interval of 30 minutes over each day of the time horizon considered.

Urgency distribution. The urgency distribution measures the percentage of patients having a urgent or a non urgent code with respect to the origin of the patients.

Service time of each ED. The service time of each ED is estimated using the information regarding the time on which the patient has been take over by the ED, and the time on which the patient has been discharged. The service time has been estimated by the code of urgency.

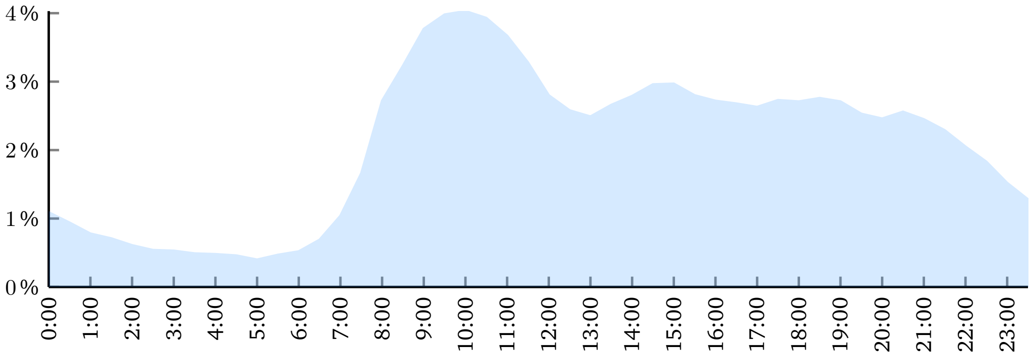
Capacity of each ED. An ED usually has a formal capacity defined a priori. On the contrary, the real practice showed that the real capacity could be different. Further, we should take into account variations in the staffing. From these considerations, we estimate the capacity of the ED by counting the maximum number of patients that are in the ED at the same time. We compute such a value for each interval of three hours in a day, for each day in the time horizon. The capacity of each interval of three hours is finally set to the value corresponding to the 90-percentile of all the values measured in the same interval and in the same day of the week.

Patient geographical distribution. From the data of the patients, we estimate the number of the patients coming from a city identified by its postal code. We also estimate the number of patients that accessed an ED from a given city.

Although the percentage of patients transported by the EMS could be evaluated dynamically, our preliminary analysis showed that such a parameter currently ranges in $[13.3\%, 14.2\%]$. Therefore, we decided to set this parameter in such a way to study the interplay between the EMS and the ED network varying such a percentage in Section 8.4.

Figure 8.3 reports an example of the distribution of the daily arrival of the patients derived from the about $150,000$ accesses to the ED network in July 2001. Note that the figure reports ticks of 1 hour (instead of 30 minutes) only to improve its readability.

Fig. 8.3: Distribution of the patient arrivals during the day (July 2011).

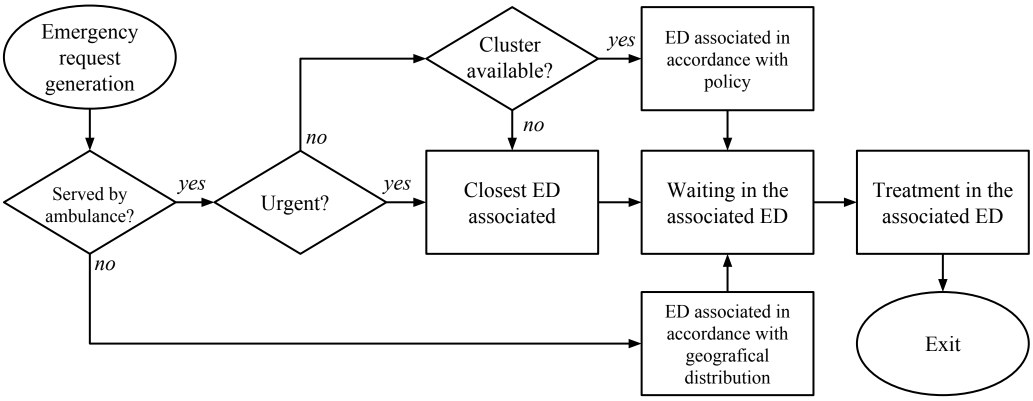


The DES model.

We propose a DES model to represent the pathway of the patient entering in the ED network. Our DES model is based on a straightforward representation of the flowchart depicted in Figure 8.4.

An emergency request of a patient is generated in accordance with the geographical distribution of the patients and the arrival distribution. At the moment of its generation, an ED is associated to the patient pursuant to the distribution of the patients accessing each ED, which usually corresponds to the closest one. Such an emergency request can be served or not by an EMS ambulance. When the request is not served by the EMS, we assume that the patient reaches – in some way – the ED previously associated. On the contrary, the transportation of the patient is in charge of the EMS. In our model, the ambulance transports the patient to the associated ED only if the urgency code is high (red or yellow in the Italian system), otherwise the EMS can decide where to transport the patient in accordance with some policies (dispatching decision for non urgent patients). After arriving at the ED, the patient will wait for the treatment, which usually lasts for a ED distributed following the service time distribution dynamically estimated. When the patient will be discharged, he/she will exit from the model.

Fig. 8.4: The flowchart representing the emergency pathway.

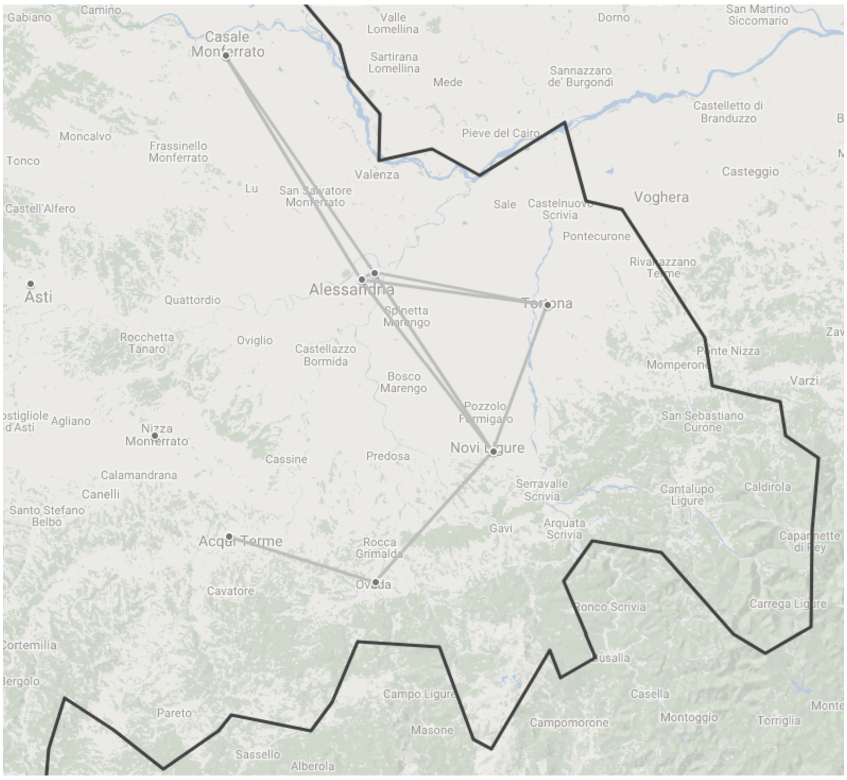

Emergency request generation
Served by ambulance?
yes
no
Urgent?
yes
no
Cluster available?
yes
no
ED associated in accordance with policy
Closest ED associated
Waiting in the associated ED
Treatment in the associated ED
ED associated in accordance with geografical distribution
Exit

The considered dispatching policies are two. The first one, say P_0, dispatches a non urgent patient to the ED associated to the patient at the moment of its generation, without any change. The second one, say P_1, dispatches a non urgent patient following a service state policy, that is, at the moment t, the patient is dispatched to the ED h having minimal ratio r_h^t

$$r_h^t = \frac{w_h^t + s_h^t}{c_h}\,, \tag{8.1}$$

in which the values w_h^t and s_h^t are respectively the number of patients waiting and receiving health care, and c_h is the estimated capacity of the ED. This policy is suggested by the fact that Piedmont region is building an ICT infrastructure to share the real-time information regarding the workload of the EDs. We would remark that the use of the number of patients in the waiting room is suggested in [184] because of it is a good trigger for ambulance diversion caused by the overcrowding.

Fig. 8.5: The clusters C_2 (province of Alessandria).



The policy P_1 does not consider all the ED network but only those belonging to a cluster of EDs. A cluster of EDs is a subset of all the EDs operating in Piedmont that can be reached in no more than 30 minutes from a given origin, as highlighted with the colored placeholders in Figure 8.2. We identified 5 different clusters in Piedmont, denoted by C_i, $i = 1, \dots 5$. The largest one C_1, composed of 20 EDs, is located in the Turin area. The clusters C_2 (province of Alessandria) and C_3 (province of Cuneo) are composed of 7 and 6 EDs, respectively. Finally, two smaller clusters, composed of 2 EDs each, are located in the area of “Valli di Lanzo” (C_4) and in the area of Alba and Bra (C_5). Please note that a cluster is not a complete graph as highlighted in Figure 8.5: for instance, this means that a patient transported to the ED of Casale Monferrato can be re-routed only to one of the two EDs in Alessandria and not to the other EDs in Tortona, Novi Ligure, Ovada, and Acqui Terme.

The proposed DES model is quite flexible: as a matter of fact, the ED network operating during the time horizon considered can be obtained by simply activating the dispatching policy P_0. Note that this also provide a tool to evaluate the ED network as a whole system, instead of having simpler measures as those reported in the “Programma Nazionale Esiti”.

Implementation details.

The dynamic estimation of the parameters has been implemented in Python 2.7. A script evaluates data concerning the time horizon of interest from the input data-set and generates an Excel file with the the parameters of the distributions described above.

Apart from the emergency demand, that has been evaluated at the regional level calculating, as mentioned before, the average number of accesses in each time interval of 30 minutes over each day of the time horizon considered, the rest of parameters takes also into account the origin of the patients and/or the related EDs. Urgency code distribution has been estimated by distinguishing for each ED four different codes (from 1 to 4). The accessing distribution has been estimated considering both the distribution of provenance of patients and the distribution of accesses of the EDs, mitigating in this way the possibility of not considering patients collected from an ED in a different location from their city of provenance. Finally the service time distribution has been estimated considering both the ED and the gravity of patients.

The DES model has been implemented using AnyLogic 7.2 [73]. At simulation start-up it takes in input the file before generated and uses it to initialize the parameters. Custom distributions have been used for the parameters above described, while specific objects (`Service` and `ResourcePool`, `Schedule` and `Agent`) have been used for the definition of the EDs, their capacities (varying pursuant to the hour of the day) and for the patients. When a patient is generated it is assigned to him a provenance, the destination hospital (pursuant to the selected policy), an urgency code and the expected service time. The routing of patients has been implemented using two matrices, associating each patient provenance to one (in case of policy P_0) or more (policy P_1) possible EDs of destination.

8.4 Quantitative Analysis

In our analysis, we considered four different months in 2011. Tables 8.1 and 8.2 provide more details about the input of our model. For each month considered, Table 8.1 reports the total number of accesses considered and their classification with respect to the urgency code (1 represents the more urgent code while 4 the less one). We would remark that the total number of accesses considers only those accesses for which at least one between the origin of the patient or the ED of destination is correctly reported in the data.

Tab. 8.1: Description of the data considered in our quantitative analysis.

	total	requests by urgency		
	accesses	3–4	2	1
Jan	126,698	107,773	17,688	1,237
Feb	116,961	99,806	16,074	1,081
Jun	132,654	113,734	17,562	1,358
Jul	123,758	106,404	15,970	1,384

For each month considered, Table 8.2 reports the total number of accesses with respect to their cluster of origin. Finally, the last column of the table reports the percentage of the accesses to an ED belonging to one of the five clusters. This means that the majority of the patients can be served by an ED belonging to a cluster. Further, the cluster C_1, composed of 20 EDs over 49, treats more than the 50% of the accesses.

Tab. 8.2: Description of the data considered in our quantitative analysis.

	requests by clusters					
	C_1	C_2	C_3	C_4	C_5	
Jan	69,773	11,480	12,467	3,701	4,201	80.21%
Feb	64,876	9,819	11,548	3,379	4,008	80.05%
Jun	70,292	11,672	12,632	3,568	4,451	77.36%
Jul	62,505	11,027	12,836	3,507	4,196	76.01%

Our quantitative analysis consists in using the two-phase DES model to solve the four instances arising from the four months in Tables 8.1 and 8.2. For each instance, a test consists in solving the instance by varying the percentage of patients transported by the EMS, denoted by p_E, in the interval $[7\%, 27\%]$ with a step of 5%. The rationale is to study the interplay between the EMS and the ED network, as discussed in Section 8.3 and suggested in [141]. Finally, the results for each solution are the average values among those obtained by running the two-phase DES model 100 times, each time starting from a different initial conditions in such a way to have independent and identically distributed repetitions.

Tab. 8.3: P_1 vs. P_0: waiting time reduction Δ_w in minutes.

		p_E	7%	12%	17%	22%	27%	**avg. Δ_w**
	all		15.51	25.91	34.70	42.16	50.79	33.82
Jan	EMS		17.75	21.63	24.73	27.74	34.37	25.24
	no EMS		15.43	26.67	37.01	46.62	57.39	36.62
	all		6.81	13.10	19.75	26.29	31.87	19.56
Feb	EMS		-4.31	1.62	6.68	11.33	15.70	6.21
	no EMS		7.67	14.70	22.48	30.60	37.99	22.69
	all		19.70	39.12	64.51	75.73	80.82	55.98
Jun	EMS		5.78	19.88	45.51	58.88	66.45	39.30
	no EMS		20.81	41.85	68.54	80.63	86.28	59.62
	all		8.27	13.19	17.64	21.15	24.23	16.90
Jul	EMS		-3.86	-2.62	0.22	3.89	7.55	1.04
	no EMS		9.20	15.35	21.20	26.03	30.43	20.44
	all		12.57	22.83	34.15	41.33	46.93	
avg. Δ_w	EMS		3.84	10.13	19.28	25.46	31.02	
	no EMS		13.28	24.64	37.31	45.97	53.02	

Table 8.3 shows the results of our quantitative analysis reporting the waiting time reduction Δ_w considering the whole network of EDs. Such values are computed as follows: for a given dispatching policy $i = 0, 1$, we compute the average waiting time w_{ij} for each ED $j = 1, \ldots, 49$, and then we set W_{P_i} equals to the average of all the values w_{ij}; finally, $\Delta_w = W_{P_1} - W_{P_0}$. Note that P_1 is better than P_0 when $\Delta_w > 0$.

The results prove a general improvements of the waiting times, which improves further as soon as the percentage of the patients transported by the EMS increases. It is worth noting that the different results for each different instances depend on the different composition of the emergency demand reported in Tables 8.1 and 8.2 (see, e.g., the last column of Table 8.2 reporting the percentage of the accesses to an ED belonging to a cluster).

Table 8.4 shows the results of our quantitative analysis reporting the waiting time reduction considering the cluster C_1, that is the bigger one in terms of both the number of EDs and the number of accesses. Although the general improvement is inferior than those for the whole network, such results confirm the comments done for the whole network.

8.5 Concluding Remarks

We presented a two-phase DES model to evaluate the dispatching policies for the regional network of emergency departments powered by the knowledge provided by the analysis of regional health care big data. The model has been tested on the case study of the Piedmont in Italy showing that there is room to improve its efficiency. Further, we observed that such an improvement is more significant as

Tab. 8.4: P_1 vs. P_0, cluster C_1: waiting time reduction Δ_w in minutes.

	p_E	7%	12%	17%	22%	27%	**avg. Δ_w**
	all	11.88	19.75	22.96	24.24	27.55	21.28
Jan	EMS	30.81	29.65	24.95	22.54	24.52	26.49
	no EMS	2.63	7.74	9.03	9.02	12.11	8.10
	all	-3.15	-1.11	1.66	4.57	6.64	1.72
Feb	EMS	6.41	5.40	5.06	5.86	7.27	6.00
	no EMS	-9.57	-9.80	-8.32	-6.37	-4.85	-7.78
	all	2.19	14.53	36.13	46.32	51.54	30.14
Jun	EMS	9.89	15.69	29.30	36.75	42.02	26.73
	no EMS	-2.19	9.17	31.60	42.18	47.85	25.72
	all	8.43	11.60	11.96	10.58	8.86	10.29
Jul	EMS	-0.76	6.17	9.52	9.87	9.47	6.85
	no EMS	12.59	16.96	18.11	17.17	15.89	16.14
	all	4.84	11.19	18.18	21.43	23.65	
avg. Δ_w	EMS	11.59	14.23	17.21	18.75	20.82	
	no EMS	0.87	6.02	12.60	15.50	17.75	

soon as the percentage of the patients transported by the EMS increases. This remark has an evident managerial implication that would not have been possible without an analysis of the entire ED network.

More generally, the results showed the effectiveness of the proposed approach in terms of the capability of modeling a whole health care system through a discrete event simulation approach, which exploits the availability of the health care big data. As discussed in [201, 202], there could be a significant difference between the formal description of the health system and the its real functioning. To overcome this modeling problem, our idea is to retrieve a picture of the system from the big data through the dynamic estimation of the parameters, which allow to fit the model on a given time horizon replicating both the patients flow and their management.

The Real Time Management of Ambulances

9

The analysis of the literature – reported in Chapter 7 – reveals an attention to real-time dispatching policies in contrast to a limited attention to the redeployment of the ambulances in real-time. From such an analysis, a list of insights can be derived, that is (i) the number of patients in the waiting room is a better trigger for ambulance diversion [184], (ii) the use of the fraction of covered calls as efficiency measures [183], (iii) incorporating equity might lead to a lower service or negative outcomes [183], (iv) priority dispatching policies can improve the performance for urgent calls at the price of a worsening of the performance for non-urgent calls [143].

The real-time management of the ambulances of an EMS is an online optimization problem in which three main decisions should be addressed to serve an emergency request, that is (1) which ambulance should be dispatched to serve an emergency request, (2) the selection of the ED facility to which the patient will be transported, and (3) where to redeploy the ambulance at the end of its service. The challenge is the definition of a proper set of Dispatching Routing and Redeployment Policies (DRRP) for the ambulance real-time management in order to guarantee a good ambulance utilization reducing their diversion, and to maximize the number of emergency requests served within the corresponding time threshold.

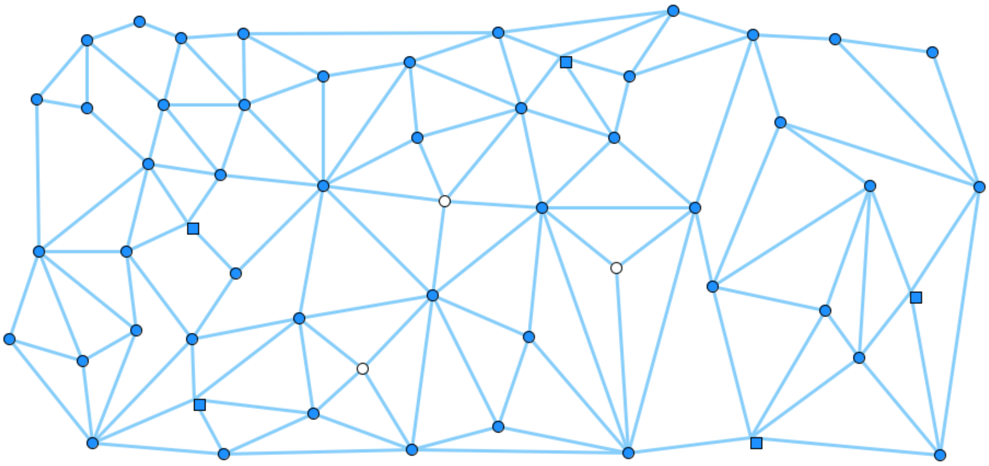
To the best of our knowledge, a comprehensive analysis of the DRRP has not yet been provided. The contribution of this chapter is twofold. The former is an ambulance simulation model capable to deal with real time management and to generate new ad hoc instances. The latter is a set of simple online algorithms to implement several DRRP. An extensive comparison among different DRRP is also provided.

The chapter is organized as follows. The instance generator is described in Section 9.1. The simulation model is presented in Section 9.2. Then a set of DRRP is proposed in Section 9.3 for the real-time management of ambulances. Then, such policies are analyzed in Section 9.4. Section 9.5 closes the chapter.

9.1 Instance Generator

An instance is a planar graph $G = (N, A)$ with n nodes and m arcs. Each node is a centroid representing a small part of the whole area served by the EMS. Each arc models the connection between two nodes. Two labels are associated to each arc: the former represents the length of the arc while the latter is the average speed of a vehicle traveling on it. The number of arcs starting from a node $u \in N$ is equal to a_u. An example is reported in Figure 9.1.

Fig. 9.1: An example of an initial random generated instance with 55 nodes and 133 arcs.



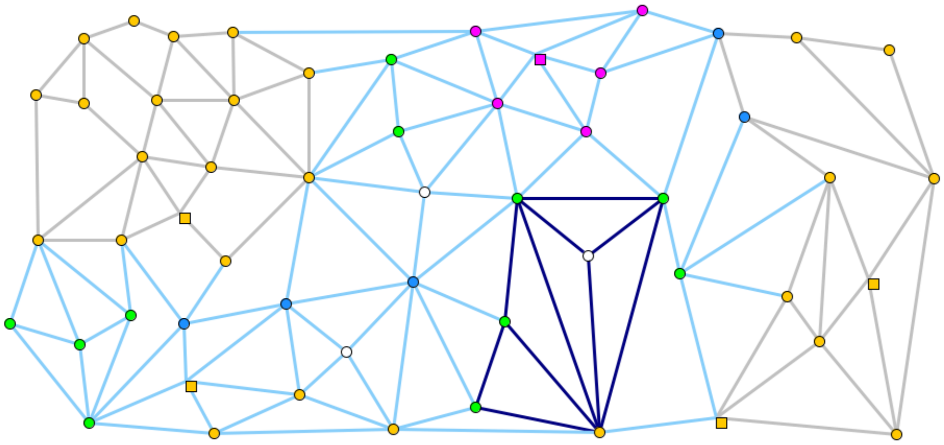
The length of an arc and, more generally, distances in the graph are euclidean. Further, a scale factor f_s determines the value of each pixels. The scale factor is useful to generate graph representing different type of areas such as urban or rural: for instance, in our settings for a urban area 1 pixel corresponds to 20 meters.

As highlighted in Figure 9.1, there are three type of nodes, that is the emergency demand generator (the colored circle), the ambulance base (the colored square), and the ED facility (the white circle). Note that an emergency request can be generated from an ambulance base node. Globally, we have n_G, n_B and n_{ED} nodes (respectively generators, bases, and ED facilities) such that $n_G + n_B + n_{ED} = n$. Let $N_G,\ N_B,\ N_{ED} \subset N$ be respectively the subsets of the $n_G + n_B$ generator nodes, the n_B bases, and the the n_{ED} ED facilities.

A generated graph can be manually adjusted adding or deleting nodes and arcs, and also moving nodes and, by consequence, all connected arcs. For urban area, it is also possible to characterize each node as residential, commercial, public utility, and offices. This classification is useful at running time to model in a proper way the generation of the emergency demand: for instance, a residential node usually generates more demand during the evening or night while a public utility node is likely to generate more demand in the morning. Further, we can change the average speed of each arc by default set to *medium* (30 km/h) to *low* (20 km/h) or *high* (40 km/h).

Figure 9.2 depicts the final version of an instance in which the yellows are the residential nodes, the greens are the commercial nodes, the light blues are the public utilities, and the purple ones are the areas with offices. Regarding the arcs, the light blue are those with medium speed while the blue and the gray arcs are the faster

Fig. 9.2: The graph in Figure 9.1 completed.



and the slower ones, respectively. At the end of the process, the graph can be saved on a file.

9.2 Simulation Model

The simulation model replicates how an EMS serves an emergency request. The EMS receives a phone call from a citizen asking for an emergency care for himself or for a third person. The operators at the EMS's operation center are in charge of answering the calls and assigning a color code to each emergency request, based on the severity of injury, through a phase called *triage*. After the triage phase the operator dispatches an ambulance following a predefined dispatching policy. Ambulance crew rescues the patient and, if necessary, transports him/her to a hospital. Note that usually the ambulance crew is in charge of the patient until he/she is handed to the hospital staff.

The model takes in input seven parameters, that is (i) the duration I of the simulation (expressed in days), (ii) the graph $G = (N, A)$, (iii) the number A_b of ambulances for each base $b \in N_B$, (iv) the maximum number A_b^{max} of ambulances that can be positioned on each base $b \in N_B$, (v) the emergency demand variation table, (vi) the workload of the ambulances, and (vii) the capacity of the ED facilities. While the first four parameters have a straightforward definition, the last three requires a detailed description, which is reported in the following. Let us denote with A the total number of ambulances, given by $A = \sum_{b \in N_B} A_b$.

9.2.1 Distances and Travel Times

The graph $G = (N, A)$ is an undirected and labeled graph. The labels on the arcs $[u, v]$ are the distance ℓ_d among u and v, and the average speed ℓ_s on that arc. We use such labels to compute both distances and/or travel times among two nodes in

G. To this end, we use an ad hoc version of the classic label-setting shortest-path algorithm (e.g., Dijkstra).

9.2.2 The Emergency Demand Variation Table

As reported by many authors (see, e.g., [152, 196]), emergency demand is not static, but, rather, fluctuates during the week, according to the day of the week, and hour by hour within a given day. The emergency demand table (Table 9.1) would model the relative demand fluctuation over the day with respect to different urban areas over the total demand (e.g. office nodes should have a higher relative demand during the business hours of the day). In accordance with the characteristics of the generator node, a negative (*low*) or positive (*high*) variation of the predefined (normal) generation rate is possible. We denote as w_u^i the generic entry of the table with respect to the time interval $i = 1, 2, 3, 4$ (morning, afternoon, evening, night) and the node u.

Tab. 9.1: The Emergency Demand Variation Table

	1: $[7-13]$	**2:** $[13-19]$	**3:** $[19-1]$	**4:** $[1-7]$
residential	normal	normal	high	low
commercial	normal	high	low	low
public utility	high	high	normal	low
offices	normal	normal	low	low
low = 0.7	normal = 1.0		high = 1.3	

9.2.3 The Workload of the Ambulances

The workload of the ambulances W^A is clearly determined by the generation rate of each node in accordance with their characteristics and the fluctuations over the day reported in Table 9.1. Our model allows to input the number of total emergency requests that should be generated for each time interval along the day: the total number of emergency requests is denoted by D which is equal to $D_1 + D_2 + D_3 + D_4$ corresponding to the number of requests to be generated during the morning, afternoon, evening, and night time intervals, respectively.

During each time interval, the D_i requests are spread to all the nodes belonging in N_G as follows: for each node $u \in N_G$, let d_u^i be the number of nodes that should be generated by u during the time interval i, and defined as

$$d_u^i = \frac{w_u^i \, D_i}{\sum_{v \in N_G} w_v^i}.$$

Then, the generation rate of the node u is equal to d_u^i divided by the duration of the time interval i.

Alternatively, the workload W^A can be fixed as a target percentage of utilization and determining – by consequence – the corresponding values of D_i, $i = 1, 2, 3, 4$. First, for each node we compute the *minimal mission time* as the minimum time required to serve an emergency request on a given node. Let $t_u^{\min}$ be such a time computed by considering the shortest time needed to follow the path starting at the closest (to node u) ambulance base, passing from u, and ending at closest ED facility, plus the average service time required at the emergency scene and at the ED. We note that the assumption under the computation of $t_u^{\min}$ is to have an ambulance always available. After that, we compute the (arithmetic) average $T_u^{\min}$ over all $u \in N_G$. The total number of emergency requests D is then computed as

$$D = A\frac{T(i)}{T_u^{\min}},$$

where $T(i)$ is the duration of the interval $i = 1, 2, 3, 4$. Finally, the value D_i are obtained from D as $D_i = D\frac{r_i}{\sum_i r_i}$ where $r_i = 1, 0.8, 0.5, 0.2$. The basic idea is to spread the daily demand over the time interval in such a way to have a peak in the morning.

Independently of its generation, the urgency code of each request is distributed in accordance with the proportion observed in [140], that is about the 10% of red codes, 50% of yellow codes, and 40% of green codes. Note that these percentages are due to the so called *over triage*, which is an over estimation of the request urgency, made by the operators answering at the emergency request phone call.

9.2.4 The capacity of the ED facilities

The capacity of each ED facility located in $u \in N_{ED}$ is derived from the total demand D plus the number of patients D' that will arrive at the ED by their own. First we compute the average service time T_S^{ED} from an estimate of the service time of each urgency codes. The minimum necessary hourly capacity of the ED located at the node $u \in N_{ED}$ is given by

$$C_u = \frac{(D + D')T_S^{ED}}{24 n_{ED}}.$$

The main assumption underlying this computation is to have patients evenly distributed among the ED facilities in such a way to have always one patient to exploit a unit of ED capacity as soon as it is released by another patient. Finally, the capacity of the ED facilities is a parameter ranging in $[1, 2]$ in such a way that the final capacity varies in $[C_u, 2C_u]$.

In our model we have therefore two sources of patients, that is those transported by the EMS and those arrived by their own. In our setting, $D' = 4D$, that is the number of patients transported by the EMS is about the 20% of the whole emergency demand at the ED facilities. This setting is consistent with the analysis in [142] and discussed in Chapter 8.

The patients arriving by their own follow the same distribution observed in our previous work [157], that is 2% of red codes, 15% of yellow codes, and 87% of green codes. To be consistent with such a distribution, the urgency code of the patients transported by the EMS are changed in such a way to obtain the same above distribution decreasing a fraction of the red code to yellow, and the yellow to green. Note that this is consistent with the common practice of an EMS in which over triage determines an overestimates of the emergency demand.

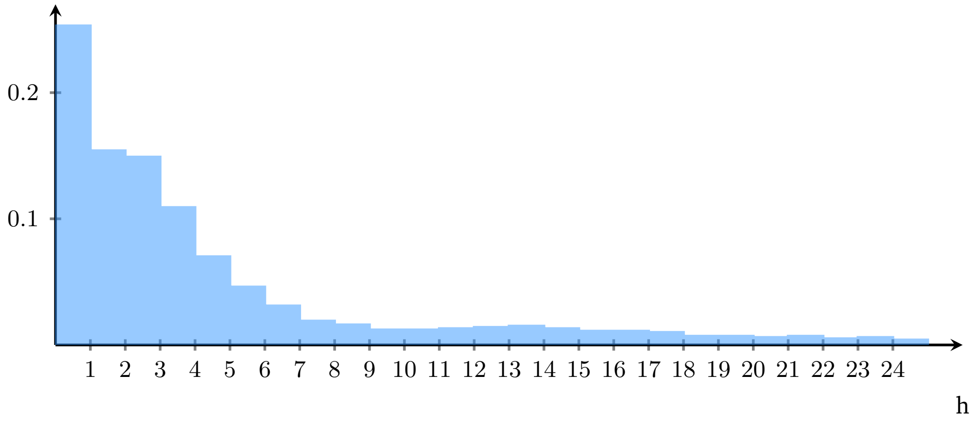
Probability distribution used to generate the total patient demand, the service times for the ambulance rescue and the weathering times are reported in Table 9.2.

Tab. 9.2: Distributions used in the simulation model (Exp=exponential, Tr=triangular).

	Distribution	Parameters	Unit of measure
Ambulance request	$\text{Exp}\left(\frac{1}{\lambda}\right)$	$\lambda = \frac{1}{6D_i}$	patients/h
Autonomous arrivals to ED	$\text{Exp}\left(\frac{1}{\lambda'}\right)$	$\lambda' = \frac{1}{6D'_i}$	patients/h
Ambulance rescue duration	$\text{Tr}(\tau_{min}, \tau_{max}, \tau_{mod})$	$\tau_{min,max,mod} = 10, 20, 15$	min
Urg. patient release at ED	$\text{Tr}(\tau^{ry}_{min}, \tau^{ry}_{max}, \tau^{ry}_{mod})$	$\tau_{min,max,mod} = 6, 10, 8$	min
Non-urg. patient release at ED	$\text{Tr}(\tau^{g}_{min}, \tau^{g}_{max}, \tau^{g}_{mod})$	$\tau_{min,max,mod} = 6, 20, 13$	min

In Figure 9.3 we report the Emergency Department Length-Of-Stay (EDLOS) a discrete distribution used to generate the treatment duration of patients within the ED. Such a distribution is obtained empirically from the real case data-set of an ED, whose details will be detailed in Chapter 10, truncating times exceeding 25 h. An interval time of 1 h is obtained through such a distribution, then a Uniform distribution with support in that interval is used to generate the precise duration in minutes. The resulting average EDLOS is 4.5 h.

Fig. 9.3: Empirical EDLOS distribution.



9.3 Real-time policies

Our main aim is to evaluate real-time policies for the management of the ambulances evaluating their impact in terms of performance of the ambulances and on overcrowding of the ED facilities. In this perspective, we recall that the ambulance real-time management is an online optimization problem in which the following three main decisions should be addressed: (1) which ambulance should be dispatched to serve an emergency request, (2) the selection of the ED facility to which the patient will be transported, and (3) where to redeploy the ambulance at the end of its service.

Before introducing the DRRP, we define an estimate of the number of ambulances needed at each base $b \in N_B$. Let $N_1^c, \dots, N_{n_B}^c$ a partition of N in such a way that each node u belongs to N_b^c (with $b = 1, \dots, n_B$) if and only if the j-th base is the closest one to the node u. Let W_b the sum of the morning weights in Table 9.1 of the nodes in N_b, that is $W_b = \sum_{u \in N_b} w_u^1$. We use morning weights since we supposed to have a peak of demand in the morning. The number A_b^e of estimated ambulance of base $b \in N_B$ is finally given by

$$A_b^e = A \frac{W_b}{\sum_{u \in N_G \cup N_B} W_u}.$$

9.3.1 Ambulance Dispatching

The most common dispatching policy is that of assigning an ambulance available at the closest base [154], which has been proven to perform, on average, uniformly better than the other dispatching rules in accordance with Larsen et al. [172]. In the following, we refer to this policy as **D-Closest**.

Alternatively, the dispatched ambulance can be selected from a list of *enough close* bases, that is those capable to reach the request within the time threshold for the urgency of that request. Let L_B be such a list of enough close bases. The **D-LLCB** policy selects the ambulance to be dispatched from the less loaded close base b in such a way that

$$\text{argmax}_{b \in L_B : A_b^a > 0} \{A_b^a - A_b^e\}$$

where A_b^a is the number of ambulances available in B at the moment of the decision. The **D-LLCB** policy is similar to those reported in [143, 164].

The cutoff priority queue (**D-CPQ**) and the smart assignment (**D-SA**) are two possible extensions of the two policies above reported. The **D-CPQ** consists in temporarily stopping to serve all the emergency requests having green urgency code when the overall number of available ambulance is less than a given threshold. The rationale here is to free up potential resources to deal with the ongoing peak of emergency demand. The **D-SA** consists in considering for dispatching not only the ambulances available in a base but even those who are in the redeployment phase, that is moving from an ED to an ambulance base. In a real setting, this means to have a sort of tracking system that allows to follow the ambulance in real time.

When **D-SA** is active, the **D-LLCB** should be slightly modified accordingly. First, we consider the list L_R of the destination basis of the redeploying ambulances that are capable to reach the request within the time threshold. This means to assign the value $A^a_b - A^e_b$ to each redeploying ambulance corresponding to the destination base. Then we select a base according to

$$\text{argmax}_{b \in L_B \cup L_R : A^a_b > 0} \{A^a_b - A^e_b\} :$$

if the selected base $b \in L_B$, we dispatch an ambulance from the base B; on the contrary, we dispatch the redeploying ambulance if $b \in L_R$. Finally, if the base $b \in L_R$ belongs also to L_B, we dispatch the closest ambulance between the redeploying ambulance and one of the those available in the base.

Both **D-CPQ** and **D-SA** are introduced and discussed by Aringhieri et al. [140] while **D-CPQ** is also analyzed by Yoon and Albert [204]. To the best of our knowledge, **D-SA** is surprisingly never cited in the literature: the closest approach we retrieved is that reported in [176] in which the centrality-based dispatching policy is improved by taking into account both idle and busy ambulances.

9.3.2 Emergency Department Facility Selection

The **H-Closest** policy selects the closest ED facility. Again, this is a common choice in the real settings. The rationale here is to provide as soon as possible a more accurate medical treatment to the patient. Anyway, the ED managers usually complain about the fact that the workload is not evenly distributed among the ED facilities of a given area. Their claim is that a more fair distribution of the workload could improve the overall efficiency of the ED facility network. Such a claim seems proved by the analysis proposed in Chapter 9.

Here, we tested two simple policies addressing the problem of reducing overcrowding at the ED facility in accordance with the remark discussed in [184], and reported at the beginning of the chapter. The first policy, say **H-SAQ** selects the ED facility having the shortest admission queue counting only those patients have same or higher urgency code. The second policy, say **H-WLB** tries to balance the workload of the ED facility taking into account the time needed to treat all patients in the admission queue. Note that **H-WLB** implements a policy from the ED point of view while **H-SAQ** implements a policy from the patient point of view. Finally, note that the two policies are applied only to those patients having yellow or green urgency code, while the **H-Closest** is applied to the red ones.

An estimate of the workload F_u of the ED facility $u \in N_{ED}$ is given by

$$F_u = t_w^r p_u^r + t_w^y p_u^y + t_w^g p_u^g$$

where t_w^r, t_w^y and t_w^g are respectively the average ED length of stay for a red, yellow and green code, while p_u^r, p_u^y and p_u^g are respectively the number of patients inside the ED facility (both waiting for admission and under treatment) for a red, yellow and green code.

To counterbalance the effect of the longer travel times in the case of ED facilities less overcrowded but far from the emergency request, the policies **H-SAQ** and **H-WLB** are applied only taking into account the ED facilities no farther than the *radius* of G, that is half of the longest travel time between a node $u \in N$ and the current ED facility.

9.3.3 Ambulance Redeployment

In the real practice, a simple policy is that of redeploy the ambulance to its original base. In the following we refer to this policy as **R-Base**. Alternatively, the aim of the EMS management should be to make an ambulance available as soon as possible redeploying it to the closest base. Therefore, the **R-Closest** policy redeploy the ambulance to the closest base at the end of the mission. This is one of the most used policy in the real settings.

A third version of the above two policies is the **R-LCBT** policy: it redeploys the ambulance to the less covered base b within a given time threshold T^R as follows

$$\text{argmin}_{b \in L_R}\{A_b^a - A_b^e\}$$

where L_R is the list of the bases that can be reached from the current ED facility within T^R. Note that T^R is a parameter introduced to counterbalance the effect of the longer travel times ad remarked in [144].

9.4 Quantitative Analysis

In this section we provide an analysis of the proposed policies in order to evaluate their impact when are used separately or together.

In order to evaluate the proposed DRRP policies, we define in Table 9.3 several performance indices taking into account the ambulance utilization and the most important aspects for the patient safety and satisfaction, which regard the waiting time from the moment of the phone call to the arrival of the ambulance and the waiting time at the ED. Observe that red code patients do not queue at the ED, then their waiting times are omitted.

Tab. 9.3: Performance indices.

Index	Definition
r	Average time to reach a request (min)
f_g	Fraction of green code patients reached within 20 min (%)
f_{ry}	Fraction of red and yellow code patients reached within 8 min (%)
u	Ambulance utilization considering only the mission time (%)
u^+	Ambulance utilization considering also the redeployment (%)
w_g	Average waiting time of green code patients at the ED (min)
w_y	Average waiting time of yellow code patients at the ED (min)
u_{ED}	ED utilization (%)

The set of configurations taken into account in our analysis is reported in Table 9.4 and defined as different combination of the DRRP policies. All the possible policy combination have been tested for each scenario, but we report the most significant for reasons of synthesis. Given a certain scenario, we start from the baseline configuration (0), in which the basic policies are used, and we change one at a time the policy for ambulance dispatching (1), for ED facility selection (2–3) and ambulance redeployment (4–5). The same policies are defined enabling the D-SA option (0s–5s). Other configurations changing more than one policy w.r.t. the baseline (6 and 7s–9s) have been chosen because they highlight interesting aspects for the analysis. For the same reason, two configuration (6t and 0st) have been selected to study the impact of the option D-CPQ.

Tab. 9.4: Configurations of DRRP set for the analysis.

id	Ambulance Dispatching	ED Facility Selection	Ambulance Redeployment
0	D-Closest	H-Closest	R-Base
1	D-LLCB	H-Closest	R-Base
2	D-Closest	H-SAQ	R-Base
3	D-Closest	H-WLP	R-Base
4	D-Closest	H-Closest	R-Closest
5	D-Closest	H-Closest	R-LCBT ($T^R = 20$ min)
6	D-Closest	H-SAQ	R-Closest
0s	D-Closest, D-SA	H-Closest	R-Base
1s	D-LLCB, D-SA	H-Closest	R-Base
2s	D-Closest, D-SA	H-SAQ	R-Base
3s	D-Closest, D-SA	H-WLP	R-Base
4s	D-Closest, D-SA	H-Closest	R-Closest
5s	D-Closest, D-SA	H-Closest	R-LCBT ($T^R = 20$ min)
7s	D-Closest, D-SA	H-SAQ	R-LCBT ($T^R = 20$ min)
8s	D-LLCB, D-SA	H-WLP	R-Base
9s	D-LLCB, D-SA	H-WLP	R-LCBT ($T^R = 20$ min)
6t	D-Closest, D-CPQ	H-SAQ	R-Closest
0st	D-Closest, D-SA, D-CPQ	H-Closest	R-Base

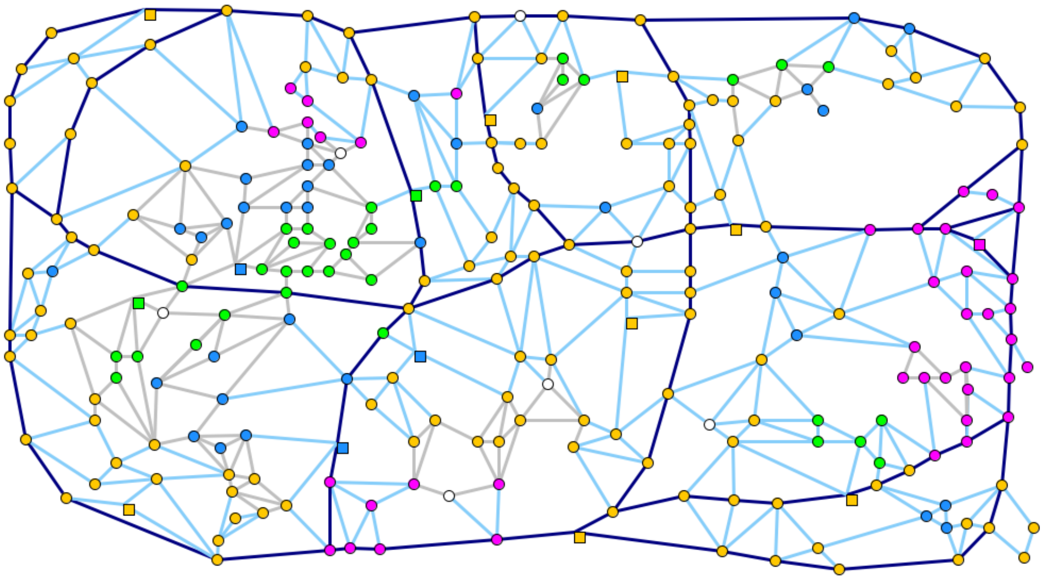
After fixing a policy configuration and a scenario through the model parameters, we execute 30 simulation runs over a time horizon of $I = 6$ days: the first day is used for the warm-up, while the values of the performance indices are collected over the other five days. The simulation model is implemented using AnyLogic 7.3.7 [73].

For the type of abstraction on which our model is based, it can not be validated through a comparison with real data. For this reason, we would check if the model is sensitive to the increase of the demand and if the ambulance and ED utilizations are consistent with the fixed ambulance and ED workloads, respectively. Observe that a total ED capacity of αC_u should cause an utilization equal to $\frac{1}{\alpha}$. To this aim, configuration 4 is used as comparison terms, because it includes the set of policies that minimize the traveling time of a single mission, supposing that an ambulance is always available on the closest base, and allows us to have reliable checking over non-crowded scenarios. Results confirm the validity of the model, as well as values of u and u_{ED} in the analysis in 9.4.1 are consistent with workloads.

9.4.1 Policy comparison

All tests are made on the graph illustrated in Figure 9.4, composed by $n = 248$ nodes and $m = 512$ arcs on a metropolitan area of 880 km^2, among which $n_{ED} = 7$ hospital and $n_B = 14$ bases are distributed approximately in a balanced way. High-speed roads have been drawn through maximum speed arc paths, while traffic areas are located in different points using minimum speed arcs.

Fig. 9.4: Graph used for the quantitative analysis.



Six different scenarios are analyzed in our analysis. Scenarios 1–3 are obtained ranging the ambulance workload W^A in $\{30\%, 40\%, 50\%\}$ and setting the total ED capacity equal to 1.5 times the minimum capacity needed to deal with the demand. Then, scenarios 4–6 are defined for the same values of W^A but keeping the same total ED capacity, which is $1.7C_u$, $1.275C_u$ and $1.02C_u$ when $W^A = 30\%$, 40% and

50%, respectively. The initial number of ambulances A_b per base is fixed equal to 2 in all scenarios, while the constraint about the maximum number A_b^{max} is relaxed in such a way to allow a higher flexibility to the redeployment policies, and to analyze their impact in the most favorable situation.

Tab. 9.5: Results: ED capacity is $1.5C_u$, that is proportional to demand.

	Scen.1 - $W^A = 30\%$					**Scen.2** - $W^A = 40\%$					**Scen.3** - $W^A = 50\%$				
id	r	f_g	f_{ry}	u	u^+	r	f_g	f_{ry}	u	u^+	r	f_g	f_{ry}	u	u^+
0	6.1	98.2	78.3	27.0	32.4	24.8	59.1	42.8	41.8	52.7	110.0	14.1	14.0	57.9	75.5
1	6.9	98.0	75.8	27.6	33.4	24.3	58.6	41.1	42.1	53.2	111.8	13.7	13.3	58.1	75.7
2	6.0	98.4	78.9	27.0	32.5	22.1	60.8	43.5	41.7	52.6	111.0	14.3	13.9	57.9	75.5
3	6.8	96.1	74.2	28.9	35.6	28.7	51.1	37.0	43.8	56.0	108.0	13.9	14.1	57.6	75.2
4	8.5	96.0	57.9	28.8	31.8	11.2	85.5	46.8	40.2	44.3	32.5	41.0	25.4	54.0	59.1
5	7.0	96.6	70.7	27.8	32.0	11.4	81.4	54.9	39.6	45.3	45.6	31.8	26.3	55.0	62.3
6	8.6	95.2	57.3	29.1	32.2	10.8	87.0	47.5	40.0	44.1	34.9	38.7	24.8	54.6	59.8
0s	5.8	99.0	79.4	25.9	31.6	7.8	94.5	65.4	37.0	42.9	20.7	58.4	37.2	52.3	56.8
1s	6.9	98.6	75.9	27.6	32.4	8.9	93.0	61.3	38.1	44.0	20.3	58.8	36.2	52.8	57.3
2s	5.8	99.1	79.0	27.0	31.8	7.9	94.4	64.6	37.4	43.4	18.7	61.7	38.3	52.0	56.7
3s	6.1	98.8	76.1	28.5	34.1	8.9	91.0	58.9	39.8	46.4	19.4	59.0	35.5	53.6	58.6
4s	8.4	96.4	58.6	28.6	31.5	9.7	91.3	51.1	39.4	42.9	19.8	59.5	32.3	52.9	55.6
5s	6.7	97.2	72.0	27.5	31.4	8.4	92.4	62.7	37.8	42.6	18.6	62.5	40.1	51.7	55.4
7s	6.7	97.1	72.6	27.6	31.5	8.2	92.8	63.8	37.6	42.5	19.1	60.7	39.8	51.9	55.5
8s	7.3	97.6	71.5	29.2	34.7	9.9	89.6	55.8	40.7	47.1	21.2	55.2	33.5	54.4	59.1
9s	7.9	95.7	67.9	29.7	33.7	9.7	92.0	51.0	39.3	42.8	21.2	56.0	35.2	54.0	57.5
	average $u_{ED} = 64.9\%$					average $u_{ED} = 65.1\%$					average $u_{ED} = 65.2\%$				

Results of Scenarios 1–3 are reported in Table 9.5 focusing on indices regarding only performance of the ambulances. As expected, the increasing of the ambulance workload W^A causes a robust lengthening of waiting times, which pass from 6 to 110 min on average for the baseline configuration, and a general worsening of the indices. Such an increasing allows us to appreciate the impact of different configurations, that is for scenarios 2 and 3. Enabling only the policy D-LLCB with respect to the baseline configuration, a slight general worsening can be observed, while H-SAQ and H-WLP provide small variations depending on the considered scenario. More significant is the impact of the redeployment policies and in particular R-Closest, which worsens the fraction f_{ry} of the urgent patients reached within 8 min by an ambulance of about 20%, but it halves the waiting times in scenario 2 and reduce them of more than 70% in scenario 3. However, a more relevant impact is observed enabling the D-SA: regardless of which policies are enabled, it always allows better performance than when it is not activated. In particular, considering the best configuration with and without enabling the D-SA, f_{ry} raises from 54.9% to 65.4% in scenario 2 and from 26.3% to 40.1% in scenario 3 confirming the results on the real case study reported in [140].

As a counter intuitive result, we observe that the configuration with the highest average waiting time r in scenario 2 (0) becomes the better just enabling the D-SA (0s). Similarly, in scenario 3 the value of r passes from 110 to 21 min (-81%) with only the contribution of the D-SA. Finally, we observe that ambulance utilizations u and u^+ are consistent with the value of the parameter W^A and little variations cause significant differences in performance. The same observation worths for the ED utilization, which have not significant variations among configurations and whose average value is reported in the last row of Table 9.5.

Tab. 9.6: Results: ED capacity is fixed, that is $1.7C_u$, $1.275C_u$ and $1.02C_u$ in scenarios 1, 2 and 3, respectively.

	Scen.4 - $W^A = 30\%$					**Scen.5** - $W^A = 40\%$					**Scen.6** - $W^A = 50\%$				
id	r	f_g	f_{ry}	w_g	w_y	r	f_g	f_{ry}	w_g	w_y	r	f_g	f_{ry}	w_g	w_y
0	6.2	97.5	77.6	1.8	0.6	21.5	61.5	44.9	41.7	4.6	110.0	13.7	13.8	230.0	5.5
2	6.0	98.2	78.7	1.4	0.7	23.2	58.5	41.1	35.0	4.8	114.4	12.8	12.9	204.2	5.6
4	6.7	96.5	74.1	0.3	0.1	26.7	50.8	37.1	36.4	5.1	69.6	27.1	18.5	199.3	5.6
6	8.5	95.8	57.4	1.9	0.8	11.9	83.2	43.3	28.6	4.8	37.1	34.5	20.5	203.7	5.6
0s	5.8	98.9	79.6	2.3	0.7	7.9	94.6	64.9	40.6	4.7	26.1	52.5	33.6	245.9	5.6
2s	5.9	99.0	79.0	1.4	0.6	8.4	93.0	61.3	26.7	4.9	21.6	52.5	33.6	188.9	5.6
4s	6.2	98.6	75.7	0.3	0.1	9.4	89.6	56.6	28.4	4.9	20.3	56.9	35.1	177.2	5.8
9s	9.4	92.6	52.1	0.2	0.1	11.3	85.4	43.3	29.5	4.9	22.7	51.9	31.9	178.4	5.8
	average $u_{ED} = 57.1\%$					average $u_{ED} = 76.2\%$					average $u_{ED} = 93.2\%$				

In Table 9.6 we summarize the most significant configuration to analyze the impact of ED facility selection on the waiting time for the ambulance arrival and the admission at the ED. Scenarios 4–6 are obtained ranging the value of W^A as well as for scenarios 1–3 but fixing the total ED capacity, which is equal to $1.7C_u$, $1.275C_u$ and $1.02C_u$, respectively. Such scenarios allow us to analyze the trade-off between the indices regarding the ambulance performance and those about the ED performance. The impact on the ED waiting times is evident: policy H-SAQ give a significant decreasing, but H-WLP is the best policy, up to -13% in scenarios 2 and 3. Both policies perform better when used enabling the D-SA, although this is not its goal. However, these improvement are not always concurrently with the best solution for the fast arrival of ambulance. For instance, in scenario 2 configuration 0s have better values of r, f_g and f_{ry} than configuration 2s, but green code patients have a waiting time 52% higher at the ED. In some cases, a good compromise can be found combining policies, as happens in scenario 2 for configuration 6 with respect to other configuration without H-DA. In the last row of Table 9.6, we can observe the model consistency about the fixed ED workload parameters. The trade-off between the average time to reach an emergency request and the average waiting time at the ED confirms the fact that incorporating equity might lead to a lower service or negative outcomes as reported in [183].

The impact of the cut-off on the ambulance dispatching is studied in Table 9.7, where results of 2 configurations (with and without D-SA) with a good trade-off between ambulance and ED indices are reported. The threshold parameter of the policy D-CPQ ranges between 5% and 50% of the total number of ambulances. We observe that waiting times of urgent and non-urgent patients are sensitive to the D-CPQ: at the increasing of the threshold, the formers have an improving at the expense of the latters. However, in this scenario the use of the D-CPQ seems to be inadvisable, because of the possible high negative impact on green code patients to obtain a slight time saving for the other patients. This confirms the fact that priority dispatching policies can improve the performance for urgents at the price of a worsening those of non-urgents, as reported in [143].

Tab. 9.7: Results: impact of the D-CPQ policy varying the threshold (Scenario 2).

id	threshold	r	f_g	f_{ry}	u	u^+	w_g	w_y
6t	5%	11.0	86.3	47.1	40.2	44.4	7.6	2.4
6t	15%	11.2	85.3	46.6	40.4	44.5	6.6	2.0
6t	25%	11.3	84.7	48.2	40.2	44.3	6.8	2.3
6t	30%	11.7	81.8	47.5	40.4	44.5	6.7	2.1
6t	35%	13.6	75.2	49.3	40.1	44.2	7.1	2.3
6t	40%	14.3	70.6	50.3	39.9	44.1	6.9	2.2
6t	45%	23.4	54.0	50.2	40.8	44.9	6.0	2.1
6t	50%	40.2	40.5	52.0	40.6	44.7	6.3	2.0
0st	5%	8.1	93.3	64.4	37.7	43.7	10.6	2.3
0st	15%	8.1	93.8	63.7	37.8	43.8	10.7	2.4
0st	25%	8.1	92.0	64.7	37.4	43.3	10.5	2.3
0st	30%	8.4	89.9	64.8	37.6	43.6	9.5	2.2
0st	35%	9.5	82.2	65.4	37.9	43.8	9.1	2.3
0st	40%	11.4	73.7	64.9	38.3	44.2	9.4	2.2
0st	45%	16.3	57.5	65.9	38.8	44.5	8.3	2.1
0st	50%	30.7	34.3	67.0	39.8	45.5	8.8	2.1

9.5 Concluding Remarks

In this Chapters, several DRRP have been presented and analyzed for the ambulance real-time management. We provided a comprehensive analysis of the EMS system that allows us to make an extensive comparison among different policies.

In particular, we provided a general analysis of the smart assignment policy, which confirms the significant results reported by Aringhieri et al. [140] for the EMS of Milano, Italy. The impact of such a policy on sufficiently crowded scenario is huge and allows us to have performance better than using any other combination of policies.

Regarding the other policies, results shown a trade-off among their impact on a fast arrival of the ambulance and the waiting times for the admission in the ED confirming the insight reported by McLay and Mayorga [183] for which incorporating equity might lead to a lower service or negative outcomes.

More generally, the trade-off among the outcomes of the different policy combinations justify the need of a modeling approach to support decision making in the EMS management.

Further works could investigate the impact of the policies in graphs with different characteristics, such as representing a rural area, positioning the ambulance bases in an unbalanced way with respect to demand and distances, or limiting the capacity of the ambulance bases.

10

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

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 modeling the ED behavior is to replicate such different paths. Moreover, in order to implement online optimization algorithms to deal with overcrowding to intervening on bottlenecks, models capable of making predictions on the patient paths evolution would be useful. 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, as reported in the literature review in Chapter 7.

In this chapter 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. Such a process model and its ability to make prediction is used in Chapter 11 to optimize the resource allocation in the ED. We apply our new framework to a real case study arising at *Ospedale Sant'Antonio Abate di Cantù*, Italy.

The chapter is structured as follows. The case study is reported in Section 10.1 describing the population of the patients and the ED organization, also providing a simple retrospective analysis. After describing how to pre-process our data-sets, in Section 10.2 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 Section 10.3 testing its replicability and its robustness over a new data-set. Concluding remarks are discussed in Section 10.4.

10.1 The case study

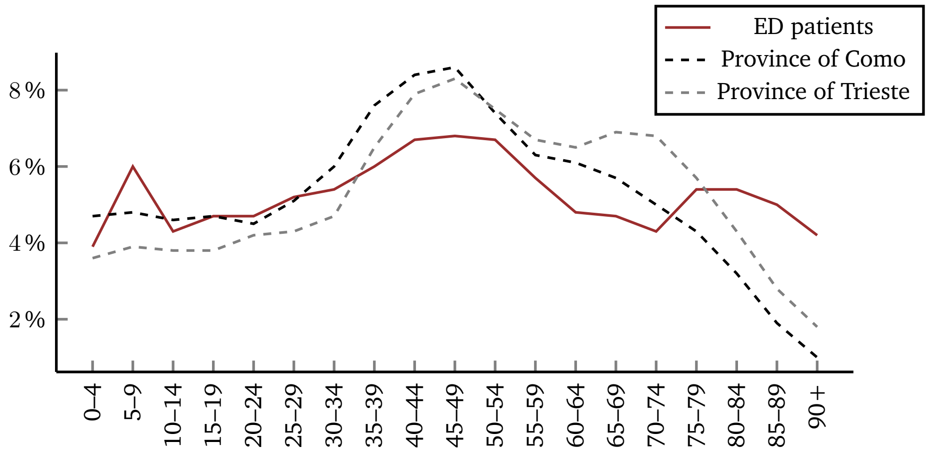
We present a real case study concerning the ED sited at *Ospedale Sant'Antonio Abate di Cantù*, which is a medium size hospital in the region of Lombardy, Italy. The ED serves about 30 000 patients per year.

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.

10.1.1 Patient population

Thanks to the collaboration with the ED, we have available all the data concerning all the 88 272 accesses made in the years 2013–2015. Such data contains sex (male 52.7% or female 47.3%) and age of the patient, type of access (autonomously 79.9% or with a rescue vehicle 20.1%), 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%, hospitalization 10.6% and abandonment 7.4%).

Fig. 10.1: Comparison between territorial and patient age distributions.



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 five-year age classes with the demographic distribution. As shown in Figure 10.1, the almost uniform distribution of accesses among the 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 EDLOS, as we will see below.

10.1.2 Organization 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 10.1.

Tab. 10.1: Urgency codes: description and frequency over 2013–2015

number	color	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%

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 data-set. 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 computerized tomography, an ecography or a specialist visit. Then, there are two different SSO, both requiring a glssso bed unit and the supervision of nurses and physicians: the first is the ordinary glssso for medical reason, while the second is the pre-hospitalization SSO, that is when patient need to be hospitalized 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 hospitalization or the discharge for patients needing non-urgent investigations.

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 (or 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 pediatric 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 hospitalized.

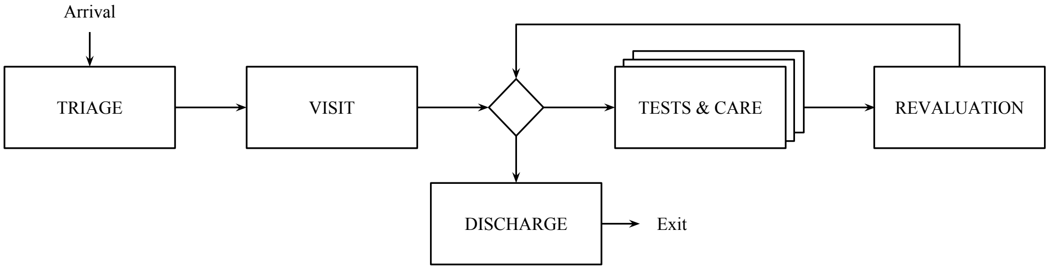
Table 10.2 resumes all the activities that could be performed by a patient within the ED. The first and the second columns indicate respectively an identifier for each activity and its description. Then, we classify the activities into 5 classes called *Triage*, *Visit*, *Tests & Care*, *Revaluation* 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 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.

Tab. 10.2: 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	Computerized tomography	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

Figure 10.2 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.

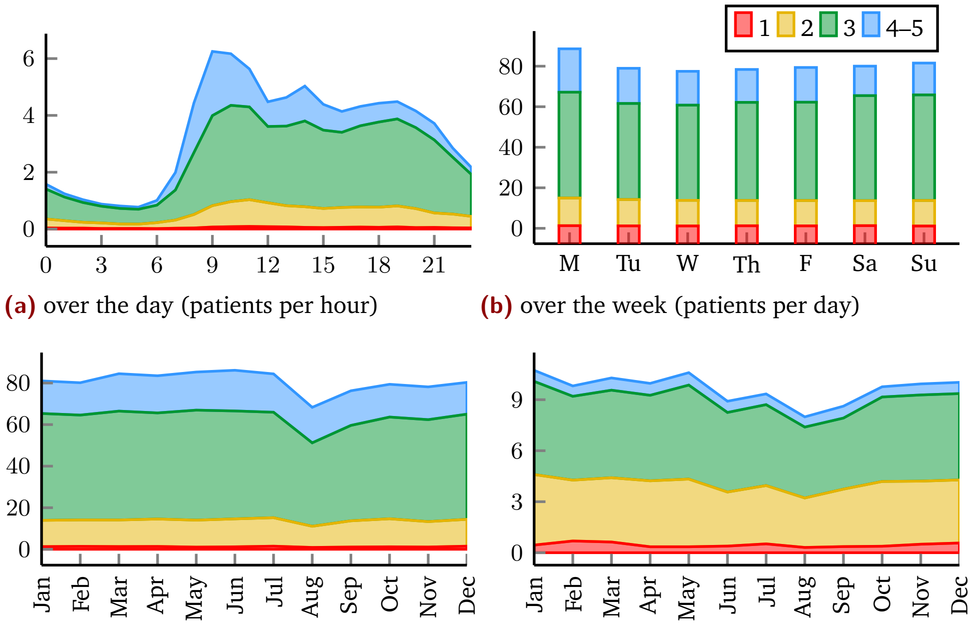
Fig. 10.2: A general path for a patient within the ED.



10.1.3 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. [165], 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.

Fig. 10.3: Patient accesses divided for urgency code



(a) over the day (patients per hour)

(b) over the week (patients per day)

(c) over the year (patients per day)

(d) number of patients concurrently treated

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 Figure 10.3(a), 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 10.3(b). 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 Figure 10.3(c) 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 Figure 10.3(d) 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.

Tab. 10.3: Waiting times, LWBS and statistics on the treatment of patients

urgency code	average wait time	percentage of LWBS	average EDLOS	perc. of SSO	average SSO duration	perc. of hosp.
1	15 min	0.2%	9 hours	28.4%	19 hours	60.3%
2	34 min	0.5%	7 hours	19.5%	20 hours	28.9%
3	65 min	3.3%	2.5 hours	5.4%	19 hours	7.8%
4–5	68 min	11.1%	1 hour	0.5%	17 hours	1.1%

The statistics in Table 10.3 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 hospitalizations. 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, that is patients leaving after the medical visit but without finishing the treatment, is similar for all the urgency classes.

10.2 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 (i) to replicate properly the possible patient paths, and (ii) to predict the next activities and the required resources of patients on the basis of their characteristics and their activities performed until that moment.

10.2.1 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.

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 10.2.

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 analyzed and reported. However, we have to take into account several noise factors that may be present in the data-set provided by the emergency room. 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;

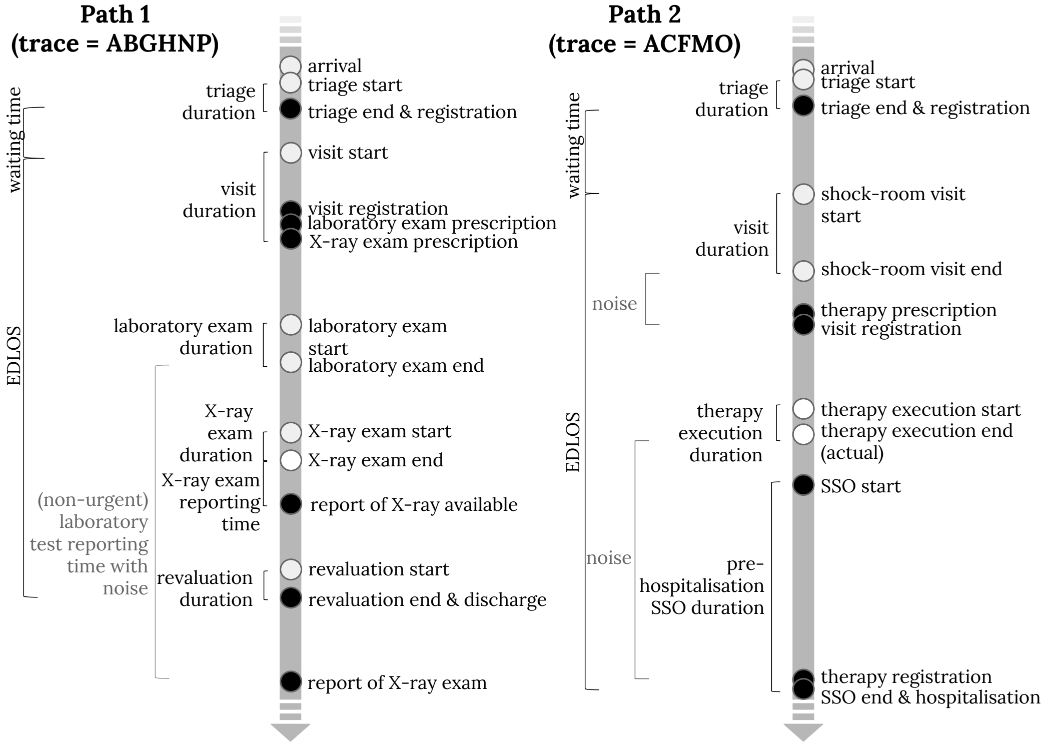
$\mathfrak{N}_4$ – ***fake or missing revaluation visit:*** sometimes the revaluation record can refer to the passage of the medical record between two physicians for the change of work shift, while other times a revaluation visit could be performed without to be recorded if the patient is discharged at once (but from ED suggestions we know that always a revaluation is performed between tests and discharge);

$\mathfrak{N}_5$ – ***fake medical visit:*** pediatric visits are performed in the Pediatric Department but that are activities also recorded in the data-set 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 analyzed and reported after the patient discharge.

In Figure 10.4 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 10.4. 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 data-set, while for the therapy execution we suppose that the start time is immediately after the prescription by the physician.

Fig. 10.4: Example of the activities for two different paths with the corresponding timestamps (black dots) and other significant times (white dots).



The pre-processing algorithm has been implemented as follows:

1. Start time and end time of each activity are estimated in accordance with Table 10.4 (noise $\mathfrak{N}_0$).

Tab. 10.4: 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

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.

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 one minute 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 one minute after the triage time (noise $\mathfrak{N}_1$);

   - if it is not a triage and it precedes the visit time, then it is moved one minute 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 minutes, 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 seconds for the whole data-set. Excluding LWBS and the pediatric fast-tracks, corresponding to the trivial traces AQ and AE, the remaining $66\,551$ cases generated $7\,868$ different traces of length ranging in $[3, 31]$, with an average value of 5.5. 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.

10.2.2 Standard process discovery

We report a summary of the analysis of process discovery techniques from the literature conducted in Duma and Aringhieri [157]. Models and results have been updated after a more accurate pre-processing in accordance with the suggestions by the ED staff.

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 [148]: fitness, precision, generality and simplicity. Fitness indicates how much of the observed behavior 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 behavior 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.

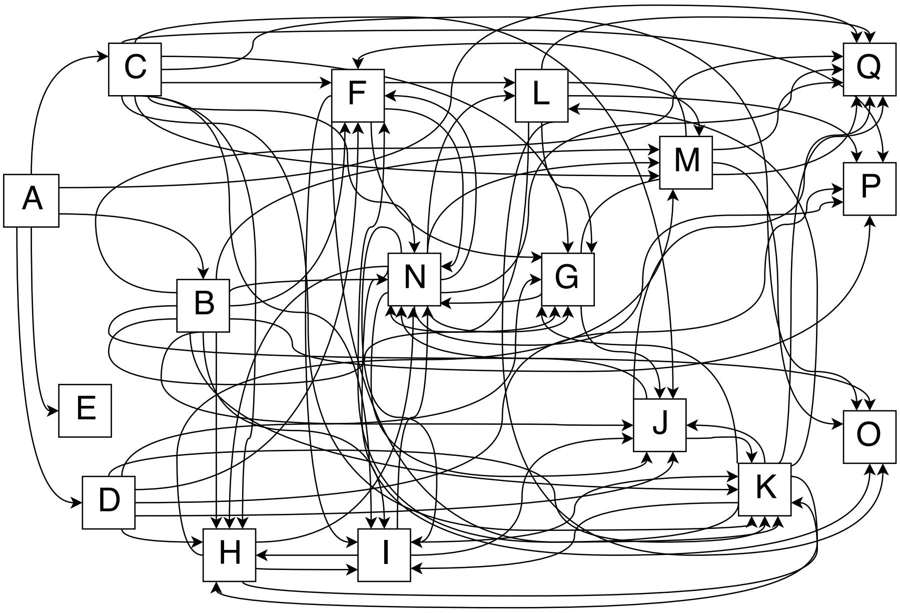
The huge number of traces suggests the use of discovery techniques that deal with low frequent behavior and noise. We focus on two different process miners, the *HeuristicMiner* (HM) [199] and the *Inductive Miner – infrequent* (IMi) [177], both based on the control-flow perspective.

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 splitted into sub-logs using these sets. The IMi uses the same approach but filters a fixed percentage of traces representing infrequent behavior 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 provided by ProM 6.6 are shown in Figures 10.5 and 10.6.

Fig. 10.5: Process model mined with the HM: model $\mathcal{H}$ (heuristic 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 behavior 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.

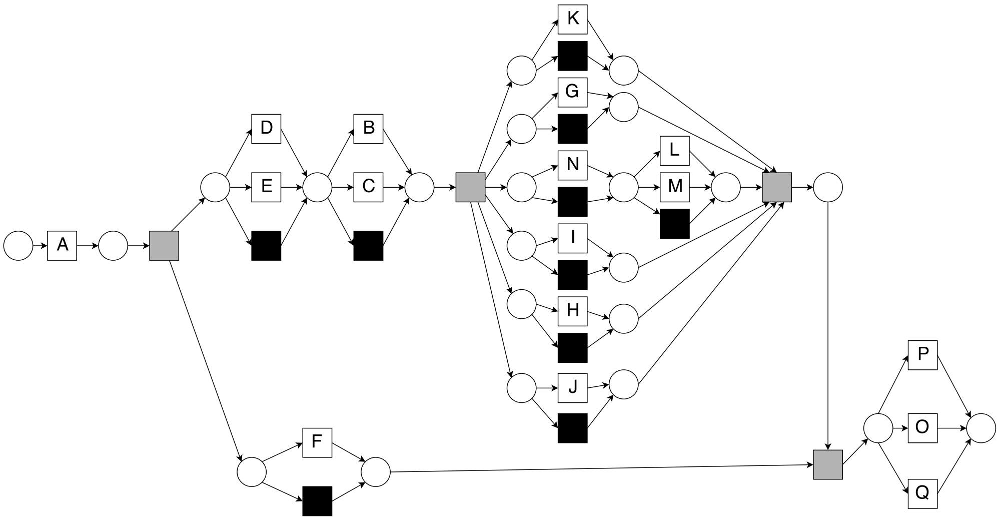
The model $\mathcal{I}$ has been obtained varying the noise parameter of the IMi in order to have a good precision, avoiding or limiting infrequent behavior. However, we observed very slight deviations among the models ranging the noise percentage, that has been fixed to 20%. Contrariwise to $\mathcal{H}$, this model is very simple but not precise: the parallelisms among activities allowed by $\mathcal{I}$(represented by the grey boxes in the Figure 10.6) imply additional behavior 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 behavior 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 [136] for an ED case study to show the main highway paths for patients to gain insights into bottlenecks and resource utilization. However the Fuzzy Miner is not suitable to our purpose, because the level of granularity necessary to implement a process model to analyze 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 $\mathcal{I}$.

10.2.3 Ad hoc process discovery model

Starting from the remarks in Section 10.2.2, we would like to design a model with a better compromise between fitness, precision, generalization 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

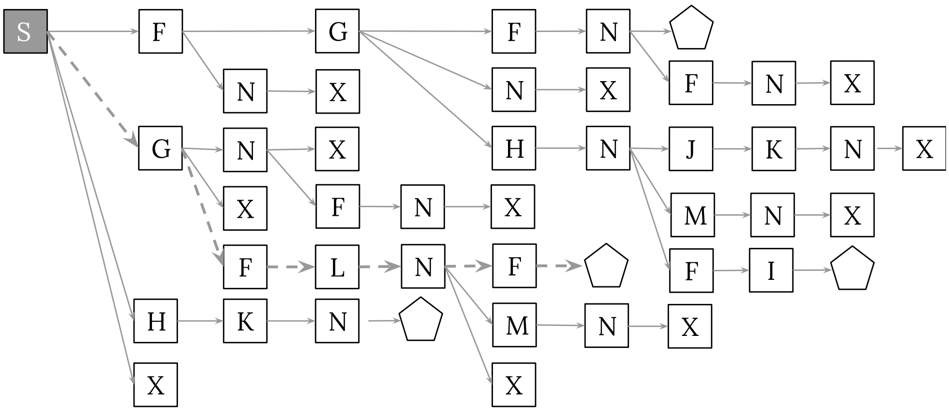
Fig. 10.6: Process model mined with the IMi: model $\mathcal{I}$ (Petri net).



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 Figure 10.7. 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 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.

Fig. 10.7: Example of process model with a tree-structure. Dashed edges highlight the possible path SGFLNF.



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 labeling 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 generalization requirement – and generating a high number of nodes. To overcome these limitations, we summarize infrequent branches with graphs, in which we do not keep track of the past activities.

A possible path is highlighted with dashed edges in Figure 10.7, whose trace starts with a node labeled with G and followed by other tree nodes labeled 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 Figure 10.8.

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 Figure 10.2 drawn together with the staff and consistent with the previously obtained models.

Excluding the cases of the LWBS and the pediatric 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 optimize 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 orthopedist 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.

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

Fig. 10.8: Example of a sub-process model with a graph-structure

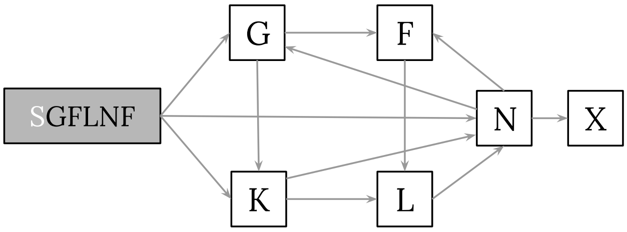
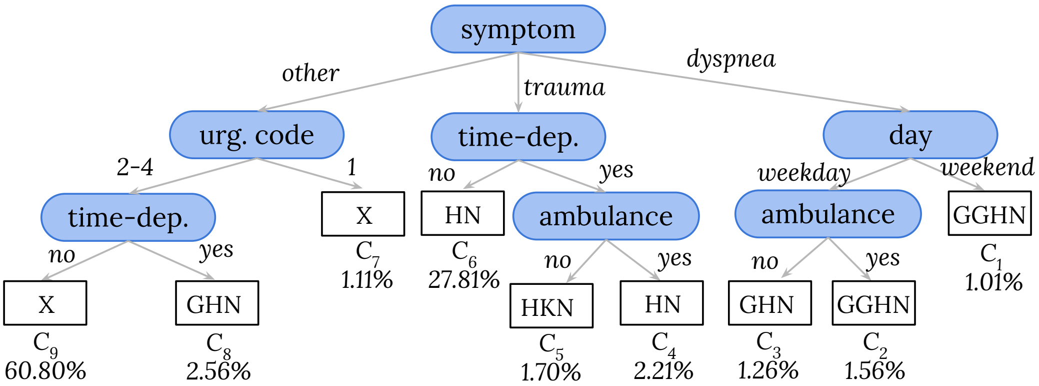


Fig. 10.9: Decision tree with gain parameter set to 0.25 and obtained clusters $C_1 - C_9$



specific symptoms), arrival day (Monday–Sunday), type of arrival day (weekday or weekend), month of arrival (January–December), arrival time slot (60 minutes period) and type of first medical visit (ordinary, shock-room or MCA).

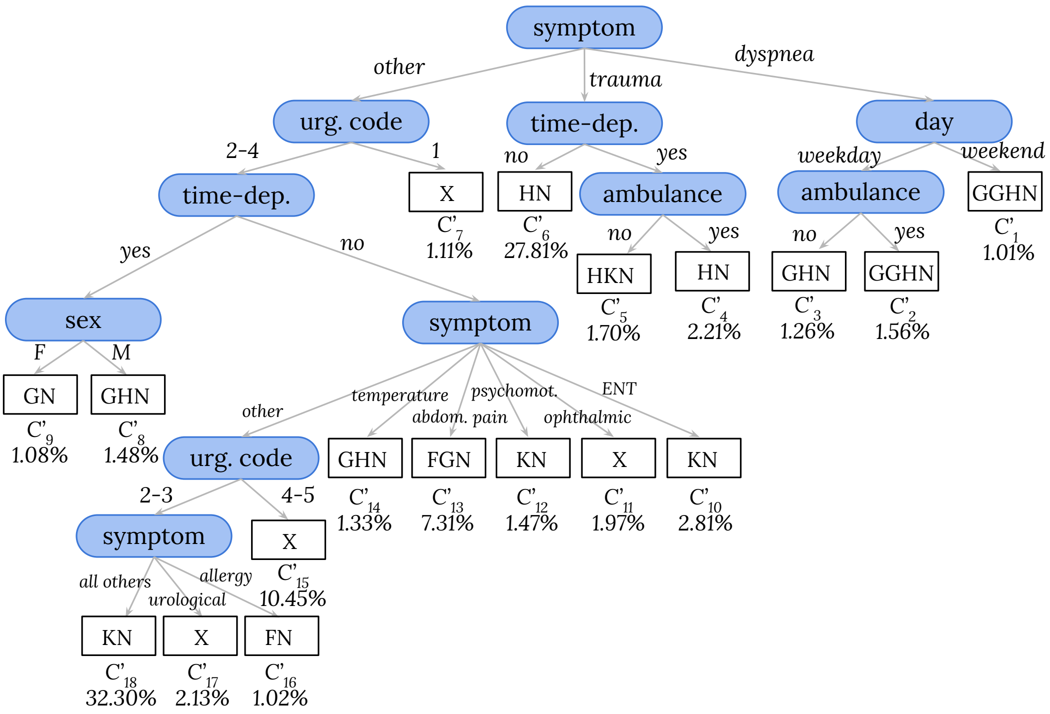
The DT approach requires the following parameters. We use the criterion called "gain ratio", that is used to reduce a bias towards multi-valued attributes by taking the number and size of branches into account when choosing an attribute. We fixed a confidence equal to 0.25 and imposed a minimum leaf 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 (Figure 10.9) and 18 (Figure 10.10), respectively.

We denote with $\{C_i\}_{i=1,\ldots,9}$ and $\{C_i'\}_{i=1,\ldots,18}$ 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 $i = 1, \ldots, 7$, $C_8 = C_8' \cup C_9'$, and $C_9 = C_{10}' \cup \ldots \cup C_{18}'$. The clusters obtained through the data 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.

Phase 2: Process Modelling

For each cluster defined in the first phase of our ad hoc approach, we model the behavior 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 labeled with the ED activities (those in Table 10.2) and $\mathbb{T}$ is a set of oriented edges indicating possible transitions between nodes and labeled with a weight $f \in (0, 1]$ equal to the relative frequency of that transition. We remark that different nodes can be labeled with the same activity: each of them represents the execution of such an activity after the execution of different activity sequences.

Fig. 10.10: Decision tree with gain parameter set to 0.2 and obtained clusters $C'_1 - C'_{18}$



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, labeled with X, or the starting node of a graph, called Sub-Tree Activity Graph (STAG), which is used to model infrequent behavior (indicated with a pentagon in Figure 10.7). 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 labeled 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 data-set 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 $\wp$ labeled 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.

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 labeled with a certain activity in τ, then a unique node is inserted in γ with that label;
   - 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.

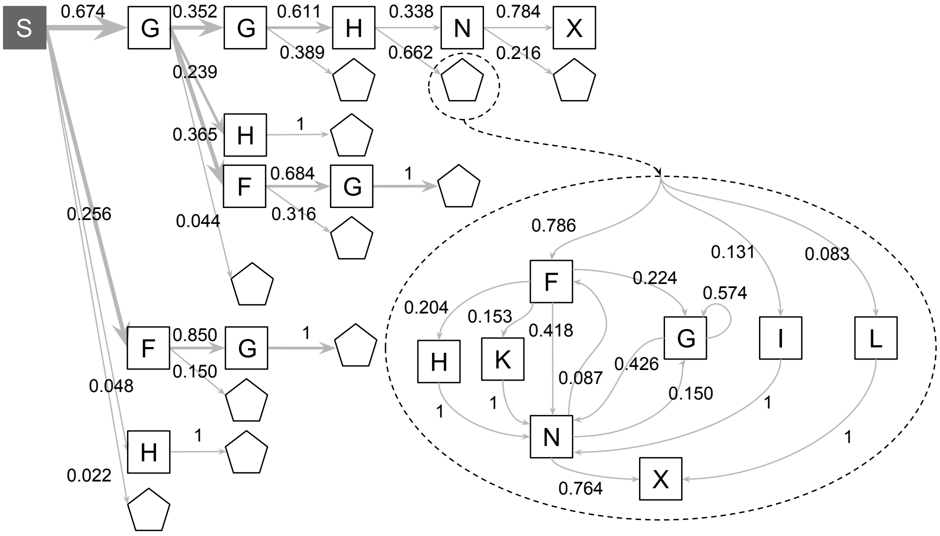
We call Hybrid Activity Forest (HAF) a set of HATs that model the behavior of different clusters $C_1, \ldots, C_l$ of patients. Let $\Gamma = \{C_1, \ldots, C_9\}$ and $\Gamma' = \{C'_1, \ldots, C_{18}\}$ be the sets of the partitions obtained through the two clusterings performed in the phase 1 of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

Table 10.5 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 behavior of the process, but could be used without problems of computational

Tab. 10.5: Characteristics of the HAFs using different clusters and values of ℓ.

			average number	mined traces totally		comp.
			of pure tree nodes	replicated on tree nodes		time
name	clustering	ℓ	in a single HAT	(number)	(percentage)	(secs)
$\mathcal{F}_1$	Γ	1	5 311	66 551	100.0%	3.8
$\mathcal{F}'_1$	Γ'	1	2 884	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

Fig. 10.11: HAT of cluster $C_2 = C'_2$ fixing $\ell = 30$



efficiency because of the tree structure, which avoid cycles and allows a simple calculation of frequency of a certain event.

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 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 generalization. The HATM required always less than 4 seconds of computational time for each parameters combination. Figures 10.11–10.14 show the differences of using different values of the parameter ℓ, for two clusters that are equals for both clustering Γ and Γ'.

In Figures 10.11 and 10.12 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 $\ell = 100$ (Figure 10.12) is too high to have a significant process model, because of the low number of patients in this cluster ($1\,041$ 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 (depicted in the figure). On the contrary, for $\ell = 30$ (Figure 10.11) the most common paths or the initial parts of them are easy deductible and different 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 (Figures 10.13 for $\ell = 30$ and Figure 10.14 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 suppose they performed the activity sequences ABGGH, ABH and ABFH, respectively. This means that the π_1 is positioned before the unique node labeled with N in Figure 10.11, π_2 is before the node labeled with N at the top of Figure 10.13, 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.

Fig. 10.12: HAT of cluster $C_2 = C_2'$ fixing $\ell = 100$

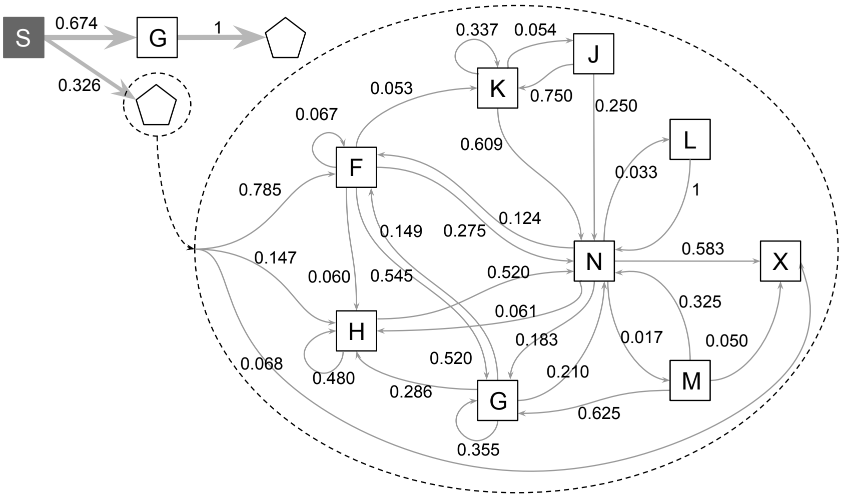
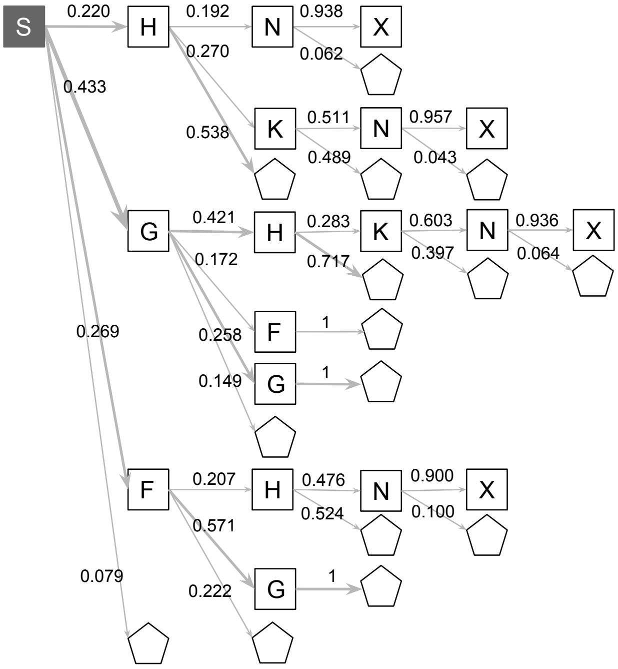


Fig. 10.13: HAT of cluster $C_4 = C_4'$ fixing $\ell = 30$



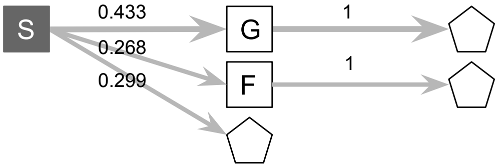
10.3 Conformance checking

In this Section, we analyze the quality of the process models discovered in Section 10.2.3 in the perspective of replicating the paths of patients that are not in the mined data set using the HAFs, which is the standard conformance checking (Section 10.3.1), and predicting the occurrence of an event, for example the execution of a particular ED activity in a certain phase of the path, that is the analysis of the robustness of our models with respect to the frequencies of the HAFs (Section 10.3.2).

10.3.1 Replicability

In order to perform a conformance checking of the HAFs discovered using the two clustering Γ and Γ' and the value 1, 30 and 100 for the parameter ℓ of the HATM algorithm using the event log of the period 2013–15, we implemented a conformance checking algorithm that, given as input a new event log E and a HAF model $\mathcal{F}$ returns the conformance index c, defined as follows:

Fig. 10.14: HAT of cluster $C_4 = C_4'$ fixing $\ell = 100$



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

We used the event log obtained applying the pre-processing algorithm (discussed in Section 10.2.1) to the ED data-set of the $29\,155$ patients arrived at the ED during the year 2016. Table 10.6 reports the conformance index c for the 6 discovered process models.

Tab. 10.6: Percentage of traces of the year 2016 replicable on the models discovered from the data of the period 2013–15.

model	clustering	ℓ	traces fully replicated	c
$\mathcal{F}_1$	Γ	1	26 289	90.17%
$\mathcal{F}'_1$	Γ'	1	25 895	88.82%
$\mathcal{F}_{30}$	Γ	30	28 517	97.81%
$\mathcal{F}'_{30}$	Γ'	30	28 353	97.25%
$\mathcal{F}_{100}$	Γ	100	28 913	99.17%
$\mathcal{F}'_{100}$	Γ'	100	28 828	98.88%

As expected, models $\mathcal{F}_1$ and $\mathcal{F}'_1$ have the worst conformance because of the overfitting of the event log used for the process discovery without adding any generalization for other behavior. Increasing the value of ℓ, we obtain better conformance indices, close to the 100% when $\ell = 100$, while using the two clustering Γ and Γ' there is not a significant difference.

10.3.2 Robustness

In Table 10.7 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 2013–15 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 2013–15 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 2013–15 and 2016 are very similar, with an average difference of

Tab. 10.7: Comparison between the frequencies of several events in 2013–15 using the HATs of $\mathcal{F}_1 - \mathcal{F}'_1$ 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

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 2013–15 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 10.7. Finally, no relevant differences in robustness for the clustering Γ and Γ' have been emerged from this analysis.

10.4 Concluding Remarks

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 modeling without making significant assumptions. Such simplifications significantly impact on the replicability of the simulation model used to identify bottlenecks and to analyze policies to alleviate the overcrowding.

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 need of implementing a simulation model for the evaluation of the real time allocation of the resources. Then, we would discover the patient flow to a high level of granularity, which make challenging the discovering of a model satisfying the four main quality criteria, that is fitness, precision, generalization and simplicity.

The model mined with the application of standard process discovery approach to the data-set of our case study does differ a lot from the 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 behavior. 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 our requirements and the process discovery criteria. Clustering gives important insights to identify different behavior 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 data-set, almost the 100% of its traces are replicable. Furthermore, fixing $\ell = 1$, the frequency of several analyzed events in our models is consistent in accordance with the paths of the such a data-set.

From the conformance analysis, we are able to implement 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 behavior 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.

11 A model for the online resource allocation of an Emergency Department

In Chapter 10, we studied how to process the data-set of an ED in such a way to replicate the behavior of the patient flow and to predict the evolution of the emergency pathways. In this chapter, we propose a simulation modeling approach for the evaluation of several online resource allocation methods for the ED management, which are based on the prediction of the next activities provided by the HAF and on the current state of the ED, which is given by the available and/or critical resources and demand characteristics (i.e. volume, type of patients and activities to be performed) in that moment.

To the best of our knowledge, an analysis of online approaches based on prediction have never been studied in literature. Further, the detail level of the simulation methodology required for this purpose should be remarkably high, because it is based on the replication of single activities in accordance with the patient pathways and the interdependence between their activities and the resulting occupation of resources.

To model the patient flow through the ED, we use a DES methodology, in which events occur at a particular instant in time and marks a change of 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 chapter is organized as follows. The simulation model is presented in Section 11.1, defining the agent types that interact in the ED. The implementation of the real case resource allocation and the proposal of an online optimization approach are provided in Section 11.2. The model validation and a quantitative analysis for comparing different allocation policies is reported in Section 11.3. Section 11.4 closes the chapter.

11.1 The simulation model

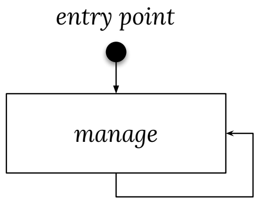
The ABS semantics allows to track the behavior of each agent acting in the simulated environment [162], and a set of rules (usually a statechart) describes the agent behavior and its interaction [163]. Therefore, we illustrate the proposed model through the description of the agents composing the model.

The first type of agent is the patient, whose statechart reproduces the general pathway structure, in accordance with the diagram in Figure 10.2 of Chapter 10. 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, a fifth agent called *decision-maker* is implemented to synchronize the other agents, managing the resource allocation and assigning tasks to the medical staff under policies that are explained in Section 11.2.

The behavior of the agents is implemented as follows.

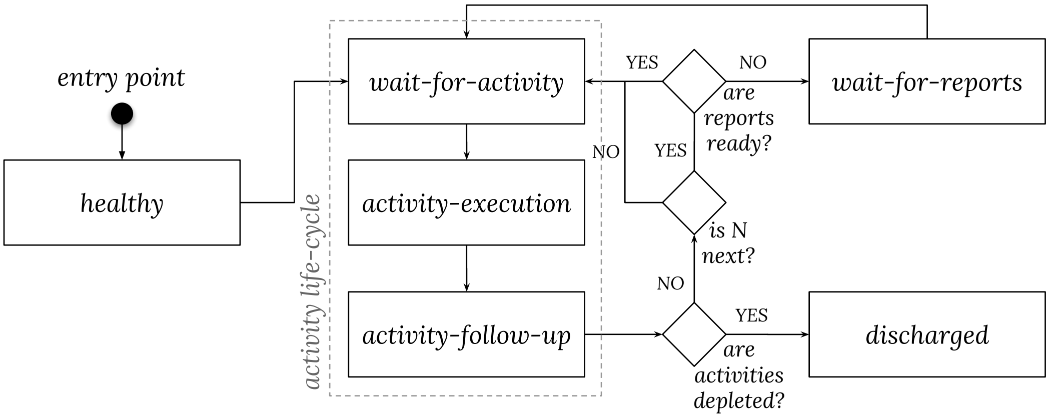
Decision-maker. When a patient need the execution of an activity, the agent is informed by a message and such a request is inserted in a prioritized queue recording the patient ID, the request timestamp, the set of resources needed, the urgency code c and a priority class or index defined by a certain rule. The agent scans the queue in real time and choose the patients to whom to assign the resources available at that moment. Then, the agent update 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 composed of a single state, and a transition that at each unit of time allows the agent to update information and to take decisions, as shown in Figure 11.1.

Fig. 11.1: Statechart of the agent *Decision-maker*



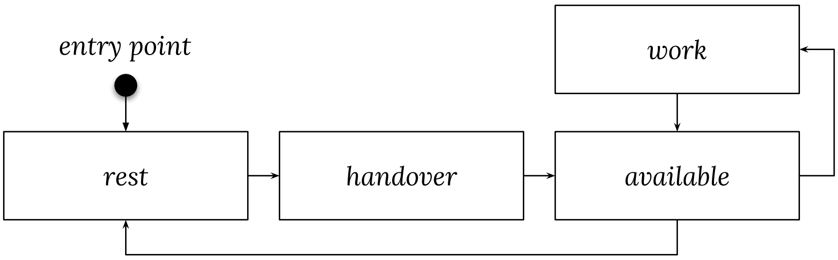
Patient. The patient population is reproduced from the event log: an agent is created for each access to the ED from the data-set and relevant information for the replication of its path (i.e. urgency code c, trace, arrival time and several activity durations) are assimilated as agent attributes. Each agent progresses in their path within the ED in accordance with its trace, following the statechart shown in Figure 11.2. The agent is on the *healthy* state until the arrival time. Then it moves on the *wait-for-activity* state, which is the first of the 3 states representing the general life-cycle of each activity and consists in sending a message to the *decision-maker* and waiting for the reply indicating the allocation of the needed resources. The second state is the *activity-execution*, which has duration defined by estimations provided by the ED staff or mined by the ED data-set, and depending on the activity type as reported in Table 10.4 of Chapter 10. A timeout passes the agent on the third state, which is the *activity-follow-up* and 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, taking into account particular cases in which it is estimable 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 the duration of one or more report times of previous activities is not finished: in such a case the patient lies in the *wait-for-report* state until the expiration of a timeout.The last situation includes all the other cases, for which the patient passes on the *wait-for-activity* state and is ready to execute the next activity of the trace.

Fig. 11.2: Statechart of the agent *Patient*



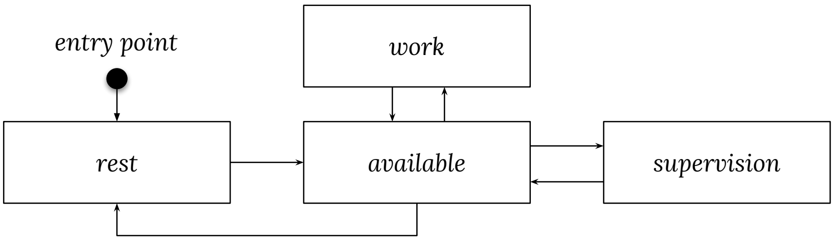
Physician. Each physician shift is represented by an agent with an attribute that indicates its competence (visit rooms, MCA, or SSO). First and revaluation visits are performed by the agents with competence visit rooms or MCA, depending on the type of visits that it executes, while SSO competence indicates 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 physician 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 he 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} min 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 Figure 11.3.

Fig. 11.3: Statechart of the agent *Physician*



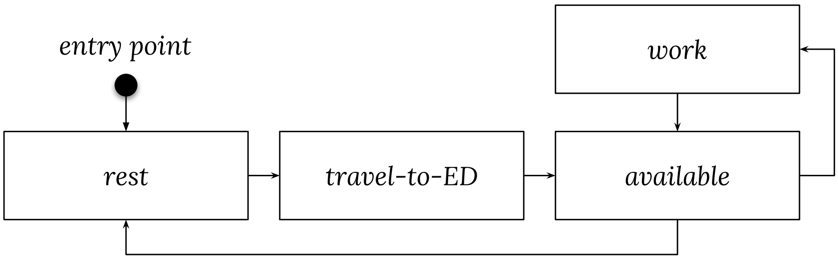
Nurse. The agent is implemented as well as the physician. An attributed indicates the competence (triage, SSO, MCA or general). Triage and SSO competence indicate that such agents could execute only the tasks of triage and supervision of the SSO units, respectively. The MCA and general competence for includes several tasks, such as first and revaluation visits, therapies, test collection for examination and assistance in other exams or specialist visits for patients with ambulation difficulties. The difference between the MCA and general competence consists in the patients assigned, which are those that are visited in the MCA or the visit rooms, respectively. The moving time of the nurse to move from the execution place of an activity to another one is considered 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 min and have duration equals to $\frac{n_{pat}^{\mathcal{A}} d_{sup}}{n_{nur}^{\mathcal{A}}}$, where d_{sur} is the average duration for assisting a patient during the supervision task, while $n_{pat}^{\mathcal{A}}$ 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 corridors or SSO units. The statechart is shown in Figure 11.4.

Fig. 11.4: Statechart of the agent *Nurse*



X-ray technician. The agent is implemented similarly to the other medical agents, but having only competence on the X-ray 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 Figure 11.5.

Fig. 11.5: Statechart of the agent *X-ray technician*



The ABS semantic allows us to model the continuity of the care process, which is allowed by the ability to identify individual resources (i.e., single physician and nurses) and to simulate their interactions: the same physician is always assigned to a patient for the activities that follow its first medical visit, that is revaluation visits and discharge; furthermore, if the assigned physician ends its shift before the completion of the care process, the activities are performed by another physician. Another important aspect represented by the model is the simulation of the behavior of the human resources during the beginning and the ending of their shift, which are the critical moments that cause a slowdown in the flow of patients.

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 a certain resources restricts the use of other 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).

11.2 Online resource allocation

In this section we present several policies for the problem of the resource allocation of the ED. In Section 11.2.1 we implement the policy used in the real case, expressing with quantitative criteria the practical sense of the ED staff in taking decisions. Then, in Section 11.2.2 we propose an online method for the online optimization of the resource allocation, which exploits the knowledge acquired by the Process Mining approach presented in Chapter 10.

In order to avoid patient paths that are not conformed to the medical guidelines and representative of the reality, all the policies presented in this section always take into account the order of activities within the patient trace without attempt to modify those sequences.

11.2.1 The real case approach

After an interview with the medical staff, we have compiled a list of criteria for the resource allocation. The main criterion is the urgency code c. Patients with $c = 1$ have always the priority on patients with $c > 1$. Then, patients with $c = 2$ and $c = 3$ have usually, but not always, priority on patients with $c > 2$ and $c > 3$, respectively, because a strong priority of these patients would result in a *starvation* situation for less urgent codes in the more crowded hours of the day. Finally, patients with code $c = 4$ and $c = 5$ (minor codes) have the less priority, but it is actually the same for the two urgency codes.

In order to replicate the real case approach, in which the common sense of the ED staff allows patients with less urgency codes to be moved forward patients with higher urgency codes, we give to each patients a priority index Φ that is equal to $\min\{c, 4\}$ at the moment of its insertion in the list of patients waiting for activities managed by the *decision-maker*. The decision-maker checks and update the priority of patients in real time as follows:

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

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 are selected. If two or more patients with urgency code $c \leq 2$ are in S_Φ, resources are assigned in such a way to perform before the execution of (i) shock-room visits, (ii) first visits, (iii) revaluation visits, and (iv) other activities. Otherwise, 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 perform the same activity, the patients with the higher waiting time is served before.

11.2.2 An online optimization approach

We propose a greedy algorithm that is simply based on a ranking of the patients that are waiting for resources. As well as the real case approach reported in Section 11.2.1, in which priority is given by the urgency codes and by the waiting times, we define a priority among patients but we also take into account which are the critical resources at that moment and we try to optimize their utilization and to lower the level of crowding. The idea is to promote the execution of those activities involving patients that occupy a certain critical resource (e.g. a stretcher) with high probability to release that resource after the activity execution.

As explained in Chapter 10, the whole trace of a patients it 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 visit (see Figure 10.2). Therefore, when patients are waiting for an activity of the visit class, we could estimate the probability of performing a certain activity after the next visit using computing its frequency on the HAT of its cluster. To this end we use:

- a HAT $\mathbb{H}_{check}$ 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 $\mathcal{P}^F$ the set of patients for which can be such checking is successful and has failed in the correspondence of the next visit class activity, respectively;

- a HAT $\mathbb{H}_{comp}$ 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 online optimization approach that we propose is divided into two phases. In the first phase, we take into account the urgency code c and the waiting time w of patients since the insertion in the *decision-maker list*. For each patient with urgency c, the following waiting index $\Psi \in \mathbb{N}_0$ is computed to have a fair allocation among patient with different urgency codes:

$$\Psi = \left\lfloor \frac{w}{m_c} \right\rfloor, \tag{11.1}$$

where m_c is a constant time fixed to normalize the waiting times with respect to c, with $m_1 < m_2 < m_3 < m_4 = m_5$.

In the second phase the *decision-maker* take into account the subsets of patients S_ψ, in increasing order of Ψ. Patients in S_ψ are then ordered in decreasing order of a score $\Sigma \in [0, 1]$ that is computed on the basis of the kind of patients or the critical resource on which has been decided to act. We propose 5 different scoring based on the probability of the occurrence of a certain activity, which is estimated using the function $\mathbb{P}$, if the next activity $\mathcal{A}$ is the first or a revaluation visit ($\mathcal{A} \in \{B, C, D, N\}$).

Let $X, Y \in \mathbb{A}$ activities using the notation reported in Table 10.2 of Chapter 10. We extend the function $\mathbb{P}$ as follows:

- $\mathbb{P}(X \vee Y)$ indicates the probability to perform ad least an activity between X and Y;
- $\mathbb{P}(X < Y)$ indicates the probability to perform both activities X and Y, with X appearing at least one time before the first occurrence of Y.

The scores for the prioritization are defined as follows.

Σ_E **– *Exit score.*** Patients have a priority that is equal to the probability to exit after the next activity, in order to lower the ED EDLOS and the number of patients in the corridors and in the SSO area of the ED. Observe that such patients slow the work of the nurses because they raise 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. Let X be the next activity of the patient p, then its score is defined as follows:

$$\Sigma_E = \mathbb{P}(O \vee P \vee Q). \tag{11.2}$$

Σ_X **– *Extra score.*** X-ray exams and *extra-ED* activities, that is activities that are not competence of the ED staff, such as X-ray exams, computerized tomography, echographies and specialist visits, can be executed by non-urgent patients only during a time frame during the day, that is from the early morning to the late evening. The idea is to promote the activities of these patients when the time frame is ending, in such a way that they can complete their path before the closing time of the hospital areas dedicated to such exams and specialist visits, which means to reduce significantly the EDLOSs. Conversely, if a patients needs one of these activity is in the ED during the night, then also the other activities can be made leisurely, without to lengthen its EDLOS. The score of the patient p is defined as follows:

$$\Sigma_X = \begin{cases} \max\{\mathbb{P}(H \vee I \vee J \vee K), 0.5\} & \text{if } 0 \leq f_{end} - t \leq \delta, \\ \min\{1 - \mathbb{P}(H \vee I \vee J \vee K), 0.5\} & \text{if } f_{start} - t > 0, \\ 0.5 & \text{otherwise,} \end{cases} \tag{11.3}$$

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, and 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.

Σ_O **– *Observation score.*** SSO units are a critical resource due to the limited amount of units; further, when they are all busy, patients have to be observed on stretchers in the corridors or visit rooms, slowing down the other activity of the ED. 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 start time, while the ending of the activity M depends on exogenous factors. Then, we can act on patients that have to execute the activity L, trying to start as soon as possible such an execution 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^-. In other words, the score is defined as follows:

$$\Sigma_O = \begin{cases} \mathbb{P}(L) & \text{if } u \geq u^+, \\ 1 - \mathbb{P}(L) & \text{if } u \leq u^-, \\ 0 & \text{otherwise.} \end{cases} \tag{11.4}$$

Σ_S – ***Stretcher score.*** We would to promote the release of stretchers by patients that need to execute an activity X belonging to the visit class and have an higher probability to be discharged or to occupy a SSO unit after that. Then, the score is defined as follows:

$$\Sigma_S = \begin{cases} \mathbb{P}(L \vee M \vee O \vee P \vee Q) & \text{if } p \in \mathcal{P}_{str}, \\ 0 & \text{otherwise,} \end{cases} \tag{11.5}$$

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

Σ_H – ***Hospitalization score.*** We deal with patients that do not occupy a stretcher and need a SSO unit before the release of a bed in a hospital ward to be hospitalized. Such patients could be treated leisurely, because they have a usually long EDLOS that does not depends on the rapidity of the ED activities, rather a fast pre-hospitalization in the SSO units could worsen the treatment of other patients. The score is defined as follows:

$$\Sigma_H = \begin{cases} 1 - \mathbb{P}(M < N) & \text{if } p \notin \mathcal{P}_{str}, \\ 1 & \text{otherwise.} \end{cases} \tag{11.6}$$

When a patient does not need to execute a visit as next activity ($\mathcal{A} \notin \{B, C, D, N\}$), the sequence of activities to be performed before the next visit is known and it is reasonable to base decisions on such a sequence. Let Λ be the set of the next activities, then the scores are trivially assigned as follows:

$$\Sigma_E = 0, \tag{11.7}$$

$$\Sigma_X = \begin{cases} 1 & \text{if } 0 \leq f_{end} - t \leq \delta \text{ and } \Lambda \cap \{H, I, J, K\} \neq \emptyset, \\ 0 & \text{if } f_{start} - t > 0 \text{ and } \Lambda \cap \{H, I, J, K\} \neq \emptyset, \\ 0.5 & \text{otherwise}, \end{cases} \tag{11.8}$$

$$\Sigma_O = \begin{cases} 1 & \text{if } u \geq u^+ \text{ and } L \in \Lambda \\ 0 & \text{otherwise}, \end{cases} \tag{11.9}$$

$$\Sigma_S = \begin{cases} 1 & \text{if } p \in \mathcal{P}_{str} \text{ and } \Lambda \cap \{L, M, O, P, Q\} \neq \emptyset, \\ 0 & \text{otherwise}, \end{cases} \tag{11.10}$$

$$\Sigma_H = \begin{cases} 0 & \text{if } p \notin \mathcal{P}_{str} \text{ and } M \notin \Lambda, \\ 1 & \text{otherwise}. \end{cases} \tag{11.11}$$

11.3 Quantitative Analysis

In this section we perform a quantitative analysis to study the impact of the online allocation approaches presented in Section 11.2.1 on several indices taking into account the perspectives of patients with different urgency. To this purpose we replicate exactly the patient accesses of the whole year 2016 recorded by the ED Sant'Antonio Abate of Cantù described in Chapter 10. Instead, the data-set of the period 2013–2015 is used for making predictions using the HAT, as shown in the same chapter. Therefore, the time horizon of the simulation is one year, of which the first 7 days are used as transitional period, while statistics are collected on the remaining 359 days.

The simulation model presented in this chapter has been implemented using AnyLogic 7.2 [73]. All the results reported in this section are the average values over 30 simulation runs starting from different initial conditions. The average time required for a single simulation run over the whole time horizon is 17.4 sec.

11.3.1 Validation

The activity durations are replicated from data when they are available (i.e. for first visits of urgent patients, specialist visits and SSOs), otherwise we use the average durations suggested by the ED staff. Furthermore, in modeling we introduced other time parameters that are not deductible from the data-set, as summarized in Table 11.1. For this reason we validate the model tuning such parameters in such a way to obtain the maximum fitness of the waiting times and the EDLOSs with respect to the real data, that is for the values reported in the second column of the table.

Tab. 11.1: 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 11.2.1)

The real case statistics and the results of the validation test are compared in Table 11.2. While waiting times of the simulation model are very close to the real case, EDLOSs 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. Furthermore, they are justified by a large lack of information and noises in the ED data-set, as discussed in Chapter 10. For these reasons, the purpose of our analysis is not to have a detailed estimation of the performance but to prove the sensitiveness of the model to online allocation approaches. Therefore, we can be satisfied by this level of fitness.

Tab. 11.2: Model validation.

	waiting times (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	-	18.6	82.1	79.9	66.7	14.5	8.0	6.3	3.7	6.2

11.3.2 Results

In this section, we test the online optimization approaches for the resource allocation proposed in Section 11.2 on the model validated with the parameter values reported in Table 11.1. The value of the normalization constants for the waiting times is fixed as follows:

$$m_1 = 1 \text{ min}, \; m_2 = 10 \text{ min}, \; m_3 = 120 \text{ min}, \; m_4 = m_5 = 180 \text{ min}. \tag{11.12}$$

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.

In order to make prediction in the online optimization, we use a data-set regarding the accesses in the period 2013–2015. A HAF over the patient clustering Γ' with $\ell = 30$ is generated a priori as explained in Chapter 10. During the simulation, such a HAF is used to estimate the probabilities defined in Section 11.2.2, in accordance with the patient cluster that is identified at the triage.

Waiting time and EDLOS are used as indices to analyze the performance using different scores of the proposed online approaches. We compare these approaches with a baseline configuration, that is the one used in the validation to replicate the resource allocation in the real case. We choose to analyze the waiting times also in 6 different 4-hours frames of the day because of the high variability of demand: frame $F = 1$ is 0:00–4:00, frame $F = 2$ is 4:00–8:00, etc. Furthermore, we observe what happen on a set of 46 days, that we call *overcrowding days*, which are the days that in the baseline configuration have a maximum length of the admission queue (i.e. patients waiting for the first visit) longer than 10 patients.

Results in Table 11.3 prove the effectiveness of the proposed optimization method using any score. Waiting times are significantly reduced for each urgency class, especially for the most numerous class $c = 3$ for which the 51% of the time is saved. The impact on the EDLOSs 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 stay of the other patients decreases from 55 to 145 min. On average, both waiting times and EDLOSs have a significant improvement, that is maximized using the *hospitalization score* Σ_H.

Tab. 11.3: Comparing online policies: waiting times and EDLOSs.

	waiting times (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)	-	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

In Table 11.4 we compare our optimization method with the baseline configuration focusing on waiting times 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 11.4, average waiting times over all days are reported. An even stronger impact of optimization in peak hours is shown: waiting times of patients with $c \leq 3$ are reduced up to the 71% using the *exit score* Σ_E. In the second part of Table 11.4 we focus only on the *overcrowding days* identified by the baseline configuration. 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.

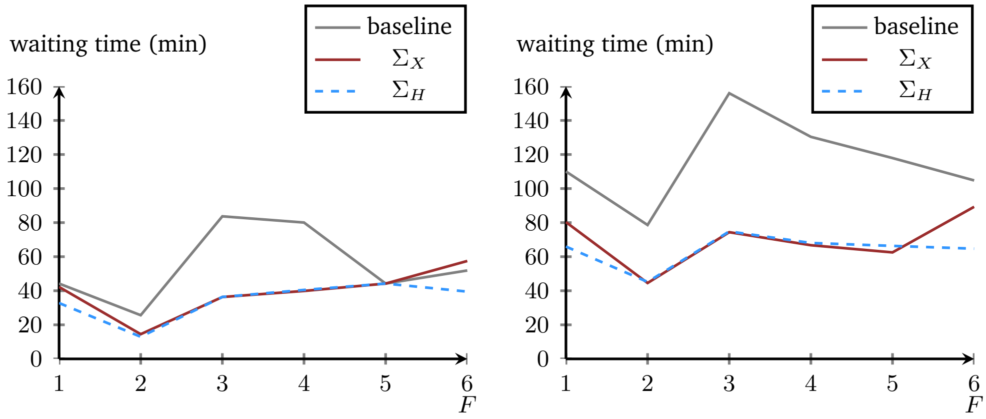
The variation of the average waiting time during the six frames of the day is illustrated in Figure 11.6. The impact of the online optimization is evident in almost all the frames, especially in the peak hours and in the overcrowding days. Scores Σ_X and Σ_H give similar waiting times. The former is slightly better in the central hours of the day, that is when the patients needing an extra-ED activity and have a

Tab. 11.4: Waiting times (min) in the frame from 8:00 to 12:00.

	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)	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

higher priority. A side-effect can be observed after the closing time of those activity, when the accumulated work to promote those patients causes an increasing of the waiting times and better performance can be obtained using Σ_H. In general, the online approach seems to deal with success peaks on demand, flattening the average waiting times over the day.

Fig. 11.6: Average waiting times of patients in the frames $F = 1, \ldots, 6$ of the day.



(a) All days

(b) Overcrowding days

In order to detect the type of impact of the proposed online approaches on the queues and on the resources, we introduce in Table 11.5 an additional set of performance indices. The impact of the different scores on such indices can demonstrate if the prediction performed using the HAF is effective. 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).

Tab. 11.5: Performance indices.

Param.	Definition
q	average number of patients in the pre-admission queue
r	average number of patients on treatment
u_{str}	stretcher utilization
γ	number of *overcrowding days*

In Table 11.6 we compare the different configuration on the indices introduced in Table 11.5, reporting the average values over all the year in the first part and a focus on the frame $F = 3$ in the more crowded day in the second part. As expected, the *exit score* Σ_E is able to reduce significantly the queue in the pre-admission waiting room, but a counter intuitive aspect can be observed in correspondence of the hours of overcrowding, where the *exit score* Σ_X and the *exit score* Σ_H have lower values of q. Such indices minimize 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 online optimization can reduce the days of overcrowding.

Tab. 11.6: Impact on queues and resource utilization.

	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)	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

In Figures 11.7 and 11.8 we report the trend of the number of patients in the waiting list queue and on treatment in the ED during a week, in which we have one overcrowding day (Thursday) for the baseline configuration. In this example, the *exit score* Σ_E is able to reduce the level of crowding in the ED.

Fig. 11.7: Length of the pre-admission waiting list in the ED during the week 1-7 April.

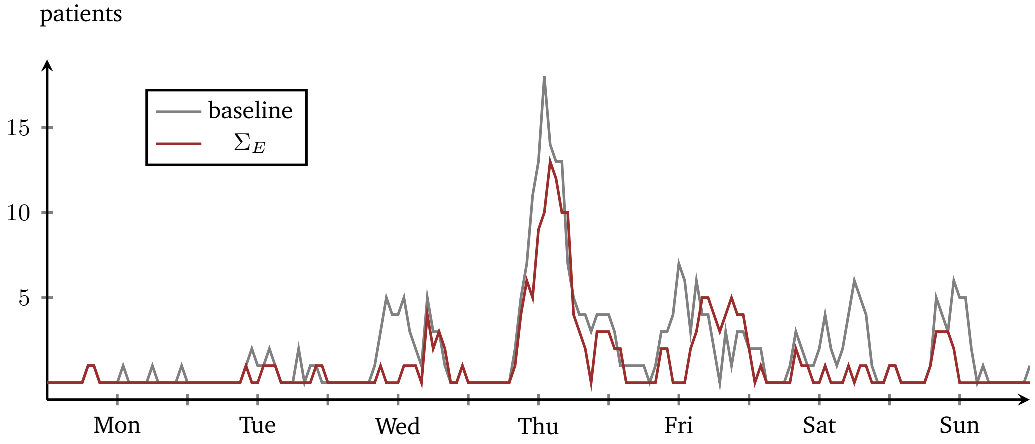
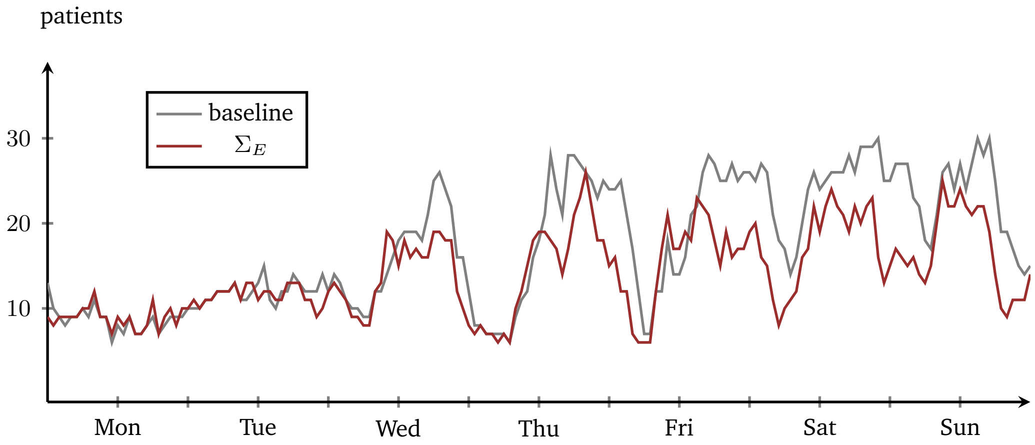


Fig. 11.8: Number of patients on treatment in the ED during the week 1-7 April.



11.4 Concluding remarks

In this chapter we provided a simulation model for the analysis of online approaches for the resource allocation of an ED. We proposed alternative solutions based on the state of the ED and the characteristics of patients, exploiting the prediction of the next activity given by the HAF.

Results prove the adequacy of online optimization for the resource allocation of the ED. Using simple policies that exploit prediction provided by the HAF, 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 scores have the impact on the queues or resources that we wanted to optimized, the effectiveness of predicting paths using the HAF is demonstrated. A further analysis could be performed with the aim to combine the scores in such a way to obtain additional improvements.

One of the main features of our approach is the ability of representing the path of a certain group of patients. Exploiting such a feature, the proposed model 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 malfunction of the primary care system can penalize the ECDS.

12 Conclusions

The use of Operations Research in health care delivery has developed considerably over the years. The current development of the health care delivery is aimed to recognize the central role of the patient as opposed to the one of the health care providers. In this context, the attention from a single health benefit can be shifted to the whole health care chain thanks to Clinical Pathways (CPs), which are specifically tailored to stimulate continuity and coordination among the treatments. Therefore, CPs are suitable for implementing simulation frameworks based on a patient-centered approach. The definition of CP allow us to lends it to the translation of the care process into a Discrete Event Simulation (DES), which is a flexible tool for quantitative analysis.

Many decision problems regarding the management of health services deal with unpredictable demand, events or variables, which make them more challenging. In this thesis, we addressed such an issue with online optimization, which is characterized by the development of algorithms whose decisions are based only on partial information that becomes available over time. Online optimization methodology takes into account the partial information obtained from the past and exploits the concept of lookahead, that is a limited overseen amount of future input data: such information can be derived by the knowledge of the CP or through predictive approaches.

The general framework proposed in this thesis for the analysis on online optimization approach is based on: (i) a DES model that replicates the considered CP including stochastic aspects, and (ii) a set of online optimization methods embedded in the DES to deal with unpredictability as information is made available over time.

The two parts of the thesis have been dedicated to different type of CPs, that is the Surgical Pathway (SP) and the Emergency Care Pathway (ECP). In Part I we dealt with problems arises in the context of Operating Room Planning (ORP), which are characterized by a well-structured but complex pathway, and by several sources of uncertainty such as the arrival of unattended patients to be operated on, and the activity durations. In Part II two different problems regarding the ECP has been addressed. The first is the Emergency Medical Service (EMS) management, which is characterized by a well-structured but simple pathway, and by several sources of uncertainty such as the arrival of unattended requests to be served as soon as possible by an ambulance. The second problem is Emergency Department (ED) management, which is characterized by a not-structured but complex pathway, and several sources of uncertainty such as the arrival of unattended patients to be served as soon as possible, and the path evolution.

The ORP and the EMS management are *lasagna processes*: the sequence of activities to be performed is known at the beginning of the CP and the possible path evolutions are limited. For this reason, online optimization approaches have been used exploit the solid knowledge of the CPs. On the contrary, the ED management is a *spaghetti process*: a large variety of path evolutions are possible and the sequence of activities to be performed is part itself of the lack of information taken into account by online optimization. To deal with this challenging aspect, in Chapter 10 we used an ad hoc process mining approach to extract information from historical data for predicting the possible path evolutions on the basis of the few information available, such as the past activities and the characteristics of the patient.

In Part I, we proposed an online optimization methodology for the Real Time Management (RTM) of operating rooms. Given an OR schedule, the RTM consists in a sort of centralized surveillance system whose main task is to supervise the execution of such a schedule and to take the more rational decision regarding elective and/or non-elective patients when unpredictable events occur. Quantitative analysis provided in Chapters 3–6 demonstrates the capability and the flexibility of the proposed framework to deal with different OR settings. Although online optimization does not exploit sophisticated mathematical approaches, the competitive analysis reported in Chapter 4 suggested its capability to deal with the stochastic aspects of a problem whenever such aspects are embedded into a well-structured optimization problem. Results indicated that the dynamic sharing of resources is the direction to follow for improving their utilization and the patient optimization. For instance, the computational experience suggested that significant improvement can be achieved making possible (i) to all the operating room sessions to draw from a shared overtime, and (ii) to non-elective patients to be inserted in operating rooms shared with elective patients.

In Part II, the effectiveness of the online optimization methodology has been proved for the ECP in both the EMS and the ED management. For the former, in Chapters 8 and 9 we shown its capability in other types of well-structured processes. Our simulation and optimization framework allows us to evaluate online policies for the real-time management of ambulances, of which few attempts can be found in literature because of the difficulties in developing a simulation model for this problem with such a general purpose. In particular, it is possible to have a significant lowering of the waiting times adopting the dynamic re-allocation of ambulances during the repositioning phase (Smart Assignment policy), which provided an overall improvement that is significantly greater than using and combining the standard dispatching and repositioning policies. For the latter, the support of prediction given by a process mining approach provided in Chapter 10 allowed us to relieve the lack of information in taking decisions in planning, as shown by the analysis in Chapter 11.

There are several common lessons learned through the computational experience performed to compare the several policies proposed for the different addressed problems. The same problems in different operative contexts usually obtain the best performance improvement using different method configurations, which is a further proof of the necessity of a decision support tool for the management problems in healthcare. Optimization approaches seem to be effective in well-structured processes or, alternatively, with the support of a predictive methods. Finally, the more flexible the operating context in which online optimization is applied (e.g. sharing resources among different patient classes) the greater the contribution of optimization in improving performance. Sometimes a large flexibility is not possible because of the lack of an adequate decision support tool, such as those proposed in this thesis, that manages the greater complexity that derives from a flexible context. Therefore, the proposed methodology could also provide insights to the health services managers for changing in their organization settings.

In conclusion, our framework in which DES is combined with online optimization represented a powerful tool for decision support as it allows to evaluate algorithm performance in very different contexts. All the analysis along the thesis suggested that online optimization can be a suitable methodology, highlighting the usefulness in using online optimization in the management of the majority health care services.

Further works could be developed on the direction of the lookahead concept. As seen in [35] predicting no-show behavior for the appointment scheduling, and in Chapters 10 and 11 predicting ED patient paths, predictive tools can be useful to compensate for the lack of information about future events empowering also the care pathway knowledge. For this reason, methodological approaches – as process mining – deserve to be further studied to reinforce the capability of online optimization to deal with health care delivery problems in the lookahead perspective. Furthermore, future research avenues could provide a comparison of the online optimization approach with the other alternative well-known methodologies taking into account uncertainty, such as stochastic programming and robust optimization.

Bibliography

[1] B. Addis, R. Aringhieri, E. Tànfani, and A. Testi. "Clinical pathways: Insights from a multidisciplinary literature survey". In: *Proceedings ORAHS 2012*. ISBN 978-90-365-3396-6. 2012 (cit. on p. 2).

[2] A. Ahmadi-Javid, P. Seyedi, and S.S. Syam. "A survey of healthcare facility location". In: *Computers and Operations Research* 79 (2017), pp. 223–263 (cit. on p. 5).

[3] A. Ahmadi-Javid, Z. Jalali, and K.J. Klassen. "Outpatient appointment systems in healthcare: A review of optimization studies". In: *European Journal of Operational Research* 258.1 (2017), pp. 3–34 (cit. on p. 5).

[4] R. Aringhieri, E. Tànfani, and A. Testi. "Operations research for health care delivery". In: *Computers and Operations Research* 40.9 (2013), pp. 2165–2166 (cit. on pp. 2, 5, 13).

[5] R. Aringhieri, V. Knight, and H. Smith. "ESI XXXI – OR applied to Health in a Modern World". In: *Operations Research for Health Care* 8 (2016), pp. 22–23 (cit. on p. 2).

[6] R. Aringhieri, V. Knight, and H. Smith. "ESI XXXI: OR applied to health in a modern world". In: *Health Systems* 5 (2016), 163–165 (cit. on p. 2).

[7] H. Balasubramanian, S. Biehl, L. Dai, and A. Muriel. "Dynamic allocation of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments". In: *Health Care Management Science* 17.1 (2014), pp. 31–48 (cit. on p. 5).

[8] S. Bayer. "Simulation modelling and resource allocation in complex services". In: *BMJ Quality and Safety* 23.5 (2014), pp. 353–355 (cit. on p. 5).

[9] A. Beaudry, G. Laporte, T. Melo, and S. Nickel. "Dynamic transportation of patients in hospitals". In: *OR Spectrum* 32.1 (2010), pp. 77–107 (cit. on p. 5).

[10] L. Becchetti, S. Leonardi, A. Marchetti-Spaccamela, G. Schäfer, and T. Vredeveld. "Average-case and smoothed competitive analysis of the multilevel feedback algorithm". In: *Mathematics of Operations Research* 31.1 (2006), pp. 85–108 (cit. on p. 5).

[11] S. Ben-David and A. Borodin. "A new measure for the study of on-line algorithms". In: *Algorithmica* 11.1 (1994), pp. 73–91 (cit. on p. 5).

[12] Bjorn P. Berg and Brian T. Denton. "Fast Approximation Methods for Online Scheduling of Outpatient Procedure Centers". In: *INFORMS Journal on Computing* 29.4 (2017), pp. 631–644 (cit. on p. 5).

[13] P. Bhattacharjee and P.K. Ray. “Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: A review and reflections”. In: *Computers and Industrial Engineering* 78 (2014), pp. 299–312 (cit. on p. 5).

[14] M. Bloin, S.O. Krumke, W. de Paepe, and L. Stougie. “The online-TSP against fair adversaries”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 1767 (2000), pp. 137–149 (cit. on p. 5).

[15] J. Boyar, K.S. Larsen, and M.N. Nielsen. “The accommodating function: A generalization of the competitive ratio”. In: *SIAM Journal on Computing* 31.1 (2001), pp. 233–258 (cit. on p. 5).

[16] J. Boyar, L.M. Favrholdt, K.S. Larsen, and M.N. Nielsen. “Extending the accommodating function”. In: *Acta Informatica* 40.1 (2003), pp. 3–35 (cit. on p. 5).

[17] B.D. Bradley, T. Jung, A. Tandon-Verma, et al. “Operations research in global health: A scoping review with a focus on the themes of health equity and impact”. In: *Health Research Policy and Systems* 15.1 (2017) (cit. on pp. 1, 5).

[18] S. Brailsford and J. Vissers. “OR in healthcare: A European perspective”. In: *European Journal of Operational Research* 212.2 (2011), pp. 223–234 (cit. on pp. 1, 2).

[19] H.H.R. Campbell, N. Bradshaw, and M. Porteous. “Integrated care pathways”. In: *British Medical Journal* 316.133-144 (1998) (cit. on p. 2).

[20] M.E. Captivo, I. Marques, and M. Moz. “ORAHS 2014 - for better practices in health care management”. In: *Operations Research for Health Care* 7 (2015), pp. 1–2 (cit. on p. 2).

[21] T. Cayirli, M.M. Gunal, E. Gunes, and L. Ormeci. “The 39th international conference of the EURO working group on operational research applied to health services: ORAHS 2013 special issue”. In: *Health Care Management Science* 18.3 (2015), pp. 219–221 (cit. on p. 2).

[22] S. Cevik Onar, B. Oztaysi, and C. Kahraman. “A Comprehensive survey on healthcare management”. In: *International Series in Operations Research and Management Science* 262 (2018), pp. 23–51 (cit. on p. 5).

[23] L. De Bleser, R. Depreitere, K. De Waele, et al. “Defining pathways”. In: *Journal of Nursing Management* 14.553-563 (2006) (cit. on p. 2).

[24] F. Dexter, E. Marcon, and X. Xie. “Operational research applied to health services 2007 special issue”. In: *Health Care Management Science* 12.2 (2009), pp. 117–118 (cit. on p. 2).

[25] R. Dorrigiv and A. López-Ortiz. “Closing the gap between theory and practice: New measures for on-line algorithm analysis”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4921 LNCS (2008), pp. 13–24 (cit. on p. 5).

[26] F. Dunke and S. Nickel. “A general modeling approach to online optimization with lookahead”. In: *Omega (United Kingdom)* 63 (2016), pp. 134–153 (cit. on pp. 5, 7, 18).

[27] F. Dunke and S. Nickel. “Evaluating the quality of online optimization algorithms by discrete event simulation”. In: *Central European Journal of Operations Research* 25.4 (2017), pp. 831–858 (cit. on p. 3).

[28] M. Fakhimi and J. Probert. “Operations research within UK healthcare: A review”. In: *Journal of Enterprise Information Management* 26.1 (2013), pp. 21–49 (cit. on p. 5).

[29] A. Fiat and G.J. Woeginger, eds. *Online Algorithms. The State of the Art*. Lecture Notes in Computer Science. Springer, 1998 (cit. on p. 3).

[30] C. Fiegl and C. Pontow. “Online scheduling of pick-up and delivery tasks in hospitals”. In: *Journal of Biomedical Informatics* 42.4 (2009), pp. 624–632 (cit. on p. 5).

[31] C. Fikar and P. Hirsch. “Home health care routing and scheduling: A review”. In: *Computers and Operations Research* 77 (2017), pp. 86–95 (cit. on p. 5).

[32] M. Grotschel, S. Krumke, and J. Rambau. *Online optimization of large scale systems*. Ed. by Springer. Springer, 2001 (cit. on p. 5).

[33] S. Hahn-Goldberg, M.W. Carter, J.C. Beck, et al. “Dynamic optimization of chemotherapy outpatient scheduling with uncertainty”. In: *Health Care Management Science* 17.4 (2014), pp. 379–392 (cit. on p. 5).

[34] E.W. Hans and I.M.H. Vliegen. “Special Issue of the 2012 conference of the EURO working group Operational Research Applied To Health Services (ORAHS)”. In: *Operations Research for Health Care* 3.2 (2014), p. 47 (cit. on p. 2).

[35] S.L. Harris, J.H. May, and L.G. Vargas. “Predictive analytics model for healthcare planning and scheduling”. In: *European Journal of Operational Research* 253.1 (2016), pp. 121–131 (cit. on p. 155).

[36] E. Koutsoupias and C.H. Papadimitriou. “Beyond competitive analysis”. In: *SIAM Journal on Computing* 30.1 (2000), pp. 300–317 (cit. on p. 5).

[37] M. Lagergren. “What is the role and contribution of models to management and research in the health services? A view from Europe”. In: *European Journal of Operational Research* 105.2 (1998), pp. 257–266 (cit. on p. 5).

[38] C. Laine and F. Davidoff. “Patient-Centered Medicine. A Professional Evolution”. In: *Journal of the American Medical Association* 275.2 (1996), pp. 152–156 (cit. on p. 1).

[39] A. Legrain, M.-A. Fortin, N. Lahrichi, and L.-M. Rousseau. “Online stochastic optimization of radiotherapy patient scheduling”. In: *Health Care Management Science* 18.2 (2015), pp. 110–123 (cit. on p. 5).

[40] J. Lin, K. Muthuraman, and M. Lawley. “Optimal and approximate algorithms for sequential clinical scheduling with no-shows”. In: *IIE Transactions on Healthcare Systems Engineering* 1.1 (2011), pp. 20–36 (cit. on p. 5).

[41] F. Mallor, S. Brailsford, M. Rauner, and C. Azcarate. “Operational research applied to health services: Finding better health-care decisions in new oceans of health data”. In: *Operations Research for Health Care* 17 (2018), pp. 1–3 (cit. on p. 2).

[42] William P. Millhiser and Emre A. Veral. “A decision support system for real-time scheduling of multiple patient classes in outpatient services”. In: *Health Care Management Science* (2018) (cit. on p. 5).

[43] International Alliance of Patients’Organizations. *What is Patient-Centred Healthcare? A Review of Definitions and Principles*. 2004 (cit. on p. 1).

[44] M. Panella, S. Marchisio, and F. Stanislao. “Reducing Clinical Variations with Clinical Pathways: Do Pathways Work?” In: *International Journal for Quality in Health Care* 15 (2003), pp. 509–521 (cit. on p. 2).

[45] E. Pérez, L. Ntaimo, C.O. Malavé, C. Bailey, and P. McCormack. “Stochastic online appointment scheduling of multi-step sequential procedures in nuclear medicine”. In: *Health Care Management Science* 16.4 (2013), pp. 281–299 (cit. on p. 5).

[46] A. Rais and A. Viana. “Operations research in healthcare: A survey”. In: *International Transactions in Operational Research* 18.1 (2011), pp. 1–31 (cit. on pp. 5, 18).

[47] M.S. Rauner and J.M.H. Vissers. “OR applied to health services: Planning for the future with scarce resources”. In: *European Journal of Operational Research* 150.1 (2003), pp. 1–2 (cit. on p. 2).

[48] T. Rotter, L. Kinsman, E. James, et al. “Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs (review)”. In: *The Cochrane Library* 7 (2010) (cit. on p. 2).

[49] M. Samorani and L.R. Laganga. “Outpatient appointment scheduling given individual day-dependent no-show predictions”. In: *European Journal of Operational Research* 240.1 (2015), pp. 245–257 (cit. on p. 5).

[50] Steven S. Seiden. “On the Online Bin Packing Problem”. In: *J. ACM* 49.5 (Sept. 2002), pp. 640–671 (cit. on p. 3).

[51] D.D. Sleator and R.E. Tarjan. “Amortized Efficiency of List Update and Paging Rules”. In: *Commun. ACM* 28.2 (1985), pp. 202–208 (cit. on p. 4).

[52] P.-F.J. Tsai and G.-Y. Teng. “A stochastic appointment scheduling system on multiple resources with dynamic call-in sequence and patient no-shows for an outpatient clinic”. In: *European Journal of Operational Research* 239.2 (2014), pp. 427–436 (cit. on p. 5).

[53] B. Vieira, E.W. Hans, C. Van Vliet-Vroegindeweij, J. Van De Kamer, and W. Van Harten. “Operations research for resource planning and -use in radiotherapy: A literature review”. In: *BMC Medical Informatics and Decision Making* 16.1 (2016) (cit. on p. 5).

[54] J. Volland, A. Fügener, J. Schoenfelder, and J.O. Brunner. “Material logistics in hospitals: A literature review”. In: *Omega (United Kingdom)* 69 (2017), pp. 82–101 (cit. on p. 5).

[55] J. Wang and R.Y.K. Fung. “Adaptive dynamic programming algorithms for sequential appointment scheduling with patient preferences”. In: *Artificial Intelligence in Medicine* 63.1 (2015), pp. 33–40 (cit. on p. 5).

[56] W.-Y. Wang and D. Gupta. “Adaptive appointment systems with patient preferences”. In: *Manufacturing and Service Operations Management* 13.3 (2011), pp. 373–389 (cit. on p. 5).

[57] X. Xie, S. Gallivan, A. Guinet, and M. Rauner. “Operational research applied to health services: A special volume dedicated to the international conference ORAHS’2007”. In: *Annals of Operations Research* 178.1 (2010), pp. 1–4 (cit. on p. 2).

[58] C. Zacharias and M. Pinedo. “Appointment scheduling with no-shows and overbooking”. In: *Production and Operations Management* 23.5 (2014), pp. 788–801 (cit. on p. 5).

[59] B. Zeng, A. Turkcan, J. Lin, and M. Lawley. “Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities”. In: *Annals of Operations Research* 178.1 (2010), pp. 121–144 (cit. on p. 5).

Part I

[60] I. Adan, J. Bekkers, N. Dellaert, J. Jeunet, and J. Visserd. “Improving operational effectiveness of tactical master plans for emergency and elective patients under stochastic demand and capacitated resources”. In: *European Journal of Operational Research* 213.1 (2011), pp. 290–308 (cit. on pp. 12, 54).

[61] A. Agnetis, A. Coppi, M. Corsini, et al. “Long term evaluation of operating theater planning policies”. In: *Operations Research for Health Care* 1.4 (2012), pp. 95–104 (cit. on p. 73).

[62] R. Aringhieri and D. Duma. “A hybrid model for the analysis of a surgical pathway”. In: *Proceedings of the 4th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (HA-2014)*. 2014, pp. 889–900 (cit. on p. 12).

[63] R. Aringhieri and D. Duma. “The optimization of a surgical clinical pathway”. In: *Simulation and Modeling Methodologies, Technologies and Applications*. Ed. by M.S. Obaidat et al. Vol. 402. Advances in Intelligent Systems and Computing. Springer, 2016, pp. 313–331 (cit. on p. 12).

[64] R. Aringhieri and D. Duma. “Patient–Centred Objectives as an Alternative to Maximum Utilisation: Comparing Surgical Case”. In: *Optimization and Decision Science: Methodologies and Applications (ODS 2017)*. Vol. 217. Springer Proceedings in Mathematics & Statistics. 2017, pp. 105–112 (cit. on pp. 11, 32).

[65] R. Aringhieri, P. Landa, P. Soriano, E. Tànfani, and A. Testi. “A two level Metaheuristic for the Operating Room Scheduling and Assignment Problem”. In: *Computers & Operations Research* 54 (2015), pp. 21–34 (cit. on pp. 11, 19, 74).

[66] R. Aringhieri, P. Landa, and E. Tanfani. “Assigning surgery cases to operating rooms: A VNS approach for leveling ward beds occupancies”. In: *The 3rd International Conference on Variable Neighborhood Search (VNS’14)*. Electronic Notes in Discrete Mathematics. 2015, pp. 173–180 (cit. on pp. 11, 32).

[67] R. Aringhieri, P. Landa, and S. Mancini. “A Hierarchical Multi-objective Optimisation Model for Bed Levelling and Patient Priority Maximisation”. In: *Optimization and Decision Science: Methodologies and Applications (ODS 2017)*. Vol. 217. Springer Proceedings in Mathematics & Statistics. 2017, pp. 113–120 (cit. on pp. 12, 32).

[68] C. Banditori, P. Cappanera, and F. Visintin. “A combined optimization–simulation approach to the master surgical scheduling problem”. In: *Journal of Management Mathematics* 24 (2013), pp. 155–187 (cit. on p. 73).

[69] Isabelle Beaulieu, Michel Gendreau, and Patrick Soriano. “Operating rooms scheduling under uncertainty”. In: *Advanced Decision Making Methods Applied to Health Care*. Ed. by Elena Tànfani and Angela Testi. Vol. 173. International Series in Operations Research & Management Science. Springer Milan, 2012, pp. 13–32 (cit. on pp. 11, 33).

[70] J. Beliën and E. Demeulemeester. “Building cyclic master surgery schedules with leveled resulting bed occupancy”. In: *European Journal of Operational Research* 176.2 (2007), pp. 1185–1204 (cit. on p. 73).

[71] J. Beliën, E. Demeulemeester, and B. Cardoen. “Building cyclic master surgery schedules with levelled resulting bed occupancy: A case study”. In: *European Journal of Operational Research* 176 (2007), pp. 1185–1204 (cit. on p. 31).

[72] Z. Bing-hai, Y. Meng, and L. Zhi-qiang. “An improved Lagrangian relaxation heuristic for the scheduling problem of operating theatres”. In: *Computers & Industrial Engineering* 101 (2016), pp. 490–503 (cit. on p. 11).

[73] A. Borshchev. *The Big Book of Simulation Modeling. Multimethod Modeling with AnyLogic*. ISBN 978-0-9895731-7-7. 2013 (cit. on pp. 19, 95, 109, 146).

[74] M.E. Bruni, P. Beraldi, and D. Conforti. “A stochastic programming approach for operating theatre scheduling under uncertainty”. In: *IMA Journal of Management Mathematics* 26.1 (2015), pp. 99–119 (cit. on p. 11).

[75] P. Cappanera, F. Visintin, and C. Banditori. “Comparing resource balancing criteria in master surgical scheduling: A combined optimisation-simulation approach”. In: *International Journal of Production Economics* 158 (2014), pp. 179–196 (cit. on pp. 11, 12).

[76] P. Cappanera, F. Visintin, and C. Banditori. “Addressing conflicting stakeholders’ priorities in surgical scheduling by goal programming”. In: *Flexible Services and Manufacturing Journal* (2016), pp. 1–20 (cit. on pp. 11, 12).

[77] B. Cardoen, E. Demeulemeester, and J. Beliën. “Sequencing surgical cases in a day-care environment: An exact branch-and-price approach”. In: *Computers & Operations Research* 36.9 (2009), pp. 2660–2669 (cit. on p. 11).

[78] B. Cardoen, E. Demeulemeester, and J. Beliën. “Operating room planning and scheduling: A literature review”. In: *European Journal of Operational Research* 201 (2010), pp. 921–932 (cit. on pp. 12, 25, 31).

[79] N. Dellaert and J. Jeunet. “A variable neighborhood search algorithm for the surgery tactical planning problem”. In: *Computers & Operations Research* (2017), pp. 1–10 (cit. on p. 11).

[80] F. Dexter. “A strategy to decide whether to move the last case of the day in an operating room to another empty operating room to decrease overtime labor costs”. In: *Anesthesia and Analgesia* 91.4 (2000), pp. 925–928 (cit. on p. 13).

[81] F. Dexter and J. Ledolter. “Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data”. In: *Anesthesiology* 103.6 (2005), pp. 1259–1267 (cit. on p. 35).

[82] F. Dexter, T. C. Smith, D. J. Tatman, and A. Macario. “Physicians’ perceptions of minimum time that should be saved to move a surgical case from one operating room to another: Internet-based survey of the membership of the Association of Anesthesia Clinical Directors (AACD)”. In: *Journal of clinical anesthesia* 15.3 (2003), pp. 206–210 (cit. on p. 13).

[83] F. Dexter, A. Willemsen-Dunlap, and J.D. Lee. “Operating room managerial decision-making on the day of surgery with and without computer recommendations and status displays”. In: *Anesthesia and Analgesia* 105.2 (2007), pp. 419–429 (cit. on p. 13).

[84] F. Dexter, R.H. Epstein, E.O. Bayman, and J. Ledolter. “Estimating surgical case durations and making comparisons among facilities: Identifying facilities with lower anesthesia professional fees”. In: *Anesthesia and Analgesia* 116.5 (2013), pp. 1103–1115 (cit. on p. 35).

[85] F. Dexter, J. Ledolter, V. Tiwari, and R.H. Epstein. “Value of a scheduled duration quantified in terms of equivalent numbers of historical cases”. In: *Anesthesia and Analgesia* 117.1 (2013). cited By 6, pp. 205–210 (cit. on p. 35).

[86] D. Duma and R. Aringhieri. “An online optimization approach for the Real Time Management of operating rooms”. In: *Operations Research for Health Care* 7 (2015), pp. 40–51 (cit. on p. 12).

[87] D. Duma and R. Aringhieri. “The management of non-elective patients: shared vs. dedicated policies”. In: *Omega* In Press (2018) (cit. on p. 12).

[88] D. Duma and R. Aringhieri. “The Real Time Management of Operating Rooms”. In: *Operations Research Applications in Health Care Management*. Ed. by C. Kahraman and I. Topcu. Vol. 262. International Series in Operations Research & Management Science. Springer International Publishing AG, 2018, pp. 55–79 (cit. on p. 12).

[89] E. Erdem, X. Qu, J. Shi, and S. Upadhyaya. “A mathematical modeling approach for rescheduling elective admissions upon arrival of emergency patients”. In: *61st Annual IIE Conference and Expo Proceedings*. 2011 (cit. on p. 13).

[90] E. Erdem, X. Qu, and J. Shi. “Rescheduling of elective patients upon the arrival of emergency patients”. In: *Decision Support Systems* 54.1 (2012), pp. 551–563 (cit. on pp. 13, 39).

[91] J.T. van Essen, E.W. Hans, J.L. Hurink, and A. Oversberg. “Minimizing the waiting time for emergency surgery”. In: *Operations Research for Health Care* 1.2–3 (2012), pp. 34 –44 (cit. on pp. 54, 56).

[92] H. Fei, C. Chu, N. Meskens, and A. Artiba. “Solving surgical cases assignment problem by a branch-and-price approach”. In: *International Journal of Production Economics* 112 (2008), pp. 96–108 (cit. on p. 11).

[93] Y. Ferrand, M. Magazine, and U. Rao. “Comparing two operating-room-allocation policies for elective and emergency surgeries”. In: *Proceedings of the 2010 Winter Simulation Conference*. 2010, pp. 2364–2374 (cit. on p. 12).

[94] Y.B. Ferrand, M.J. Magazine, and U.S. Rao. “Partially Flexible Operating Rooms for Elective and Emergency Surgeries”. In: *Decision Sciences* 45.5 (2014), pp. 819–847 (cit. on p. 12).

[95] A. Ghoniem and H.D. Sherali. “Defeating symmetry in combinatorial optimization via objective perturbations and hierarchical constraints”. In: *IIE Transactions (Institute of Industrial Engineers)* 43.8 (2011), pp. 575–588 (cit. on p. 46).

[96] F. Guerriero and R. Guido. “Operational research in the management of the operating theatre: a survey”. In: *Health Care Management Science* 14 (2011), pp. 89–114 (cit. on p. 12).

[97] E. Hans, G. Wullink, M. van Houdenhoven, and G. Kamezier. “Robust surgery loading”. In: *European Journal of Operational Research* 185 (2008), pp. 1038–1050 (cit. on p. 11).

[98] Erwin W. Hans and Peter T. Vanberkel. “Operating Theatre Planning and Scheduling”. English. In: *Handbook of Healthcare System Scheduling*. Ed. by Randolph Hall. Vol. 168. International Series in Operations Research & Management Science. Springer US, 2012, pp. 105–130 (cit. on p. 13).

[99] M. Heng and J. G. Wright. “Dedicated operating room for emergency surgery improves access and efficiency”. In: *Canadian Journal of Surgery* 56.3 (2013), pp. 167–174 (cit. on p. 12).

[100] W.L. Herring and J.W. Herrmann. “A stochastic dynamic program for the single-day surgery scheduling problem”. In: *IIE Transactions on Healthcare Systems Engineering* 4 (2011), pp. 213–225 (cit. on p. 17).

[101] W.L. Herring and J.W. Herrmann. “The single-day surgery scheduling problem: sequential decision-making and threshold-based heuristics”. In: *OR Spectrum* 34 (2012), pp. 429–459 (cit. on p. 11).

[102] M. Lamiri, X. Xie, A. Dolgui, and F. Grimaud. “A stochastic model for operating room planning with elective and emergency demand for surgery”. In: *Journal of Operational Research Society* 185 (2008), pp. 1026–1037 (cit. on p. 12).

[103] M. Lamiri, X. Xie, A. Dolgui, and F Grimaud. “Optimization methods for a stochastic surgery planning problem”. In: *International Journal of Production Economics* 120 (2009), pp. 400–410 (cit. on p. 17).

[104] P. Landa, R. Aringhieri, P. Soriano, E. Tànfani, and A. Testi. “A hybrid optimization algorithm for surgeries scheduling”. In: *Operations Research for Health Care* 8 (2016), pp. 103–114 (cit. on p. 11).

[105] G. Latorre-Nunez, A. Luer-Villagra, V. Marianov, et al. “Scheduling operating rooms with consideration of all resources, post anesthesia beds and emergency surgeries”. In: *Computers & Industrial Engineering* 97 (2016), pp. 248 –257 (cit. on p. 11).

[106] Xiangyong Li, N. Rafaliya, M. Fazle Baki, and Ben A. Chaouch. “Scheduling elective surgeries: the tradeoff among bed capacity, waiting patients and operating room utilization using goal programming”. In: *Health Care Management Science* 20.1 (2017), pp. 33–54 (cit. on p. 11).

[107] J.M. Magerlein and J.B. Martin. “Surgical demand scheduling: A review”. In: *Health Services Research* 13 (1978), pp. 418–433 (cit. on pp. 11, 13).

[108] I. Marques and M.E. Captivo. “Different stakeholders’ perspectives for a surgical case assignment problem: Deterministic and robust approaches”. In: *European Journal of Operational Research* 261.1 (2017), pp. 260–278 (cit. on pp. 11, 12).

[109] I. Marques, M.E. Captivo, and M.V. Pato. “An integer programming approach to elective surgery scheduling”. In: *OR Spectrum* 34 (2012), pp. 407–427 (cit. on p. 11).

[110] S. Martello and P. Toth. *Knapsack Problems: Algorithms and Computer Implementations*. Wiley-Intersci. Ser. Discrete Math. Optim., John Wiley and Sons, 1990 (cit. on p. 22).

[111] J.H. May, D.P. Strum, and L.G. Vargas. “Fitting the lognormal distribution to surgical procedure times”. In: *Decision Sciences* 31.1 (2000), pp. 129–148 (cit. on p. 34).

[112] N. Meskens, D. Duvivier, and A. Hanset. “Multi-objective operating room scheduling considering desiderata of the surgical teams”. In: *Decision Support Systems* 55 (2013), pp. 650–659 (cit. on p. 11).

[113] R. M’Hallah and A.H. Al-Roomi. “The planning and scheduling of operating rooms: A simulation approach”. In: *Computers and Industrial Engineering* 78 (2014), pp. 235–248 (cit. on p. 13).

[114] Y.A. Ozcan, E. Tànfani, and A. Testi. “A Simulation-based modeling framework to deal with Clinical Pathways”. In: *Proceedings of the 2011 Winter Simulation Conference*. Ed. by S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu. 2011, pp. 1190–1201 (cit. on pp. 17, 26, 33–35).

[115] D.N. Pham and A. Klinkert. “Surgical case scheduling as a generalized job shop scheduling problem”. In: *European Journal of Operational Research* 185 (2008), pp. 1011–1025 (cit. on pp. 11, 13).

[116] A. Riise and E.K. Burke. “Local search for the surgery admission planning problem.” In: *Journal of Heuristics* 17.4 (2011), pp. 389–414 (cit. on p. 11).

[117] B. Roland, C.D. Martinelly, F. Riane, and Y. Pochet. “Scheduling an operating theatre under human resource constraints”. In: *Computers & Industrial Engineering* 58 (2010), pp. 212–220 (cit. on p. 11).

[118] M. Samudra, C. Van Riet, E. Demeulemeester, et al. “Scheduling operating rooms: achievements, challenges and pitfalls”. In: *Journal of Scheduling* 19.5 (2016), pp. 493–525 (cit. on p. 12).

[119] C. L. Siqueira, E.F. Arruda, L. Bahiense, G. L. Bahr, and G. R. Motta. “A stochastic model for operating room planning with elective and emergency demand for surgery”. In: *European Journal of Operational Research* (2016), pp. 1–14 (cit. on p. 12).

[120] W.E. Spangler, D.P. Strum, L.G. Vargas, and H.M. Jerrold. “Estimating Procedure Times for Surgeries by Determining Location Parameters for the Lognormal Model”. In: *Health Care Management Science* 7 (2004), pp. 97–104 (cit. on p. 34).

[121] P.S. Stepaniak, C. Heij, G.H. Mannaerts, M. de Quelerij, and G. de Vries. “Modeling procedure and surgical times for current procedural terminology-anesthesia-surgeon combinations and evaluation in terms of case-duration prediction and operating room efficiency: a multicenter study”. In: *Anesthesia and Analgesia* 109.4 (2009), pp. 1232–1245 (cit. on p. 59).

[122] P.S. Stepaniak, C. Heij, and G. De Vries. “Modeling and prediction of surgical procedure times”. In: *Statistica Neerlandica* 64.1 (2010), pp. 1–18 (cit. on p. 59).

[123] D.P. Strum, J.H. May, and L.G. Vargas. “Modeling the Uncertainty of Surgical Procedure Times: Comparison of lognormal and normal models”. In: *Anesthesiology* 92.4 (2000), pp. 1160–1167 (cit. on p. 34).

[124] K. Stuart and E. Kozan. “Reactive scheduling model for the operating theatre”. In: *Flexible Services and Manufacturing Journal* 24.4 (2012), pp. 400–421 (cit. on p. 13).

[125] E. Tànfani and A. Testi. “A pre-assignment heuristic algorithm for the Master Surgical Schedule Problem (MSSP)”. In: *Annals of Operations Research* 178.1 (2010), pp. 105–119 (cit. on pp. 16, 25).

[126] E. Tànfani, A. Testi, and R. Alvarez. “Operating Room Planning considering stochastic surgery durations”. In: *International Journal of Health Management and Information* 1.2 (2010), pp. 167–183 (cit. on p. 17).

[127] A. Testi, E. Tànfani, and G.C. Torre. “A three-phase approach for operating theatre schedules”. In: *Health Care Management Science* 10 (2007), pp. 163–172 (cit. on p. 11).

[128] V. Tiwari, F. Dexter, B.S. Rothman, J.M. Ehrenfeld, and R.H. Epstein. “Explanation for the near constant mean time remaining in surgical cases exceeding their estimated duration, necessary for appropriate display on electronic white boards”. In: *Anesthesia and Analgesia* 117 (2013), pp. 487–493 (cit. on p. 13).

[129] J.M. Van Oostrum, M. Van Houdenhoven, J.L. Hurink, et al. “A master surgical scheduling approach for cyclic scheduling in operating room departments”. In: *OR Spectrum* 30.2 (2008), pp. 355–374 (cit. on p. 73).

[130] C. Van Riet and E. Demeulemeester. “Trade-offs in operating room planning for electives and emergencies: A review”. In: *Operations Research for Health Care* 7 (2015), pp. 52–69 (cit. on pp. 12, 37, 53, 54).

[131] R.E. Wachtel and F. Dexter. “Reducing tardiness from scheduled start times by making adjustments to the operating room schedule”. In: *Anesthesia and Analgesia* 108.6 (2009), pp. 1902–1909 (cit. on p. 13).

[132] B. Wang, X. Han, X. Zhang, and S. Zhang. “Predictive-reactive scheduling for single surgical suite subject to random emergency surgery”. In: *Journal of Combinatorial Optimization* 30.4 (2015), pp. 949–966 (cit. on p. 54).

[133] G. Wullink, M. Van Houdenhoven, E.W. W. Hans, et al. “Closing emergency operating rooms improves efficiency”. In: *Journal of medical systems* 31.6 (2007), pp. 543–546 (cit. on p. 12).

[134] B. Yang and J. Geunes. “Predictive-reactive scheduling on a single resource with uncertain future jobs”. In: *European Journal of Operational Research* 189.3 (2008), pp. 1267–1283 (cit. on p. 54).

[135] M.E. Zonderland, R.J. Boucherie, N. Litvak, and C.L.A.M. Vleggeert-Lankamp. “Planning and scheduling of semi-urgent surgeries”. In: *Health Care Management Science* 13.3 (2010), pp. 256–267 (cit. on p. 54).

Part II

[136] Waleed Abo-Hamad. “Patient Pathways Discovery and Analysis Using Process Mining Techniques: An Emergency Department Case Study”. In: *Health Care Systems Engineering*. Ed. by Paola Cappanera, Jingshan Li, Andrea Matta, et al. Springer International Publishing, 2017, pp. 209–219 (cit. on pp. 86, 124).

[137] L. Aboueljinane, E. Sahin, and Z. Jemai. “A review on simulation models applied to emergency medical service operations”. In: *Computers & Industrial Engineering* 66.4 (2013), pp. 734–750 (cit. on p. 83).

[138] L. Aboueljinane, E. Sahin, and Z. Jemai. “A review on simulation models applied to emergency medical service operations”. In: *Computers & Industrial Engineering* 66 (2013), pp. 734–750 (cit. on p. 86).

[139] T. Andersson and P. Varbrand. “Decision support tools for ambulance dispatch and relocation”. In: *Journal of the Operational Research Society* 58.2 (2006), pp. 195–201 (cit. on p. 83).

[140] R. Aringhieri, G. Carello, and D. Morale. “Supporting decision making to improve the performance of an Italian emergency medical service”. In: *Annals of Operations Research* 236 (2016), pp. 131–148 (cit. on pp. 90, 103, 106, 110, 112).

[141] R. Aringhieri, M.E. Bruni, S. Khodaparasti, and J.T. van Essen. “Emergency Medical Services and beyond: Addressing new challenges through a wide literature review”. In: *Computers and Operations Research* 78 (2017), pp. 349–368 (cit. on pp. 82, 83, 96).

[142] R. Aringhieri, D. Dell’Anna, D. Duma, and M. Sonnessa. “Evaluating the dispatching policies for a regional network of emergency departments exploiting health care big data”. In: *International Conference on Machine Learning, Optimization, and Big Data*. Ed. by G. Nicosia, P. Pardalos, G. Giuffrida, and R. Umeton. Vol. 10710. Lecture Notes in Computer Science. Springer International Publishing, 2018, pp. 549–561 (cit. on p. 104).

[143] D. Bandara, M.E. Mayorga, and L.A. McLay. “Priority dispatching strategies for EMS systems”. In: *Journal of the Operational Research Society* 65.4 (2014), pp. 572–587 (cit. on pp. 83, 84, 99, 105, 112).

[144] T. van Barneveld, C. Jagtenberg, S. Bhulai, and R. van der Mei. “Real-time ambulance relocation: Assessing real-time redeployment strategies for ambulance relocation”. In: *Socio-Economic Planning Sciences* 62 (2018), pp. 129–142 (cit. on pp. 85, 107).

[145] V. Bélanger, A. Ruiz, and P. Soriano. “Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles”. In: *European Journal of Operational Research* (2018) (cit. on p. 83).

[146] S.L. Bernstein, V. Verghese, W. Leung, A.T. Lunney, and I. Perez. “Development and validation of a new index to measure emergency department crowding”. In: *Academic Emergency Medicine* 10.9 (2003), pp. 938–942 (cit. on p. 82).

[147] S.C. Brailsford, V.A. Lattimer, P. Tarnaras, and J.C. Turnbull. “Emergency and on-demand health care: Modelling a large complex system”. In: *Journal of the Operational Research Society* 55.1 (2004), pp. 34–42 (cit. on p. 88).

[148] J.C.A.M. Buijs, B.F. van Dongen, and van der Aalst W.M.P. “Quality dimensions in process discovery: the importance of fitness, precision, generalization and simplicity”. In: *International Journal of Cooperative Information Systems* 23.1 (2014), pp. 1440001/1–39 (cit. on p. 122).

[149] M. van Buuren, R. van der Mei, K. Aardal, and H. Post. “Evaluating dynamic dispatch strategies for emergency medical services: TIFAR simulation tool”. In: *Proceedings of the 2012 Winter Simulation Conference*. Berlin, 2012 (cit. on p. 85).

[150] Emilie J. B. Calvello, Morgan Broccoli, Nicholas Risko, et al. “Emergency Care and Health Systems: Consensus-based Recommendations and Future Research Priorities”. In: *Academic Emergency Medicine* 20.12 (2013), pp. 1278–1288 (cit. on p. 81).

[151] G.M. Carter, J.M. Chaiken, and E. Ignall. “Response Areas for Two Emergency Units”. In: *Operations Research* 20.3 (1972), pp. 571–594 (cit. on p. 83).

[152] N. Channouf, P. L’Ecuyer, A. Ingolfsson, and A.N. Avramidis. “The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta”. In: *Health Care Management Science* 10.1 (2007), pp. 25–45 (cit. on p. 102).

[153] Marta Cildoz, Fermin Mallor, Amaia Ibarra, and Cristina Azcarate. “Dealing with Stress and Workload in Emergency Departments”. In: *Health Care Systems Engineering*. Ed. by Paola Cappanera, Jingshan Li, Andrea Matta, et al. Cham: Springer International Publishing, 2017, pp. 297–298 (cit. on p. 81).

[154] R.A. Cuninghame-Greene and G. Harries. “Nearest-neighbour rules for emergency services”. In: *Zeitschrift fur Operations Research* 32.5 (1988), pp. 299–306 (cit. on p. 105).

[155] R.W. Derlet. “Overcrowding in emergency departments: Increased demand and decreased capacity”. In: *Annals of Emergency Medicine* 39.4 (2002), pp. 430–432 (cit. on p. 86).

[156] R.W. Derlet and J.R. Richards. “Overcrowding in the nation’s emergency departments: Complex causes and disturbing effects”. In: *Annals of Emergency Medicine* 35.1 (2000), pp. 63–68 (cit. on pp. 81, 86).

[157] D. Duma and R. Aringhieri. “Mining the patient flow through an Emergency Department to deal with overcrowding”. In: *3rd International Conference on Health Care Systems Engineering*. Vol. 210. Springer Proceedings in Mathematics and Statistics. To appear. Springer International Publishing AG, 2017 (cit. on pp. 86, 104, 122).

[158] Y.-Y. Feng, I.-C. Wu, and T.-L. Chen. “Stochastic resource allocation in emergency departments with a multi-objective simulation optimization algorithm”. In: *Health Care Management Science* 20.1 (2017), pp. 55–75 (cit. on p. 86).

[159] J.A. Fitzgerald and A. Dadich. “Using visual analytics to improve hospital scheduling and patient flow”. In: *Journal of Theoretical and Applied Electronic Commerce Research* 4.2 (2009), pp. 20–30 (cit. on p. 86).

[160] M. Gendreau, G. Laporte, and F. Semet. “The Maximal Expected Coverage Relocation Problem for Emergency Vehicles”. In: *Journal of the Operational Research Society* 57.1 (2006), pp. 22–28 (cit. on p. 85).

[161] F. George and K. Evridiki. “The effect of emergency department crowding on patient outcomes”. In: *Health Science Journal* 9.1 (2015), pp. 1–6 (cit. on p. 81).

[162] N. Gilbert. *Agent-Based Models*. Vol. 153. Quantitative Applications in the Social Sciences. SAGE Publications, Inc, 2008 (cit. on p. 138).

[163] N. Gilbert and P. Terna. “How to build and use agent-based models in social science”. In: *Mind & Society* (2000), pp. 57–72 (cit. on p. 138).

[164] A. Haghani, Q. Tian, and H. Hu. “Simulation Model for Real-Time Emergency Vehicle Dispatching and Routing”. In: *Transportation Research Record: Journal of the Transportation Research Board* 1882.1 (2004), pp. 176–183 (cit. on pp. 85, 105).

[165] N.R. Hoot, C. Zhou, I. Jones, and D. Aronsky. “Measuring and Forecasting Emergency Department Crowding in Real Time”. In: *Annals of Emergency Medicine* 49.6 (2007), pp. 747–755 (cit. on pp. 82, 117).

[166] U. Hwang and J. Concato. “Care in the emergency department: How crowded is overcrowded?” In: *Academic Emergency Medicine* 11.10 (2004), pp. 1097–1101 (cit. on p. 81).

[167] C.J. Jagtenberg, S. Bhulai, and R.D. van der Mei. “An efficient heuristic for real-time ambulance redeployment”. In: *Operations Research for Health Care* 4 (2015), pp. 27–35 (cit. on p. 85).

[168] C.J. Jagtenberg, P.L. van den Berg, and R.D. van der Mei. “Benchmarking online dispatch algorithms for Emergency Medical Services”. In: *European Journal of Operational Research* 258.2 (2017), pp. 715–725 (cit. on p. 84).

[169] S. Kanchala, M.E. Mayorga, and L.A. McLay. “Recommendations for dispatching emergency vehicles under multitiered response via simulation”. In: *International Transactions in Operational Research* 21.4 (2014), pp. 581–617 (cit. on p. 85).

[170] Melik Koyuncu, Ozgur M. Araz, Wes Zeger, and Paul Damien. “A Simulation Model for Optimizing Staffing in the Emergency Department”. In: *Health Care Systems Engineering*. Ed. by Paola Cappanera, Jingshan Li, Andrea Matta, et al. Cham: Springer International Publishing, 2017, pp. 201–208 (cit. on p. 86).

[171] Y.-H. Kuo, J.M.Y. Leung, and C.A. Graham. “Simulation with data scarcity:Developing a simulation model of a hospital emergency department”. In: *Proceedings - Winter Simulation Conference*. 2012 (cit. on p. 86).

[172] A. Larsen, O. Madsen, and M. Solomon. “Partially Dynamic Vehicle Routing-Models and Algorithms”. In: *The Journal of the Operational Research Society* 53.6 (2002), pp. 637–646 (cit. on p. 105).

[173] S. Lee. “The role of preparedness in ambulance dispatching”. In: *Journal of the Operational Research Society* 62.10 (2011), pp. 1888–1897 (cit. on p. 83).

[174] S. Lee. “The role of centrality in ambulance dispatching”. In: *Decision Support Systems* 54.1 (2012), pp. 282–291 (cit. on p. 84).

[175] S. Lee. “Centrality-based ambulance dispatching for demanding emergency situations”. In: *Journal of the Operational Research Society* 64.4 (2013), pp. 611–618 (cit. on p. 84).

[176] S. Lee. “Role of parallelism in ambulance dispatching”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44.8 (2014), pp. 1113–1122 (cit. on pp. 85, 106).

[177] S.J.J. Leemans, D. Fahland, and W.M.P. van der Aalst. “Discovering block-structured process models from event logs containing infrequent behaviour”. In: *Lecture Notes in Business Information Processing* 171 LNBIP (2014), pp. 66–78 (cit. on p. 122).

[178] R. Luscombe and E. Kozan. “Dynamic resource allocation to improve emergency department efficiency in real time”. In: *European Journal of Operational Research* 255.2 (2016), pp. 593–603 (cit. on p. 86).

[179] R.S. Mans, W.M.P. Van Der Aalst, R.J.B. Vanwersch, and A.J. Moleman. “Process mining in healthcare: Data challenges when answering frequently posed questions”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7738 LNAI (2013), pp. 140–153 (cit. on p. 86).

[180] M. S. Maxwell, S. G. Henderson, and H. Topaloglu. “Ambulance redeployment: An approximate dynamic programming approach”. In: *Proceedings of the 2009 Winter Simulation Conference (WSC)*. 2009, pp. 1850–1860 (cit. on p. 85).

[181] M.S. Maxwell, M. Restrepo, S.G. Henderson, and H. Topaloglu. “Approximate Dynamic Programming for Ambulance Redeployment”. In: *INFORMS Journal on Computing* 22.2 (2010), pp. 266–281 (cit. on p. 85).

[182] L.A. McLay and M.E. Mayorga. “A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities”. In: *IIE Transactions* 45.1 (2012), pp. 1–24 (cit. on pp. 83, 84).

[183] L.A. McLay and M.E. Mayorga. “A Dispatching Model for Server-to-Customer Systems That Balances Efficiency and Equity”. In: *Manufacturing & Service Operations Management* 15.2 (2013), pp. 205–220 (cit. on pp. 84, 99, 111, 112).

[184] A.R. Nafarrate, J.W. Fowler, and T. Wu. “Bi-criteria analysis of ambulance diversion policies”. In: *Proceedings of the 2010 Winter Simulation Conference*. Baltimore, 2010, pp. 2315–2326 (cit. on pp. 85, 94, 99, 106).

[185] A.A. Nasrollahzadeh, A. Khademi, and M.E. Mayorga. “Real-Time Ambulance Dispatching and Relocation”. In: *Manufacturing & Service Operations Management* Published Online: April 11, 2018 (2018) (cit. on p. 85).

[186] E.C. Ni, S.R. Hunter, S.G. Henderson, and H. Topaloglu. “Exploring bounds on ambulance deployment policy performance”. In: 2012 (cit. on p. 85).

[187] S.A. Paul, M.C. Reddy, and C.J. Deflitch. “A systematic review of simulation studies investigating emergency department overcrowding”. In: *Simulation* 86.8-9 (2010), pp. 559–571 (cit. on pp. 81, 86).

[188] N.C. Proudlove, S. Black, and A. Fletcher. “OR and the challenge to improve the NHS: Modelling for insight and improvement in in-patient flows”. In: *Journal of the Operational Research Society* 58.2 (2007), pp. 145–158 (cit. on p. 88).

[189] A. Ramirez-Nafarrate, A.B. Hafizoglu, E.S. Gel, and J.W. Fowler. “Comparison of ambulance diversion policies via simulation”. In: *Proceedings of the 2012 Winter Simulation Conference*. Berlin, 2012, pp. 1–12 (cit. on p. 85).

[190] T.J. Reeder and H.G. Garrison. “When the safety net is unsafe: Real-time assessment of the overcrowded emergency department”. In: *Academic Emergency Medicine* 8.11 (2001), pp. 1070–1074 (cit. on p. 82).

[191] T.J. Reeder, D.L. Burleson, and H.G. Garrison. “The overcrowded emergency department: A comparison of staff perceptions”. In: *Academic Emergency Medicine* 10.10 (2003), pp. 1059–1064 (cit. on p. 82).

[192] Melanie Reuter-Oppermann, Pieter L. van den Berg, and Julie L. Vile. “Logistics for Emergency Medical Service systems”. In: *Health Systems* 6.3 (2017), pp. 187–208 (cit. on p. 83).

[193] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro. “Process mining in healthcare: A literature review”. In: *Journal of Biomedical Informatics* 61 (2016), pp. 224–236 (cit. on p. 86).

[194] V. Schmid. “Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming”. In: *European Journal of Operational Research* 219.3 (2012), pp. 611–621 (cit. on p. 83).

[195] V. Schmid and K.F. Doerner. “Ambulance location and relocation problems with time-dependent travel times”. In: *European Journal of Operational Research* 207.3 (2010), pp. 1293–1303 (cit. on p. 85).

[196] H. Setzler, C. Saydam, and S. Park. “EMS call volume predictions: A comparative study”. In: *Computers & Operations Research* 36.6 (2009), pp. 1843–1851 (cit. on pp. 90, 102).

[197] D. Sinreich, O. Jabali, and N.P. Dellaert. “Reducing emergency department waiting times by adjusting work shifts considering patient visits to multiple care providers”. In: *IIE Transactions (Institute of Industrial Engineers)* 44.3 (2012), pp. 163–180 (cit. on p. 86).

[198] S. Vanderby and M.W. Carter. “An evaluation of the applicability of system dynamics to patient flow modelling”. In: *Journal of the Operational Research Society* 61.11 (2010), pp. 1572–1581 (cit. on p. 88).

[199] A.J.M.M. Weijters and J.T.S. Ribeiro. “Flexible heuristics miner (FHM)”. In: *IEEE SSCI 2011: Symposium Series on Computational Intelligence - CIDM 2011: 2011 IEEE Symposium on Computational Intelligence and Data Mining* (2011), pp. 310–317 (cit. on p. 122).

[200] S.J. Weiss, R. Derlet, J. Arndahl, et al. “Estimating the Degree of Emergency Department Overcrowding in Academic Medical Centers: Results of the National ED Overcrowding Study (NEDOCS)”. In: *Academic Emergency Medicine* 11.1 (2004), pp. 38–50 (cit. on p. 82).

[201] E. Wolstenholme. “A patient flow perspective of U.K. Health Services: Exploring the case for new "intermediate care" initiatives”. In: *System Dynamics Review* 15.3 (1999), pp. 253–271 (cit. on pp. 88, 98).

[202] E. Wolstenholme, D. Monk, D. McKelvie, and S. Arnold. “Coping but not coping in health and social care: Masking the reality of running organisations beyond safe design capacity”. In: *System Dynamics Review* 23.4 (2007), pp. 371–389 (cit. on pp. 88, 98).

[203] J.-Y. Yeh and W.-S. Lin. “Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department”. In: *Expert Systems with Applications* 32.4 (2007), pp. 1073–1083 (cit. on p. 86).

[204] Soovin Yoon and Laura A. Albert. “An expected coverage model with a cutoff priority queue”. In: *Health Care Management Science* (2017) (cit. on p. 106).

List of Acronyms

Notation	Description
ABS	Agent-Based Simulation.
BE	Based on EOTs.
BII	Break-In-Interval.
BILLS	Break-In Layout Local Search.
BIM	Break-In-Moment.
BL	Based on Lengths.
CP	Clinical Pathway.
DES	Discrete Event Simulation.
DOA	Dedicated Overtime Allocation.
DOR	Dedicated Operating Room.
DRG	Diagnosis Related Group.
DRRP	Dispatching Routing and Redeployment Policies.
DT	Decision Tree.
ECDS	Emergency Care Delivery System.
ECP	Emergency Care Pathway.
ED	Emergency Department.
EDLOS	Emergency Department Length-Of-Stay.
EMS	Emergency Medical Service.
EOO	Elective-Oriented Optimization.
EOT	Estimated Operating Time.
FOA	Flexible Overtime Allocation.
FS	Flexible Scheduling.
HAF	Hybrid Activity Forest.
HAT	Hybrid Activity Tree.
HATM	Hybrid Activity Tree Miner.
HCBD	Health Care Big Data.
IP	Investigations Process.
LOS	Length Of Stay.
LWBS	Leaving Without Being Seen.
MCA	Minor Codes Ambulatory.
MSS	Master Surgical Schedule.

Notation	**Description**
MTBT	Maximum Time Before Treatment.
NEDOCS	National Emergency Department Overcrowding Scale.
NEIC	Non-Elective Insertion Criterion.
NERTI	Non-Elective Real Time Insertion.
NEW-Fit	Non-Elective Worst-Fit.
NOO	Non-elective-Oriented Optimization.
OR	Operating Room.
ORP	Operating Room Planning.
ROT	Real Operating Time.
RTM	Real Time Management.
SCAP	Surgical Case Assignment Problem.
SD	System Dynamics.
SOR	Shared Operating Room.
SP	Surgical Pathway.
SSO	Short-Stay Observation.
STAG	Sub-Tree Activity Graph.
URG	Urgency Related Group.

List of Figures

List of Tables

List of Publications

Publications related to the topics of the thesis

R. Aringhieri and D. Duma. *A hybrid model for the analysis of a surgical pathway*. In Proceedings of the 4th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (HA-2014), pages 889-900, August 2014. ISBN 978-989-758-038-3. Winner of the Best Paper Award.
Seminal paper for Part I

D. Duma and R. Aringhieri. *An online optimization approach for the Real Time Management of operating rooms*. Operations Research for Health Care, 7:40-51, 2015.
Basis for Chapter 3

R. Aringhieri and D. Duma. *The optimization of a surgical clinical pathway*. In M.S. Obaidat et al, editor, Simulation and Modeling Methodologies, Technologies and Applications, volume 402 of Advances in Intelligent Systems and Computing, pages 313-331. Springer International Publishing, 2015. Invited Chapter.
Basis for Chapter 3 and Chapter 6

D. Duma and R. Aringhieri. *Mining the patient flow through an Emergency Department to deal with overcrowding*. In 3rd International Conference on Health Care Systems Engineering, volume 210 of Springer Proceedings in Mathematics and Statistics, pages 49-59. Springer, Cham, 2017.
Basis for Chapter 10

D. Duma and R. Aringhieri. *The Real Time Management of Operating Rooms*. In C. Kahraman and I. Topcu, editors, Operations Research Applications in Health Care Management, volume 262 of International Series in Operations Research & Management Science, pages 55-79. Springer, 2018.
Basis for Chapter 4

D. Duma and R. Aringhieri. *The management of non-elective patients: shared vs. dedicated policies*. Omega, 2018. Advance online publication 14 March 2018.
Basis for Chapter 5

R. Aringhieri, D. Dell'Anna, D. Duma, and M. Sonnessa. *Evaluating the dispatching policies for a regional network of emergency departments exploiting health care big data*. In G. Nicosia, P. Pardalos, G. Giuffrida, and R. Umeton, editors, International Conference on Machine Learning, Optimization, and Big Data, volume 10710 of Lecture Notes in Computer Science,

pages 549-561. Springer International Publishing, 2018.
Basis for Chapter 8

D. Duma and R. Aringhieri. *An ad hoc process mining approach to discover patient paths of an Emergency Department*. Submitted to Flexible Services and Manufacturing. Currently in evaluation after a minor revision.
Basis for Chapter 10

R. Aringhieri, S. Bocca, L. Casciaro, and D. Duma. *A simulation and online optimization approach for the real time management of ambulances*. Accepted for the publication to the proceedings of the Winter Simulation Conference 2018.
Basis for Chapter 9

Finally, the materials reported in Chapters 3, 4, and (partially) 5 have been presented at the *EURO Summer Institute on Online Optimization* in which the author was one of the 18 delegates admitted (see `https://www.euro-online.org/media_site/reports/ESI32_Report.pdf`).

Other publications during the Ph.D period

R. Aringhieri and D. Duma. *Patient-centred objectives as an alternative to maximum utilisation: comparing surgical case solutions*. In Optimization and Decision Science: Methodologies and Applications. ODS 2017, volume 217 of Springer Proceedings in Mathematics and Statistics, pages 105-112. Springer, Cham, 2017.

R. Aringhieri, D. Duma, A. Grosso, and P. Hosteins. *Simple but effective heuristics for the 2-Constraint Bin Packing Problem*. Journal of Heuristics, 24(3), pp. 345-357, 2017.

R. Aringhieri, D. Duma, and V. Fragnelli. *Modeling the rational behavior of individuals on an e-commerce system*. Operations Research Perspectives, 5(1):22-31, 2018.

R. Aringhieri, D. Duma, and E. Faccio. *Ex post evaluation of an operating theatre*. In Joint EURO/ALIO International Conference 2018 on Applied Combinatorial Optimization, Electronic Notes in Discrete Mathematics, 69:157-164, 2018.

R. Aringhieri, G. Bonetta e D. Duma. *Reducing overcrowding at the emergency department through a different physician and nurse shift organisation: a case study*. In: ODS2018 - International Conference on Optimization and Decision Science. AIRO Springer Series. To appear. Springer, 2018.

R. Aringhieri, D. Duma e F. Polacchi. *Integrating mental health into a primary care system: a hybrid simulation model*. In: ODS2018 - International Conference on Optimization and Decision Science. AIRO Springer Series. To appear. Springer, 2018.

Colophon

This thesis was typeset with $\LaTeX\,2_{\varepsilon}$. It uses a slightly adapted version of the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.