Operations optimization in emergency departments

Karim Ghanes

NNT : 2016SACLC038

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PARIS-SACLAY
PRÉPARÉE À
"CENTRALESUPÉLEC"

ÉCOLE DOCTORALE N° 573
Interfaces : approches interdisciplinaires / fondements, applications et innovation

Spécialité de doctorat : Sciences et technologies industrielles

Par

Karim GHANES

Optimisation des opérations dans les services d'urgence

Thèse présentée et soutenue à Châtenay-Malabry, le « 29/04/2016 » :

Composition du Jury :

M. Vincent AUGUSTO, École des Mines de Saint-Étienne	Rapporteur
M. Slim HAMMADI, École Centrale de Lille	Rapporteur
M. Jean-Charles BILLAUT, Université de Tours	Examinateur (président)
M. Vincent MOUSSEAU, CentraleSupélec	Examinateur
M. Zied JEMAI, CentraleSupélec /École Nationale d'Ingénieurs de Tunis	Directeur de thèse
M. Oualid JOUINI, CentraleSupélec	Co-encadrant de thèse
M. Ger KOOLE, VU University Amsterdam	Co-encadrant de thèse
M. Mathias WARGON, Hôpital Saint Camille	Invité
M. Romain HELLMANN, Agence Régionale de santé d'Ile de France	Invité

Titre : Optimisation des opérations dans les services d'urgence

Mots clés : Services d'urgence, opérations, évaluation de performance, optimisation, indicateurs clés de performance, simulation, files d'attente

Résumé : Un Service d'urgence (SU) est le service hospitalier responsable de la prise en charge d'une grande diversité de patients, 24 heures sur 24, 7 jours sur 7. Les SU de par le monde sont actuellement confrontés à un problème de surcharge, qui résulte de l'inadéquation entre la capacité et la demande en soins. Ce problème entraîne plusieurs effets négatifs tels que des durées de passage excessivement longues, une insatisfaction des patients, un environnement de travail stressant et l'augmentation de la fréquence des erreurs médicales. L'objectif principal de cette thèse est de développer des solutions internes permettant d'améliorer la performance des SU, à l'aide de méthodes issues de la Recherche Opérationnelle. Nous abordons trois catégories de questions de recherche. La première catégorie comprend des questions prospectives portant sur les indicateurs clés de performance ainsi que sur les différents facteurs contribuant à la congestion des urgences. La deuxième catégorie correspond au dimensionnement de la capacité des ressources humaines et à l'optimisation des emplois du temps. La troisième catégorie de questions porte sur l'optimisation du processus, où nous analysons des modifications et des alternatives innovantes dans le parcours du patient. De manière générale, cette thèse aborde des questions de recherche innovantes, et fournit aux managers des recommandations et des outils permettant d'améliorer la performance des SU. Elle ouvre également la voie pour de futurs axes de recherche liés à l'optimisation des opérations dans les SU.

Title : Operations optimization in emergency departments

Keywords : Emergency departments, operations, performance evaluation, optimization, key performance indicators, simulation, queueing systems

Abstract : Emergency Department (ED) is the service within hospitals responsible for providing care to a wide variety of patients over 24 hours a day, 7 days a week. As a result to the existing mismatch between available caring capacity and patients demand, EDs are currently facing a worldwide problem, namely overcrowding. ED overcrowding or congestion may result in several negative effects such as long patient stays and waiting times, dissatisfaction of patients, high levels of stress, and increased medical errors. The objective of the present thesis is to develop internal and cost-effective solutions to alleviate overcrowding in EDs and improve their performance, using Operations Research methods. We address three categories of research questions. The first category includes prospective questions about ED Key Performance Indicators and about the diverse factors contributing to overcrowding. The second category is associated to the dimensioning and shift-scheduling of ED human resource capacity. The third category of questions deals with process-related issues where we investigate potential alternative and innovative ED patient flow designs. Roughly speaking, this thesis addresses innovative OM research questions for EDs. It provides decision makers with recommendations and tools in order to improve ED performance. It also highlights various avenues for future research related to the optimization of ED operations.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

Contents

Chapter 1

Introduction

In this chapter, we give a general introduction of the thesis. First, a broad view on the background and the motivations of this research work is provided. Second, the research context is introduced and positioned with respect to the emergency department literature. Third, research objectives are identified and the thesis main contributions are highlighted. Finally, a graphical representation of the structure of the manuscript is given.

1.1 Background and motivation

Few things affect the quality of life more than health, so few issues might be more important than healthcare (Hopp and Lovejoy, 2012). In recent decades, health spending has dramatically increased due to several factors, such as demographic trends or the widespread diffusion of expensive technological advances in medical practices (White, 2007). For instance, the total healthcare expenditures in the United States (US) reached a larger proportion of the gross domestic product, 17.7% in 2011 compared to 13.6% in 2000. Similar trends are also observed in Europe. This proportion increased in the Netherlands from 8% in 2000 to 11.9% in 2011, and from 10.1% to 11.6% in France (World Health Organization, 2014).

Hospitals and emergency systems are two crucial actors of the public healthcare system. Hospitals are central to the healthcare delivery process, and constitute a significant percentage of the total healthcare spending (Hopp and Lovejoy, 2012). Emergency systems represent another major element of the healthcare system which includes emergency medical services (EMS) and emergency departments (EDs). The mission of an EMS is to provide timely out-of-hospital acute medical care, in response to an emergency call. It is also in charge of patient transportation to an appropriate care facility, generally an emergency department at a hospital, where the patient is handed over. Hence, emergency department is at the crossroads between hospitals and emergency systems, and plays a key role in patient safety and public health. Emergency department (ED) is the service within hospitals responsible for providing unscheduled care to a wide variety of patients (life-threatening and other emergency cases) over 24 hours daily, 7 days a week. It is the main entrance to a hospital, through which about half of non-obstetrical admissions occur in the US (Pitts et al., 2008).

The number of patients visiting EDs is in continuous increase while the number and the capacity of EDs are both decreasing (Hoot and Aronsky, 2008; Niska et al., 2010; Harrison and Ferguson, 2011; Abo-Hamad and Arisha, 2013). According to the National Center for Health Statistics (2012), between 1995 and 2010, the annual number of ED visits in the US increased by 34% (from 97 million to 130 million visits), whereas the number of hospital EDs decreased during this same period by about 11% (from 4,160 to 3,700). The reader may refer to Hsia et al. (2011) for more literature about the factors associated with EDs closures in the US. In France, 10.6 millions of patients have visited a hospital ED in 2012 (which represents about one sixth of the French population), sometimes more than once during the same year. A total number of visits of 18 millions has been recorded in 2012, which represents an increase of 30% in ten years (IRDES, 2015). There is a direct correlation between this increased usage of emergency services on the one hand and the aging of a population on the other (George et al., 2006). Similar

trends are intensifying pressure on EDs around the globe. Many surveys report that more than half of worldwide EDs do not have sufficient capacity to support the patients flow in optimal conditions and without prolonged waiting times (Pateron, 2012; American Hospital Association, 2010). As a result of this mismatch between available caring capacity and patients demand, EDs are currently facing a recurrent worldwide problem, namely overcrowding.

Emergency department overcrowding or congestion is a worldwide crisis that may result in several negative effects. The phenomenon manifests itself through different ways (Paul et al., 2010). For instance, an excessive number of patients present in the ED, long patient stays and waiting times, and treatment in hallways, are all overcrowding signs. Congestion in emergency departments leads to decreased physician productivity, miscommunication between working staff, diversion of ambulances (Paul et al., 2010; Solberg et al., 2003), and dissatisfaction of patients who may sometimes leave without treatment (Liao et al., 2002). Moreover, overcrowding leads to high levels of stress, physical violence and verbal abuse toward emergency nurses (Emergency Nurses Association, 2011) and decreased morals among the staff (Public Health and Injury Prevention Committee, 2011; Paul et al., 2010). Overcrowding is also related to increased medical errors and mortality rates (Spirivulis et al., 2006; Carmen and Van Nieuwenhuyse, 2014), high staff turnovers and unnecessarily high costs (Trzeciak and Rivers, 2003; Kuo et al., 2012; Solberg et al., 2003). For all these reasons, EDs became a central concern for health administrators and experts, politicians and media. Moreover, addressing the problem of overcrowding has become a critical challenge for both healthcare emergency practitioners and researchers in operations research and operations management (Hopp and Lovejoy, 2012). Given the increasing demand, high operating costs (Sinreich and Marmor, 2005; Warner, 2013) combined to budgetary limitations (Carmen and Van Nieuwenhuyse, 2014; Abo-Hamad and Arisha, 2013), there is an urgent need for cost-effective improvement solutions to address the current inefficiencies in emergency departments. The present thesis falls within this context.

The present research work is conducted in collaboration between the public French Regional healthcare Agency (Agence Régionale de Santé-ARS Ile de France) and the Industrial Engineering Laboratory (Laboratoire Génie Industriel, LGI) at Ecole Centrale Paris.

1.2 Research context

Through the last decades, the importance of EDs and the increasing need to improve their operations efficiency is being accompanied by an extensive and growing literature. The scientific disciplines dealing with EDs are numerous: medicine, statistics, operations research, industrial

engineering, as well as psychology, sociology, architecture, and finance. The present thesis is pertaining to the operations research/operations management (OR/OM) domain. OR/OM techniques have significantly helped in improving the performance of various parts of hospitals (and especially their EDs) in the last decades (Saghafian et al., 2015; Hopp and Lovejoy, 2012).

Problem types

The OR/OM literature dealing with the improvement of ED performance is diverse. In general, it can be categorized into the following three different streams of interventions, according to the nature of the employed improvement lever:

- Resource-related interventions: deal with the dimensioning of ED resource capacity (staffing, shift-scheduling and rostering).
- Process-related interventions: deal with the modification of some protocols and organizational rules in ED patient-flow (ED process).
- Environment-related interventions: aim at modifying some characteristics of ED external environment, mainly those concerning demand and admission services.

Environment-related or external interventions correspond to interventions that must be undertaken outside the ED while involving external actors such as other hospital services or alternative facilities like alternative EDs, etc. It must be noted that this last category falls out of the scope of this thesis. Our focus is to provide ED decision makers with managerial insights and solutions that could be implemented autonomously and independently from external actors that are beyond ED perimeter and responsibility. Some external interventions are highlighted in the prospective chapters (Chapters 2 and 3), yet the main concern of this thesis is ED internal interventions. Consequently, experimental chapters will solely focus on resource-related and process-related issues.

Method types

An ED is a highly complex system with heterogeneous patients and various types of resources that evolve within a sophisticated process. The analysis methodology must be carefully selected so as to comply with both academic opportunities and industrial expectations. The main OR/OM tools for modeling and improving ED patient flow include simulation and analytical methods (queueing theory, Markov models, etc.). Both have legitimate advantages and drawbacks. The main benefit of analytical models is that they are more transparent (Kolker,

2008), require less data, have shorter model development time and provide more generic results than simulation (Wang et al., 2013; Saghafian et al., 2015). However, they include less details and represent simplified versions of the ED because major simplifying transformations are required for mathematical convenience (Wang et al., 2013), while simulation models can capture most details of the system without requiring major hypotheses (Kolker, 2008). Since realistic ED models are intractable analytically (Zeltyn et al., 2011), we resort to simulation for an appropriate framework. Simulation is an important systems analysis tool which provides great flexibility in testing scenarios, policies and re-engineering ideas in healthcare (Paul et al., 2010). The need for high impact solutions motivates us to use discrete-event simulation (DES). In using DES for ED operations management, we are following a longstanding practice (Paul et al., 2010; Günal and Pidd, 2010).

In order to explore a large set of feasible solutions, simulation-optimization is used in some of the experiments. We also test intuitive what-if scenarios when performing sensitivity analysis. Statistical methods are used throughout this thesis either to provide statistical distributions as inputs for the simulation model or to identify correlations between variables. We also use mathematical programming and continuous time Markov chains methods. Moreover, some of the addressed issues present research gaps in both medical and OR/OM literature (process-related issues) where limited information is available. Therefore, we resort to field surveys, that are carried out in collaboration with medical experts, in order to define proper frameworks for our analysis.

1.3 Objectives, research questions and contributions

The primary objective of this thesis is to provide ED managers with internal and cost-effective solutions and insights so that to alleviate overcrowding and improve ED performance. Achieving this objective requires responding to a series of research questions (essential and subsidiary questions) that were identified through a logical order. This thesis has practical implications thanks to a close collaboration with the emergency department of Saint Camille hospital. Thus, our research questions were defined in a way to comply with both industrial and academic perspectives, and they are organized as follows.

Preliminary questions

Before investigating the different methods to improve ED performance, some essential prospective questions are answered. The first question focuses on how to properly measure the ED performance. The second one deals with the understanding of the overcrowding phenomenon.

Question 1: *What are the most relevant ED key performance indicators and how to choose them according to the study context?*

To answer this question, a detailed literature review is provided in Chapter 2 on the commonly used key performance indicators (KPIs) from an OR/OM perspective. The review summarizes the advantages and drawbacks of each KPI and provides several useful insights. For instance, each KPI measures something different in the ED, and we underline the value of combining different KPIs to complement one another. This chapter gives also an overview of the OR/OM ED literature and introduces useful notions and concepts for the rest of the thesis. It also serves as a basis for the appropriate selection of KPIs in the next chapters.

Question 2: *What are the factors contributing to overcrowding and long delays in EDs, and how can they be addressed?*

This question is addressed in Chapter 3. Using real data from two hospitals, we perform a series of statistical tests among several potential influencing factors (represented by variables) in order to identify the ones currently affecting ED performance. A thorough interpretation of results is conducted, which helped identifying the factors leading to the obtained dependencies between ED performance and some variables in practice. Moreover, we provide for each influencing factor the corresponding relevant remedial measures (interventions) existing in practice and in the literature. The outcomes of this chapter represent a departure point for the research questions that will be addressed in the next ones.

Resource-related questions

In order to alleviate congestion, ED managers and the general management of Saint Camille hospital intend to invest in human resources staffing in order to improve performance. The objective is to find the most rational and efficient increase in staffing budget. Hence, a first step to address here is the modeling of the ED. Moreover, when addressing the first research question in Chapter 2, the insights derived from the identified research gaps have pushed us to include two different major KPIs in our experiments, and assess the impact of such a combination. This all gave birth to the following research question:

Question 3: *By how much should the current staffing budget be increased and how should*

this additional budget be used in the allocation of human resources?

This question is addressed in Chapter 4. We build a realistic ED model using discrete-event simulation. It has been decided that the use of this tool should not be limited only to this research work, but should be generalized as a decision aid tool useful to other French EDs. Thus, most essential structural and functional characteristics of EDs, at least in France, are taken into consideration thanks to a close collaboration with practitioners. Consequently, we point out a set of important ED characteristics that are frequently ignored in the related literature. Using simulation-optimization, we focus in our experiments on human staffing levels. We want to minimize the patients average length of stay (LOS), by integrating a staffing budget constraint and a constraint securing that the most severe incidents will see a doctor within a specified time limit. The obtained results allowed us to provide useful insights to managers on how the budget impacts ED performance, and how investments should be allocated among resources. We also highlight and explain an important managerial insight about the effect of combining two major KPIs on the solutions.

While dealing with the improvement of staffing levels, an additional research question is also identified. Saint Camille ED uses a daily shift pattern composed of only two shifts, which we use as such to address the previous questions. It is clear that such a division of the day allows very little flexibility given the patient arrival pattern that changes on a hourly basis. The problem of shift definition was rarely addressed in the literature, since researchers generally use predetermined shifts, designed intuitively by practitioners. The wide majority of studies address the question of how to efficiently fill those predetermined shifts, instead of how to define them in a way to best match demand profile. Yet, we believe that if the question of how to divide the day properly into different shifts is answered, it may provide managers with a cost effective and simple way to improve ED performance:

Question 4: *How to define an appropriate shift pattern that matches better the arrival pattern of patients in EDs?*

This question is dealt with in Chapter 5 where we propose a method of shift definition that optimizes the allocation of available resources without increasing costs, while respecting the main constraints encountered in practice. The method includes simulation-optimization and linear programming. The simulation model supplies the linear program with the staffing levels (performance standards). The linear model determines the shift-scheduling of all employees

with the use of the minimum cost. Finally, we propose a heuristic that combines the results of the above models and secures that the budget constraint will be met by the final staff allocation.

Process-related questions

The objective is to investigate alternative ED patient flow designs (with fixed budget). The identified research gaps in the literature combined with the outcomes of Chapter 3 enabled us to formulate two major process-related research questions that are exposed hereafter.

Typically in current ED practices, each patient is assigned to a single physician who will be exclusively responsible of her during all stages of the ED process. We refer to the aforementioned rule as the "Same Patient Same Physician (*SPSP*)" rule. The objective is to investigate another strategy (that we call collaborative strategy) which consists in ignoring the *SPSP* rule. The intuition behind assessing the removal of *SPSP* rule is the well-known inefficiency of forcing customers/patients to wait for their assigned server to become free, even if another server is idle (Song et al., 2013; Saghafian et al., 2012). We are not aware of any work that deals with this research question, neither in the medical domain nor in OR/OM literature:

Question 5: *Should a patient be handled by the same physician during all stages of the ED process?*

The question can also be formulated as follows: *Is it beneficial for the ED performance to remove the same patient same physician rule?* This issue is tackled in Chapter 6. We conduct a survey which confirms that *SPSP* stands as the standard practice in most EDs worldwide. The survey reveals that removing *SPSP* rule is very controversial among practitioners because of human considerations (related to both patients and practitioners). From a quantitative point of view, the collaborative strategy would suffer from a time extension in the tasks that are performed by a different physician. From this appears the necessity of a risk/benefit analysis. We introduce the two system processes as complexity-augmented $Erlang - R$ queueing networks and show through simulation that the relevancy of removing *SPSP* depends on the system load. There is a certain threshold under which the collaborative strategy outperforms *SPSP*, and above which its application becomes detrimental. We further confirm the obtained insights under realistic conditions using simulation. The potential performance improvement stands as a strong argument against the widespread reluctance of practitioners towards the collaborative strategy.

The second issue is an anticipation method involving the triage process. The common pro-

tocol in EDs is that the triage nurse cannot order diagnostic tests. She is essentially responsible of making a first assessment of patients state and categorizing them into different acuity levels. The decision of requiring diagnostic tests or not comes after. It is traditionally under the responsibility of the physician. However, it has been revealed in the medical literature that giving the triage nurse the possibility to initiate diagnostic tests, without waiting for the initial consultation of the physician, may improve patients satisfaction and possibly decrease their length of stay (Rosmulder et al., 2009; Cheung et al., 2002). Triage nurse ordering (TNO) appears to be a promising approach that does not require any resource investment. It could be achieved using existing triage nurses with little additional training (Rowe et al., 2011). However, this issue was not addressed from an OR/OM perspective and there is a real need to conduct studies that will legitimize the use of TNO in EDs in terms of LOS reduction (Robinson, 2013; Rowe et al., 2011):

Question 6: *Is it beneficial in terms of LOS to allow triage nurse ordering diagnostic tests?*

This question is discussed in Chapter 7. The conducted survey reveals that the majority of experts consider TNO as a potential relevant practice in general. However, there is a wide variety of diagnostic tests and the feasibility of applying TNO varies greatly from one test type to another. For each diagnostic test, the survey provides the practical reasons about the possibility to apply TNO or not. We model the new patient path and assess its efficiency on the ED performance through simulation, while considering the length of stay as the key indicator. We examine the impact of the key elements (triage nurse ability, system load and triage time extension) on the benefits that might be derived from triage nurse ordering.

1.4 Dissertation organization

In this section, we present the structure of the manuscript. We describe the dissertation organization which consists of 8 chapters, and give their corresponding published or working papers. Given the diversity of the addressed issues, each chapter comprises a specific literature review. The organization of chapters is illustrated through Figure 1.1.

Chapter 2: We conduct a detailed literature review on the commonly used KPIs from an OR/OM perspective. The review summarizes the advantages and drawbacks of each KPI and provides several useful insights. The paper version of this chapter (Ghanes et al., 2014a) is under second round revision in the journal *IIE Transactions on Healthcare Systems Engineering*.

Chapter 3: A series of statistical analysis are performed in the purpose of identifying the main influencing factors of performance. This Chapter is based on Vegting et al. (2015) which is published in *The Netherlands Journal of Medicine*.

Chapter 4: A realistic ED discrete-event simulation model is proposed. We provide useful insights to managers about the impact of the budget on performance and how investments should be allocated among resources, as well as the effect of combining two different major KPIs. The paper versions of this chapter (Ghanes et al., 2015c, 2014b) are published in the journal *SIMULATION*, and the proceedings of the *2014 Winter Simulation Conference* held in Savannah, USA.

Chapter 5: We propose a heuristic for the optimization of the shifts of human resources. The method combines simulation-optimization and linear programming. The paper version of this chapter (Ghanes et al., 2015a) is published in the proceedings of the *45th International Conference on Computers and Industrial Engineering (CIE45)* held in 2015, in Metz, France.

Chapter 6: We investigate the relevancy of the *SPSP* rule. We carry out a field survey which shows that this issue is very controversial among practitioners, mainly because of human considerations. We use discrete-event simulation to gain insights into the behaviors of systems using or not *SPSP*.

Chapter 7: We model the triage nurse ordering (TNO) process and assess its efficiency on ED performance as a function of key parameters. The paper corresponding to this chapter (Ghanes et al., 2015b) is published in the proceedings of the *IEEE International Conference on Industrial Engineering and Systems managements (IESM)* held in 2015, in Sevilla, Spain.

Chapter 8: This chapter gives general concluding remarks of the thesis and highlights a number of possible directions for future research.

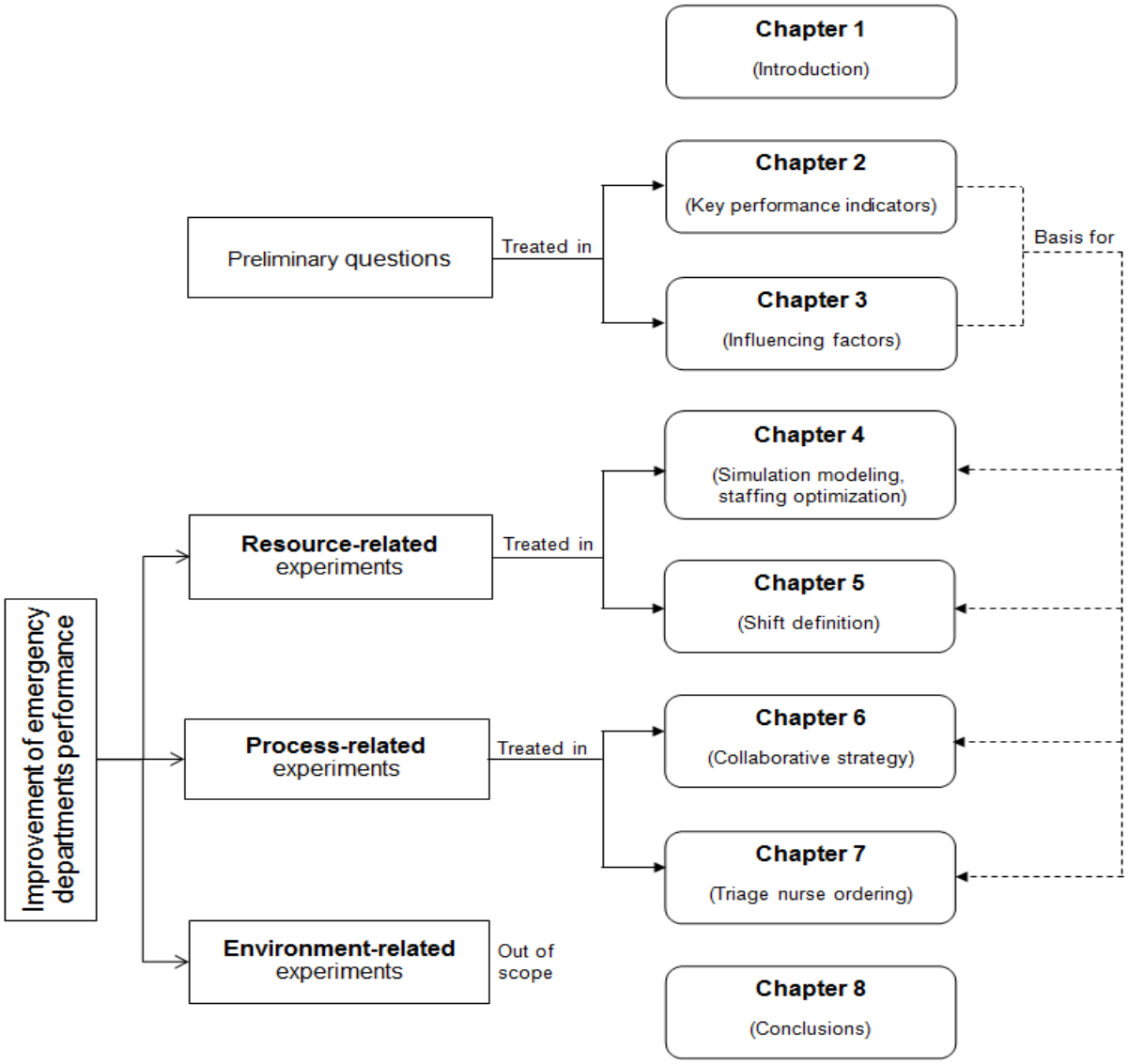

Figure 1.1: Dissertation organization

Chapter 2

Key performance indicators: A survey from an operations management perspective

In this chapter, we discuss the most relevant key performance indicators in the emergency department and review the related operations research and operations management literature. It is well known that ED overcrowding, a phenomenon referring to a deteriorated performance such as long waiting times, affects hospitals worldwide. An important stream of the operations research and operations management literature focuses on improving the ED performance in order to alleviate this congestion. A first required step is then to define the performance indicators. In this survey, we discuss the relevancy of each metric in order to provide researchers with a support to select those that best match with a given study context (environment, type of patients, objective, etc.).

The paper version of this chapter (Ghanes et al., 2014a) is under second round revision in the journal *IIE Transactions on Healthcare Systems Engineering*.

2.1 Introduction

Improving the performance of any system requires first to know how to measure its performance properly. The most commonly found ED key performance indicators (KPIs) in the literature include the total length of stay (LOS), the door-to-doctor time (DTDT), ambulance diversion (AD), the rate of patients that leave without being seen by a physician (LWBS), etc. These KPIs are strongly correlated to congestion and patient satisfaction. The selection of the appropriate KPIs has always been a controversial subject, for which the whys and wherefores remain unclear. An ED is a large and complex system (Smith and Feied, 1999) and each one of the available metrics measures something different (Hwang et al., 2011). Neither the scientific community nor practitioners are able to decide about the most appropriate KPI, as each indicator presents at the same time benefits and drawbacks.

In this survey, we review the existing literature by enumerating the used KPIs and describing how researchers propose to improve them from an OR/OM perspective. It should be noticed that the medical literature includes surveys on ED metrics. Sorup et al. (2013) perform a review that analyzes the use of several ED metrics in medical papers. Welch et al. (2006, 2011) present the definitions of the metrics used in the medical literature. Hwang et al. (2011) conduct a systematic review of all existing crowding measures and compare between them in terms of their validity. The book by Hopp and Lovejoy (2012) provides guidance for applying the appropriate metrics to measure EDs and hospitals performance.

The main contributions of this chapter can be summarized as follows. We review the most used KPIs. For each KPI, we explain its relevancy in order to provide researchers with a support to select the KPI(s) that best match with their study context (environment, type of patients, objective, etc.). For example, AD and the rate of LWBS depend on external factors on which the ED has no control. They cannot be then used as a reference to compare between different EDs. For each KPI, we also highlight its advantages and drawbacks. For instance, DTDT is a crucial KPI for critical patient acuity levels, but it does not give any information about the system state during other important stages of the process (beyond the first consultation). As for LOS, it gives an overview on the entire system performance but does not allow to figure out local strengths and weaknesses. We therefore review studies that combine different KPIs. We discuss relevant combinations of KPIs and highlight potential interdependency between them (for example the correlation between DTDT and the rate of LWBS). Finally, we point out some universal quantitative measures of crowding. These have received a poor attention from the OR/OM community while they are employed and recognized by medical practitioners. Although the commonly used KPIs in the OR/OM literature (LOS, DTDT, etc.) are correlated

with crowding, strictly speaking, they do not allow to say whether an ED is overcrowded or not, so, improving them does not necessarily mean reducing overcrowding.

The rest of the chapter is organized as follows. Section 2.2 provides a brief background on EDs. In Section 2.3, the most used KPIs in the OR/OM literature are discussed and relevant related papers are reviewed. Section 2.4 summarizes the main findings and highlights avenues for future research. The chapter ends with concluding remarks.

2.2 ED background

We give a background on EDs and the patient path in an ED. We also provide the list of the KPIs that are analyzed in this survey.

An ED is a large system involving several resources and heterogeneous patient types that follow a complex process with specific rules and protocols. Human resources consist of physicians, nurses (triage or ordinary nurses), junior physicians (residents or medical students) and patient transporters (also called hospital porters or stretcher-bearers). Other resources present in the ED are examination rooms (also called cubicles or boxes), shock rooms (also called resuscitation rooms) for life-threatening cases, waiting rooms and stretchers. Some EDs also have an observation unit that admits short stay patients in order to wait for an inpatient bed or for further control before being released.

A typical ED process can be described as follows. After registering at the main entrance of the ED, the patient is assessed in the triage station, in most cases by a nurse that diagnoses the severity of the situation. The patient is assigned a severity code (an acuity level) and proceeds to the waiting room. Patients are often divided into five acuity levels according to a triage method (Tanabe et al., 2007; Abo-Hamad and Arisha, 2013). After triage, the consultation starts as soon as the adequate physician becomes available. The physician makes a first assessment and may decide, if necessary, to request one or more ancillary tests (radiology and/or laboratory tests) in order to confirm or refine the diagnosis. If not, the patient is released. Once all the tests are completed, the physician responsible for the patient examines the results, makes an interpretation and chooses the appropriate process outcome for the patient. Finally, the patient can be admitted to another service of the hospital, transferred to another hospital, admitted to the observation unit or discharged. All the stages described above are separated by waiting times (WT) that depend on the availability of the required resources. An illustration of the patient path is given in Figure 2.1.

The OR/OM literature considers KPIs that are defined on the ED environment, but particularly at the patient path stages. The most used KPIs in the literature are:

- *Length of stay (LOS):* The time period spent by the patient in the ED from the entrance until the discharge from the system or the admission to an Internal Unit (IU).
- *Door-to-doctor time (DTDT):* The time interval between the arrival in the ED and the first consultation by a physician.
- *Left without being seen (LWBS):* The percentage of patients that leave the ED after the process of triage and before the initial consultation.
- *Ambulance diversion (AD):* The amount of time that ambulances are signaled to seek for an alternative ED because of overcrowding.

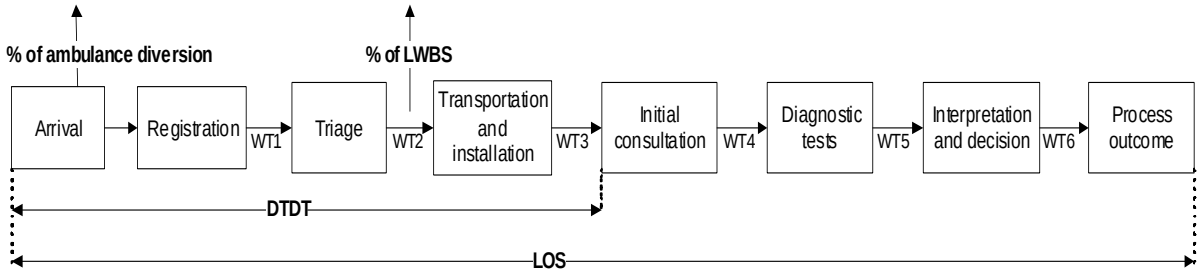

Figure 2.1: Typical stages of the patient path

There is an abundant OR/OM literature that deals with the improvement of ED performance. This literature can be categorized into different streams. According to the improvement that authors chose to focus on, we can broadly categorize the studies into resource-related studies (deal with staffing and shift-scheduling) and process-related studies (consist in modifying some protocols and organizational rules in the process). Concerning the used OR tools, we can distinguish between the two main categories : simulation (Paul et al., 2010; Günal and Pidd, 2010; Ghanes et al., 2015c) and analytical models like queueing and Markov chains (Huang et al., 2012; Green et al., 2006; Saghafian et al., 2012, 2014). Less used tools are mathematical programming (Beaulieu et al., 2000b; Centeno et al., 2003) and game theory (Hagtvedt et al., 2009; Deo and Gurvich, 2011). For further discussion on the OR/OM techniques used for the ED analysis, we refer the reader to the literature reviews provided by Wiler et al. (2011); Bhattacharjee and Ray (2014); Saghafian et al. (2015). It should be mentioned that there is a growing stream of empirical studies that are published in OR/OM journals (Batt and Terwiesch, 2015). The interest in field experiments stems from the necessity to better understand human behavior aspects, from the patient perspective such as abandonment (Bolandifar et al., 2014; Batt and Terwiesch, 2015), or from the ED staff perspective like state-dependant service times (Batt and Terwiesch, 2014).

2.3 KPI analysis

We review articles that propose to improve ED performance as measured by a subset of the aforementioned KPIs using OR and OM concepts and tools. The way we synthesize the literature is by classifying the papers based on the used KPIs. We first start by the KPIs individually: for each KPI, we review the papers that use this KPI, the considered objective, the used methodology, the results, the advantages and disadvantages of the KPI, etc. We next consider combination of KPIs, and show how existing studies combine complementary KPIs.

In order to identify the most relevant OR/OM studies (assess and improve ED performance) that were published or available on-line between 1991 and 2015, we relied on a specific and systematic search strategy in the databases of Web of Science, SSRN, JSTOR and ScienceDirect. Titles and abstracts were screened using different sensitivities (i.e., and/or/not) on the keywords: emergency department, crowding, performance indicators, metrics, operations research, operations management, optimization, performance evaluation, staffing, waiting time, etc. Given the large amount of obtained articles and in order to retain the most relevant ones to this study, we undergo a second filter based on citations, the journal or the conference. This has led to about 160 referenced articles.

Table 2.1 briefly summarizes relevant OR/OM papers according to the considered KPI to improve. In addition to the commonly used KPIs, we report a number of other less used metrics in the table such as ED time intervals other than DTDT (such as time to triage and transfer duration), multidimensional scores, patient throughput, resource utilization, etc. Papers focusing on combinations of KPIs can be seen on the lines with more than one bullet.

2.3.1 Length of stay (LOS)

The LOS, also referred to as *throughput time* (Ruohonen et al., 2006; Komashie and Mousavi, 2005) or *time to completion* (Vegting et al., 2011), is the most widely used metric in the OR/OM literature and also in practice. It measures the total duration of time spent by the patient in the ED. Sometimes policy makers set a maximum limit of the patient LOS. The most known example is the 4 hour target in the UK, which states that 98% of patients must be discharged, transferred or admitted in an internal unit within 4 hours (Mayhew and Smith, 2008; Izady and Worthington, 2012). Another example of the administratively suggested LOS target was published by the Canadian Association of Emergency Physicians (CAEP). It suggested that 95% of levels 1, 2 and 3 should be seen within 6 hours. For levels 4 and 5 patients, LOS should not exceed 4 hours in 95% cases. We should mention that setting an LOS target might lead to some inconvenience in the treatment procedure and downgrade the quality of service (Orr,

Table 2.1: KPIs studied in OR/OM papers

Reference	LOS	DTDT	LWBS	AD	Other ED time intervals	Others
Abo-Hamad and Arisha (2013)	•	•	•			•
Ahmed and Alkhamis (2009)		•				•
Alavi-Moghaddam et al. (2012)	•				•	
Allon et al. (2013)				•		
Armony et al. (2011)						•
Ashour and Kremer (2013)	•				•	•
Batt and Terwiesch (2014)	•	•	•		•	
Batt and Terwiesch (2015)			•			
Broyles and Cochran (2007)			•			
Burström et al. (2012)	•	•	•			
Chan et al. (2005)	•	•	•			
Chonde et al. (2013)	•	•				
Cochran and Roche (2009)	•	•	•			
Cooke et al. (2012)		•				
Deo and Gurvich (2011)				•		
Dobson et al. (2013)						•
Duguay and Chetouane (2007)	•				•	
Ferrin et al. (2007)	•		•	•		
Garcia et al. (1995)	•					
Ghanes et al. (2015c)	•	•				
Gorelick et al. (2005)	•					
Green et al. (2006)			•			
Hagtvedt et al. (2009)				•		
Hoot et al. (2008)	•	•		•		
Huang et al. (2012)	•	•				•
Jones and Evans (2008)		•				
Kelen et al. (2001)			•	•		
Khare et al. (2009)	•					
Kolker (2008)	•			•		
Komashie and Mousavi (2005)	•					•
Kuo et al. (2012)					•	
Lin et al. (2013)		•				
Mandelbaum et al. (2012)						•
McGuire (1994)	•					
Powell et al. (2007)	•		•			
Ramirez-Nafarrate et al. (2014)				•		
Roche and Cochran (2007)	•		•			
Rossetti et al. (1999)	•					
Saghafian et al. (2012)	•	•				
Saghafian et al. (2014)	•					•
Samaha et al. (2003)	•					
Sinreich et al. (2012)	•	•				
Song et al. (2013)	•					
Vilke et al. (2004)				•		
Wang et al. (2012)	•					
Wang (2013)	•					
Weng et al. (2011)						•
Wiler et al. (2013)			•			
Xu and Chan (2013)		•	•			
Yankovic and Green (2011)	•					•
Zayas-Caban et al. (2013)			•			•
Zetlyn et al. (2011)	•	•				•

2008). Employees are likely forced to discharge patients just before the LOS threshold and thus distort the clinical proprieties and put patient safety in jeopardy (Montimore and Cooper, 2007). In contrast to public objectives in terms of percentiles, most of the existing papers focus on the average of LOS.

LOS represents a combination of many different steps of the patient flow through the entire process from registration to discharge (Kolker, 2008). It gives an overview of the entire system performance. However, it does not allow to figure out some eventual local strengths and weaknesses. For instance, while considering two different systems or two different situations of the same system, the average LOS could be similar but with different duration combinations for the different stages of the process: DTDT, diagnostic tests durations, boarding time (duration between hospitalization decision to actual transfer), etc. Moreover, LOS is impacted by exogenous variables that are out of the control of EDs (Pines et al., 2012). Such external factors include the visit volume, the case mix (acuity level, age, specialty needed, etc.) and the hospital bed access known as the boarding effect problem (Forster et al., 2003). Boarding time is a key contributor to ED overcrowding worldwide. Yet, it depends on general wards (also called internal wards) of the hospital (Shi et al., 2014). Note that some papers such as Armony et al. (2011) exclude boarding time from the measure of LOS. Therefore, LOS should be used with caution, for example, when comparing performance between institutions (Olshaker and Rathlev, 2006). Moreover, It seems that a very small percentage of severe incidents have a major impact on the mean value of LOS observed in hospitals (LaCalle and Rabin, 2010; Freitas et al., 2012). These outliers show that it may be useful to add medians and percentiles in the statistical analysis of LOS (Ding et al., 2010).

Several papers in the literature focus on improving the ED performance in terms of LOS, using different quantitative methods such as queueing analysis, mathematical programming, dynamic programming and simulation.

Song et al. (2013) study the potential negative effects of queue pooling on ED performance. They focus on the Kaiser Permanente South Sacramento ED. The authors propose to modify the traditional pooling based triage. In the latter, nurses assign a severity index to patients. The highest acuity level patients proceed directly to the resuscitation room, acuity levels 2 and 3 are treated in the main area (main ED) and the lowest acuity level patients (levels 4 and 5) are treated in the fast track. In the main ED, patients wait in a pooled queue to be served by a physician from a pooled set of physicians under a first come first served (FCFS) discipline. Nurses are considered as resources shared by the different physicians. The new triage approach consists in assigning a patient to a physician-nurse team that work exclusively together. The

motivation for the new approach is that it could reduce social loafing and contribute to a more distributed utilization of shared resources. This new approach is tested on a sample of 234,334 patients. In contrast to results predicted by queueing analytical models, the study shows that moving from a pooled system to a dedicated system reduces LOS by about 9%. This corresponds to a reduction in LOS by 25 minutes for a medium severity patient served by a mean performing physician.

An intervention in triage is also addressed in Gorelick et al. (2005). The question is whether in-room registration of patients has an effect on their LOS, or not. The study is performed in a pediatric ED that serves annually approximately 45,000 patients. The authors suggest to apply an in-room registration process. In this way, patients are placed directly into a room after triage, and the registration process is completed after physician consultation. The results indicate that in-room registration has an effect on LOS, reducing it by an average of 18.6 minutes or 9.3%. The reader is referred to Oredsson et al. (2011) and references therein for studies focusing on triage-related interventions.

Wang (2013) addresses an ED staffing problem in two steps. In the first step, she focuses on the optimal scheduling of patients using a separated continuous linear programming approach. The author proposes an alternative way to examine the LOS. She divides the ED into 3 stages: the time spent in the waiting room, the period waiting for an examination and the time spent to see the physician again after the examination. Given that the treatment and the examination procedure durations cannot be reduced, the author considers an objective function comprising only the remaining parts of the LOS which can be minimized. In the second step, she focuses on the optimization of the ED staffing levels which minimizes the ED operating costs.

Rossetti et al. (1999) consider the ED of Virginia Medical Center, which has close to 60,000 visits per year. They use simulation for the problem of shift-scheduling of physicians. Their goal is to minimize the total LOS. They use four different scenarios as solution approaches. Scenario 1 is based on the ED manager experience and intuition. Scenario 2 is determined by the arrival rate (data collection process of 1,175 patients). Scenario 3 consists of adding an additional shift to the pre-existing shift-scheduling solution. Scenario 4 is the same as Scenario 2, but the changes can only be applied on weekdays. The results indicate that scenario 2 is the best option. By adding a physician in the peak hours (10 a.m. to 6 p.m.), the average LOS decreases by 14.5 minutes. In general in the literature, simulation methods have been widely used to evaluate possible alternatives to reduce the LOS. Some references include McGuire (1994); Samaha et al. (2003); Khare et al. (2009); Wang et al. (2012); Zetlyn et al. (2011), and references therein.

2.3.2 Door-to-doctor time (DTDT)

DTDT measures the time interval between the patient arrival to the ED and the first consultation by a physician. Triage and the waiting time for a doctor are part of the door-to-doctor time (Vegting et al., 2011). This is an important metric: "Reductions in "Door-to Doc" times are frequently at the forefront of ED quality improvement initiatives" (Jones and Evans, 2008). DTDT is also called *time to physician* (Wiler et al., 2010; Burström et al., 2012), time to first treatment (Saghafian et al., 2012; Chonde et al., 2013), or simply waiting time in some papers (Duguay and Chetouane, 2007; Ahmed and Alkhamis, 2009; Oredsson et al., 2011). The latter term is confusing since after the initial consultation, many other procedures also imply waiting times.

Low acuity level patients (levels 4 and 5) have a small probability of undergoing ancillary tests (Robinson, 2013). Their average DTDT is therefore in general close to their average LOS. For those patients, the two metrics can be used indifferently. In case of severe incidents, EDs must be able to respond immediately. Guttmann et al. (2011) report that mortality in EDs is particularly associated to the initial waiting time. DTDT is therefore one of the most significant metrics for critical patients, though it only characterizes a small part of the process and ignores the performance of other important care stages. Moreover, as highlighted by Wiler et al. (2010), patient satisfaction is strongly correlated to timeliness of care, with time to be seen by a physician having the most important association. It is common that EDs use DTDT targets that depend on the patient triage level. For instance, the Canadian government published its own acuity guideline in 1998 (revised in 2004 and in 2008) as shown in Table 2.2.

Table 2.2: DTDT target per triage level (Beveridge et al., 1998)

Triage level	Expected waiting time to see a physician
I: Resuscitation	Immediate
II: Emergent	<15 min
III: Urgent	<30 min
IV: Less Urgent	<60 min
V: Non Urgent	<120 min

Cooke et al. (2012) conduct experiments in an ED in the UK using a sample of 13,606 patients. They introduce a separate stream for minor accident injuries in order to reduce DTDT. The retrospective analysis is based on a 10 weeks trial, 5 weeks examined with the regular triage system and 5 weeks with the application of the new stream. The results show that the percentage of patients having a DTDT less than 30 minutes increases by 24.3%. Similarly, Cochran and

Roche (2009) use a split patient flow approach in order to improve the access to the ED. They set a DTDT target and then try to find which operations (e.g. bed capacity in each stage) are required in order to achieve the objective. Other related studies include Lau and Leung (1997); Miró et al. (2003); Subash et al. (2004).

Lin et al. (2013) focus on the allocation of ED and inpatient unit (IU) resources in a hospital. They develop a queueing model to estimate the waiting time of patients to access the ED as well as the necessary amount of resources to achieve the wait time targets for each priority class. This queueing model consists of two connected queues: one upstream queue for the patients entering the ED and one downstream queue for the patients transferred from the ED to the IU. It is reported that there is an optimal IU capacity and whenever the LOS in the IU or the arrival rate to the ED is uncertain, it is preferable to increase the resources in the IU rather than in the ED. Adding resources is also preferable whenever the IU LOS is fixed and the arrival of patients in the ED is fluctuating. Therefore, the DTDT strongly depends on the boarding effect, so, adding resources in the IU can be a way to optimize DTDT. Lin et al. (2013) also include the analysis of the benefits of a fast-track on DTDT. Although the total DTDT of patients is reduced, there is an increase in the waiting times of the high severity patients. Analytical results are verified through Monte-Carlo simulations.

Jones and Evans (2008) develop an agent based simulation model in order to evaluate different ED physician staffing schedules. The authors focus exclusively on DTDT and neglect other stages of the process like diagnosis tests. In order to determine whether the designed tool is capable of providing accurate estimations of DTDT, they compare the observed and simulated distributions using the non-parametric Wilcoxon rank-sum test. Using discrete-event simulation, Medeiros et al. (2008) develop and implement a new approach called *provider directed queueing* where an emergency physician is placed at triage. The method is similar to *team triage* (Subash et al., 2004; Burström et al., 2012) and the *Triage-Treat-and-Release* program that motivated the work by Zayas-Caban et al. (2013). It is applied on low risk patients (Emergency Severity Index -ESI- levels 3 to 5) during the busiest part of the day. The authors report a reduction of 35% in DTDT (from 93 minutes to 60 minutes). Other examples of studies using simulation to reduce DTDT include Connelly and Bair (2004); Duguay and Chetouane (2007); Laskowski et al. (2009); Ajami et al. (2011); Sinreich et al. (2012), and references therein.

2.3.3 Left without being seen (LWBS)

While waiting for the first examination by a physician, a patient can abandon and is then considered as LWBS. Patients decide to leave the ED because they consider that their waiting

time is too long compared to what they are willing to wait. Batt and Terwiesch (2015) explain LWBS by the fact that customers underestimate their actual waiting time.

Patients can abandon the ED at any time and thus similar abandonment metrics could be defined (when waiting for test completion, when waiting for admission bed, etc.). However, the large majority of research focuses on LWBS. The so called *walkaway* patients do not only present a safety issue but also contribute toward lost revenue. According to Carmen and Van Nieuwenhuyse (2014), ED occupancy and waiting times are the main factors that influence LWBS rates. DTDT is a primary driver for patients LWBS (Rowe et al., 2006; Cochran and Roche, 2009). The most effective way to decrease LWBS is rapid assessment which means the reduction of DTDT (Batt and Terwiesch, 2015; Fernandes et al., 1997; Welch, 2009).

The rate of LWBS differs widely from a triage level to another but it also varies with countries, regions (the access to other care facilities in the area), social levels, ages and the day of the week. For this reason, the percentage of LWBS is considered as a bad metric to compare between the performance of different EDs. Table 2.3 provides an illustration on the percentages of LWBS as reported in the literature. Empirical analyses of abandonment are often confounded by censored or missing data (Batt and Terwiesch, 2015). Unfortunately, it is very difficult to measure times before leaving. The staff realizes that a patient has left only when this patient is called by the nurse to be brought to the examination room.

In an ED in Torrance, California, Baker et al. (1991) state that about the half (46%) of the patients that are LWBS were judged to require immediate medical attention, and about the third (29%) of them would require medical care within one or two days. Therefore, it is important to understand and characterize abandonments. The system manager may then control the capacity to deliver better quality of service to patients. One way is to reduce the waiting time before the first examination to the detriment of the waiting durations for next stages.

Other studies address the analysis of the factors influencing the LWBS feature and try to propose solutions. Skaikh et al. (2012) focus on how long patients LWBS would be willing to wait for the first examination. The results show that half of the patients were willing to wait up to two hours. Concerning psychological responses to waiting, prior literature has generally found that people are more willing to wait when they are kept informed of why they are waiting and how long the wait will last (Hui and Tse, 1996; Batt and Terwiesch, 2015). By interviewing LWBS patients, Arendt et al. (2003) report that 85% of them would have liked to be updated over time on how long they should have to wait, and 70% would have preferred an immediate temporary treatment.

Green et al. (2006) consider an urban hospital in New York with about 25,000 patients

Table 2.3: Data on LWBS

Reference	Country	Sample	% of LWBS
Stock et al. (1994)	USA	92,570	4.20
Arendt et al. (2003)	USA	20,494	0.83
McMullan and Vesser (2004)	USA	18,664	3.37
Vieth and Rhodes (2006)	USA	11,743	9.00
Pitts et al. (2008)	USA	119,191,000	2.03
Johnson et al. (2009)	USA	11,147	1.10
Pham et al. (2009)	USA	289,079	1.70
Guttmann et al. (2011)	USA	13,934,542	4.43
Batt and Terwiesch (2015)	USA	180,000	6.50
Fernandes et al. (1994)	Canada	23,933	1.40
Monzon et al. (2005)	Canada	10,808	3.57
Rowe et al. (2006)	Canada	15,660	4.50
Mohsin et al. (2007)	Australia	14,471	8.60
Tropea et al. (2012)	Australia	1,829,854	11.23
Liao et al. (2002)	Taiwan	74,485	0.10
Goodacre and Webster (2005)	UK	76,843	7.20
Armony et al. (2011)	Israel	>1,000,000	3-5
Parekh et al. (2013)	Guyana	3,027	5.70
Grosgurin et al. (2013)	Switzerland	57,645	4.18
Fayyaz et al. (2013)	Pakistan	38,762	13.12

per year. Using a queueing modeling, they study the shift-scheduling problem, in order to improve the ED efficiency in terms of the percentage of LWBS. They propose an alternative staff scheduling, implement it during 39 weeks, and compare the results with those during a previous 39 weeks period. Although the number of admissions increases between the two time periods by 1,078 patients, the number of patients LWBS decreases by 258 units. The probability to abandon then decreases by 22.9%.

Batt and Terwiesch (2015) perform an empirical study of queue abandonment in EDs. The authors use a detailed time-stamp data of 180,000 patient visits in order to examine the queue abandonment behavior of patients. Many papers address the problem of queue abandonment in many areas but there is still a limited empirical work studying how queue status information affects customers. This paper focuses on the impact of what patients observe and experience during their wait on their abandonment decisions. The authors consider that waiting patients observe and consider two types of variables: stock variables (such as the total number of patients, the total number of patients with a higher priority, or the total number of patients with a later arrival time) and flow variables (such as the number of arrivals and departures in the last hour). The study provides useful insights on patient abandonment behavior. For instance, the observed flow of patients in and out of the waiting room has an effect on abandonment, with

arrivals leading to increased abandonment and departures leading to decreased abandonment. It is also shown that patients respond differently to the flow of more and less severe patients. Thus, allocating separate waiting rooms for different triage levels may reduce abandonment.

2.3.4 Ambulance diversion (AD)

Ambulance diversion (AD) consists in re-routing ambulances from the closest ED to other neighboring EDs if they are willing to accept additional patients. The overcrowded hospital declares his "diversion status" to the local emergency medical service (EMS) who advises ambulances on better destinations. AD is an ED operations practice which is commonly used in North America while its application in Europe is still rare. AD is a coordination policy that aims to balance capacity and demand within a network of EDs (Do and Shunko, 2013; Deo and Gurvich, 2011). The most common reasons for diversion are high number of patients, no appropriate facilities or trained personnel (e.g. scanner, neurosurgeons, etc.) and no appropriate inpatient beds (Allon et al., 2013; Pham et al., 2006). Burt et al. (2006) report that the number of ambulances diverted each day in the U.S could be as high as 1,886 with almost half of all EDs (44.9%) experiencing ambulance diversion periods.

Kolker (2008) indicates that the percent of time when ED is on diversion is an important performance indicator. The importance of AD can be underlined by examining the case of a high severity patient that requires immediate treatment. If the closest hospital to this patient is applying AD, then the crucial transportation duration to an ED will be increased. However, using AD as a KPI might not be straightforward for all situations. AD is a useful measure in the case of large cities where several EDs are usually available which allows the re-routing of ambulances. Yet, for small cities, where only one ED is available, such a practice is not likely applicable (Ospina et al., 2006; Allon et al., 2013).

Allon et al. (2013) study the impact of size and occupancy of the hospital (inpatient and emergency departments) on the extent of AD. They propose a two-station queueing model to describe the patient flow between the ED and the IU. Using diffusion and fluid approximations, the analysis leads to the following results: *i)* The capacity of the IU is negatively correlated to AD hours; *ii)* The threshold of unused beds below which the ED applies AD policy is positively correlated to the fraction of time spent on diversion. Increasing the threshold of unused beds would harm other performance indicators, such as LOS and DTDT; *iii)* The fraction of time on diversion increases with the number of hospitals in its neighborhood.

Other studies focus on the AD metric without dealing directly with internal ED operations. They focus on the relationship between different EDs. Hagtvedt et al. (2009) focus on cooperative

strategies as a way of reducing AD. Their study uses different approaches, such as Markov chains, simulation and game theory. The authors show that agent-based simulation is unlikely to be applicable in reality. The game theory approach explains, in turn, how cooperation between hospitals can be achieved. The prisoner's dilemma for only two hospitals shows that each hospital will try to divert ambulances before it would have actually need to. This characterizes the existing rivalry and individualism between hospitals within an AD cooperation. Hospitals are therefore forced to adopt solutions that are non-optimal. It is natural that a larger number of hospitals will face even greater difficulties when trying to manage diversion. The authors conclude that the cost of diverting ambulances should be adequately high in order to promote cooperation between different EDs. The study also suggests that a centralized planner agent (i.e., EMS) is necessary to enable regulation of AD strategies between providers.

Vilke et al. (2004) propose an approach to decrease AD hours. Their study focuses on two neighboring hospitals (A and B) in San Diego, California, that serve in total about 84,000 patients per year. Since the two EDs are the only ones in a radius of 5 miles, there can be a correlation between their AD hours. The authors observe that whenever Hospital A goes on diversion, Hospital B diverts ambulances after a small period of time. This observation motivates them to study the performance of each hospital for 3 weeks. In the first and third weeks the already available resources are used in each hospital, whereas during the second week, Hospital A works with supplementary resources in order to eliminate ambulance diversion phenomena. The results show that both EDs manage to reduce the hours of ambulance diversion, even though Hospital B did not increase its capacity during the experiment.

Deo and Gurvich (2011) compare between centralized and decentralized AD from a network perspective. They use queueing and game theory in order to develop a model that explains the difference between the two AD methods. As seen previously in Allon et al. (2013), hospitals apply AD under a threshold policy on the number, say K of occupied IU beds. In the decentralized method, each ED aims to maximize its own utility function, where the optimization refers to the reduction of the waiting time for each ED separately. The game shows that the optimal solution is to set $K = 0$ for each ED, meaning that they would signal that they are always on diversion. However, legislation has set the ADND (All on Diversion, Nobody on Diversion) guide, which stipulates that whenever all hospitals are on diversion, then the initial itinerary of each ambulance will be maintained. Using decentralized AD, EDs try to optimize their utility function separately, preventing the pooling benefits that AD could lead to. Contrariwise, the centralized method consists of a decision made by a social planner on when each ED should go on diversion. This method thus optimizes a holistic utility function, as it accounts for pooling

effects. The authors provide numerical experiments illustrating that centralized AD is preferable.

An AD optimization study is addressed in Ramirez-Nafarrate et al. (2014). The authors study optimal AD control policies using a Markov decision process (MDP) formulation that minimizes the average expected tardiness of care. Tardiness of care is defined as the time that patients wait beyond their recommended safety time threshold (RSTT). In other words, the objective is to minimize the average non-negative difference between DTDT and RSTT. The model assumes that the time to start treatment at the neighboring facility is known. The authors show that the optimal AD policy follows a threshold structure, and explore the behavior of optimal policies under different scenarios. They analyze the value of information on the time to start treatment in the neighboring hospital, and show that optimal policies depend strongly on the congestion experienced by the other facility. Using a discrete-event simulation model under more realistic assumptions, they demonstrate that the optimal policies obtained using the MDP model outperforms the simple heuristics used in practice.

2.3.5 Combination of KPIs

Each one of the above KPIs provides a particular and restricted information on performance. Different KPIs can be then combined to complement one another. According to the way that KPIs are used, two categories of papers can be distinguished: a "descriptive use" and a "proactive use". The first category refers to papers where KPIs are only used to measure and assess the effect of some introduced changes in the ED (Abo-Hamad and Arisha, 2013; Duguay and Chetouane, 2007). It generally consists of medical papers with empirical experiments, and also in simulation studies using intuitive what-if scenarios. The second category includes papers where KPIs have a central role in the optimization model, expressed in the objective function or in a constraint (Saghafian et al., 2012; Ghanes et al., 2015c). The proactive use is often found in analytical and simulation-based optimization studies. In general, we can state that a combination of KPIs is actually performed only when they are used in a proactive way.

Descriptive use of KPIs. Abo-Hamad and Arisha (2013) develop an interactive simulation-based decision support framework to improve planning and efficiency of healthcare processes in a large university Hospital in Dublin. The model is used to investigate the impact of decisions and alternatives (i.e., what-if scenarios) on system performance. Scenarios were developed by varying both human and space capacities and by introducing a new policy where patients are dismissed when the LOS exceeds 6 hours. The comparison between seven different scenarios using an important set of KPIs (LOS, LWBS, DTDT, resource utilization, etc.) provided hospital managers with helpful insights on the appropriate strategies to adopt.

Hoot et al. (2008) use a sample of 13,248 patients in order to forecast ED crowding. The authors try to predict several KPIs using simulation. Their aim is to conclude whether this simulation can lead to valuable forecasts. The results show that the level of accuracy of predictions is not the same for different KPIs. Boarding time forecasts are not accurate, as the model seems to predict less hours of boarding. Nevertheless, predictions for the remaining indicators, such as LOS, DTDT and AD, seem to be more reliable, especially for a prediction made up to 4 hours ahead. As expected, they also show that the nearest the forecasted period is, the more accurate the simulation results are. Forecasting the workload in an ED has been the subject of several studies. Some references include Wargon et al. (2009); Chase et al. (2012); Xu and Chan (2013); Plambeck et al. (2014). Accurate forecasts are important in particular for staffing optimization problems. Related references include Xiao et al. (2010); Al-Najjar and Husain Ali (2011); Yankovic and Green (2011); Green et al. (2013).

Kelen et al. (2001) conduct a ten-week experiment on a sample of 10,871 patients. The authors introduce a supplementary acute care unit (ACU) that serves the most severe incidents of the ED. The ACU is entirely supplied with resources of the ED and serves for several procedures, such as primary evaluation and admission processing. Results show that their intervention reduces LWBS by about 5% and AD by about 4 hours/100 patients·week, compared to two weeks prior to the study.

Burström et al. (2012) use data from 3 big Swedish EDs (147,579 patient records), with different patient receptions. The authors use a statistical analysis to study the effect of staff allocation in triage on LOS, DTDT and LWBS. They compare the three different triage models: *i)* Physician-led team triage, where the physician is the head of a smaller team that consists of a junior doctor and a nurse; *ii)* Nurse/Emergency physician triage, where a nurse performs triage and an emergency physician deals with the patient treatment; *iii)* Nurse/Junior triage, where a nurse performs triage and a junior physician examines the patient. The authors show that Physician-led team triage significantly outperforms the other alternatives. The average DTDT improves, compared to the second and third scenarios, by 56.5% and 49.5%, respectively. The average LOS improves by 15.7% and 3.1%, respectively. Also, the rate of patients that LWBS was 3.1% for physician-led team triage, 5.3% for nurse/emergency physician, and 9.6% for nurse/junior physician triage. Similarly, Han et al. (2010) propose triage performed by a physician. Their intervention leads to improvement in LOS, AD and LWBS.

Intervention in triage is also the subject in Chan et al. (2005). The authors propose a method called "ED REACT". In the latter, physicians are able to initiate the treatment of a patient (e.g. laboratory examinations), even if no beds are available. The study uses statistical analysis of

two six-month periods (pre-REACT and post-REACT) in order to demonstrate the effect of the intervention proposed. The results show significant improvement in 3 KPIs: LWBS is reduced by 3.2%, DTDT and LOS are reduced on average by 24 and 31 minutes, respectively. Therefore, their intervention manages to increase the number of patients treated (reduced LWBS) and simultaneously to improve performance in terms of waiting durations.

Proactive use of KPIs. In the context of a highly congested ED, Huang et al. (2012) address the question of whether the physician should choose a patient that will be assessed for the first time (right after triage) or a patient that has been already seen by a doctor and returns to her after the completion of an examination (in-process, IP). The objective is to minimize a waiting cost function (related to LOS) subject to deadline constraints for triage patients (related to DTDT). The authors prove that a threshold policy that selects between the two types of patients is asymptotically optimal. As a case study, the authors consider a context with additional elements of advanced triage, such as the prediction of whether a patient will be admitted in the hospital or discharged. They compare between three levels of information: no information, partial information (where only the number of in-process phases is known) and full information (also the patient outcome is predicted). The results show that partial information and full information improve the objective function by 18% and 27%, respectively. Other papers in the literature address similar decision-making issues but for slightly different KPIs (Zayas-Caban et al., 2013; Dobson et al., 2013). For example, in order to penalize patient abandonment, Zayas-Caban et al. (2013) attributes rewards for completing each phase of service, while no rewards are perceived for patients who abandon the system. The authors in Dobson et al. (2013) analyze, in turn, the throughput optimal workflow decisions.

Saghafian et al. (2012) propose patient streaming as a mechanism for improving responsiveness in EDs. They use a combination of analytic (MDP) and simulation models to analyze this streaming policy. The authors focus on patients with ESIs 2 and 3 (which account for approximately 80% of all patients). They introduce a supplementary triage element: a prediction of whether a patient will be admitted in the hospital or not after the ED. Patients that will be admitted in the hospital (A patients) require a small DTDT, as safety is the most important factor for severe cases and they thus need a quasi-immediate medical treatment. On the other hand, patients that will be discharged (D patients) after the treatment in the ED typically require a small LOS. The authors compare between the three following policies : *i)* Simple pooling, where all patients form one single group of people waiting for treatment; *ii)* Streaming, where patients are separated into two groups depending on the prediction of admission; *iii)* Virtual streaming, which is streaming without the practical constraints of the separate paths (for example available

resources such as physicians or beds of one path can be used in order to serve the other path in case of high demand). The authors conclude that although pooling is more efficient than streaming, virtual streaming is the best method. Virtual streaming allows to balance the need for low DTDT for A patients and low LOS for D patients in a better way than pooling does.

Using data from an academic hospital, Saghafian et al. (2014) propose a complexity-augmented triage as a way that can improve patient safety and increase operational efficiency. This triage method was earlier discussed in Hopp and Lovejoy (2012). In the proposed triage method, the patient path depends on both the patient urgency and complexity. Saghafian et al. (2014) focus on LOS and the risk of adverse events (ROAE). The latter is related to DTDT since the probability of having an adverse event is much higher before the initial consultation rather than while waiting for examination results. Using an MDP approach, the authors develop a threshold policy that determines the optimal patient selection by physicians. Using simulation, they also show that their method improves LOS and ROAE by 21.3% and 18.0%, respectively.

Xu and Chan (2013) propose a pioneering approach of diversion applied to patients that are visiting the ED as walk-ins (not being brought by an ambulance). They propose a proactive method including walk-in diversions, based on arrival predictions. They study the threshold on waiting patients above which the ED should apply the new diversion policies. The model focuses on DTDT, and uses diversion as a control variable. The trade-off between diversion and LWBS is further analyzed. The authors state that patients that are not examined in the ED are the ones that are diverted and the ones that abandon (LWBS). With the application of their method, the sum of diverted and LWBS remains the same while the average DTDT of patients is reduced by approximately 8%. Using MDP and simulation, Helm et al. (2011) propose a framework for improving the patient flow in the hospital. It consists of controlling admission by postponing scheduled admission when the ED is highly occupied, and treating them when the ED is less occupied.

2.3.6 Other KPIs

We give here further metrics that have received little attention in the ED literature, but hold however a growing importance in the ED medical literature or also in practice.

Measure of crowding. The common used KPIs in the OR/OM literature (LOS, DTDT, etc.) do not quantify strictly speaking crowding, though an evident correlation with crowding. Improving these indicators does not necessarily mean reducing overcrowding since one KPI could be improved on the detriment of another. For this reason, multidimensional scores were developed by experts in order to measure the degree of congestion, the most important of

which are the National ED Overcrowding Scale (NEDOCS) and the Emergency Department Work Index (EDWIN). "Although emergency physicians have an intuitive sense of when an ED is becoming crowded, before EDWIN and NEDOCS, there was no universally accepted quantitative index of ED crowding" (Bernstein et al., 2003).

EDWIN has been shown to be correlated with impression of crowding by doctors and nurses (Weiss et al., 2004). NEDOCS is more commonly used by the medical industry (Weng et al., 2011) and is calculated with a linear regression model that associates several operational variables (waiting time, amount of sickbeds, number of hospitalized patients, total number of patients, etc.) with the degree of crowding assessed by physicians and nurses. It is a simple tool that can be used easily and quickly to determine the degree of overcrowding at an academic institution (Weiss et al., 2004). The higher is the value of this variable, the higher is the degree of congestion. A NEDOCS score above 100 means a crowding state. A NEDOCS score under 100 means that ED is below the congestion level (Weng et al., 2011).

In the OR/OM literature, the use of NEDOCS and EDWIN is rare. An exception is Weng et al. (2011). Using simulation-optimization, the authors address the problem of resource allocation in EDs considering NEDOCS as a metric. The analysis shows that a new resource allocation can improve the NEDOCS value from 126.79 to 116.63.

Fairness. Fairness (justice and equity are alternative terms) in an ED is related to both patients and employees. The ability to secure fairness between patients and between employees might stand as an alternative way to improve efficiency (SoRelle, 2002). In contrast to the employee perspective, fairness has been extensively studied from the patient perspective. The reader is referred to Tseytlin (2009) and references therein for papers related to patient fairness. Tseytlin (2009) investigates different configurations such as a single queue versus multiple queues or FCFS versus other queueing disciplines. In general, fairness between patients has been widely addressed in the literature based on the logic that the most severe cases must be prioritized, which has produced a number of triage methods, such as Canadian Triage and Acuity Scale (CTAS), Manchester Triage System (MTS), the Emergency Severity Index (ESI), Australian Triage scale (ATS), etc. It is agreed that FCFS policy is essential for justice perception within a queue, i.e., a triage category. Since clinical priority dominates FCFS justice, waiting in the multi-queueing ED system produces a sense of lack of fairness, even though prioritization of a queue over another is justified (Mandelbaum et al., 2012). Batt and Terwiesch (2015) propose to allocate separate waiting rooms for different triage levels in order to reduce patient abandonment.

The objective of fairness from an ED staff point of view means that each nurse/doctor should have similar workload as others (Mandelbaum et al., 2012). Unfair policies toward staff could

internalize inefficiencies (Mandelbaum et al., 2012) because faster servers work more, which gives them an incentive to slow down - an undesirable result for the overall system. Using data from a large Israeli hospital with approximately 75,000 patients hospitalized yearly, Armony et al. (2011) conduct an empirical research and discuss fairness toward staff. High workload tends to cause personnel burnout especially if the routing of patients is perceived as unfair. The authors demonstrate that the most efficient resources are subject to the highest load. Based on data from the same hospital, Mandelbaum et al. (2012) study the fair routing of patients from an ED to internal wards. The incentive for this study stems from data observations: one of the five wards of the hospital was experiencing a very high patient per bed ratio compared to the other four wards. This deviation is explained by the difference in efficiency between employees. Using a queueing analysis with heterogeneous server pools, where the pools represent the wards and servers are the beds, the authors propose routing policies in order to minimize the deviation of work rate between employees. Note that fairness toward staff could alter operational efficiency, because routing jobs to the fastest capacity is better. This is obviously unfair toward the fast care providers (which get "punished" for being fast by working more) (Mandelbaum et al., 2012).

2.4 Discussion

We summarize the key points analyzed in the survey. We also highlight some limitations encountered by researchers like data collection issues, and suggest possible future research opportunities. Table 2.4 briefly summarizes the main ideas for the selection of relevant KPI.

LOS and DTDT are the two most used KPIs in the literature. LOS is the most used in practice because it provides to managers an overview of the entire system performance. However, it does not allow to figure out eventual local strengths and weaknesses of the system. LOS depends strongly on the patient mix (acuity level, age, specialty needed, etc.). Thus, it should be used with caution when comparing the performance between institutions (Olshaker and Rathlev, 2006). DTDT is one of the most significant metrics in EDs since it is the most associated to patient satisfaction and is correlated with the mortality rate of critical patients. However, this KPI measures the performance of a small part of the process while ignoring other important stages. Thus, as done in Saghafian et al. (2012) and Ghanes et al. (2015c), combining these two KPIs with the use of the overall average LOS and the DTDT for urgent cases seems relevant. Note that for non-urgent patients, DTDT or LOS can be chosen indifferently since these two values are relatively close on average for this type of patients. However, DTDT is a primary driver for patients LWBS, which could make its use also relevant for non-urgent patients,

Table 2.4: KPI selection

KPI	Recommended	Risks
LOS	- Overview of the entire system performance - The average is the most commonly used - Could be combined with DTDT	- Do not allow to figure out local strengths and weaknesses in the ED process - Setting an LOS target might lead to perverse effects - Use with caution when comparing performance between different institutions - Using the average could deteriorate the performance for critical patients
DTDT	- Critical patients - With the objective to reduce LWBS (dependency) - Linked to patient satisfaction	- For low acuity levels, LOS or DTDT can be used indifferently - Ignores the performance of other important ED stages
LWBS	- To improve both patient safety and ED revenue - Use DTDT as a lever - People are more willing to wait when they are kept informed	- Differs widely from a triage level to another - Varies with countries, regions, social levels, ages and even the day of the week - A bad metric to compare between the performance of different EDs - Time before leaving is difficult to measure in practice
AD	- In the case of large cities where several EDs are usually available	- Commonly used in North America but rarely in Europe - Not applicable in small cities when ambulances have only one alternative
Multidimensional scores	- To measure the degree of crowding	- Validated with the subjective sensation of crowding felt by the ED staff
Fairness	- Equity between employees/patients	- FCFS acuity-based rule could produce a sense of lack of fairness between patients - Applying fairness toward staff could harm the system operational efficiency

when LWBS is an issue for the ED management.

LWBS and AD have also been studied extensively, but in a smaller extent relatively to DTDT and LOS. The patient abandonment time and rate vary with countries, regions (the access to other care facilities in the area), social levels, ages and even the day of the week which makes the rate of LWBS a bad metric to compare the performance between different EDs. AD is a useful measure in a large, inner-city institution, but of no value to a regional hospital that is the only choice for ambulance personnel (Ospina et al., 2006). Moreover, this KPI is quite common in North America but rarely used in European countries.

The importance of universal measures of ED crowding like EDWIN and NEDOCS should be highlighted. They represent the first standardized scale developed to determine whether an ED is overcrowded or not (Bernstein et al., 2003; Weiss et al., 2004). They are calculated by converting a simple data set into a score that correlates accurately with the degree of overcrowding as perceived by the staff. Although EDWIN and NEDOCS are extensively addressed in medical papers and increasingly used in practice, their use remains rare in the OR/OM literature. The

validation of the multidimensional scores is mainly made through comparison with the subjective sensation of crowding felt by the staff. We suggest, as an avenue for future research, the validation of these scores using OR/OM tools, such as simulation.

ED performance is affected by external factors (Shi et al., 2014) that are out of control like visit volume, case mix, interactions with internal wards and other institutions (other hospitals and emergency medical services), etc. For instance, boarding time which is defined as the time from admission order to departure from the ED (Olshaker and Rathlev, 2006), and diagnostic test time are some of the longest stages in the process and yet depend on other services of the hospital. EDs are compared in practice directly through common KPIs as addressed in this chapter. This information is useful to the public to compare quality and is useful to payers to reward better performance. Given the strong impact of external factors on ED performance (Forster et al., 2003), the results of such comparisons should be considered with caution. There is a real need to create more appropriate measures that consider exogenous factors and allow a fairer comparison between EDs (Pines et al., 2012).

Concerning the KPIs used, there are two main issues that require further investigation. The first issue is focusing on domains that have not yet been extensively investigated, such as walk-in diversion and fairness. More specifically for the latter, Mandelbaum et al. (2012) propose a pioneering approach on increasing productivity levels of employees. Therefore fairness can be a tool that might be taken into consideration in resource optimization models. The second concerns the importance of combining KPIs. Focusing on one single metric might harm other important metrics of the system.

According to the issue being addressed, the OR/OM literature can be divided into three types: resource-related, process-related and environment-related experiments. Existing studies focus recurrently on the two first categories of issues (resource optimization and improvement of the patient path). Concerning resource optimization, simulation is the tool that is mostly used. The improvement of the patient path is mainly a result of a modification in the process and analytical methods are generally used. They rely, in particular, on queueing and Markov models. Analytical models require a set of hypotheses and represent simplified versions of the ED while simulation models can capture most details of the system without requiring major hypotheses. However, simulation results are in general "tailor-made" solutions that are useful only for the system examined and could not be generalized to others (Paul et al., 2010), whereas analytical models are more convenient to provide general guidelines. The literature that proposes process-related interventions mentions that the proposed methods might by biased. New protocols might indeed face resistance to change by employees that could prefer a convenient existing method

for them rather than a new one that improves the system performance (van Dyke et al., 2011; Jahangirian et al., 2015).

We also identify a growing stream of empirical studies that are published in OR/OM journals. The interest in field experiments stems from the necessity to better understand human behavior aspects, from the patient and ED staff perspectives. In the existing literature, the problem of data collection was often mentioned (Armony et al., 2011), as it is difficult to collect data in such a complex system. For example, it is feasible to count the number of patients LWBS, as it is the difference between triaged and examined patients. However, it is rather difficult to collect data recording when these patients had left the ED and for what reason. Processing times can be collected using on-site observations but this method is also difficult and time consuming. Therefore, researchers generally make assumptions on missing data. The problem of data scarcity is a neglected area in the literature with the exception of some papers. Kuo et al. (2012) propose a method to estimate the distribution of simulation parameters when data are incomplete. Green et al. (2007) focus on the estimation of abandonment times in call centers. They propose a method to estimate them using a hazard-rate function. The method can be applied to emergency departments where data are also censored. Data about arrivals are relatively easy to collect since they are often recorded in databases. Nevertheless, the ED arrival pattern that varies with the day of the week, the hour of the day and even the period of the year is often simplified which could compromises the robustness of the obtained solutions in practice.

Finally, there are some studies that include prediction made by ED staff based on their experience (e.g. admission prediction). The quality of prediction is decisive for the result of the method. Therefore, researchers could use decision making criteria in order to study the threshold of accuracy of predictions above which prediction is worth using. The above could be useful for studies such as Burström et al. (2012), Song et al. (2013) and Saghafian et al. (2014).

2.5 Conclusion

This chapter reviews the literature on the commonly used key performance indicators of emergency departments from an operations research and operations management perspective. It explores their characteristics as well as their selection approach. The study reveals that each KPI is used to measure specific ED aspects, and thus the choice of the appropriate KPI to be optimized is important. It also highlights the value of combining complementary KPIs to provide relevant solutions in practice. Finally, this chapter underlines some lacks in the OR/OM literature of studies related to fairness, universal measures of crowding, etc.

Chapter 3

Statistical analysis of factors influencing crowding

In this chapter, the primary aim is to investigate which factors currently contribute to overcrowding and LOS longer than four hours in emergency departments. The second purpose is to deduce appropriate remedial measures that would alleviate the influence of these factors on delays. We use detailed data from two hospitals in the Netherlands to examine statistically which factors contribute to a longer stay in EDs. The study reveals that multiple factors lead simultaneously to longer delays, and that ED congestion is a multifactorial phenomenon. A thorough interpretation of results is conducted in order to highlight the factors leading to the obtained dependencies (between ED performance and the assessed variables) in practice. We also provide for each influencing factor the corresponding relevant remedial measures (interventions) existing in practice and in the literature. The conclusions of this chapter and the research avenues that are derived represent common concerns that could be generalizable to the French context. This chapter provides a basis to define the issues (or confirm their relevancy) that will be addressed in the next chapters.

The paper version of this chapter is published in *The Netherlands Journal of Medicine* (Vegting et al., 2015).

3.1 Introduction

Overcrowding and long emergency department (ED) completion times can occur when the maximum available care capacity does not meet increasing demands. As explained in Chapter 2, length of stay (LOS) is a key measure of ED throughput and a marker of overcrowding (Yoon et al., 2003; Trzeciak and Rivers, 2003). It has been demonstrated that long stay on the ED was associated with negative outcomes, such as increased risk of hospital admission within seven days and in-hospital mortality (Hong et al., 2013), preventable medical errors, poor pain control, longer hospital stay and decreased patient satisfaction (Liew and Kennedy, 2003; Hwang et al., 2006). Therefore, optimizing ED patient flow is an important and frequently discussed topic.

In the past, increased congestion with long waiting times in emergency departments (EDs) in the United Kingdom (UK) was frequently noticed (Audit commission, 2001). With the aim of reducing this congestion, the National Health Service in the UK set a target which prescribed that 98% of patients presenting at the ED should be examined, treated, admitted or discharged (LOS) in less than four hours (Locker et al., 2005; Mayhew and Smith, 2008; Izady and Worthington, 2012). This resulted in a tremendous improvement in the LOS. Although congestion with long waiting times is frequently noticed in some EDs in the Netherlands, no target for LOS is defined or enforced. At the VU University Medical Center (VUmc) of Amsterdam, an academic tertiary care center, and St. Anotonius hospital, a large community hospital in Nieuwegein, it was noticed in the past years that the LOS exceeded four hours in many patients. However, reasons for these delays were unclear and the exact percentage of patients spending more than four hours in the ED was unknown (Vegting et al., 2011).

In this chapter, the primary aim is to examine statistically which factors currently contribute to overcrowding and LOS longer than four hours in EDs. The second purpose is to discuss and explain the practical causes leading to these correlations, and then to identify appropriate remedial measures that would alleviate the influence of these factors. This analysis serves as a basis for the definition of the issues that will be addressed in the following chapters. Among numerous measures that we deduce from the analysis, we de not retain, for the rest of this thesis, those going beyond the scope and the responsibility of the ED. This kind of measures, which we call external or environment-related measures, is likely to involve external actors (other services of the hospital, other EDs, etc.) that may have a contradictory interests with the ED, and thus jeopardize the success of the measure. Instead, we will examine in the next chapters some internal measures that could be implemented autonomously by ED managers. In practice, administrative data and observational studies generally does not provide sufficient information. For example, in an ED, administrative data might track a patient total length of stay and basic

patient information, but might not include detailed time-stamps because it is intrusive and time-consuming (Campello et al., 2013). This is the case for our French ED collaborator (Saint Camille). We decided to conduct the study on the two different above-mentioned hospitals in the Netherlands because they managed to collect data with a rare level of detail, and in order to obtain generalizable insights. Note that the perverse effect of the four-hour target (mentioned in Section 2.3.1) is not a concern here because it is not employed by the two studied EDs in practice. It is solely used as a reference variable in our statistical analysis.

The rest of the chapter is structured as follows. In Section 3.2, we present the study design. We describe the two studied EDs and how data were collected, we introduce the statistical tests used in the next section and provide some useful definitions for the rest of the chapter. In Section 3.3, we conduct a statistical data analysis to identify which ED variables have a dependency with ED length of stay. We interpret these results and discuss them in Section 3.4 in order to identify the practical inefficiencies and problems corresponding to these variables from the one hand, and which remedial measures could be undertaken to address them from the other hand. We conclude in Section 3.5.

3.2 Materials and methods

3.2.1 Emergency departments description

This prospective study was performed in the EDs of the VUmc and St. Antonius Hospital. VUmc is an academic urban level 1 trauma centre in Amsterdam with approximately 29,000 ED visits per year. During the study period, there were 11 residents in emergency medicine, including seven fellows of emergency medicine and four non-trainees working in shifts. Residents were supervised by four qualified emergency physicians (EPs) and one surgeon. At the ED of the VUmc, all patients presenting themselves without a referral from a general practitioner are seen by emergency medicine residents and qualified EPs. Depending on the needs of the patient, the EP can consult the medical specialists. If a patient needs more specialized care or needs to be admitted to the ward, the necessary specialism is consulted and the patient is handed over to the specialist for further treatment. Referred patients are seen by (non) trainee residents of various medical specialities under the supervision of medical specialists belonging to the particular department. St. Antonius Hospital is a large community medical center with approximately 23,000 ED visits per year. There were seven trainee residents in emergency medicine working in shifts. Non-referred patients were seen by EP residents and supervised by qualified EPs and referred patients were seen by residents of a specific speciality supervised by

the medical specialist. However, senior EPs were able to admit a patient for a specialism directly to the ward after a phone consultation with the specialist on call.

3.2.2 Selection of participants and data collection

In the VUmc, the study was conducted during a four-week period from 8 October until 4 November 2012. At St. Antonius Hospital, this was divided into two periods of two weeks each from 21 November until 5 December 2012, and from 11 February until 24 February 2013. For all patients visiting the ED in these aforementioned weeks, the following time moments were registered: ED arrival, triage, first contact with a physician, and discharge from the ED, in addition to information on triage level, type of referral, ordering of radiological and diagnostic testing, discharge disposition, first and last consulting medical speciality and the total number of consultations. At VUmc, these data were extracted from paper forms filled in by nurses and physicians. At St. Antonius Hospital, data were retrieved from a computer system called Intracis. In addition, other relevant data were collected by trained observers (medical students under the supervision of an internal medicine resident and a specialist) on a selected sample of patients older than 18 and triaged to Emergency Severity Index (ESI) level 2 or 3 at VUmc, and Manchester Triage System (MTS) category orange or yellow at St. Antonius Hospital. This selection was based on the previous measurement, demonstrating that these categories had longer LOS. The additional data collection included timestamps for the ordering, conduction and evaluation of radiological and diagnostic testing and the request, conduction and ending of a medical consultation. Also data on the time physicians arrived at their final diagnostic conclusions on the ED and when the nurses were informed that the patient could leave the ED were noted.

Note that the triage systems of hospitals were different, which can introduce bias. However, in the Netherlands both triage systems are frequently used and are largely comparable in determining the severity of the condition of the patient. Furthermore, the measuring period was not at the same time in the two hospitals. Seasonal influence may alter the situation. However, the benefit of measuring in both hospitals one after another is that we had the same team of researchers, using the same technique during both study periods. The main characteristics of all patients in both hospitals are summarized in Table 3.1.

3.2.3 Definitions

Door-to-doctor time. We defined door-to-doctor time as the time that elapsed between registration and the first visit of a physician. Triage and the waiting time for a physician are

Table 3.1: Patient characteristics

	Site, No. (%)					
Variable	**VUmc (n=2,272)**			**St. Antonius Hospital (n=1,656)**		
Age	0-17 years:	423	19%	0-17 years:	183	11%
	18-64 years:	1420	62%	18-64 years:	923	56%
	65+ years:	429	19%	65+ years:	550	33%
Triage category	ESI 1:	112	4.9%	Red:	26	1.6%
	ESI 2:	113	5.0%	Orange:	346	21%
	ESI 3:	1000	44.0%	Yellow:	698	42%
	ESI 4:	894	39.3%	Green:	581	35%
	ESI 5:	153	6.7%	Blue:	5	0.3%
Arrival	Ambulance	531	23%	Ambulance	225	28%*
	Traumahelicopter	4	0.2%			
Discharge destination	Home	1737	76.5%	Home	1025	61.9%
	Hospital admission	535	23.5%	Hospital admission	631	38.1%

*Data only known for the patients on the ED between February 11th until February 24th in 2013.

part of the door-to-doctor time. For more details, refer to Chapter 2.

Diagnostic tests. To get some insight into the role of diagnostic tests in the length of the ED stay, we divided the total time spent at the ED into three subprocess.

- Prediagnostic tests: Time from arrival at the ED until the first request for a diagnostic test. For example: taking a blood sample and sending it to the laboratory, a request for an X-ray or CT scan, or a request for any other kind of diagnostic test.
- Diagnostic tests: Time between the request for the first diagnostic test until the results of the last diagnostic test are available. This also includes waiting times between different diagnostic tests.
- Time after diagnostic tests: Time from the last result of the diagnostic tests until discharge.

3.2.4 Methods for statistical data analysis

Data from the VUmc and St. Antonius Hospital are analyzed separately. Exceeding a length of stay of four hours is selected as the primary endpoint. Patients are split into two groups: patients with an ED LOS of less than four hours or an ED LOS of more than four hours. For statistical analysis, two types of statistical tests are used, depending on the type of the tested variable.

$Pearson's\ chi-square$ test is used to assess the association between the variable "exceeding or not exceeding the four-hour target", and the nominal (categorical) variables such as age

category, triage level, the medical speciality, the hour of the day and the number of consultations. The null hypothesis, which represents an independence between the two variables, is rejected if the $p-value$ is lower than 0.05 (significant dependency). The $Mann-Whitney$ test, also called $Wilcoxon$ or $rank-sum$ test, is performed to compare the two populations of patients (exceeding and not exceeding the four-hour target) in terms of some duration variables (quantitative). This test allows to determine whether a particular population tends to have larger values than the other (in terms of quantitative variables such as LOS, DTDT, sub-processes durations, etc.). If the $p-value$ is lower than 0.05, the null hypothesis that the distributions are similar is rejected, which means that the two distributions are significantly different and there is a significant dependency between exceeding or not exceeding the four-hour target and the chosen variable. Table 3.2 summarizes the tested variables, the used statistical test and the $p-values$ obtained in the different tests performed in the next section.

Table 3.2: Summary of the statistical tests conducted in the analysis

			$p-value$	
Variable 1	**Variable 2**	**Statistical test**	**VUmc**	**St. Antonius**
Hour of the day	Four-hour target	$chi-square$	0.020	0.011
Day of the week	Four-hour target	$chi-square$	0.054	0.162
Triage level	Four-hour target	$chi-square$	$7.56*10^{-29}$	$1.98*10^{-13}$
Medical speciality	Four-hour target	$chi-square$	$2.86*10^{-13}$	$4.04*10^{-18}$
Number of specialities involved	Four-hour target	$chi-square$	$4.66*10^{-48}$	$8.55*10^{-33}$
Age group	Four-hour target	$chi-square$	$2.36*10^{-10}$	$5.71*10^{-17}$
Door-to-doctor time	Four-hour target	$Mann-Whitney$	0.7	$3.15*10^{-04}$
Time before diagnostic tests	Four-hour target	$Mann-Whitney$	0.12	0.022
Diagnostic duration	Four-hour target	$Mann-Whitney$	$1.24*10^{-14}$	$2.77*10^{-09}$
Time after diagnostic test	Four-hour target	$Mann-Whitney$	$3.02*10^{-11}$	$4.80*10^{-30}$
Undergoing or not Radiology	Four-hour target	$chi-square$	$1.54*10^{-24}$	$1.47*10^{-19}$
Undergoing or not X-rays	Four-hour target	$chi-square$	0.0017	$4.11*10^{-07}$
Undergoing or not CT scan	Four-hour target	$chi-square$	$6.54*10^{-49}$	$8.76*10^{-34}$
Discharge destination	Four-hour target	$chi-square$	$1.66*10^{-29}$	$3.59*10^{-32}$
Age group	Triage level	$chi-square$	$3.88*10^{-23}$	$2.71*10^{-15}$
LOS	Which hospital	$Mann-Whitney$	$4.71*10^{-38}$	
Age group	Which hospital	$chi-square$	$5.53*10^{-29}$	
Triage level	Which hospital	$chi-square$	$2.99*10^{-57}$	

3.3 Statistical data analysis

In this section, we perform a series of statistical tests among several potential influencing factors represented by variables in order to identify the ones currently affecting LOS. In the VUmc, 2,272 patients were seen at the ED between 8 October and 4 November 2012, a total of four weeks. A subgroup of 372 ESI 2 and ESI 3 patients was followed closely by researchers to obtain more detailed information (sub-processes). In the St. Antonius Hospital there were 1,656 patients of which a total of 492 orange- and yellow-triaged patients were closely observed for detailed information. The reason for this relatively small group is that it is time consuming to record all steps in the processes on the ED due to lack of an electronic tracking system.

3.3.1 Length of stay

The length of stay (LOS) was significantly different between the two hospitals, $p < 0.001$. In the VUmc, 89% of the patients had a LOS less than four hours. The average LOS (n = 2,262) was 2:10 hours, (median 1:51 hours, range: 0:05-12:08). In the St. Antonius Hospital, 77% of patients had a LOS shorter than four hours (n = 1,656). The average completion time in hours (n = 1655) was 2:49 (median 2:34, range: 0:08-11:04). Figure 3.1 demonstrates the cumulative distribution of completion times for both hospitals. The next analysis will provide some information that help explain these longer LOS in St. Antonius Hospital (patient characteristics).

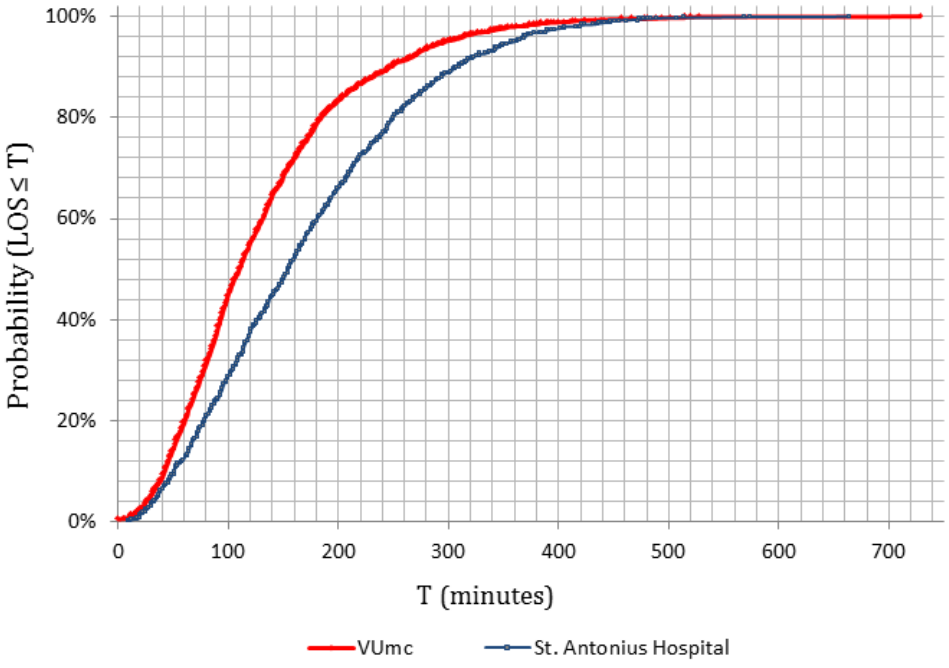

Figure 3.1: Cumulative distribution of Length of stay in both hospitals

3.3.2 Arrival pattern

Most patients arrived between 9.00 and 23.00 hours. An association was found for both VUmc ($p = 0.02$) and St. Antonius Hospital ($p = 0.01$) between arrival time and the four-hour target (Figure 3.2). No significant differences were found in exceeding the four-hour target between ED visits on different days of the week: VUmc ($p = 0.054$), St. Antonius Hospital ($p = 0.162$).

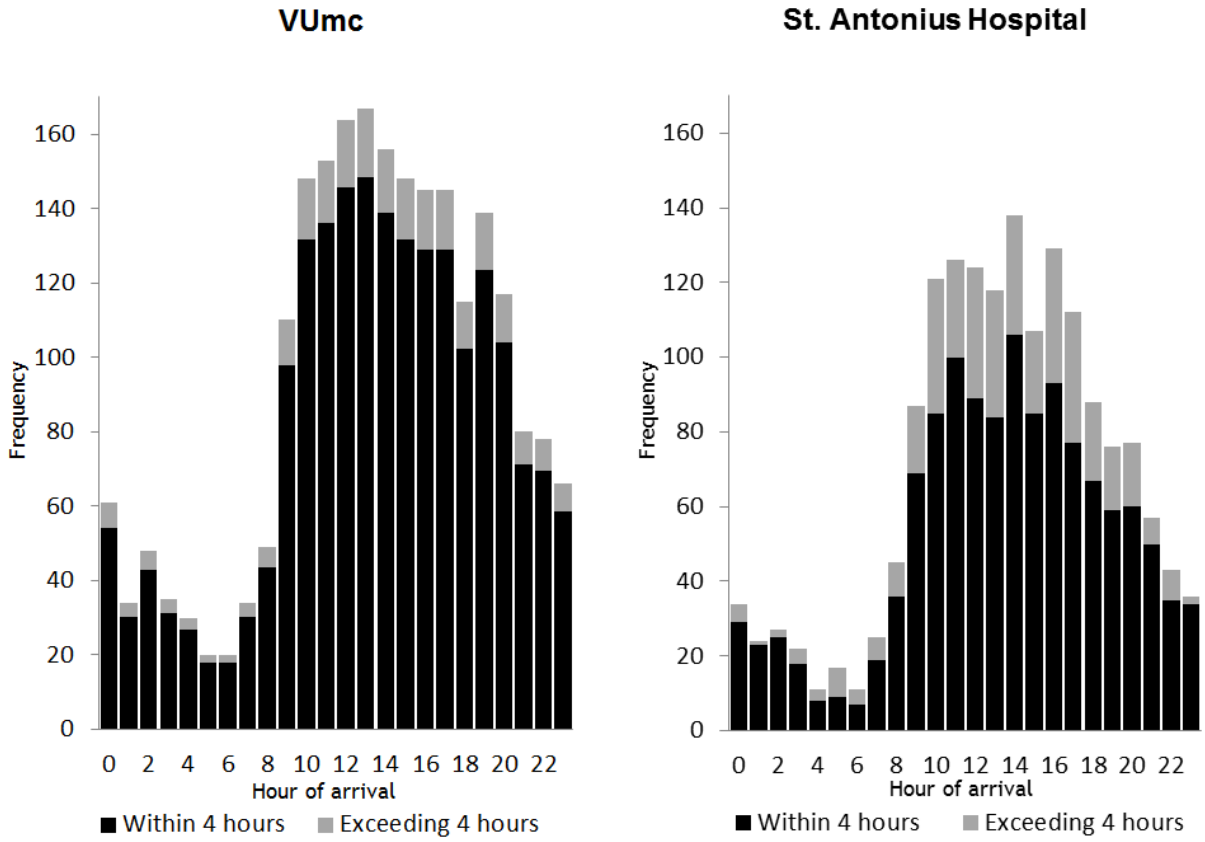

Figure 3.2: Four-hour target and time of arrival in both hospitals

3.3.3 Patient triage level

The distribution of patients over the five triage levels was significantly different in the two hospitals ($p < 0.001$). In the VUmc, a higher percentage of ESI 1 patients were seen compared with the number of red-triaged patients in the St. Antonius Hospital, due to the fact that the VUmc is a level 1 trauma centre. However, more orange-triaged patients were seen in the St. Antonius Hospital compared with ESI 2 patients in the VUmc, probably because acute cardiology patients (mostly ESI 2) are not presented to the ED in the VUmc but to the cardiology department. In the VUmc, most patients were categorized as ESI 3 (44%) and ESI 4 (39%) (Table 3.1). In St. Antonius Hospital, most patients were categorized as yellow (42%) and green

(35%).

In the VUmc, a larger percentage of ESI 1, 2 and 3 patients did not achieve the four-hour target (14%, 20% and 19%) compared with ESI 4 and 5 patients (2.7% and 0%), $p < 0.001$. At the St. Antonius Hospital, a greater percentage of orange and yellow categorized patients exceeded the four-hour target (32% and 28%) compared with red (8%), green (13%) and blue (0%), $p < 0.001$ (see Figure 3.3).

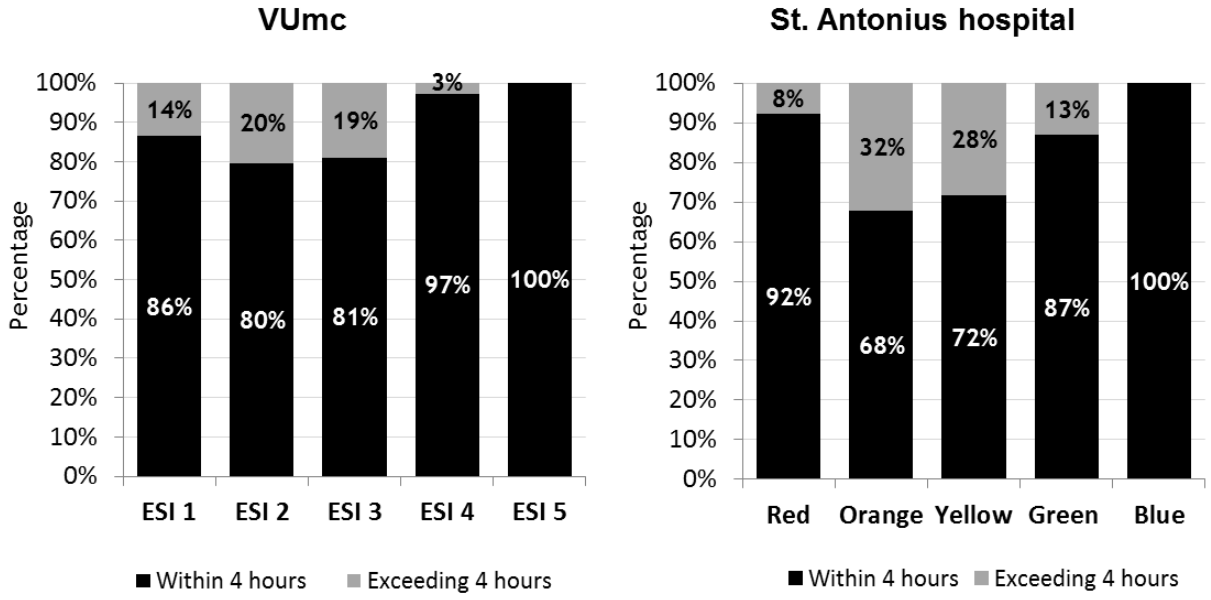

Figure 3.3: Realization of the four-hour target per triage level in both hospitals

3.3.4 Age

The patients age group distribution of the two hospitals was significantly different ($p < 0.001$). The average age of patients in the VUmc was 40 years (standard deviation 24.1); this was significantly higher in the St. Antonius Hospital with an average age of 50 years (standard deviation 23.6). In both hospitals, patients age group has a significative impact on whether their LOS is within or exceed four hours ($p < 0.001$) (see Figure 3.4). Figure 3.5 demonstrates the average LOS per age group. Moreover, there is a significant association between patients age group and triage level in both hospitals ($p < 0.001$). This would explain why St. Antonius patients were both older and sicker.

3.3.5 Door-to-doctor time

In the VUmc, the door-to-doctor time was not significantly different between patients who did or did not exceed the four-hour target, $p = 0.07$, while in St. Antonius Hospital, there was a

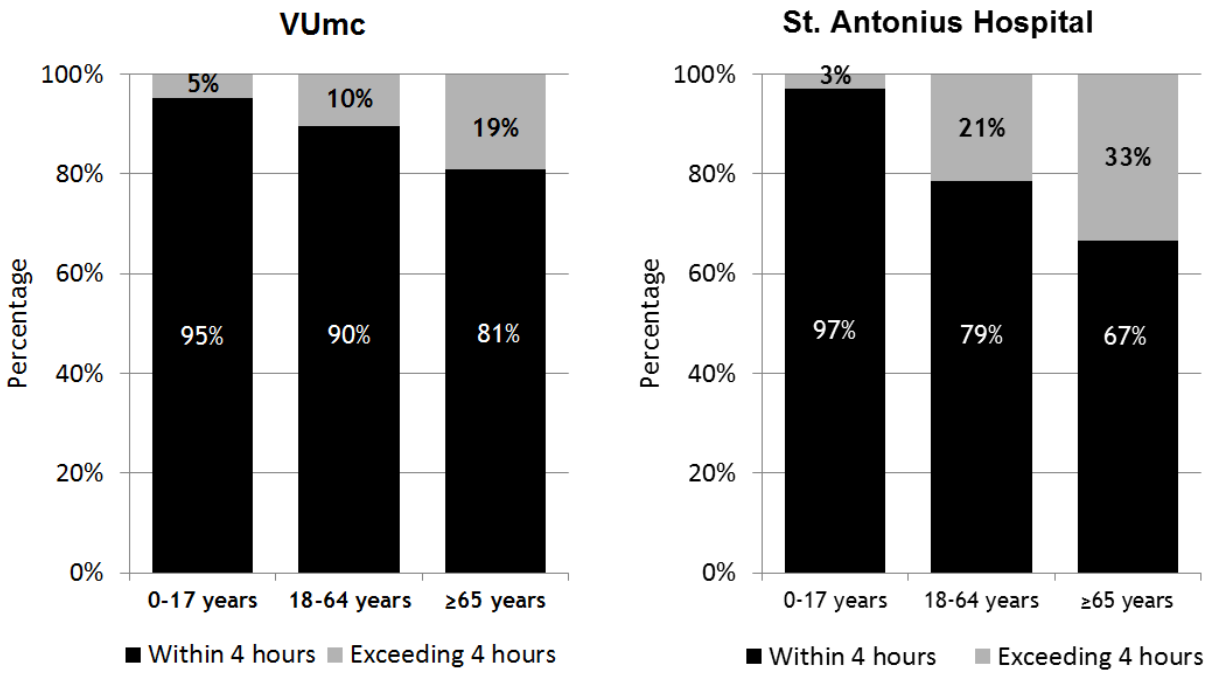

Figure 3.4: Realization of the four-hour target per age group in both hospitals

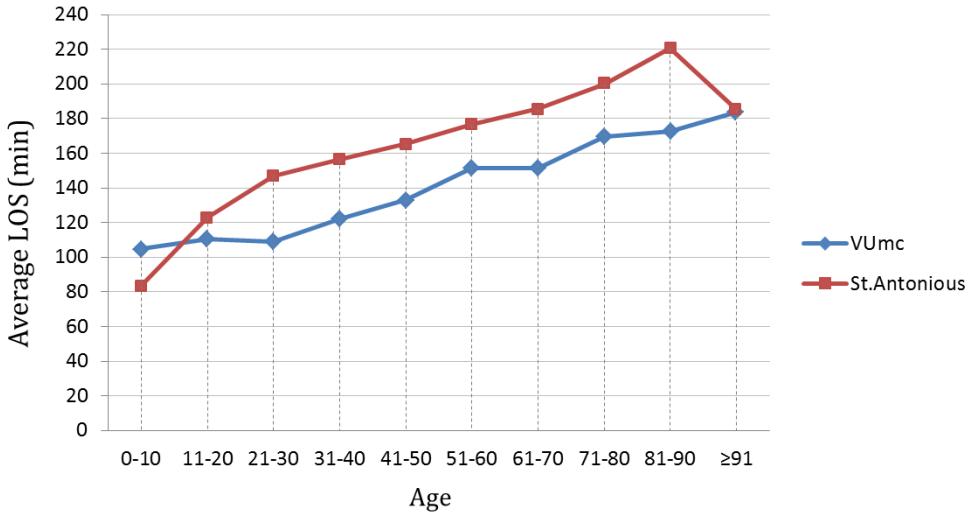

Figure 3.5: Average LOS per age in both hospitals

significant correlation for this analysis, $p < 0.001$ (Figure 3.6).

3.3.6 Medical speciality and the number of specialities involved in the care

In both hospitals, a significant dependency was found between speciality and exceeding the four-hour target ($p < 0.001$). The responsible medical speciality is the one corresponding to the first consultation. If necessary, other specialities could be involved for further consultations. In the VUmc, the average number of additional consultations (additional specialities involved) per

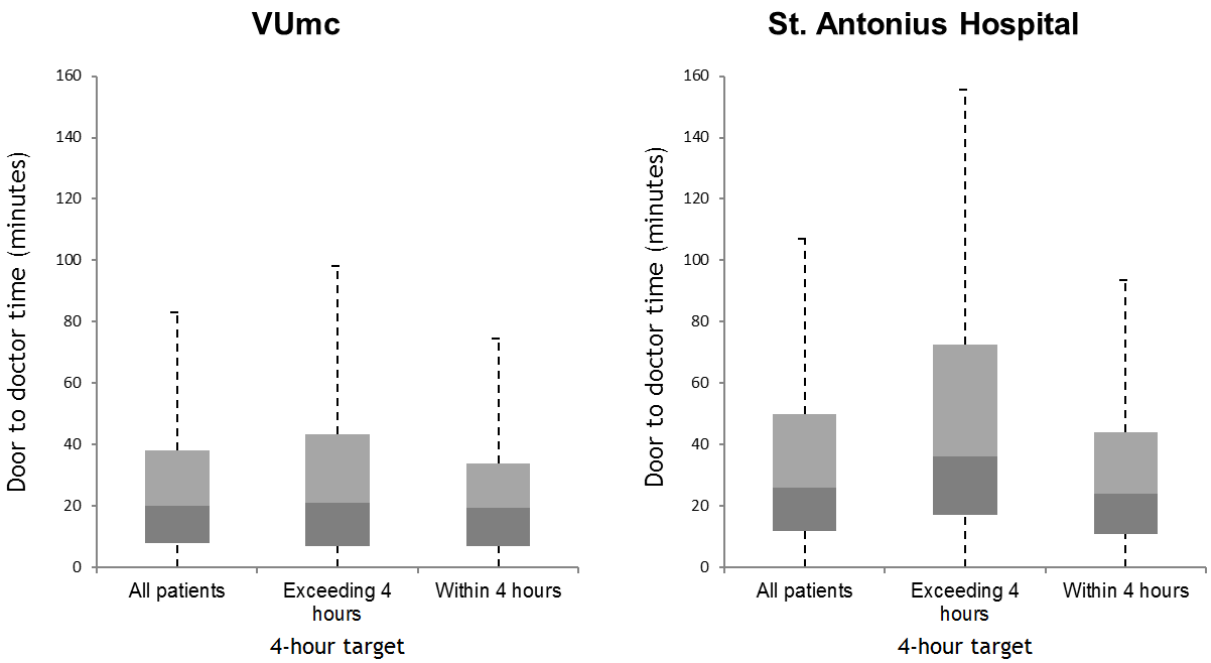

Figure 3.6: Boxplots of the Door-to-doctor time according to the four-hour target in both hospitals

patient was 0.306, this was 0.155 in St. Antonius. For both hospitals there was a significant dependency between exceeding the four-hour target and the number of additional specialities, $p < 0.001$ (Figure 3.7).

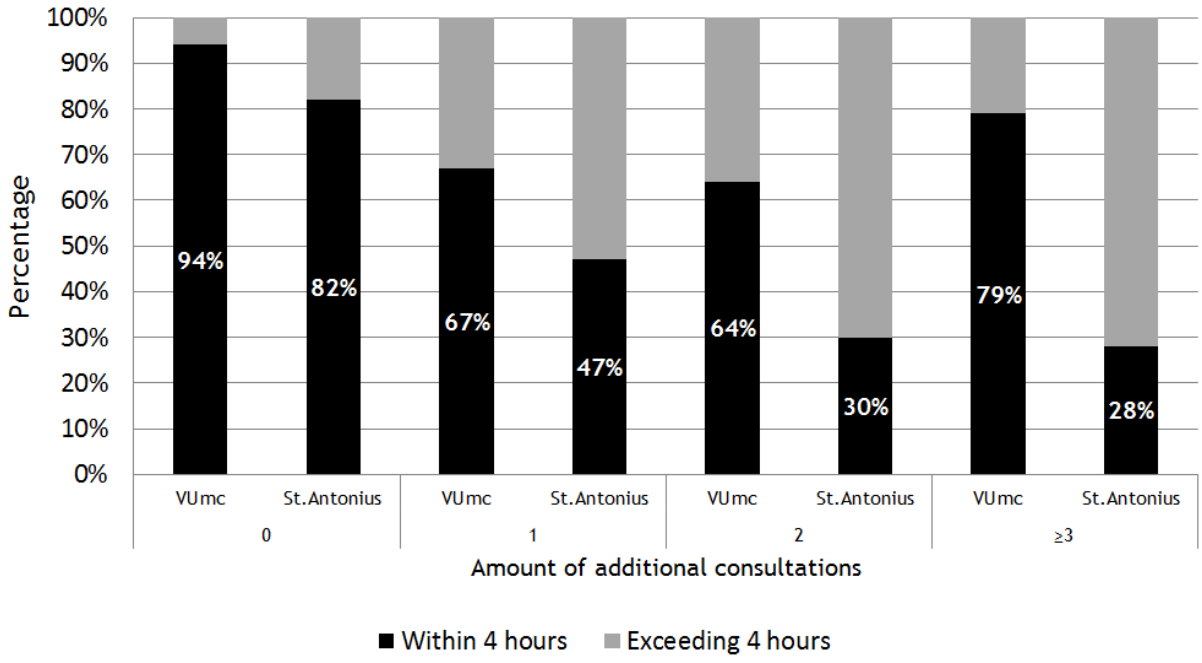

Figure 3.7: Number of additional consultations and the four-hour target in both hospitals

3.3.7 Diagnostic tests

In the VUmc, data of 283 detailed patients were useful (i.e., complete and not aberrant) for analyzing diagnostic tests, as illustrated in Figure 3.8. No significant difference in duration of "prediagnostic tests" was found for patients who did or did not exceed the four-hour target (p = 0.12). For "diagnostic tests" and "time after diagnostic tests" there was a significant difference (both $p < 0.001$). In the St. Antonius Hospital there was a significant difference in the duration of all the sub-processes for patients ($n = 349$) who did or did not exceed the 4 hour-target (Figure 3.8).

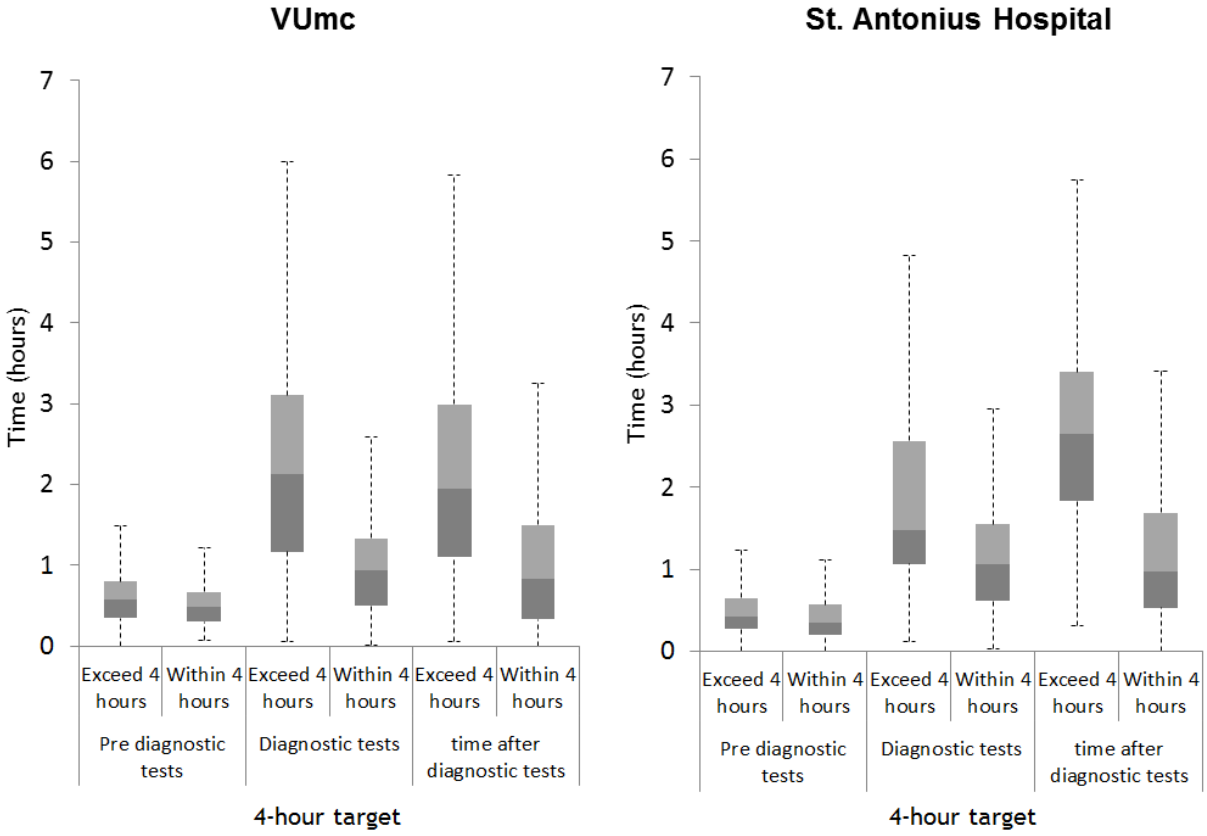

Figure 3.8: Boxplots of the durations of the sub-processes: prediagnostic tests, diagnostic tests and time after diagnostic tests for patients who did or did not exceed the four-hour target in both hospitals

3.3.8 Radiology

In the VUmc, 34% of patients underwent an X-ray, followed by CT scan (11.4%), Ultrasound (8%) and MRI (0.4%). In the St. Antonius Hospital, 49% of patients underwent an X-ray, followed by CT scan (15%), ultrasound (7.9%) and MRI (0.4%). All radiology tests were correlated with a significantly higher chance to exceed the four-hour target. The patients in the VUmc

who did not undergo any radiological tests had a chance of 4.9% of exceeding the four-hour target. This chance to exceed the target increased to 8.5% in patients only undergoing X-ray(s) ($p = 0.002$), and to 35.3% for patients only undergoing CT scan(s) ($p < 0.001$) and 33.3% for patients undergoing only ultrasound(s) ($p < 0.001$). In the St. Antonius Hospital the chance to exceed the four-hour target was 11% for those who did not have radiological tests. This chance increased to 22% for patients having only X-rays(s) ($p < 0.001$), to 49% for patients undergoing only CT scan(s) ($p < 0.001$) and to 45% for only undergoing ultrasound(s) ($p < 0.001$). For both hospitals there was a significant correlation for the number of radiology tests and exceeding the four-hour target, $p < 0.001$ (Figure 3.9).

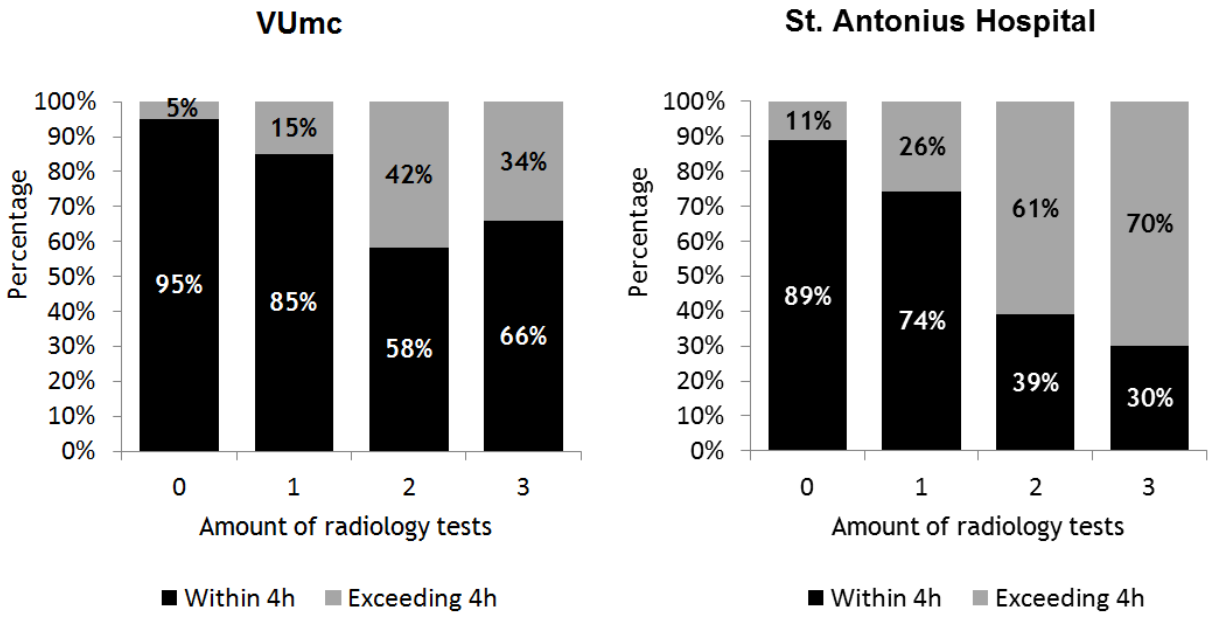

Figure 3.9: Realization of the four-hour target and the amount of radiology tests in both hospitals

3.3.9 Discharge destination

In both hospitals, most ED visits did not result in a hospital admission (Table 3.1). Patients who were admitted or transferred elsewhere were more likely to exceed the four-hour target in the VUmc (25% and 29% of exceeding) compared with those who were discharged home (7%) ($p < 0.001$). In the St. Antonius Hospital 37.5% of admitted patients and 57.1% of transferred patients exceeded the four-hour target compared with 11.5% of released patients ($p < 0.001$).

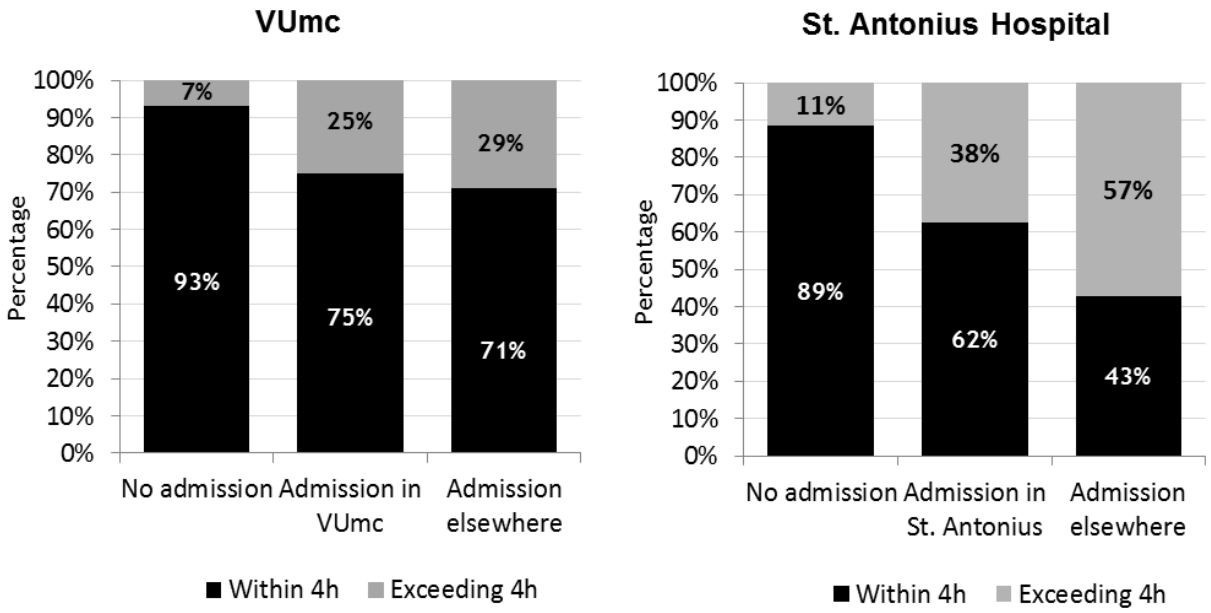

Figure 3.10: Realization of the four-hour target and the discharge destination in both hospitals

3.4 Results interpretation and discussion

In this section, we summarize and discuss the results of the statistical analysis. We explain how each influencing factor contribute to a longer stay in practice, and highlight some potential remedial measures to alleviate this influence. Influencing factors could be classified into two large categories: internal and external factors. Each of which might be addressed by internal or external remedial measures, depending on the nature of the lever used. This discussion will serve as basis for the definition of the issues that will be investigated in the next chapters.

3.4.1 Internal factors

Type and amount of specialities involved

Statistical tests revealed that the number and type of specialities involved in the patient care have a significant impact on their LOS. Patients in triage categories ESI 2/3 and orange/yellow are relatively old and frequently have multiple comorbidities demanding the expertise of more than one specialist. In contrast to ESI 1 and red category, they are not initially seen by a team of specialists. Consultations occurred consecutively in these patients contributing to a longer LOS in both hospitals. Brick et al. (2014) also concluded that multiple consultations and advanced age were significantly associated with a longer stay on the ED. Consulting physicians tend to treat the patient individually, one after the other, instead of working as a team. This fragmented delivery of care increases the LOS and may thereby lead to complications and reduced

patient satisfaction. The proposed solution for this problem is the introduction of "assessment teams" for these patients. Especially in old patients with multiple comorbidities, it was decided that specialities such as internal medicine, neurology, surgery or emergency physicians should be called upon to examine the patients together as a team at the outset so that multiple, consecutive consultations could be avoided. However, note that in contrast to the two studied EDs in the Netherlands which contain physicians from several medical specialities (VUmc in particular), most French EDs use mainly polyvalent emergency physicians, with the possibility to call specialists from other departments when needed.

ED sub-processes

We also analyzed some ED sub-processes to discover which processes contributed most to a longer time to completion.

Prediagnostic tests duration and Door-to-doctor time. Triage and the waiting time for a physician are part of the door-to-doctor time. The door-to-doctor time is a part of the sub-process that we called "prediagnostic tests" duration. In the VUmc, these two periods did not contribute to a longer patients LOS. However, in the St. Antonius Hospital, there was a significant difference in these durations between patient who did or did not exceed the 4 hour-target, to a greater extent for door-to-doctor time. Moreover, the duration of "prediagnostic tests" and door-to-doctor time in particular are frequently at the forefront of ED quality improvement initiatives (Jones and Evans, 2008) since they are particularly associated to mortality, abandonment and satisfaction of patients (see Chapter 2). Consequently, we will address in this thesis the question of how to reduce the delay of ED pre-diagnostic periods. To this end, we assess in Chapter 7, an ED intervention called *triage nurse ordering*, which consists in allowing triage nurses to order some diagnosis tests right after triage, instead of waiting for a physician. In addition, we address in Chapter 4 the optimization of ED staffing levels while taking door-to-doctor time into consideration.

After diagnostic tests. The elapsed time between receiving all diagnostic results and admission/discharge had a big influence on the LOS in both hospitals. This period include the waiting time for the physician who will further make an interpretation of the results and take a decision about the process outcome, as well as the organization of the admission/transfer when required. The latter will be discussed later in the section addressing external factors (boarding time), because it is related to the availability of external resources from the ED point of view (internal beds of the hospital).

Although this was not tested in our study, it was proposed that another cause for this delay is

the delay in decision-making, because of the lack of direct supervision on the ED. Residents often see patients alone on the ED and telephone their supervisor after finishing anamnesis, physical examination and first diagnostic tests. Especially during late hours when the supervisor is no longer in the hospital, they tend to collect necessary information for all patients before they call her for advice, so that she would not be disturbed too many times during sleep. In addition, during the daytime, supervisors are not always directly available to discuss a case because they are busy with multiple patients. The two hospitals are in the process of increasing the number of emergency physicians to cover all the shifts 24/7. The working hours of senior doctors have been adjusted to cover the busiest moments at the ED. This more direct contact between supervisors and residents might help to quicken the process of decision-making, after all diagnostic tests are performed. Another reason for delay is the lack of communication. Sometimes the doctor is simply not aware of the fact that the diagnostic tests have already been performed.

In order to reduce the duration of after diagnostic tests, and also waiting times for physicians in general, we investigate in Chapter 6 a modification in the current practices of operating diagnostic tests interpretation.

3.4.2 External factors

Significant dependencies were found between EDs performance and external factors. These factors are related to the ED environment and are uncontrollable from an ED perspective. Such external factors are either related to patient characteristics or to external resources having interactions with the ED (admission beds and diagnostic resources).

Volume and mix of patients

The case mix characteristics (triage level, age, specialty needed, etc.) were identified as influencing factors. Most of the patients who stayed longer than four hours in both EDs were old and vulnerable patients (higher triage categories). In addition, there were patients who stayed much longer than the expected four hours. There is a dependency between these different characteristics (such as age and triage levels). The mix of patients in St. Antonius was significantly different from VUmc (older and more critical) resulting in longer LOS. This is because these patients with complex pathologies require longer interactions with practitioners and more diagnosis tests. Note that in contrast to other patients, the most acute category of patients (ESI 1 or category red) are treated in the shock room by a team of specialists directly after arrival on the ED with the opportunity to perform radiological testing at the bedside, resulting in a relatively short completion time on the ED. The visit volume of patients and its variability during the

hours of the day had also a significant influence on patients stay.

Some interventions which seeks to modify the patient flow to the ED exist in the literature and in practice. A first stream consists in refusing patients with minor problems and divert them to outpatient clinics in order to reduce unnecessary ED use (Lowe et al., 1994), which would cause ED overcrowding. In France, alternative structures called "Maisons Médicale de Garde" were created in the last decade in order to receive non-urgent patients reoriented from ED (Gentile et al., 2009). However, the notion that non-urgent patients are a major cause of the ED overcrowding crisis has been abandoned in the US (Trzeciak and Rivers, 2003) because non-urgent visits cause extremely crowded waiting rooms but reportedly do not cause crowding in the ED treatment areas (Trzeciak and Rivers, 2003; Vertesi, 2004). The other stream consists in diverting ambulances to reduce arrival rates when the ED is overcrowded, in metropolitan areas where multiple hospitals are available to serve the population (Burt et al., 2006). As highlighted in Chapter 2, there is an extensive literature addressing ambulance diversion. However, diverting patients to external facilities fall out of the scope of this thesis, which solely focuses on interventions within the ED. Instead of modifying the patient demand, it is primordial to adjust the ED capacity in accordance to the demand. It is necessary to best match the amount of available resources in the ED with patient arrivals, through appropriate staffing levels and adapted resources allocations. This issue will be addressed in Chapters 4 and 5.

Factors related to exogenous resources

ED does not operate as an isolated unit but interacts with other actors in the context of the larger hospital system. They exert a significative influence on the ED. Examples of these actors are: the services where patients are sent to undergo diagnosis tests, and admission beds in other services of the hospital (or even other hospitals sometimes) where patients are transferred.

Diagnostic tests duration. The duration and the amount of "diagnostic tests" was demonstrated to be an influencing factor in both hospitals. However, these diagnostic tests are performed using facilities which are outside the ED and typically handle many other patients besides those from the ED (Saghafian et al., 2014). Radiological tests (CT scan, X-ray, Ultrasound and MRI) are performed in the radiological department. After sampling in the ED, biological tests (blood and urine) are performed in the laboratory. Furthermore, the use of diagnostic procedures such as CT scans has increased in the last decade, as they improve diagnostics and therapeutic decision-making, but on the other hand they also take up a long completion time (Kocher et al., 2012). In this study, all radiological tests were associated with a longer LOS on the ED, and CT scan especially. It is known that it takes time before all the images of the CT scan are uploaded

and available to interpret.

Being limited to modification within the ED, the reduction of these delays fall out of the scope of this thesis. However, several interventions have been applied to shorten the process of laboratory testing (Oredsson et al., 2011) such as faster transportation to the laboratory, and faster reporting systems. A solution called Point-of-care testing (POCT) appears to be an effective approach to reduce diagnostic tests turnaround time. POCT consists in decentralizing biological tests (blood and urine) and simple imaging by performing them inside the ED with the use of special devices. POCT devices make the test results available immediately allowing more rapid decision making by physicians. Several studies were conducted in the medical literature (Lee-Lewandrowski et al., 2003; Murray et al., 1999; Fermann and Suyama, 2002), as well as in the OR/OM one (Hanna et al., 1974; McGuire, 1994) and showed that POCT has the potential to significantly shorten LOS in the ED.

Boarding effect. For both hospitals, admitted patients to the hospital and transferred patients to other hospitals were more likely to exceed the four hour target. Besides, time after diagnostic tests was longer for admitted/transferred patients compared to patients who were discharged home. This is probably caused by the limited availability of hospital beds which leads to a time-consuming search for a bed or transfers to other hospitals.

This hospital bed access issue is known as the "boarding effect" or "bed-block" problem (Forster et al., 2003). Boarding time which is defined as the time from admission order to departure from the ED (Olshaker and Rathlev, 2006) is a key contributor to ED overcrowding worldwide (Forster et al., 2003; Derlet and Richards, 2000). Bed-block refers to situations in which ED patients who need to be hospitalized cannot be transferred to their inpatient units (internal units or internal wards) due to lack of bed availability (Shi et al., 2014; Forster et al., 2003). In some healthcare funding policy contexts, bed-block could also be due to the reluctance of internal wards to accept old patients with multiple pathologies (long and costly stays), because they are the less profitable ones in terms of revenue, in addition to their competition with scheduled admissions (Bonastre et al., 2013; Potel et al., 2005). Boarding causes the ED to be filled beyond capacity with the highest acuity patients (Trzeciak and Rivers, 2003). Boarded patients block ED beds and prevent from seeing new patients. Decreasing boarding times has been found to be a major lever for reducing LOS (Saghafian et al., 2015). Despite the importance of boarding effect, we will not address this issue because the source of the problem comes from beyond the ED responsibility. However, an avenue for future research is highlighted in Chapter 7. The latter is an anticipation method which consists in allowing triage nurse to initiate search for admission beds earlier.

3.5 Conclusions

Through statistical analysis, we examined which factors contribute to a longer stay in EDs. We used detailed data from two hospitals with different work procedures and different patient populations, in order to obtain generalizable insights. Both hospitals were facing largely the same problems. This study revealed that multiple factors lead simultaneously to ED longer delays. ED congestion is a multifactorial phenomenon. Therefore, the improvement of ED performance require a series of different remedial measures each focusing on a distinct influencing factor. Thanks to the result interpretation and discussion, several remedial measures were derived in order to reduce delays in EDs. In coherence with our thesis framework and purpose, we divide these interventions into two types: interventions inside the ED (internal interventions), and interventions in the ED environment (external or environment-related interventions). The second category of issues falls out of the scope of this thesis because we aim to provide ED decision makers with solutions that could be implemented autonomously, and independently from external actors that are beyond the ED responsibility. Examples of these external interventions are: To master the patients demand (volume, mix and variation) by refusing or diverting patients to external facilities, the reduction of diagnostic tests duration which are mainly performed outside the ED using resources common to all the hospital, the addition of hospital admission beds and the transfer optimization to alleviate the ED boarding effect.

Several relevant internal interventions have been derived. The following internal measures correspond to those selected to be addressed in the remaining chapters of this thesis. The influence of the patients mix and demand fluctuation requires to rationalize resource utilization. It is necessary to best match the amount of available resources in the ED with patient arrivals, through appropriate staffing levels and adapted resource allocation. This is addressed in Chapters 4 and 5 in the context of the so-called resource-related experiments. This chapter also revealed the importance of reducing the delay of ED pre-diagnostic periods. To this end, we include in Chapter 4 the door-to-doctor time in the optimization of ED staffing levels. In order to quicken the pre-diagnostic delays, we model and analyze in Chapter 7 an ED process modification called *triage nurse ordering*, which allows triage nurses to order diagnostic tests right after triage, instead of the standard procedure, i.e., waiting for the physician to examine the patient and order tests. Another process-related issue is investigated in Chapter 6 in order to reduce after-diagnostic tests durations, and also waiting times for physicians throughout the process. It consists in assessing a modification in the current protocols of operating diagnostic tests interpretation. Typically in current ED practices, each patient is assigned to a single physician for the whole process ("Same Patient Same Physician", *SPSP* rule). We assess the

relevancy of removing the *SPSP* restriction.

In addition, we came out in collaboration with the two EDs staff with a series of organizational recommendations. Consecutive consultations by different specialists, in patients with complex pathology, was one of the main reasons for extreme delays. The different specialities tended to work individually and not as a team. "Team assessment" with multiple specialities was recommended to reduce this lack of coordination of care. In France, there is a growing trend to use organizations composed of polyvalent emergency physicians instead of specialists. Still, it is possible to call a specialist when necessary, but the emergency physician always remains responsible of the patient. This is the case in Saint Camille ED. The lack of direct supervision on the two Dutch EDs was also a concern as well as some lack of communication concerning the completed diagnostic tests. In order to quicken the process of decision-making when diagnosis tests are completed, we recommended a more direct contact between supervisors and residents, and to improve the communication by alarming physicians as soon as tests results are ready.

Note that the identified factors influencing longer ED delays and the research avenues that have been derived in this chapter, were validated by our collaborators in France as common concerns. In addition, some French studies show that they are generalizable to the context of French EDs (Le Spegaque et al., 2006). The different issues addressed in the rest of this thesis and the conducted experiments were performed under a close collaboration with the French ED of Saint Camille hospital.

Chapter 4

Resource-related experiments: Simulation-based optimization of ED staffing levels

In this chapter, we use discrete-event simulation to model and analyze a real-life emergency department. Our approach relies on the appropriate integration of most real-life ED features to the simulation model in order to derive useful practical results. Data is supplied from the ED of the urban French hospital Saint Camille. Our purpose is to optimize the human resource staffing levels. We want to minimize the patient average length of stay ($\overline{LOS}$), by integrating a staffing budget constraint and a constraint securing that the most severe incidents will see a doctor within a specified time limit. The second constraint allows to avoid the perverse effect of only considering the $\overline{LOS}$ metric that would delay the treatment of the most urgent patients. We use simulation-based optimization, in which we perform a sensitivity analysis expressing $\overline{LOS}$ as a function of the staffing budget and also the average door-to-doctor time for urgent patients ($\overline{DTDT}$). We show that the budget has a diminishing marginal effect on the problem solution. Due to the correlation between $\overline{LOS}$ and $\overline{DTDT}$, we also observe that the $\overline{DTDT}$ constraint may significantly affect the feasibility of the problem or the value of the optimal solution.

The paper versions of this chapter (Ghanes et al., 2015c, 2014b) are published respectively in the journal *SIMULATION*, and the proceedings of the *2014 Winter Simulation Conference* held in Savannah, USA.

4.1 Introduction

As demonstrated in Chapter 3, the performance of EDs is significantly influenced by patient demand variability. Under a difficult economic context, ED managers are trying to improve performance by minimizing the mismatch between this demand and supply. However, an ED is a complex environment with various types of heterogeneous patients and resources where most of the parameters are uncertain. Healthcare practitioners have therefore resorted to researchers in operations management and operations research in order to develop scientific approaches for the performance optimization of EDs. As mentioned in Chapter 1, the used tools can be divided into two main categories: analytical methods and simulation. In this case study, the need for high impact solutions motivates us to use discrete-event simulation (DES). This allows to capture most of the realistic features in an ED. In using simulation for ED operations management, we are following a longstanding practice. Rossetti et al. (1999), Komashie and Mousavi (2005), Duguay and Chetouane (2007), Ahmed and Alkhamis (2009) and Abo-Hamad and Arisha (2013) conduct simulation studies for the analysis of EDs in Virginia (USA), London (Britain), Moncton (Canada), Kuwait and Dublin (Ireland) respectively. They address the problem of resource staffing optimization. Sinreich and Marmor (2005) lay the foundation for developing a simulation tool to analyze the ED performance. For a background on simulation models for EDs, we refer the reader to the surveys by Paul et al. (2010) and Günal and Pidd (2010).

The simulation model proposed in this study is based on a comprehensive understanding of the real-world functioning of emergency departments. A field study was conducted for this purpose through a close collaboration with the ED of Saint Camille hospital. Saint Camille hospital is a teaching hospital situated in an Eastern suburb of Paris. Real data and expert judgments are both used for the construction of the model. For the validation, the model outputs are compared to historical data and judged by experts. In order to alleviate congestion, ED managers and the general management of Saint Camille hospital intend to invest in human staffing. Their objective is to improve the ED performance by investing in human resources. The question we are facing here is: By how much should the current staffing budget be increased and how should this additional budget be used in the allocation of human resources?

As explained in Chapter 3, the selection of a KPI for ED optimization has always been a controversial subject. Neither the scientific community nor practitioners are able to decide about the most appropriate KPI, as each indicator presents at the same time benefits and drawbacks. As a reminder, the most known and used KPI is the average length of stay ($\overline{LOS}$). LOS is the sum of the sojourn times in all subsections of the ED. It is the KPI on which EDs are generally

judged in practice, because it allows to approach the ED in a holistic way. It is abundantly used in the literature as well. Some references include Huang et al. (2012), McGuire (1994), Centeno et al. (2003), Saghafian et al. (2014), Gorelick et al. (2005), Wang et al. (2012), and Song et al. (2013). However focusing only on $\overline{LOS}$ could have important drawbacks. It gives an overview of the entire system performance but doesn't allow to figure out local strengths and weaknesses. Besides, the impact could be in the non-urgent cases, or worst, the non-urgent cases could be benefited on behalf of prolonging the waiting time of the urgent ones. From this appears the necessity to take another ED KPI into consideration, which is the average door-to-doctor time ($\overline{DTDT}$). $DTDT$, also called time to first treatment or time to physician, describes the time between the patient arrival and the first handling by a physician. $DTDT$ measures the most crucial element for seriously ill patients because they need urgent attention. For non urgent patients, the average $DTDT$ is generally close to the entire $\overline{LOS}$ and thus the latter is sufficient as a KPI for this kind of patients. There are references in the literature that consider $DTDT$ as the sole performance indicator for the analysis of EDs. Examples include Cooke et al. (2012), Cochran and Roche (2009) and Lau and Leung (1997). Only rare papers such as Saghafian et al. (2012) and Burström et al. (2012) consider both indicators, as we do in this chapter.

The main contributions of this chapter can be summarized as follows. We propose a simulation model that is based on a comprehensive understanding of the ED functioning. Most common structural and functional characteristics of EDs, at least in France, are taken into consideration thanks to a close collaboration with Saint Camille ED. Based on the above, we point out a set of important ED features that are frequently ignored in the related literature. The model is close to the real system and is then appropriate to be used to address some operations management issues. We focus on the simulation-based optimization of staffing levels of the various human resource types involved in the ED. We study the effect of the staffing budget on $\overline{LOS}$, and show that it has a diminishing marginal effect. For instance, an increase of 10%, 20% and 30% in the staffing budget can generate an improvement of 33%, 44% and 50% in the optimal $\overline{LOS}$, respectively. We also show the effect of including a $\overline{DTDT}$ constraint for urgent patients in the model. We investigate how this additional constraint affects the optimality and the feasibility of the staffing problem solution. The results point out the fact that considering $\overline{DTDT}$ in addition to $\overline{LOS}$ involves a trade-off that managers should be aware about. We also derive useful insights about which type of resource to prioritize according to the available budget and the $\overline{DTDT}$ target. We surprisingly find that additional investments should be allocated in priority to doctors, which is counterintuitive to ED practitioners. Although the modeling is based on a specific ED, qualitative conclusions hold for other ED frameworks.

The rest of the chapter is organized as follows. In Section 4.2.1, we describe how the ED characteristics are implemented in the simulation model and the way data is collected. In Section 4.2.2, we validate the simulation model using historical data and expert judgments. Furthermore, we highlight the detailed level of modeling and compare it with the existing literature. In Section 4.3, we conduct simulation-based optimization experiments for the ED staffing problem. In Section 4.4, we give concluding remarks and highlight some future research.

4.2 Emergency department modeling

In this section, we provide the building of the simulation model as well as its validation.

4.2.1 Simulation model

We use Saint Camille hospital ED as a main reference to build our model. In this section we give an overview of the service with its resources and processes as well as the necessary data to construct the simulation model.

Saint Camille hospital has approximately 300 beds and covers most of the medical and surgical specialties. Its ED is operating 24 hours per day and serves more than 60,000 patients per year. Within the ED, we consider the following different zones:

- The external waiting room for walk-in patient arrival
- The registration and triage zone
- A shock room (SR) for acute ill patients
- Examination rooms (ER) also called boxes or cubicles
- An internal waiting room with stretchers for lying patients
- An internal waiting room for sitting patients
- The Observation Unit (OU)

In addition, the ED includes an ambulance arrival area and a central operation room where all the tasks that do not require the presence of the patient are made, such as reporting on computer, interpretation of diagnostic tests, discussions between medical staff, preparation of equipments, etc.

Patients arriving to the ED cover a big range of severity levels. At the beginning of the process, patients are categorized by a triage nurse according to their condition into five degrees

of severity, known as Emergency Severity Index (ESI), where ESI 1 are the most severe patients and ESI 5 the least severe ones (Tanabe et al., 2007). There are several different types of resources. The resources are also splited into dedicated groups for the ESIs, with different staffing levels for each group (see Appendix A.6). A physician for instance can be either senior or junior. A junior physician can be responsible only for a combination of ESI 3, 4 and 5 patients, while seniors can treat all categories. There are also two different types of nurses: The first one, referred to as triage nurse, is dedicated to the triage. The other nurses are inside the ED and are in charge of in-process patients. Moreover, ESIs 1, 2 and 3 belong to a group of patients referred to as long circuit (LC) and are treated by dedicated physicians and nurses. ESIs 4 and 5 are part of a group called short circuit (SC) and are also treated by resources dedicated to them. The shock room is dedicated to ESI 1 patients and a part of ESIs 2 and 3 patients. The shock room is also known as trauma and resuscitation room (Kuo et al., 2012; Saghafian et al., 2014). Examination rooms are also assigned to certain ESIs but with a different subdivision: medium boxes for ESIs 2 and 3, general boxes for ESI 4, and a fast track for ESI 5. Other resources such as stretcher bearers are not dedicated to any specific patient type. The reason for not including some resources in our model, such as janitorial staff, is that they do not really affect the system performance in terms of patient waiting times.

Similarly to Rossetti et al. (1999), Centeno et al. (2003) and Duguay and Chetouane (2007), our methodology is based on assessing the effect of staff changes on key performance indicators. We consider human and space resources in the model. Human resources are considered as control variables. The model development is performed using Arena simulation software provided by Rockwell Automation. During their sojourn, patients go over several stages that involve various types of limited resources, and then various patient waiting durations. The optimization of $\overline{LOS}$ involves the optimization of the sum of these durations. Processing times such as physician examinations or diagnostic tests are considered as exogenous variables, and thus they are not to be optimized. The main waiting durations of the simulation model are given in Figure 4.1.

The patient path in the ED comprises a series of assessments that constitutes the ED process, as synthesized in Figure 4.1. Patients have different severity levels. Therefore, the process varies from one patient type to another. However, the typical complete patient stay in an ED can be divided into five main parts (see Figure 4.2), as described below.

(1) From arrival to triage: Upon arrival to the ED, the patient is first registered at the reception desk and she is then triaged by the triage nurse in a dedicated box at the entry of the ED, based on the ESI triage system. The severity determines the priority of the patient over others (Tanabe et al., 2007) and how she will be routed to the appropriate resources throughout the

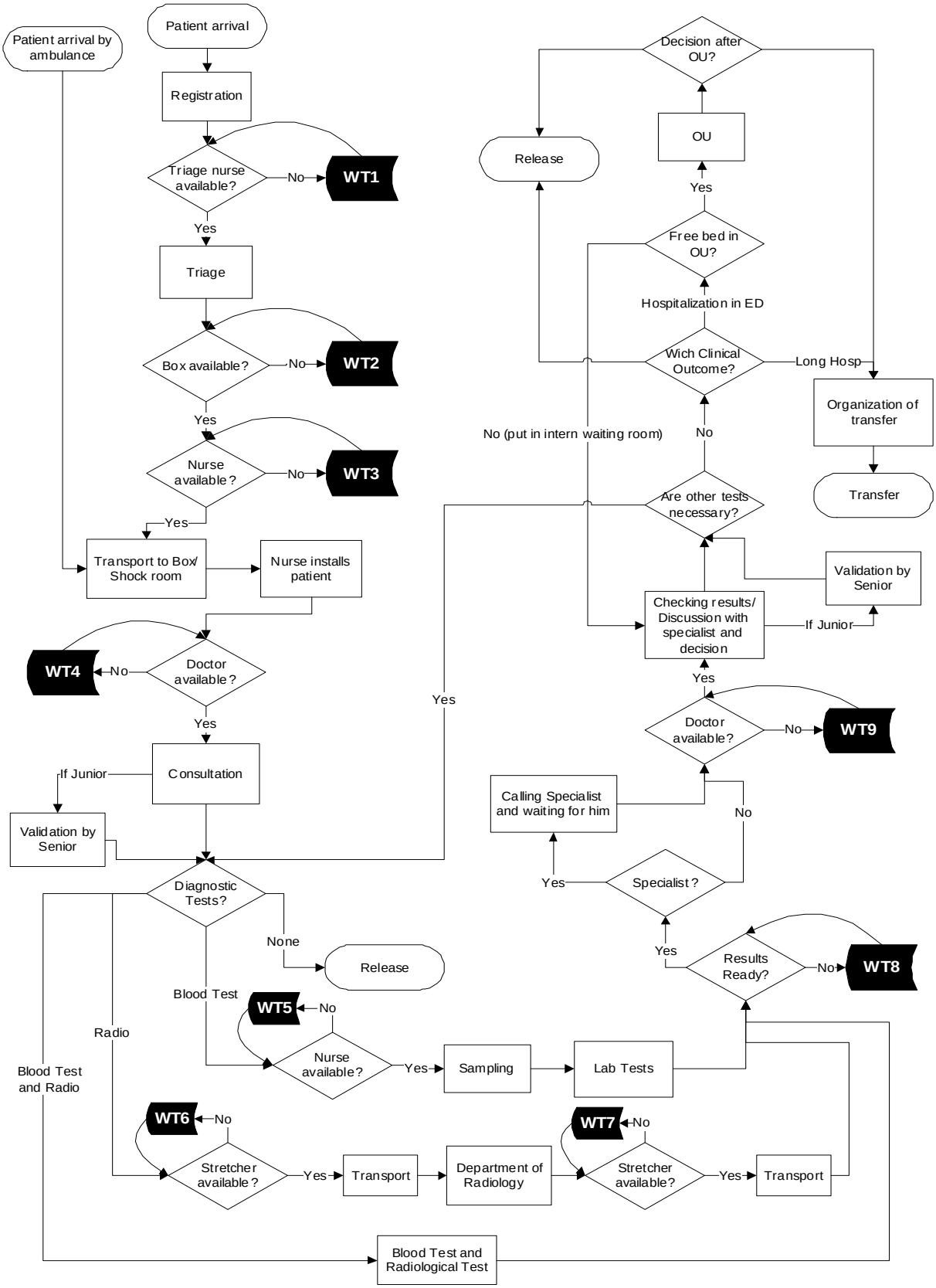

Figure 4.1: The conceptual model

Figure 4.2: The five typical stages of an ED process

process. When the triage nurse is busy, patients must wait in the external waiting room. The red code patients (ESI 1) generally arrive by ambulance. They must be stabilized immediately and skip triage.

(2) The initial consultation: After completing the triage process, the patient goes to the waiting room (sitting or on a stretcher depending on the severity) until an appropriate box becomes available. Then, she is transported and installed in the box by an appropriate nurse except ESI 5 patients who can do it themselves. The consultation starts once a doctor that is responsible for the patient category becomes available. The doctor makes a first assessment and may request tests in order to confirm or refine her diagnosis. In case there is no examination required, the patient is discharged from the system. After the consultation, the doctor reports the diagnosis and the decisions made in the information system. Moreover, some important organizational aspects in the model are to be mentioned:

- Each decision made by a junior doctor must be validated by a senior one,
- Each patient must be treated by the same doctor and the same nurse all along the process. The "same patient-same staff" constraint, mentioned in Saghafian et al. (2012) and Saunders et al. (1989), is a strong constraint with a significant impact on the system behavior,
- Among any given ESI level and for any doctor, arriving patients have the priority over in-process ones.

(3) Diagnosis tests: According to the decision made by the doctor, there is a large variety of diagnosis tests that can follow the consultation. The doctor can order an electrocardiogram which is generally performed by a nurse in the box. Blood tests can be ordered; the nurse is responsible for the sampling in the box. Then, the sample is sent to the laboratory to be analyzed. During this time, the patient can wait in her box or can be put in an internal waiting room (if possible) in order to release the box and make it available for other patients. This decision depends on the patient condition and we integrate it in our model by using a certain

probability for each ESI. The duration of blood tests starts at that moment and finishes as soon as the results are ready. It represents one of the longest delays in the ED. Radiology tests can be also ordered with different combinations of X-Ray, CT scan, Echo and MRI. Note that LC patients must be transported by a Stretcher Bearer. When both tests are ordered, radiology and lab tests periods generally overlap. Analgesics can also be requested by the doctor. In the case of a perfusion, it will be done at the same time with the sampling (if any). It requires however an additional delay because a preparation beforehand is needed.

Diagnosis tests are undergone by resources located in another department and shared with other services of the hospital. Therefore, the durations that we fit do not represent only processing times, but the total wait for the results. We include in this duration waiting times outside the ED. Consequently, reducing waiting times for external activities (radiology and laboratory) falls out of the scope of this study. They are considered as incompressible.

(4) Result interpretation and decision of the outcome: Once all the tests are completed, the doctor responsible for the patient evaluates the results, makes an interpretation and decides how the treatment procedure will be continued. In several cases, the doctor asks the patient to undertake supplementary examinations or even to redo some already taken examinations. The doctor can also request the opinion of a specialist from the hospital, a scenario that we model with a certain probability. Since the specialist belongs to another department, her intervention implies three additional durations: The time that the ED doctor spends to call the specialist by phone, the time necessary for the specialist to arrive, and the discussion with the ED doctor once she arrives. The duration is longer when the ED doctor is a junior one due to the lack of experience and her interest in learning.

(5) The process outcome: After the completion of the treatment procedure, the patient can be transferred to another service of the hospital, transferred to another hospital, admitted in the observation unit (OU) or discharged. When a patient is transferred to another department to be hospitalized, the responsible doctor must organize the transfer by phone. Then, the stretcher bearer is responsible for the transportation and the installation of the patient to the destination department. When a patient is transferred to another hospital, the responsible doctor must also call the hospital to organize the transfer. In this case, the transportation to the ambulance is done by the ambulance crew.

The OU is the area of the ED that hosts patients for a short stay before a transfer to another unit that could be the ward of the hospital or another hospital, or when the patient situation requires an additional observation before being released (Broyles and Cochran, 2011). The beds are the critical resources of the OU. It has a limited capacity of beds and it admits and releases

patients only during some specific periods of the day. Observation units are generally neglected in ED modeling in the literature, and yet it is very important to include them because they interact with the rest of the ED and have an impact on its performance. In Saint Camille ED, when the OU is full, patients supposed to be admitted are kept in the ED, laid in boxes or in the internal waiting room. In this case, a nurse from the ED must control these patients regularly, as described also in Weng et al. (2011).

It is well known that the quality of output data relies on the accuracy of input parameters. Therefore, data collection and analysis are undertaken carefully. The first step consists of the collection of the different types of data. In the second step, we model the data with statistical distributions in order to use them as input parameters for the model. Our simulation model requires three types of data: arrival pattern, processing times and routing probabilities. Depending on their type, ED data are more or less easy to collect. Thus we relied on the wide variety of data sources commonly used in similar studies and summarized in Paul et al. (2010): records from databases, interviews with experts and decision makers, and on-site observations; in addition to comparison with other EDs (VUmc and St. Antonius databases, and some input data provided by similar studies (Khare et al., 2009; Centeno et al., 2003; Weng et al., 2011)). Arrival pattern and some routing probabilities are relatively easy to collect since the corresponding data is systematically recorded and stored in the ED database. On the other hand, processing times and some process information are not recorded. For the above we used on-site observations and interviews with experts.

Arrival pattern: Similarly to Yom-Tov and Mandelbaum (2014) and Ahmed and Alkhamis (2009), we assume that arrivals follow a non-homogenous Poisson process. The time dependent arrival pattern is quite typical for most EDs in the world (Zayas-Caban et al., 2013). Monday is usually the day that records the most arrivals, whereas higher arrival rates are found in the period between 10 am and 10 pm for any given day. Arrivals are modeled by using an average arrival rate $\widehat{\lambda}(t)$ for each hour of the week (7 days $\times$ 24 hours = 168 rates). These 168 rates are estimated from the database of Saint Camille ED for 103 consecutive weeks, starting from September 2011 and ending in September 2013 (Figure 4.3).

Processing times: There are 26 different service times that we modeled with statistical distribution fits, using the package Input Analyzer in Arena software. The processing times for each step of the process depend on the resource type (junior doctors are slower than seniors) as well as patient category (critical patients require more time).

Routing probabilities: These probabilities depend on the patient ESI and represent the

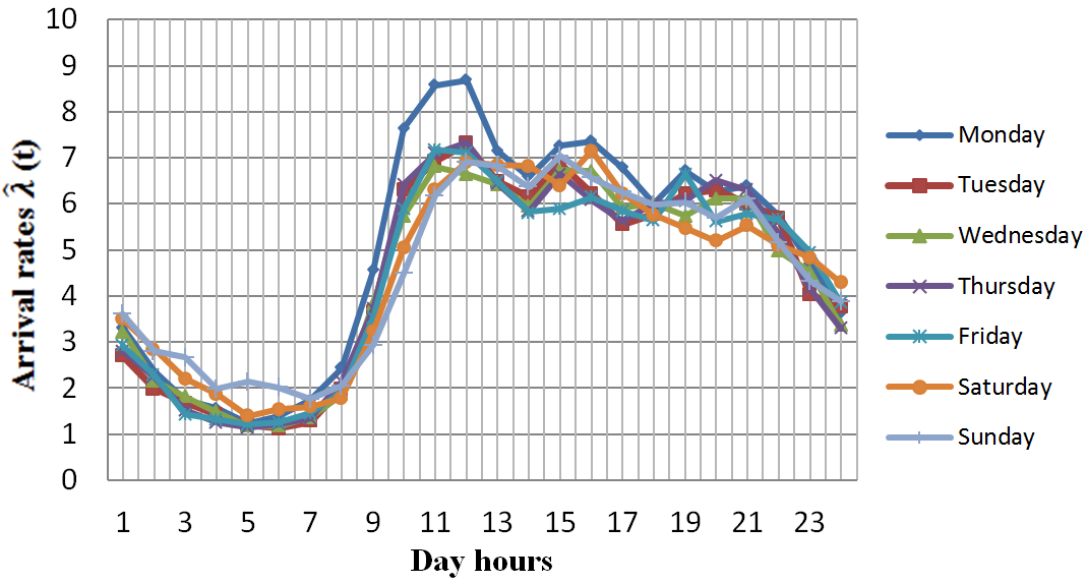

Figure 4.3: Estimated hourly patient arrival rates $\widehat{\lambda}(t)$ per day

chance for a patient to experience or not a certain stage of the process. The probabilities needed in our model correspond for instance to diagnosis tests, the mix of these tests (imaging, lab test, none or both), imaging mix (X-ray, scan, echo or MRI), patient abandonment, the need for specialist opinion, the clinical outcome, Remaking tests, observation unit outcome, etc.

4.2.2 Model validation

Law and McComas (2001) explain that if the model is not a close approximation of the real system, any conclusions derived from the model are likely to be erroneous and may result in costly and ineffective decisions. Simulation models need to be built in a very precise way in order to represent the real environment as realistically as possible. The completion of our simulation model was a long procedure that contained many iterations; each step of the conceptual model had to be validated by experts in order to secure that it is an accurate representation of the system.

Exhaustivity: Concerning the granularity of simulation models, researchers have stated in the past that EDs are such complex systems that it is impossible to take all their features into consideration. Robinson (1994) has shown that in most cases, 80% of model accuracy is obtained from only 20% of the model detail. However, ED models in the literature generally use many assumptions where important characteristics of the system are neglected. In most cases, such simplifications are more frequent in models using analytical methods, but they still exist in simulation models as well.

Building a realistic and useful simulation model requires an appropriate selection of the model level of detail. Table 4.1 synthesizes some of the important features included in our model, and compares that with the existing studies. For instance, the feature Resources Subdivisions refers to the differentiation of the staff members. As explained in Sinreich and Marmor (2005), some EDs distinguish between acute and ambulatory patients and allocate doctors accordingly. Another possible subdivision is the difference between seniors and juniors (generally neglected). This is included in our model where processing times are function of both the expertise and the patient category.

Comparison with real data: To validate the simulation model, we compare between LOS given by our model and that obtained from the ED data using descriptive statistics.

We consider a steady-state type simulation run with one pseudo-infinite length of time during which the system is not re-initialized. This is coherent with the real system that works without interruption (24/7). The replication length is 11 weeks (110,880 minutes), of which one week is used as a warm-up period (10,080 minutes). The choice of the warm-up duration is based on graphical inspection of the time-series of the simulation outputs. We observe that after one week the system reaches typical conditions of steady-state situations. Note that we do not use a cool-down period because the ED works 24/7 without interruption.

Figure 4.4 provides a box-plot where the real LOS of 37,986 patients is compared to the LOS given by simulation for 7,604 patients. The outliers represent less than 5% for both real and simulated values. Figure 4.4 shows that there are some differences between the two distributions. Nevertheless, the comparison between the real and simulated cumulative distributions reveals encouraging similarities (Figure 4.5). For instance, starting from $LOS = 200$ minutes, the two distributions become very close. Furthermore, we successfully confronted two other indicators with expert judgment: resources workload and the durations of the five stages of the ED process (including the corresponding waiting durations). These encouraging similarities allowed considering the model reliable and valid to support experiments.

4.3 Staffing level optimization

Investing in human staffing is one of the possible ways to improve the ED performance. We want to address the following questions: By how much should we increase the current staffing budget, and how should this additional budget be used in the allocation of human resources? The results of this study has stood as a strong argument in order to convince the Saint Camille hospital management on the usefulness of increasing the funding for ED staffing. In general,

Table 4.1: Comparison of previous works and the present study in terms of model granularity

	Centeno et al. (2003)	Komashie and Mousavi(2005)	Duguay and Chetouane (2007)	Ahmed and Alkhamis (2009)	Weng et al. (2009)	Present Study
Arrival process	Depends on day period	Depends on week day	Depends on week day	Depends on day hours	Depends on day period	Depends on week day and day hours
Patients categories	4	2	5	3	4	5
Included resources	Doctors Nurses Boxes	Doctors Nurses Boxes	Doctors Nurses Boxes	Receptionists Doctors Nurses Boxes Lab technicians Beds	Doctors Nurses Sick Beds	Stretcher Bearers Doctors Nurses Boxes Sick Beds Beds
Resources subdivision	No	Yes	No	No	Yes	Yes
Severity and/or expertise based processing times	Yes, based on severity	Yes, based on severity	Yes, based on severity	No	No	Yes, based on both
Lab tests/ radiology	Yes	No	Yes	Yes	Yes	Yes
transportation times	No	No	No	No	No	Yes, for patients
Staff shifts	Yes	No	Yes	No	No	Yes
Teaching aspects	No	No	No	No	No	Yes
Specialist	No	No	No	No	No	Yes
Abandonment	Yes	No	No	No	No	Yes
Observation unit	No	Yes	No	Yes	Yes	Yes
Experiments	Simulation-optimization	Intuitive what-if scenarios	Intuitive what-if scenarios	Simulation-optimization	Simulation-optimization	Simulation-optimization
Control variables	Nurses	All included resources	All included resources	Doctors Nurses Lab technicians	Doctors Nurses	All included human resources

similar approaches are also expected to support decision maker arbitrations.

We formulate an optimization problem that seeks to minimize the average length of stay under a budgetary constraint, and a constraint ensuring that the average $DTDT$ of LC patients ($\overline{DTDT}$) does not exceed some specified threshold. This is a hard problem, for which we use Arena OptQuest package for simulation-optimization. OptQuest is a commercial global optimizer that uses heuristics to efficiently explore the set of feasible solutions (Adenso-Diaz

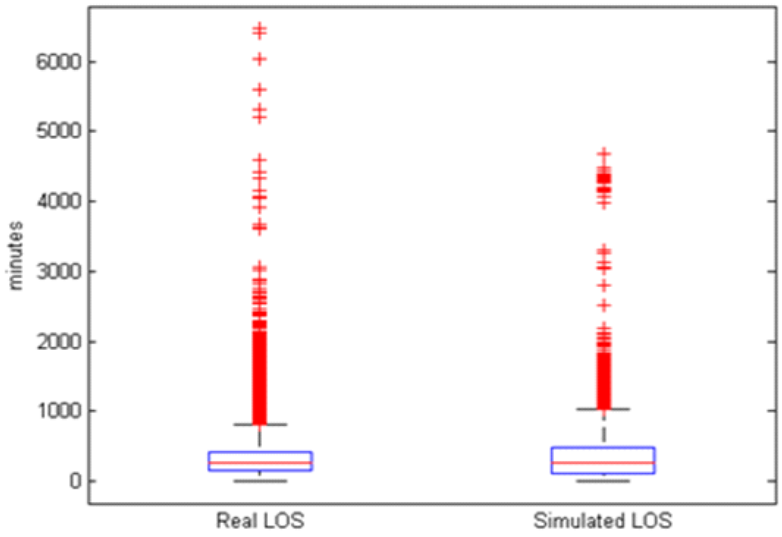

Figure 4.4: Real and simulated LOS

and Laguna, 2006; Kleijnen and Wan, 2007). The ED uses two different shifts, a first one from 9:30 am to 6:30 pm (day shift), and another one from 6:30 pm to 9:30 am (night shift). Let I = {Senior, Junior, Nurse, Triage nurse, Stretcher bearer} be the set of the considered resources with all possible subdivisions detailed in Section 4.2.1. Let J = {Day shift, Night shift} be the set of the considered shifts. The real salaries of the ED staff have been used. The control variables $X_{i,j}$ represent the amount of a certain resource i during a given shift j, which applies to the different days of the week. This is consistent with practice where resources staffing levels in Saint Camille ED, with the exception of weekends, are the same during the week. These variables are defined in Arena and used as control variables in OptQuest. For each resolution, OptQuest needs a starting solution that will serve as a starting point for exploring the set of feasible solutions. The initial parameters we choose correspond to the actual scheduling used in Saint Camille ED. Since the results of the optimization can slightly vary according to the initial solution, we made each optimization several times by varying the starting parameter values. For practical reasons, the staffing levels for doctors during weekends will remain unchanged. The problem is expressed as follows:

$$\begin{cases} \min \overline{LOS} \\ \text{subject to} \\ \sum_{i=1}^{n}\sum_{j=1}^{m} C_{i,j}X_{i,j} \leq C(1+\alpha), \text{ for } i\epsilon I, j\epsilon J \\ \overline{DTDT} \leq L, \\ X_{i,j} \geq 0, \text{ for } i\epsilon I, j\epsilon J \end{cases} \tag{4.1}$$

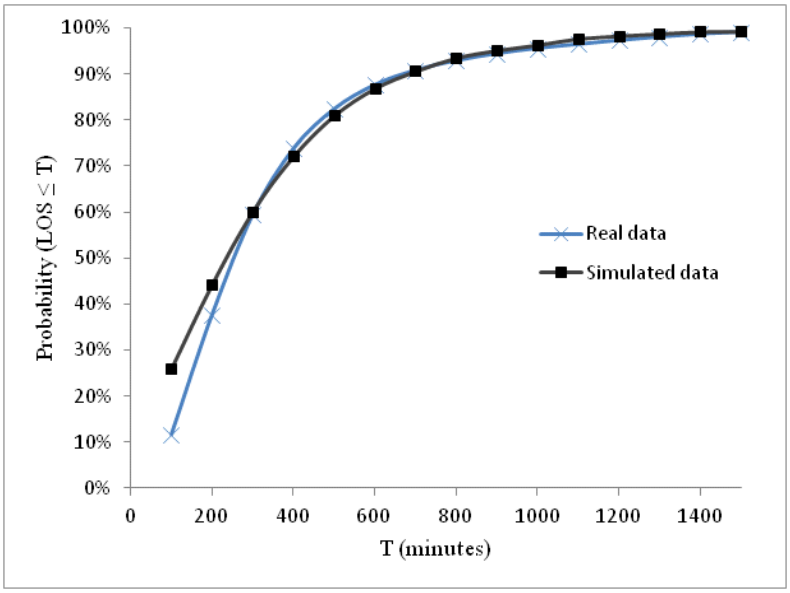

Figure 4.5: Cumulative distributions of real and simulated LOS

where

$\overline{LOS}$ = Average length of stay in the system,

$X_{i,j}$ = Amount of resource i during shift j,

$C_{i,j}$ = Salary for resource i during shift j,

C = Current staffing budget,

α = Percentage of additional staffing budget,

$\overline{DTDT}$ = Average door-to-doctor time for LC patients (ESIs 1, 2 and 3),

$L = \overline{DTDT}$ limit.

The first constraint represents the staffing budget constraint. The budget limit is expressed as a function of a coefficient α that is the percentage of additional staffing budget. The second constraint secures that the average door-to-doctor time for LC patients does not exceed a predetermined threshold L. Although the differences in staffing requirements for LC patients (junior doctors), we do only consider one single $DTDT$ constraint for all LC patient types. In practice, the most important point, with regard to $DTDT$, is the classification LC/SC and not the resource type allocations.

We perform a sensitivity analysis by varying at the same time α and L. Table 4.2 gives the results obtained by simulation-optimization. Cells containing INF indicate that the combination of the budget and $\overline{DTDT}$ constraints can not produce a feasible solution. The remaining values are the achieved $\overline{LOS}$, measured in minutes for an arbitrary patient. It should be mentioned that when the limit L is higher than 57 minutes, which is the value obtained in the initial

simulation model with no supplementary budget, then the constraint is relaxed.

Table 4.2: Numerical experiments for the optimal $\overline{LOS}$

Additional Staffing Budget (α)	Current $\overline{DTDT}$=57	$\overline{DTDT}\leq$50	$\overline{DTDT}\leq$40	$\overline{DTDT}\leq$30	$\overline{DTDT}\leq$20	$\overline{DTDT}\leq$10
0%	367	485	INF	INF	INF	INF
5%	323	389	397	INF	INF	INF
10%	246	277	277	INF	INF	INF
20%	205	205	205	229	INF	INF
30%	182	182	182	182	221	INF
40%	171	171	171	171	192	INF
50%	165	165	165	165	165	INF

Table 4.3: Resource staffing for optimal solutions of the sensitivity analysis

Additional Staffing budget (α)	**Current $\overline{DTDT}$=57**	**$\overline{DTDT}\leq$50**	**$\overline{DTDT}\leq$40**	**$\overline{DTDT}\leq$30**	**$\overline{DTDT}\leq$20**	**$\overline{DTDT}\leq$10**
0%	Initial staffing	-1 senior SC night shift +1 junior ESI3 day shift +1 junior ESI45 day shift	INF	INF	INF	INF
5%	+1 senior SC day shift	+1 senior LC night shift -1 senior SC day shift +1 junior ESI345 night shift	+1 senior LC night shift -1 senior SC day shift +1 junior ESI45 day shift	INF	INF	INF
10%	+1 senior LC night shift	+1 senior LC night shift -1 senior SC night shift +1 junior ESI345 night shift +1 triage nurse day shift	+1 senior LC night shift -1 senior SC night shift +1 junior ESI345 night shift +1 triage nurse day shift	INF	INF	INF
20%	+1 senior LC night shift +2 nurses LC night shift	+2 senior LC night shift +1 junior ESI3	+2 senior LC night shift +1 junior ESI3	+1 senior LC night shift -1 senior SC day shift +1 senior SC night shift +1 junior ESI45 day shift	INF	INF
30%	+2 senior LC night shift +1 senior SC day shift +1 nurse LC night shift +1 triage nurse day shift	+2 senior LC night shift +1 senior SC day shift +1 nurse LC night shift +1 triage nurse day shift	+2 senior LC night shift +1 senior SC day shift +1 nurse LC night shift +1 triage nurse day shift	+2 senior LC night shift +1 senior SC day shift +1 nurse LC night shift +1 triage nurse day shift	+2 senior LC night shift -1 senior SC day shift +2 senior SC night shift +2 triage nurses night shift	INF

From Table 4.2, we observe that the budget has a diminishing marginal effect on performance. This can be seen from the first column of the table where the $\overline{DTDT}$ constraint is relaxed. The highest marginal effect of the coefficient α on the $\overline{LOS}$ corresponds to an investment of 10% of the current budget. This result allowed the ED managers with the general management of Saint Camille hospital to take an important tactical decision that consists in increasing the current staffing budget by 10% in order to reduce the current $\overline{LOS}$ by 33%.

We also observe that the $\overline{DTDT}$ constraint affects the optimality or the feasibility of the problem for small budgets. In certain cases, the limit L cannot be met by any possible allocation of resources and therefore the problem is infeasible. In other cases, by decreasing the limit of the $\overline{DTDT}$ constraint for a certain budget, the optimal $\overline{LOS}$ increases. For example, for α=20%, any value of $L \geq 40$ leads to an optimal $\overline{LOS}$ of 205 minutes. However when $L = 30$, the optimal $\overline{LOS}$ increases to 229 minutes. For high budget levels, the $\overline{DTDT}$ constraint is automatically satisfied (staff allocation secures a low $\overline{DTDT}$), and thus the $\overline{LOS}$ is independent of this constraint to some extent. This captures the trade-off between the two performance metrics.

The explanation of the last result requires the examination of the different solutions of Table 4.2 in terms of resource staffing. Table 4.3 provides the staffing changes for each optimal solution with regard to the initial staffing solution with no additional budget (α=0%, $L = 57$).

We can observe in all cases (for all problem formulations, i.e., with or without the $\overline{DTDT}$ constraint) that the resource doctor is the most preferred one. There is always at least one additional doctor for all combinations of investment and $\overline{DTDT}$ limit. Concerning the additional doctors type, with the use of the $\overline{DTDT}$ constraint ($L \leq 50$), resources tend to be devoted to LC patients in order to reduce $\overline{DTDT}$. For instance, when α=5%, an LC doctor is added during night shift to satisfy the $\overline{DTDT}$ constraint while an SC doctor is added when this constraint is relaxed ($\overline{DTDT}$=57). This means that under the $\overline{DTDT}$ constraint, there are less available resources for the SC patients (majority of patients) which increases the overall $\overline{LOS}$. Up to a certain budget (α=10%), there is no investment on other resources such as nurses. This is consistent with the fact that senior doctors workload is the highest among all ED human resources.

When higher budgets are available, additional nurses are staffed. For instance, when α=20%, two additional nurses are added during night shift for LC patients when the $\overline{DTDT}$ constraint is relaxed. Note that the nurse type privileged to overcome the $\overline{DTDT}$ limit are triage nurses (not "in-process" nurses) because the triage stage and the corresponding waiting time is a part of the $DTDT$. For instance, when α=10%, one additional triage nurse is staffed during day shift to

satisfy the $\overline{DTDT}$ constraint. For higher budgets ($\alpha \geq 30\%$), resources are devoted independently of the $\overline{DTDT}$ constraint. This means that regardless to the $\overline{DTDT}$ constraint, there are enough resources to secure that the LC patients will be treated within the threshold L.

The main conclusions from the above observations can be summarized as follows:

- Additional investments should be allocated in priority to doctors. A restrictive quality of service in terms of $\overline{DTDT}$ will further give priority to LC doctors. This result seems surprising and counterintuitive to ED managers. As explained in Paul et al. (2010), these findings are interesting given the large amount of research focusing on optimizing nursing allocation in various parts of the hospital (Miller et al., 1976; Shuman et al., 1975; Burke et al., 2004). Only rare papers focus on the important impact of doctor scheduling (compared to that of nurse) on the ED performance (Clark and Waring, 1987; Evans et al., 1996).

- The lower is the budget, the more apparent is the correlation between $\overline{LOS}$ and $\overline{DTDT}$.

4.4 Conclusions

We have built a realistic ED model using discrete-event simulation. All common structural and functional characteristics of EDs, at least in France, were taken into consideration thanks to a close collaboration with practitioners. Based on the above, we point out a set of important ED features that are frequently ignored in the related literature. Although a simulation model can not be an exact imitation of the real system, the characteristics that we mention should be preferably taken into account in ED models, given their impact on the system performance. Our experiments focused on human staffing levels and provided useful insights to managers on the impact of the budget and $\overline{DTDT}$ constraints on $\overline{LOS}$.

We observed that the staffing budget reveals a decreasing marginal effect on performance. For instance, an increase of 10%, 20% and 30% in the staffing budget can generate respectively an improvement of 33%, 44% and 50% in the optimal $\overline{LOS}$, when the $\overline{DTDT}$ constraint is relaxed. Moreover, managers should be aware of the correlation between $\overline{DTDT}$ and $\overline{LOS}$, for a given staffing budget. In some cases, $\overline{DTDT}$ limits cannot be met with the use of several budgets, whereas in other cases meeting the $\overline{DTDT}$ limits for the most severe patients has a negative effect on the total length of stay of all patients. The explanation lies in the fact that for low $\overline{DTDT}$ targets, the budget tends to be devoted to urgent patients at the expense of non urgent patients (that represent the majority of patients) which affects the overall $\overline{LOS}$. Besides, we derived insights about the most appropriate type of resource to prioritize depending on the

available staffing budget and the $\overline{DTDT}$ target. We surprisingly find that additional investments should be allocated in priority to doctors, which is counterintuitive to ED practitioners. The results provide to managers a better understanding on how the budget can affect the system performance as well as on the interdependency between the two main ED KPIs. This may then assist them in choosing the most appropriate operational decisions.

Some limitations of the current study are as follows. One limitation is related to input data. For instance, we considered routing probabilities and processing times as a function of the patient severity. However, in practice, some of these data depend also on the patient age or the medical specialty required for her treatment. Even though some correlations between several aspects exist, such as between ESI and age (see Chapter 3), we think that this represents a shortcoming. Moreover, we used an abandonment probability for patients as input to the model, while this parameter should be an output that depends on the patient waiting time before abandonment. Unfortunately, the data about abandonment times is not reliable since it is not registered in the database when the patient leaves the ED, but only once her absence is noticed by the staff. Another limitation is related to the designed process. We assumed that the health status of a patient does not deteriorate during her sojourn in the ED, which is not the case in general. Since this may affect the in-process operations and durations, the simulation model can present a lack of accuracy.

Chapter 5

Resource-related experiments: A heuristic for definition of shifts

In this chapter, we address the question of how to define efficient work-shifts that make the best use of current resource capacity given the demand profile. The problem of shift definition was rarely addressed in the literature, and researchers generally use predetermined shifts, designed intuitively by practitioners. Yet, answering to the question of how to divide the day into different shifts properly could provide ED managers with a cost effective and simple way to improve ED performance. We propose a model that combines simulation-optimization and linear programming in order to define the shift pattern that best match the arrival pattern of patients in an emergency department. The final solution must respect a certain staffing budget and satisfy the main constraints encountered in practice. The simulation model supplies the linear programming with the staffing levels that secure the performance of the ED, expressed in terms of the average length of stay of patients. The linear model determines the shift-scheduling of all employees with the use of the minimum cost, including several constraints as experienced in practice. The model includes also a heuristic which leads to a solution that satisfies budget restrictions. The application of the developed method leads to a reduction of 8.9% in the ED average LOS with the use of the same staffing budget.

This work is published in the proceedings of the *45th International Conference on Computers and Industrial Engineering (CIE45)* held in 2015, in Metz, France (Ghanes et al., 2015a).

5.1 Introduction

One cause of inefficiency in EDs is that due to the sporadic demand, the staff are idle at times and overworked at other times (Hanna et al., 1974). Concerning staff allocation, there are several issues that ED managers have to deal with. First of all, the ED must be able to respond to the demand of patients with adequate staffing levels. Staffing involves determining the number of personnel of the required skills in order to meet predicted requirements (Burke et al., 2004). Meeting the staffing levels can be a challenging task. The latter are allocated based on shift-scheduling, which deals with the assignment of the number of employees to each shift, in order to meet demand (Ernst et al., 2004). Finally, Rostering deals with the work schedule of each employee in the ED and the shifts that this particular employee will work in for a certain period of time (usually week or month). The allocation of staff contains numerous constraints. For example, in rostering management, a certain employee cannot work more than an upper limit of hours per week and simultaneously cannot work in consecutive shifts. In our model we propose a method that determines a shift-scheduling model; rostering of the shifts falls out of the scope of this study. The performance of an ED can be measured with the use of the average length of stay (LOS) which is the KPI on which EDs are generally judged in practice, because it allows to approach the ED in a holistic way and gives an overview of the entire system performance (see Chapter 2).

The most straightforward way to alleviate crowding and improve responsiveness is by adding resources. This approach is widely spread in the literature of resource allocation (Komashie and Mousavi, 2005; Duguay and Chetouane, 2007), and we investigated how to use it rationally and efficiently in Chapter 3. However, because this is also the most expensive approach, and because of the worldwide budgetary restrictions in healthcare, it is generally not the preferred option (Saghafian et al., 2012; Carmen and Van Nieuwenhuyse, 2014). Nowadays, the number and the capacity of EDs is decreasing while the number of patients visiting EDs is continuously increasing all over the world (Derlet and Richards, 2000; Schafermeyer and Asplin, 2003; McCaig and Burt, 2004; Green et al., 2006; Hoot and Aronsky, 2008; Niska et al., 2010; Harrison and Ferguson, 2011; Abo-Hamad and Arisha, 2013). In such a context, it became crucial to explore cost effective alternatives and opportunities that optimizes EDs with limited investment or ideally, with fixed budget.

In the literature related to staff scheduling, authors generally use a preexisting shift set defined intuitively by practitioners and that might not match adequately with patient arrival pattern (for instance day shift, evening shift and night shift). Only rare papers addressed the problem of shift definition. Yet, we believe that answering to the question of how to divide the

day into different shifts properly may provide ED managers with a cost effective and simple way to improve ED performance. Shift Definition is a complex and large combinatorial problem since the shift set can vary according to the number of shifts, start times and durations of shifts, allowing shifts' overlapping or not, using the same shift pattern for all resource types or not, etc.

In this chapter, we propose a method which allows generating work-shifts that best fit the demand. In contrast to the majority of resource allocation literature, decision variables are not restricted to staff levels, but the search for optimal shifts is done simultaneously. We further demonstrate that this method can improve the system performance without any investment in resource allocation. We use the realistic discrete-event simulation (DES) model presented in Chapter 4. We formulate an optimization problem that seeks to minimize the average LOS under a budgetary constraint, using resource staffing levels as variables. We solve this problem using Arena OptQuest package for simulation-optimization. Concerning shift-scheduling, we use a linear program (LP) that we solve using Cplex. The main goal of the LP is to create shifts of minimum cost, while obeying to the performance standards (expressed in terms of staffing levels obtained from simulation-optimization) and other practical constraints discussed further in Section 5.3.2. Simulation-optimization and LP are the two tools that provide the initial solution, which might violate the staffing budget. Therefore, the developed heuristic searches for the feasible solution by decreasing the staffing costs in a way that harms the performance of the ED as less as possible.

The rest of the chapter is structured as follows. In Section 5.2, we provide a brief literature review on relevant issues concerning staff allocation in the ED and similar systems. In Section 5.3, we present our method, analyzing the simulation-optimization, the LP and the heuristic. The method is applied on a real case in Section 5.4 and the results are presented. In Section 5.5, we conclude, present the limitations of our study and propose some future work possibilities.

5.2 Literature review

In this section we present some articles with relevant work to our study. Besides the ED domain, we present some research performed in call centers, a domain that has some common characteristics with EDs, as well as some reviews on personnel staffing.

In the ED context, Centeno et al. (2003) combine a simulation model with LP in order to provide shifts that contain adequate staffing levels. However, they select between five predetermined shifts of fixed length that have different starting points and they do not include any budget restrictions in their model. An example of personnel staffing is the paper of Beaulieu

et al. (2000a), who are constructing a mathematical programming model that determines, considering predetermined sets of shifts, the way that physicians are scheduled in the ED within a specified period of time. In general, the environment in which employees work might affect their productivity and thus several researchers have taken into consideration the employees preferences in personnel scheduling problems. Yankovic and Green (2011) have used queueing theory in order to determine the staffing levels of nurses in hospitals, developing a heuristic that gives good approximations of the analytic problem. Sinreich et al. (2012) use simulation models combined with heuristics in order to allocate the predetermined 8 hour-shifts for each type of employee within the day based on the performance of the ED. The total number of each employee remains the same in the previous model, but this does not necessarily secure that the budget remains the same, as employees usually have different costs during the day.

The issue of staff allocation for approaching a performance goal has been studied by operations management researchers in the domain of call centers. Wallace and Whitt (2005) focus on a problem with skill-based routing, whereas Robbins and Harrison (2010) use stochastic programming for scheduling call-centers. Pot et al. (2008) use a two-stage method that initially determines staffing levels that are then grouped in shifts. This model is close to our method, with the main difference being that the analytical methods that determine the staffing levels in the first step (while we use simulation) cannot be applied in the ED, as ED employees have more numerous and complex tasks than servers in call-centers. Avramidis et al. (2010) propose a cutting planes method with the use of simulation in order to schedule agents in the ED. Heuristics for staffing multi-skill call centers have been used by Pot et al. (2008) and Avramidis et al. (2009). Ernst et al. (2004) and Van den Bergh et al. (2013) stand as two examples of reviews on personnel staffing.

5.3 Method

In this section we present the method used to obtain the shift schedules that optimize the ED performance, while obeying to the budget restriction. We briefly present the simulation model that is used as basis for the simulation-optimization in Section 5.3.1, and then we explain in detail the linear program (LP) model in Section 5.3.2. Finally, in Section 5.3.3, we propose a heuristic that combines the results of the above models and secures that the budget constraint will be met by the final staff allocation in the shifts.

5.3.1 Simulation-optimization to provide ED staffing levels

We use the simulation model developed in Chapter 4 as a basis for this part. The main advantage of this model is that it was validated as a good representation of the real system. It includes essential ED features which allows representing the actual performance of the ED very well. As explained previously, the patient path in the ED depends on her severity: patients of ESI 1, 2 and 3 are noted as Long Circuit (LC) patients (critical patients), whereas ESI 4 and 5 are noted as Short Circuit (SC) patients (non-urgent patients). There are 8 different employee types (senior and junior doctors, nurses, stretcher bearers, etc.) that might be assigned to one of the two categories of patients, or both. The details are shown in Table 5.1. Even though a single KPI cannot assess perfectly the performance of the system, the most suitable metric that approaches the ED from a holistic point of view is the length of stay (LOS), which measures the total average sojourn of all patients in the system (see Chapter 2). The LOS contains all the waiting times generated in each queue in which a patient has to wait during his sojourn in the ED.

Table 5.1: Types of employees in the ED

Employee	Category
Senior Doctor 1	SC
Senior Doctor 2	LC
Junior type 1	LC (ESI 3)
Junior type 2	SC (ESI 4, 5)
Junior type 3	mixed (ESI 3, 4, 5)
Nurse 1	SC
Nurse 2	LC
Triage nurse	all
Stretcher Bearer	all

The objective is to determine the staffing levels, or in other words the number of each employee type required for every hour of the day in order to minimize the total average LOS of patients in the day. To this aim, we use the optimization problem formulated in Chapter 4 that seeks to minimize the average LOS under a constraint on the staffing budget. To compute this program, we have previously used in Chapter 4 the real shifts that existed in Saint Camille ED. Here, we neglect these shifts and consider instead 24 periods (i) with a length of one hour each. The parameters obtained by simulation-optimization are $\mathbf{a_{i,l}}$ and represent the optimal amount of employee l required in period i in the ED. The subprogram selected for the above purpose is OptQuest package contained in Arena Simulation software. The staffing levels $\mathbf{a_{i,l}}$ are then used in the constraints of the next part of the model, which is a linear programming model.

5.3.2 The linear programming model to define shifts

The objective at this step is to find the shifts that are able to satisfy the staffing levels $\mathbf{a_{i,l}}$ provided by simulation-optimization, while minimizing the corresponding budget. In addition, when dealing with shift-scheduling, there are several practical constraints that should be considered. Each shift must have a minimum length, because it is not reasonable that employees go to work, for example, for only one hour. In several cases, the shifts are constrained to start only in convenient hours of the day. In our case, we will consider that employees will not start their shifts between midnight and 5 am. Furthermore, during a given day, a limited number of shifts must be scheduled for a given resource; usually in each day EDs have 2 to 3 shifts for each type of employee, but even 4 or 5 can be feasible. The number of shifts per day also depends on the total number of employees available for each day, but in our case we consider that there are enough employees to meet the shift schedules proposed. Finally, we should mention that other features, such as fixed breaks for employees, are not taken into consideration in this model as most employees usually adjust their breaks in time periods where demand is low. The complete linear programming model is given below.

$$\text{Minimize } Budget = \sum_{i=1}^{imax} \sum_{l=1}^{lmax} \mathbf{C_{i,l}} * (\mathbf{Y1_{i,l}} + \mathbf{Y2_{i,l}})$$

subject to:

$$\mathbf{Y1_{i,l}} = \sum_{j=1}^{i} \sum_{k=i-j}^{kmax} \mathbf{x_{j,k,l}}, \text{ for all } i, l \tag{5.1}$$

$$\mathbf{Y2_{i,l}} = \sum_{j=i+1}^{jmax} \sum_{k=kmax-j+i}^{kmax} \mathbf{x_{j,k,l}}, \text{ for all } i, l \tag{5.2}$$

$$\mathbf{Y1_{i,l}} + \mathbf{Y2_{i,l}} \geq \mathbf{a_{i,l}}, \text{ for all } i, l \tag{5.3}$$

$$\sum_{j=1}^{jmax} \sum_{k=1}^{kmax} \mathbf{w_{j,k,l}} \leq \mathbf{shift_l}, \text{ for all } l \tag{5.4}$$

$$\mathbf{w_{j,k,l}} \leq \mathbf{x_{j,k,l}} \leq \mathbf{M} * \mathbf{w_{j,k,l}}, \text{ for all } j, k, l \tag{5.5}$$

$$\sum_{j=1}^{jmax} \sum_{k=1}^{kmin} \mathbf{w_{j,k,l}} = 0, \text{ for all } l \tag{5.6}$$

$$\sum_{j=1}^{jmin} \sum_{k=1}^{kmax} \mathbf{w_{j,k,l}} = 0, \text{ for all } l \tag{5.7}$$

Indices:

i= the hour of the day (i=1,$imax$), $imax$=24;

j= the hour of the day when a shift starts (j=1,$jmax$), $jmax$ and $jmin$ are the maximum and the minimum shift starting hours, respectively;

k= the duration of the shift (k=1,$kmax$), $kmax$ and $kmin$ are the maximum and the minimum shift lengths, respectively;

l= the different types of employees (l=1,$lmax$), $lmax$ is the total number of employee types.

Parameters:

$\mathbf{C_{i,l}}$ = the cost paid for an employee of type l working in period i;

$\mathbf{a_{i,l}}$ = the number of employees of type l required in period i (determined by simulation-optimization);

$\mathbf{shift_l}$ = the maximum number of shifts allowed within a day for employees of type l

M= a big number.

Variables:

$\mathbf{Y1_{i,l}}$ = number of employees of type l that started working in the same day and are working in period i (integer variable);

$\mathbf{Y2_{i,l}}$ = number of employees of type l that started working in the previous day and are working in period i (integer variable);

$\mathbf{x_{j,k,l}}$ = number of employees of type l that started working in period j for a duration of length k (integer variable);

$\mathbf{w_{j,k,l}}$ = binary variable that shows if there are employees of type l that start working in period j for a duration of length k.

We should mention that $\mathbf{Y1_{i,l}}$ and $\mathbf{Y2_{i,l}}$ are redundant variables, as they could have been expressed in terms of the variable $\mathbf{X_{j,k,l}}$. However, they have been used in order to help the reader understand the model more rapidly. The model constructed in Cplex does not include these variables. The connection between simulation-optimization and linear programming is found in Constraint 5.3, where the parameters $\mathbf{a_{i,l}}$ are the results of the staffing levels determined in the previous section, and represent the performance standard. Constraint 5.4 is the constraint limiting the number of shifts per employee type. Constraint 5.5 synchronizes the two variables $\mathbf{x_{j,k,l}}$ and $\mathbf{w_{j,k,l}}$. Constraint 5.6 secures a minimum shift length, and Constraint 5.7 is the one dealing with convenient shift starting hours.

The major problem that arises from the proposed combination (simulation-optimization and LP) is that the final budget obtained from the LP is higher than the real budget restriction of the ED. In other words, the budget restriction posed in the simulation-optimization is violated in the LP. This is due to the fact that simulation-optimization leads to very discontinuous staffing levels that can widely fluctuate from one hour to another. The created shifts supply the ED with more employees than what is actually required by staffing levels. The practical Constraints 5.4, 5.6 and 5.7 are the ones that lead to this overstaffing problem. The LP would be able to schedule the precise number of employees if there were no restriction on the number of shifts per day, on shifts durations and on starting hours. This is consistent with what has been reported in Ernst et al. (2004): it is usually not possible to exactly match the staff on duty to a demand that varies on an hourly basis, when using shifts of several hours long.

For clarification, an example with a maximum number of shifts per day set to two is depicted in Figure 5.1. In this figure, the brief explanation shows that the budget appears in both models, but used differently. It is primarily used as a constraint in the simulation-optimization in order to obtain the performance standards (staffing levels $\mathbf{a_{i,l}}$). Then, it is used in the objective function of the LP, where it is not limited. As a remark, note that the maximum number of different shift types per day for a certain employee l ($\mathbf{shift_l}$) has a diminishing marginal effect. When increasing the number of possible shifts per day for an employee l, the value of the solution improves until a certain limit of $\mathbf{shift_l}$. After this limit, when increasing the number of $\mathbf{shift_l}$, the value of the solution remains the same.

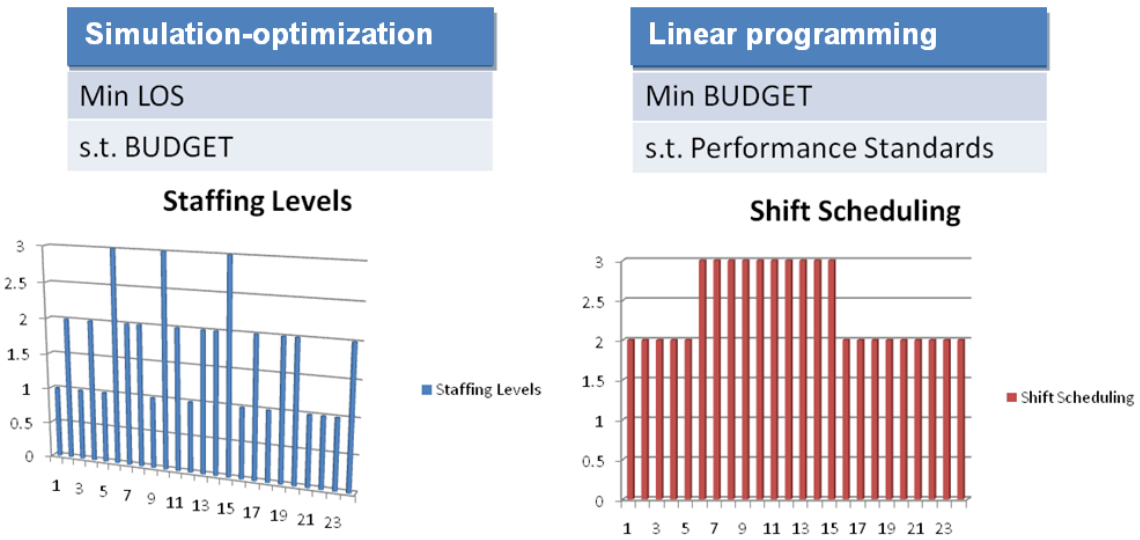

Figure 5.1: Staffing levels using simulation-optimization and shifts created using linear programming

In order to overcome this overstaffing problem, we propose a heuristic that assists ED man-

agers to determine the required shifts, while obeying to the real budget restrictions. In the LP model we replace Constraint 5.3 by 5.8. The other parts of the initial LP remain the same.

$$\mathbf{Y1_{i,l}} + \mathbf{Y2_{i,l}} = \mathbf{a_{i,l}} + \mathbf{b_{i,l}}, \tag{5.8}$$

where:

$\mathbf{b_{i,l}}$ = the difference between the number of employees of type l scheduled in period i and the staffing levels determined by simulation-optimization for the same period and same employee type.

The quantity $\mathbf{b_{i,l}}$ is an integer variable that shows the surplus of employees l in period i. It allows to detect the hours of the day where each shift is over-staffed, fact that leads to the violation of the actual budget. In the proposed heuristic, we try to make modifications in the shifts obtained from the LP model (smoothing modifications), in order to reduce the budget to the predetermined goal, while remaining as close as possible to the optimal solution given by staffing levels.

5.3.3 The heuristic

We propose a heuristic that uses the initial staffing levels and reduces the obtained cost from LP gradually while mitigating the impact on the ED performance. In the first part of the heuristic, we smoothen the staffing levels within shifts by means of transfers (from one $\mathbf{a_{i,l}}$ to another). This procedure is used until all $\mathbf{b_{i,l}}$ are either equal to 0 or 1. In the second part, the objective is to appropriately reduce the number of hours i whose $\mathbf{b_{i,l}} = 1$.

First part of the heuristic:

In the first part of the heuristic (steps 1, 2, 3, 4, 5 and 6), we try to identify the points where consecutive hours have big differences in staffing levels and therefore we try to smoothen this difference by allocating one employee from the hour that has a staffing level with too many employees to an hour with less employees. This modification will not provoke big changes in the LOS of patients, as the total number of employees l in the shift will be greater than the sum of employees given by the staffing levels.

Step 1 If Budget 2> Real Budget, then calculate all $\mathbf{b_{i,l}}$, else *Step 6*

Step 2 Identify all $\mathbf{b_{i,l}} \geq 2$, then find in each corresponding shift the highest value of $\mathbf{a_{i,l}}$, and

reduce this $\mathbf{a_{i,l}}$ by 1 unit, else *Step 6*

Step 3 Add a unit to the closest $\mathbf{a_{i,l}}$ (i.e., $i \pm 1$ then $i \pm 2$, etc.) whose respective $\mathbf{b_{i,l}} \geq 2$

Step 4 Solve the LP and gain new value for Budget 2

Step 5 Repeat steps 1, 2, 3, 4 until $\mathbf{b_{i,l}} \leq 1$ for all i, l

Step 6 Stop

Second part of the heuristic:

In the first part of the heuristic, we have made the differences in staffing levels more smooth. In the second part of the heuristic (steps 7, 8, 9), we have to deal with $\mathbf{b_{i,l}}$ that are equal to either 0 or 1. We modify the obtained shifts in order to delete as many hours as possible that contain $\mathbf{b_{i,l}} = 1$ and as less as possible that contain $\mathbf{b_{i,l}} = 0$. Shift modifications consist in sundering the shifts at overstaffing points (where $\mathbf{b_{i,l}}=1$). Examples of shift modifications are depicted in Figure 5.2 for a given resource type l. Then, for each employee we try to identify the shift modifications that reduce the cost while harming the LOS as less as possible (with the highest absolute value of $\mathbf{\Delta cost_l}/\mathbf{\Delta LOS_l}$ ratio). Shifts will either be of shorter length or entirely deleted (more details about the procedure of shift modification are given later).

Step 7 for all l, identify the shift $\mathbf{modification_l}$ that maximizes $\mathbf{score_l}$ (if there are many include them all)

Step 8 for all l, calculate the $\mathbf{\Delta cost_l}$ saved and simulate the $\mathbf{modification_l}$ to obtain the $\mathbf{\Delta LOS_l}$

Step 9 for all l, select the $\mathbf{modification_l}$ with the highest $\mathbf{\Delta cost_l}/\mathbf{\Delta LOS_l}$ ratio

Step 10 Repeat Steps 1, 7, 8, 9

Parameters:

Budget2= the budget obtained by the objective function of the LP

Real Budget= the initial budget used by the ED

$\mathbf{modification_l}$ = the best change in shift (based on step 7) that is selected for employee l

$\mathbf{score_l}$ = the number of employees with $\mathbf{b_{i,l}} = 1$ that are deleted subtracted by the number of employees with $\mathbf{b_{i,l}} = 0$ that are deleted

$\mathbf{\Delta cost_l}$ = the absolute value of the cost saved by $\mathbf{modification_l}$

$\mathbf{\Delta LOS_l}$ = the absolute increase in the value of the LOS shown by simulation after $\mathbf{modification_l}$

If the modification is performed in the interior of the shift, then two separate shifts will be generated and in this case they should not violate the constraint dealing with the maximum number of shifts scheduled per day. If the length of the new shift is less than the minimum shift length, then we consider that the whole original shift is deleted.

Table 5.2: Costs per employee

	Hourly Cost	
Type of Employee	Day	Night
Senior Doctor	1	1.3957
Junior Doctor	0.3245	0.5101
Nurse	0.429	0.4566
Stretcher Bearer	0.2842	0.3118

We select the ratio explained in step 9 because we believe that it represents the most efficient way of reducing the budget. As long as our performance goal is the patient LOS and the main restriction is the staffing budget, we use the ratio related to LOS and cost in order to find solutions that are more efficient. After the first iteration of the second part, steps 8, 9 must be performed again for all employee types in each iteration. The need for the above stems from the fact that resources in the ED are interdependent.

5.4 Application and results

The model is applied on Saint Camille ED. The unit uses two shifts for all employees every day. The first shift starts at 09.30 and finishes at 18.30 (duration of 9 hours) and the other shift covers the remaining part of the day (duration of 15 hours). We aim to determine repeatable daily shifts based on the fact that the arrival pattern seems to be similar from day to day during the weekdays (Figure 4.3). In Saint Camille ED, the minimum shift length (**kmin**) is 5 hours and the maximum shift length (**kmax**) is 24 hours. Furthermore, the maximum number of shifts ($\mathbf{shift_l}$) used in the ED will be 4 (no more than 4 types of shifts for each type of employee in a day). The initial budget of Saint Camille is calculated based on the costs of employees per day which depend on the employee type and the working time (different costs from day to night). The costs in Table 5.2 are standardized in a way that the unit corresponds to the hourly wage of a senior doctor during the day. On that basis, the weekly initial staffing budget also called "real budget" in the heuristic is equal to 110.647. Finally, we mention that we divide the day into 24 segments i of one hour length each.

We use the simulation model and more specifically simulation-optimization in order to obtain the staffing levels ($\mathbf{a_{i,l}}$). Then, we use them as parameters in the constraint as shown in Constraint 5.3 in Section 5.3.2. We solve the LP with the use of Cplex and we obtain the shifts and the values of ($\mathbf{b_{i,l}}$), which are the basis of the heuristic. The cost generated by the shifts in the LP is equal to 127.671 units. It is possible that the termination condition, which is that Budget 2 $\leq$ Real Budget, is satisfied in the first part of the heuristic. However, this is not the

case in our application, as the budget equals 120.856 after the completion of the first part. In the first part 11 variables $\mathbf{b_{i,l}}$ were greater than or equal to 2 and 4 iterations were required for them to become at most equal to 1. The number of iterations required is less than the number of variables because steps 2, 3 can provoke changes to more than 1 variable at a time.

In the second part of the heuristic, we have only $\mathbf{b_{i,l}} \leq 1$. Some of these variables that have non-zero values may be in the same shift. In Figure 5.2 we demonstrate an example of how the shift transformations can be visualized. The shift obtained from the first part of the heuristic for a given employee in this example has a length of 7 hours (from 8.00 to 15.00) and contains two $\mathbf{b_{i,l}}$ variables that are non-zero, the one at i=9 and one at i=12. The possible modifications of this shift are depicted with the cells containing the red color. Our objective is to delete as many cells as possible that contain a $\mathbf{b_{i,l}}$=1 and as less as possible that contain $\mathbf{b_{i,l}}$=0. We calculate this with the use of the score column, which is equal to the number of cells deleted that contained $\mathbf{b_{i,l}}$=1 subtracted by the number of cells deleted that contained $\mathbf{b_{i,l}}$=0. As mentioned above, the modification can either lead to a reduction of the length of the shift (modifications 1, 2 and 3) or the complete deletion of the shift (modification 4). The two constraints that should be taken into consideration are that the new shifts generated after the modification should have a minimum length of 5 hours and that the total number of shifts should not be greater than 4. This explains why modifications 2 and 3 are infeasible. Finally we select the feasible modification with the maximum score, which in this case is modification 1.

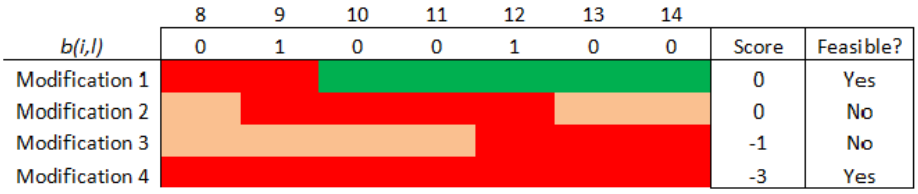

Figure 5.2: Example of shift modifications for a certain employee type l

Similarly, for each employee type l, we identify the modifications with the higher score. Only 4 out of the 8 employee types contain $\mathbf{b_{i,l}}$=1 and thus we only investigate them. In Table 5.3, we present the higher score modification for each of the 4 employee types in iteration 1. For each modification we calculate the cost saved and the LOS increase in the simulation model. We underline that ΔLOS and Δcost are expressed in absolute values. At the end of iteration 1, the modification of employee type 2 is selected because it has the ratio with the highest value. In the next iteration, we already know the modifications for the remaining employees, but we

Table 5.3: **1st** Iteration in the **2nd** part of the heuristic

Modification	ΔLOS	Δcost	Ratio
1	85	5	0.05882353
2	8.22	4.0808	**0.49644769**
3	19.25	2.1726	0.11286234
4	36.43	0.5684	0.01560253

Table 5.4: Budget and LOS of all Iterations in the **2nd** part of the heuristic

Iteration	LOS (minutes)	Budget (monetary units)
0	211.70	120.856
1	219.92	116.775
2	228.62	114.603
3	249.67	112.320
4	245.62	**110.048**

have to find the new best modification for the employee whose shift was changed. Furthermore, we must calculate again ΔLOS for all modifications, as the previous shift modification might have affected the values of other modifications as well. In fact in our case study, the ΔLOS was different for each shift modification after each iteration for all employees.

In Table 5.4 we present the Budget and the LOS in each iteration. Iteration 0 refers to the state that the heuristic is after the completion of the first part. At iteration 4 the termination condition is satisfied for the first time, as 110.048 is less than 110.647.

The resulting LOS is equal to 245.62 minutes. For comparison to the actual system we used the same budget restriction for the 2 predetermined shifts used in the ED of Saint Camille. As we have used 24 hourly slots, we started the shifts at 09.00 and 18.00 instead of 09.30 and 18.30 respectively. The optimal solution corresponding to the current shifts used in Saint Camille ED is an LOS of 269.65 minutes. Therefore the heuristic managed to reduce the LOS by 8.9% by creating more efficient shifts that respect the same staffing budget.

In Table 5.4, we can see that the LOS is reduced from iteration 3 to iteration 4. This result seems absurd at a first glance, but still an explanation exists. In iteration 4 the employee that had the highest ratio was Junior Doctor 3 (see Table 5.1) and thus the corresponding shift modification was performed. The reduction of the LOS stems from the fact that Junior Doctors usually require more time for the treatment of a patient compared to Senior Doctors, a parameter that has been taken into consideration in the simulation model (see Chapter 4). Finally, we mention that the above modification might provoke negative effects on other performance metrics, such as the door-to-doctor time of patients, because we only used LOS.

5.5 Conclusions and further extensions

This chapter is an ongoing work that contributes to the literature a new method that assists ED managers to determine efficient shift-scheduling in the ED, while satisfying staffing budget restrictions. We use staffing levels obtained from a simulation model as constraints in the linear program model that determines the schedule of shifts throughout a day in the ED. Because of constraints encountered in reality, the budget of the final solution in the linear program is higher than the real budget, which is used as a constraint in the simulation-optimization. In order to overcome this over-staffing problem, we propose a heuristic that involves the simulation model and the LP, and consider practical constraints encountered in EDs when defining shifts. However, we mention that it is a time-consuming procedure that involves several software programs, which will require to develop an automated combination of these. In addition, the final solution results from a heuristic, which means that it is not necessary the optimal solution. Nevertheless, it remains an efficient and cost-effective proposal for scheduling in the ED. The case study revealed that an improvement of 8.9% in the system performance could be made with fixed budget using the proposed method.

Chapter 6

Process-related experiments: Modeling and assessing the *Same Patient, Same Physician* rule

In this chapter, we investigate the relevancy of a new emergency department patient flow design, where the concerned modification takes place in the after-diagnostic stage. We address the question of whether a patient should be assigned or not to the same physician during all stages of the process. We carry out a field survey which shows that this issue is very controversial among practitioners, mainly because of human considerations. Since their additional complexity renders the problem hardly tractable analytically, we use discrete-event simulation to gain insights into both systems behaviors. We demonstrate that there is a threshold related to the system load separating between a region where the intervention is beneficial and another where it is detrimental. These results are further tested under realistic ED conditions using the comprehensive simulation model presented in Chapter 4.

6.1 Introduction

High salaries of doctors and high costs of medical equipments (Sinreich and Marmor, 2005; Warner, 2013) combined to budgetary restrictions (Carmen and Van Nieuwenhuyse, 2014; Abo-Hamad and Arisha, 2013) has prompted healthcare practitioners and researchers in operations management to investigate methods that improve ED operations without investing in human or physical capacity. This gave birth to a new stream of OR/OM literature that investigates alternative ED patient flow designs, in order to reduce congestion without increasing costs (Saghafian et al., 2012; Huang et al., 2012; Saghafian et al., 2014; Song et al., 2013). The present study falls into this category.

In the ED process, physicians are responsible of two major tasks. The first one is the initial consultation of "new patients", where the physician makes a first assessment of the patient state and may decide, if necessary, to request diagnosis tests. The second one is the interpretation of diagnosis tests for "internal patients", where the physician examines the results of the diagnosis tests and decides about the next steps which could be a release, an admission or a transfer. Typically in current ED practices, each patient is assigned to a single physician who will be exclusively responsible of conducting the initial consultation, and later the interpretation of test results (when performed). The aforementioned rule is referred to as the "Same Patient Same Physician ($SPSP$)" rule. The strategy that ignores the $SPSP$ rule is referred to as $\overline{SPSP}$.

In this chapter, we propose to compare between $SPSP$ and $\overline{SPSP}$. A conducted survey has led us to the conclusion that expert opinions widely diverge, making an appropriate quantitative comparison between $SPSP$ and $\overline{SPSP}$ very interesting. We are not aware of any work that deals with this research question, neither in medical nor in operations management/research literatures.

The intuition behind assessing $\overline{SPSP}$ is the well-known inefficiency of forcing customers/ patients to wait for their assigned server to become free, even if another server is idle (Song et al., 2013; Saghafian et al., 2012). Hence, the collaborative process ($\overline{SPSP}$) may benefit from the reduced waiting time derived from pooling physicians. However, not surprisingly, our field survey revealed that the duration of the interpretation step handed to the second physician is likely to increase, because the latter is not familiar with the patient situation and must "climb on the bandwagon". To synthesize this trade-off, ignoring $SPSP$ would improve the queueing performance (more pooling effect), but it would also induce a non-negligible duration for a physician to understand the health situation of a patient that has been first seen by another physician. From a modeling point of view, the $Erlang - R$ model introduced by Yom-Tov (2010); Yom-Tov and Mandelbaum (2014) (see Section 6.2) stands as the most relevant

framework. $SPSP$ and $\overline{SPSP}$ can be seen as queueing networks presenting similarities with $Erlang - R$, but with additional complexities that we highlight in Section 6.4 and Section 6.7. These additional and essential features render the problem intractable analytically. In contrast to other ED resources that could also be concerned by this question, we solely focus in this chapter on ED physicians because we found in Chapter 4 that they require special attention in the context of improving ED performance. Physicians are probably the scarcest resource (Brandeau et al., 2004) in EDs which represent the primary bottlenecks constricting patient flow (Jones and Evans, 2008; Tan et al., 2002).

The main contributions of this chapter can be summarized as follows. Since the collaborative strategy issue appears to be very controversial among practitioners, we first conduct a survey amongst experts from different EDs worldwide. The outcome of the survey provides information about their current practices, and the practical motivations and reasons for applying or not $\overline{SPSP}$. It also allowed us to capture the most important features in our model. We propose two queueing networks corresponding to $SPSP$ and $\overline{SPSP}$ and use simulation to compare their performance. We show that the effectiveness of $\overline{SPSP}$ depends on the system load, and that it performs better in lower system loads, which is counterintuitive for surveyed practitioners. We also propose an analytical approximation and highlight the complexity to capture some basic ED features mathematically. Through a case study conducted with a realistic simulation of a French hospital, we confirm the previous insights and demonstrate that the collaborative strategy would be beneficial for a wide range of overall system loads. This stands as a strong argument against the reluctance of ED managers towards the application of $\overline{SPSP}$.

The remainder of the chapter is organized as follows. Section 6.2 summarizes previous research relevant to our research question. In Section 6.3, we present the quantitative and qualitative results of the conducted survey. In Section 6.4, we introduce the two queueing networks $SPSP$ and $\overline{SPSP}$, which represent extensions of the $Erlang - R$ model. The latter are compared through simulation in Section 6.5 in order to gain insights into both systems behaviors. Section 6.6 tests the insights gained from Section 6.5 under ED-realistic conditions. As a perspective, we provide in Section 6.7 an analytical approximation for both models using continuous time Markov chains. We also highlight and explain the difficult tractability of ED analytical models when the level of details is raised (by including additional characteristics). We conclude in Section 6.8.

6.2 Literature review

In this section we highlight two categories of OR/OM papers that are relevant to this study. The first category is related to the routing of patient flow in general. The second category is related to the routing of returning patients to physicians in particular.

ED patient flow

In contrast to the vast majority of OR/OM papers addressing resource allocation in EDs, researchers are nowadays developing methods that aim at modifying some protocols and organizational rules regarding the patient path in the ED (Samaha et al., 2003; Medeiros et al., 2008). Most of these papers use analytical models. These contributions present the benefit of improving ED performance without any significant investment (Saghafian et al., 2012), which is very valuable in the current worldwide context of healthcare budgetary restriction.

The control of patient flow in EDs is addressed these last years by a number of papers. In the context of a highly congested ED, Huang et al. (2012) address the question of whether the physician should choose a new patient coming from triage (triage patient) or a patient that has already seen a doctor and returns to her after the completion of an examination (in-process, IP). They modeled the physician capacity as a queueing system with multi-class customers, where the objective is to minimize a waiting cost function for IP patients subject to deadline constraints for triage patients. The authors propose a threshold policy that chooses between the two types of patients and prove its asymptotic optimality. Similarly, Dobson et al. (2013) analyze the throughput optimal work flow decisions of an investigator, with server interruptions, that has to determine whether to prioritize seeing a new customer, or complete the work with a customer already in the system. They use a stylized queueing network in order to understand the impact of the investigator choices on system throughput. They derive recommendations on the optimal work flow decisions depending on the presence of interruptions or not. Another reference for the control of ED patient flow is Zayas-Caban et al. (2013). Using an MDP formulation, the authors investigate the optimal control policy of patients to a physician that handles both triage and treatment.

In addition to the control of patient flow, other modifications of the regular ED patient path were examined in the literature. Examples of these are Saghafian et al. (2014) that discuss a complexity-augmented triage system. This additional complexity evaluation at triage would only take a matter of seconds but can improve patient safety and increase operational efficiency. Through simulation analysis calibrated with hospital data and an MDP model, they demonstrate that ED performance can substantially benefit from complexity-augmented triage. Saghafian

et al. (2012) introduce a supplementary triage element in EDs: a prediction of whether a patient will be admitted in the hospital or not after the ED; and propose patient streaming as a mechanism for improving responsiveness. They use a combination of analytic (MDP) and simulation models, and compare between three policies; pooling, streaming and virtual streaming. The authors conclude, under the considered modeling, that although pooling is more efficient than streaming, virtual streaming is the best method. For other related studies that analyze patient flow in EDs, we refer the reader to Paul et al. (2010); Wang (2004) and references therein.

Systems with returning patients and the *SPSP* question

Yom-Tov (2010) and Yom-Tov and Mandelbaum (2014) address queueing networks with reentering customers and introduce the $Erlang - R$ model ("R" for reentrant customers or repetitive service). They widely address the time varying environments, but we will only focus on the part with constant arrival rates, which fits more with our analysis. The $Erlang - R$ model corresponds to systems where customers return for further service with probability p after a certain delay, or exit the system with probability *1-p* right after service completion (Yom-Tov, 2010; Yom-Tov and Mandelbaum, 2014). The queueing policy is FCFS. They refer to the service phase as a *needy state* and to the delay phase as a *content state*, while they are respectively called *processing step* and *external delay* in Campello et al. (2013). In order to examine the use of "speedup", Chan et al. (2014) use an $Erlang - R$ model where they consider state-dependent service times and state-dependent return probabilities. Yom-Tov (2010); Yom-Tov and Mandelbaum (2014) demonstrate analytically that, in steady state, quality measures of $Erlang - R$ (such as the probability of waiting) depend exclusively on the offered load of the Needy station. Carmen and Van Nieuwenhuyse (2014) point out the fact that the $Erlang - R$ model assumes that patients can be treated by any of the physicians; while in practice they usually receive treatment from the same physician, which represents a central element in our framework.

A few papers mention the *SPSP* rule to indicate that this feature has been included in their model (Saunders et al., 1989; Saghafian et al., 2012; Ghanes et al., 2014b). Saghafian et al. (2012) use the term "non-collaborative" to describe a service process applying *SPSP*. Ghanes et al. (2014b, 2015c) report that the "same patient-same staff" rule is a strong constraint with a significant impact on the system behavior, and yet commonly neglected in ED models. The *SPSP* rule has an implication in the model design used in some studies like Saghafian et al. (2014) and Green (2006). Since each physician is dedicated to her own slate of patients, Saghafian et al. (2014) choose in their MDP model to focus on a single physician decision of who to see next. This choice of isolating a server is clearly not adapted to our research question. Carmen and

Van Nieuwenhuyse (2014) suggest as an avenue for future research simulation as a tool to assess how this decrease in flexibility would affect the performance.

The closest contribution to our research question is the paper by Campello et al. (2013). The authors define a "case manager" as a server who is assigned multiple customers and repeatedly interacts with those customers, that they name "case". Then, they define and analyze three systems; the system (S) which is similar to our *SPSP* (assignment of patients to physicians) with a smallest-caseload routing policy, the system (R) which is the same but with a random routing policy, and the system (P) which is similar to our $\overline{SPSP}$ system (pooled physicians). They numerically show that random routing (R) and pooled (P) systems provide lower and upper bounds on the (S) system in terms of the overall delay, with (S) being consistently closer to (P). They further analyze numerically the stability limits of these different systems and show that (P) has the largest stability region. The present study differs from Campello et al. (2013) in significant aspects. They use in each model a pre-assignment queue as well as a maximum caseload for case managers (M). The numerical experiments depend on the parameter M, and the waiting time (WT) they consider is divided into a pre-assignment WT and an internal WT. In contrast, we use unbounded queue lengths as done in Yom-Tov and Mandelbaum (2014) and Chan et al. (2014). Finally, the authors consider the same service time distribution for the initial vs. subsequent interactions. The latter assumption renders the superiority of the pooled system quite predictable, in addition to not be in line with reality.

6.3 Survey results

As a first step, we aim to better understand, through a survey, the current practices and motivations related to the *SPSP* rule, and the opinions concerning the eventual removal of this obligation, in different EDs worldwide. The survey is based on 33 practitioners (ED managers and physicians) from 23 different EDs in 7 different countries: France, USA, the Netherlands, Germany, Belgium, Greece and Tunisia (see the sample composition in Appendix A.1.2).

Which method are you applying in your ED?

52% of surveyed EDs exclusively use *SPSP*. Only 9% of surveyed EDs use a collaborative strategy ($\overline{SPSP}$). The rest of the EDs (39%) say that *SPSP* stands as the reference except in some specific situations that are: when "physician in triage" is applied (Oredsson et al., 2011), in case of change in patient status, when the architecture of the ED requires a separation between the initial consultation and the rest of the ED process, when the assigned physician remains busy or absent for several hours (e.g. particular organizations where physicians can

leave the ED for ambulance interventions during their shift) and in case of unexpected peaks in a particular sector of the ED. In addition to these particular cases, there are common situations found in most EDs, where patients are transferred from one physician to another. Examples of these are the handovers from ambulance team to ED staff, transmissions of studies pending between shifts, transfers from juniors to seniors and transfers from ED physicians to specialists. The aforementioned situations falls out of the scope of our study. Instead, the addressed debate is whether the collaborative strategy could be adopted as a general rule.

Can $\overline{SPSP}$ be adopted as a general rule in EDs?

The answers are positive for 18% of surveyed clinicians, positive with condition for 15%, and negative for 67%. The main collected arguments for rejecting $\overline{SPSP}$ are the following:

- Results interpretation would take more time (see the last title of this section);
- Transmitted cases correspond to an increased error rate:
  - There is a risk of loss of information, which could lead to misdiagnosis or inappropriate decisions (e.g. patient disposition instead of admission);
  - For example, it is well know in practice that transfer of shifts represents a major source of errors, waste of time and recrimination from patients;
  - The quality of care is better when a patient is treated by the same physician.
- $\overline{SPSP}$ raises a deontological problem;
- The request of ancillary tests relies on a diagnostic assumption which is based on the first assessment and is strongly linked to the physician who made it. However, heterogeneities exist between physicians in terms of experience, education and skills. Not involving the initial physician represents a rupture in the process of establishing a medical diagnosis;
- Patients and their families do not appreciate to have different interlocutors;
- It is frustrating for a physician not to follow her patient case until the end of the process;
- It requires a high level of confidence between physicians. The first handling could be not satisfying for the second physician;
- Psychiatric patients do not want to have contact with many different persons.

15% of the clinicians approve the potential benefit of $\overline{SPSP}$ under the following conditions:

- The application must be on simple straightforward cases (stable patients requiring ancillary tests);
- The initial physician must provide a proper handover, medical records must be reliable and filled properly;
- The necessity of homogeneous physician profiles (same team, experience and education);
- $\overline{SPSP}$ could be beneficial in the periods of high demand compared to ED capacity.

Service time extension

73% of the sample state that the "interpretation and decision" handled by another physician would be longer than if it had been conducted by the same. The collected arguments justifying this service time extension are listed below:

- It is already known in practice that cases transmitted between shifts generally take more time than others;
- A diagnosis cannot be exclusively based on the results of examination tests;
- The physician is not familiar with the case, and may need some time to understand the situation of the patient. She would certainly need to ask again some essential questions (anamnestic data) or make a clinical exam in order to be sure of the decision that will be made;
- Patient files are rarely enough exhaustive to avoid asking the patient. Some patient files are not or only partially filled which may force the second physician to repeat the initial consultation.

To synthesize, "interpretation and decision" would be longer if made by a different physician. This extension correspond to the required time to read and understand the patient case (while the first physician already have it in mind), and eventually to ask some questions and make a clinical examination. The amount of this time extension depends mainly on the quality of the handover. The survey reveals that this time extension would represent a percentage of the initial consultation duration. We also asked the experts to provide an approximation of this percentage. The distribution ranges from 10% to 100% of the initial consultation, with a mean around 30%. However, experts insist on the fact that this prolongation highly depends on the quality of the handover.

We conclude from this section that experts opinions about the relevancy and the applicability of $\overline{SPSP}$ diverge widely. Our research question embodies quantitative aspects (service time extension) as well as human considerations (from the physician and also the patient perspectives). From a quantitative point of view, the collaborative strategy ($\overline{SPSP}$) could benefit from the advantage of pooling, but at the same time would suffer from service time extension. From this appears the necessity of a risk/benefit analysis that we perform in Section 6.5 and Section 6.6.

6.4 Modeling

We define here the non-collaborative ($SPSP$) and the collaborative ($\overline{SPSP}$) models. In both models, we consider a set of s identical physicians. Patients arrive according to a Poisson process with a constant average arrival rate λ. Patients have to wait for the initial consultation, after which they may undergo diagnostic tests with a probability p, or leave the system with a probability $1 - p$. We refer to patients heading to an initial consultation as "new patients". We refer to those who underwent tests and seeking for interpretation as "returning patients". We consider diagnostic tests as an exponentially distributed delay with service rate δ. In both models, we also assume no priority between new and returning patients. This choice is generally not formalized in practice and depend on each physician preference. That was addressed in the literature (Huang et al., 2012), but falls out of the scope of this study. We assume all the service durations to be exponentially distributed. In practice, the duration of the initial consultation is slightly longer than the interpretation. Similarly to Yom-Tov (2010); Yom-Tov and Mandelbaum (2014); Campello et al. (2013), we assume them to have the same service rate μ when performed by the same physician.

6.4.1 The non-collaborative model (*SPSP*)

Under $SPSP$ (see Figure 6.1), each of the s physicians has her own slate of assigned patients. There are $s+1$ different queues in the system: s queues each corresponding to a single physician and containing her own returning patients, and one common queue where new patients wait upon their arrival. When a physician becomes free, she chooses to serve a patient either from the new patients queue or her returning patients queue according to a FCFS discipline.

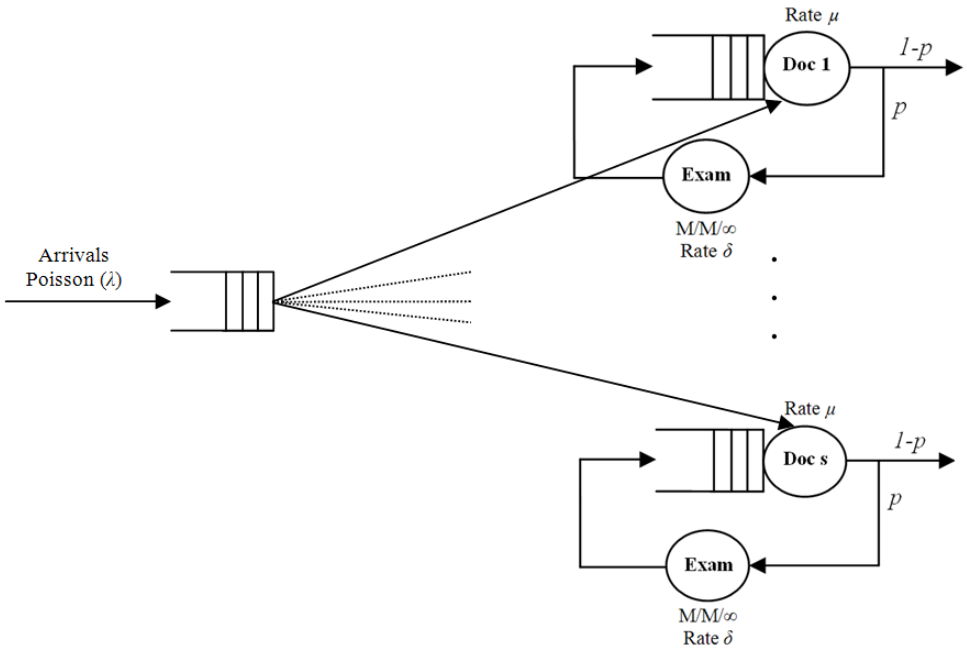

Figure 6.1: The *SPSP* system process

6.4.2 The collaborative model ($\overline{SPSP}$)

Under $\overline{SPSP}$ (Figure 6.2), patients are not assigned to any particular physician. New and returning patients wait in a common FCFS queue. Returning patients may be treated by any of the s servers. As a consequence, they may be handled either by the same physician who performed the initial consultation, or by a different one. In these two cases, the mean service rates for returning patients are respectively μ and μ', where $\frac{1}{\mu'} = (1+\alpha)\ \frac{1}{\mu}$ ($\mu > \mu'$, see Section 6.3). In sum, the service rate is μ for new patients and returning patients to the same physician. The service rate is μ' for patients returning to a physician different from the initial one.

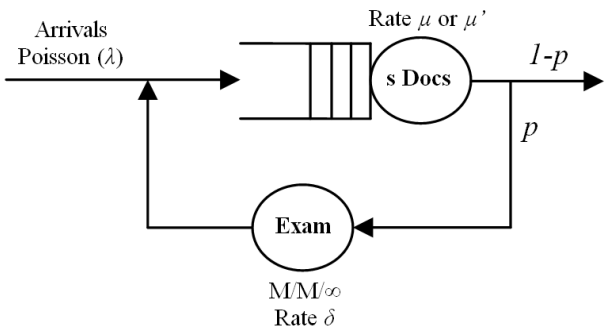

Figure 6.2: The $\overline{SPSP}$ system process

6.5 Performance comparison between *SPSP* and $\overline{SPSP}$

We focus on the performance in terms of the steady-state average waiting time in the queue (WT). The two models are hardly tractable analytically. From the one hand, $\overline{SPSP}$ has no product-form solution (because of the difference between new and returning patients service times), in contrast to the model used in Yom-Tov (2010); Yom-Tov and Mandelbaum (2014). From the other hand, *SPSP* has a general but not Poisson arrival process. Hence, we resort to simulation. We simulate *SPSP* and $\overline{SPSP}$ systems using the Arena simulation software. We consider a steady-state type simulation run with one pseudo-infinite length of time. These simulations required between one and five minutes per instance depending on the system load.

In Figure 6.3, we use as a reference a specific set of parameters ($s = 2, 1/\lambda = 10, 1/\mu = 5, 1/\delta = 60$) to compare *SPSP* and different scenarios of $\overline{SPSP}$ that differ in terms of service time extension ($\alpha = 0\%, 20\%, 40\%, 60\%, 80\%, 100\%$). In order to assess the impact of the system load on this comparison, we vary, for each scenario, the test probability p up to the system stability limit. This particular figure is important for the rest of the chapter since it will be used as a reference for comparison. Minutes will be used as a unit throughout the whole chapter.

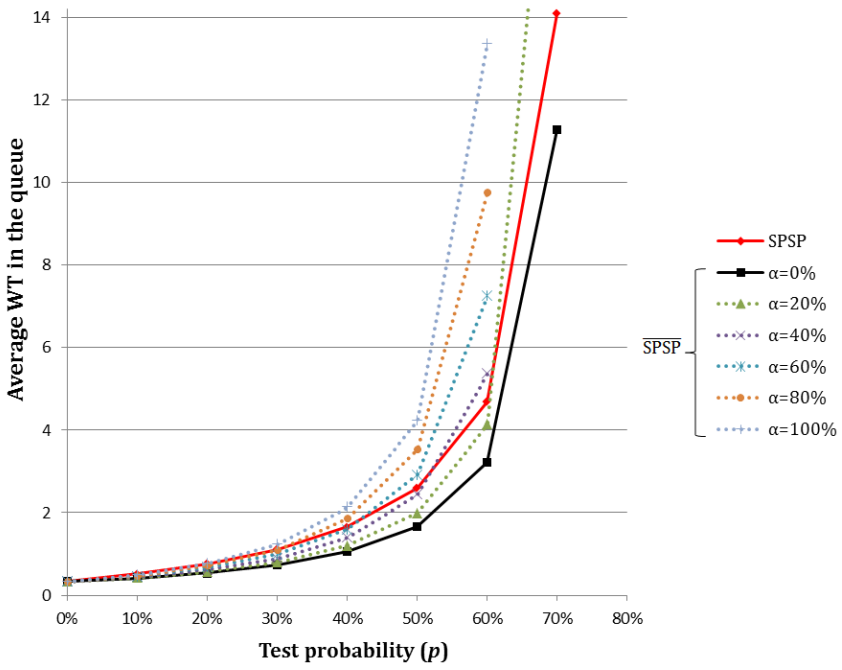

Figure 6.3: Performance comparison as a function of p and α for the reference set of parameters ($s = 2, 1/\lambda = 10, 1/\mu = 5, 1/\delta = 60$)

The figure shows that when we ignore service time extension ($\alpha = 0\%$), $\overline{SPSP}$ is obviously always better than *SPSP*. However, we observe that for any given $\alpha > 0\%$, $\overline{SPSP}$ is better

than *SPSP* for lower system loads, and from a certain threshold of p, it becomes worse. The extension α also has an influence on when this switch occurs. We observe that the higher is α in the $\overline{SPSP}$ scenario, the faster the threshold will occur. For instance, when $\alpha = 100\%$, $\overline{SPSP}$ is better than *SPSP* up to $p = 20\%$; while when $\alpha = 20\%$, this switch occurs at $p = 61\%$.

In order to confirm the previous observations and test the influence of μ, we perform the same comparisons using two extreme cases, heavily loaded (the same parameters set but with $1/\mu = 10$ instead of 5) and lightly loaded (the same parameters set with $1/\mu = 1$ instead of 5). In the heavily loaded system (Figure 6.4), $\overline{SPSP}$ is just slightly better than *SPSP* when $\alpha = 0\%$, and when $\alpha > 0\%$, $\overline{SPSP}$ becomes always detrimental, except for lowest values of α and p (least loaded scenarios). On the other hand, Figure 6.5 shows that in the system with light load, $\overline{SPSP}$ is significantly better than *SPSP* for $\alpha = 0\%$, and remains almost always better when $\alpha > 0\%$, except for highest values of α and p (most loaded scenarios).

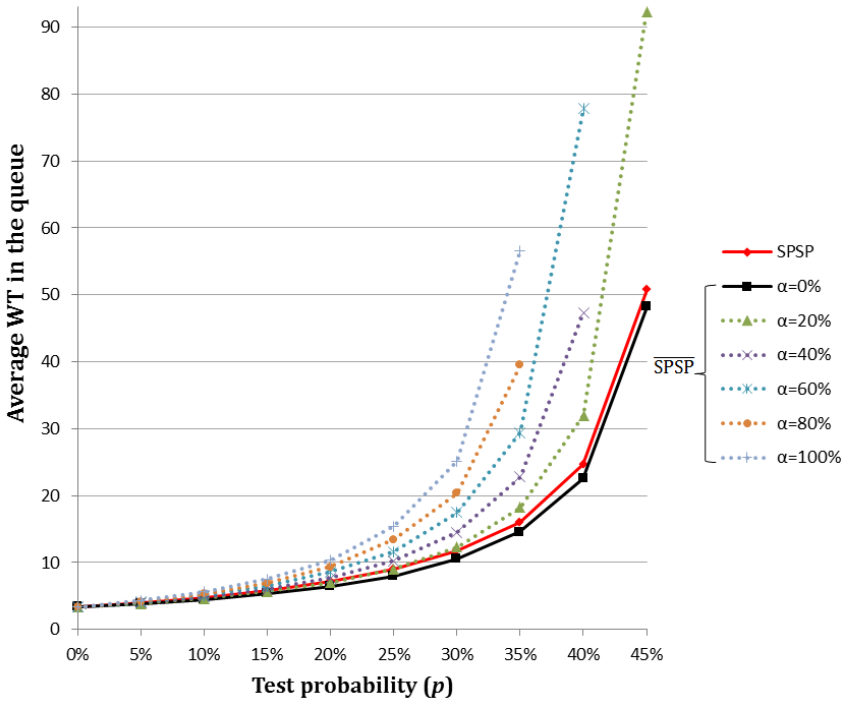

Figure 6.4: Performance comparison as a function of p and α for the highly loaded set of parameters ($s = 2, 1/\lambda = 10, 1/\mu = 10, 1/\delta = 60$)

Up to now, the experiments reveal the impact of the parameters α, p and μ on whether $\overline{SPSP}$ is beneficial or detrimental compared to *SPSP*. As a remaining driver of the system load, we also assess the effect of λ. Figure 6.6 illustrates experiments where we vary λ from 0.058 to 0.18 (which amounts to varying $1/\lambda$ from 5.5 to 17) in a specific scenario derived from Figure 6.3

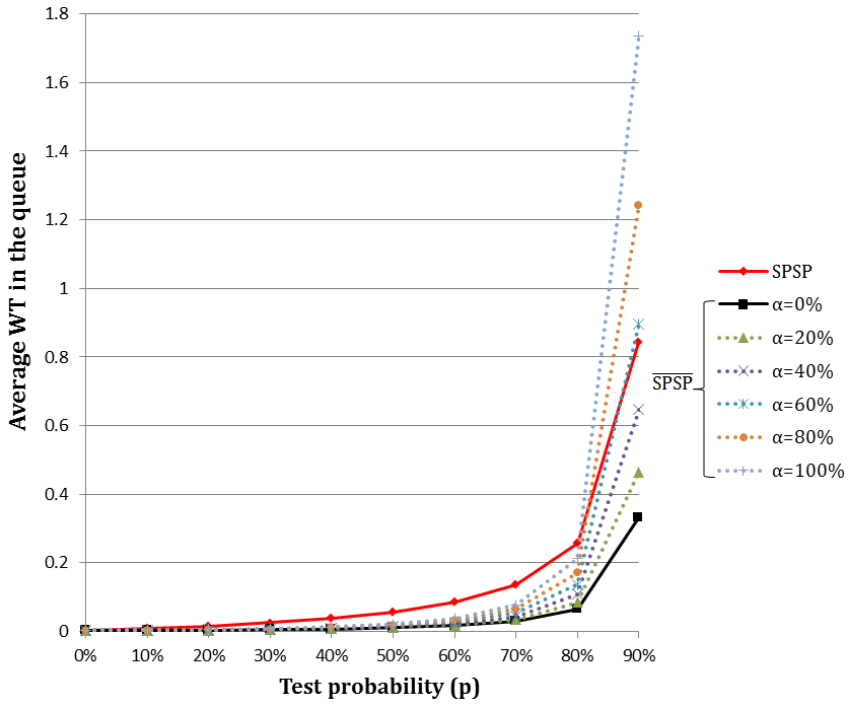

Figure 6.5: Performance comparison as a function of p and α for the lightly loaded set of parameters ($s = 2, 1/\lambda = 10, 1/\mu = 1, 1/\delta = 60$)

($p = 20\%$). The observations lead to the same previous conclusions, namely a certain threshold separating two regions. From the one hand a lower system load where $\overline{SPSP}$ is beneficial, and higher system load where it is detrimental. Note that these results contradict some physicians expectations which suggested in the survey (see Section 6.3) that the collaborative strategy would be relevant during peak hours.

The interpretation of the previous insights is that $\overline{SPSP}$ avoid situations where patients wait for their assigned physician while another physician is idle. This idling situation where pooling provides a benefit is more likely to occur in low system loads. Consequently, when the system load increases, the benefit that could be gained from $\overline{SPSP}$ pooling decreases. At the same time, the proportion of returning patients increases and the performance deteriorates because of service time extension. In sum, before reaching the threshold, $\overline{SPSP}$ dominates $SPSP$ thanks to pooling effect. Beyond the threshold, $SPSP$ prevails because $\overline{SPSP}$ pooling effect is no longer an advantage (no or rare idleness situations) from the one hand, and because of service time extension from the other hand.

We conclude from this section that the relevancy of the collaborative strategy ($\overline{SPSP}$) depends on the system load. There is a threshold, related to the system load, under which the

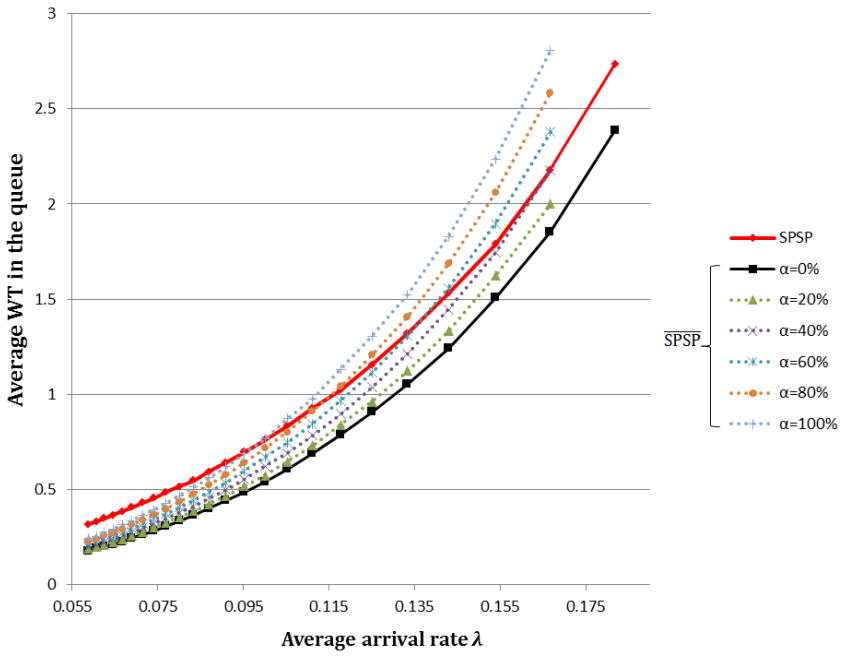

Figure 6.6: Performance comparison as a function of λ and α for the set of parameters ($s = 2, 1/\mu = 5, p = 20\%, 1/\delta = 60$)

collaborative strategy ($\overline{SPSP}$) outclasses *SPSP*, and above which its application becomes detrimental. We have demonstrated that this threshold is a function of the system load characteristics (α, p, μ and λ). Moreover, we also used simulation to confirm that using different average service times for the initial consultation and the tests interpretation does not affect the obtained results.

6.6 Realistic conditions

We consider here a case study involving the French ED Saint Camille. We check the insights gained from Section 6.5 under realistic conditions by means of the comprehensive discrete-event simulation model of the full ED developed in Chapter 4. In this model, all common structural and functional characteristics of EDs, at least in France, are taken into consideration thanks to a close collaboration with practitioners.

As already mentioned in Chapter 4, patients are sorted into 5 acuity levels called Emergency Severity index: ESI 1, 2, 3, 4 and 5 (ESI 1 being the most critical patients and ESI 5 the least). The corresponding patient mix in Saint Camille is respectively 0.22%, 12.34%, 34.27%, 40.91% and 12.25%. The probability to undergo examination tests is respectively 100%, 94%, 77%, 60% and 9% (with different mix of radiological and biological tests). In our experiments, we assess the effect of applying $\overline{SPSP}$ on the average length of stay (LOS) of patients from ESI 3, ESI

4 and ESI 5, which represent the majority. We do not apply $\overline{SPSP}$ on ESI 1 patients because their care process is generally continuous, and ESI 2 because their acuteness requires to avoid the risks related to $\overline{SPSP}$ (see Section 6.3).

6.6.1 Assessing the effect of applying $\overline{SPSP}$ at Saint Camille ED

We apply $\overline{SPSP}$ only on the following task: interpretation of diagnosis tests performed by senior doctors. We maintain the *SPSP* rule for junior doctors whose lack of experience does not allow the flexibility required in $\overline{SPSP}$. We also do not apply $\overline{SPSP}$ when an expert's opinion is required, because when the latter arrives at the ED, she must discuss the patient case with the same physician who has made the initial consultation and ordered the tests. $\overline{SPSP}$ is also not applied for other physician tasks such as the organization of the patient transfer. We assess the effect of these changes on the average LOS, for different percentage extensions of the interpretation duration (α). Table 6.1 summarizes the results. The corresponding average LOS values are summarized in the table of Appendix A.5.

Table 6.1: The percentage of the evolution of the average LOS when applying $\overline{SPSP}$ on ESI 3, 4 and 5

	$\overline{SPSP}$ applied on ESI 3, 4 and 5						
	α=0%	α=20%	α=**30**%	α=40%	α=60%	α=80%	α=100%
ESI 1	-9.33%	-5.91%	**-3.96%**	-3.08%	-2.47%	+0.50%	+2.91%
ESI 2	-5.30%	-3.75%	**-2.59%**	-2.09%	-0.66%	-0.29%	+1.44%
ESI 3	-8.24%	-6.73%	**-5.36%**	-5.19%	-3.75%	-2.63%	-0.49%
ESI 4	-9.67%	-9.13%	**-8.41%**	-8.20%	-7.12%	-6.01%	-5.48%
ESI 5	-10.48%	-8.71%	**-8.23%**	-8.04%	-7.78%	-6.97%	-6.29%
Overall	-8.24%	-7.31%	**-6.14%**	-5.99%	-4.72%	-3.80%	-2.48%

For all the extensions of the interpretation duration α, we observe from Table 6.1 a decrease in the average LOS of ESI 3, 4 and 5 and also the overall system, compared to the current average LOS in Saint Camille. As expected, The amount of this reduction decreases when the service time extension increases. Let us consider the average estimation of the expected service time extension, provided by experts in the survey of Section 6.3, which is α=30%. With the latter, the application of $\overline{SPSP}$ in Saint Camille ED would allow a reduction of 5.36%, 8.41%, 8.23%, for ESI 3, ESI 4 and ESI 5 respectively, which corresponds to an overall system reduction of 6.14% of the average LOS. Note that ESI 1 and 2 are also impacted, even though, as explained before, these particular patient categories are excluded from our experiments. This counterintuitive effect on ESI 1 and 2 patients (which are not concerned by the application of $\overline{SPSP}$) is due to the fact that in Saint Camille, there are common resources that are shared between ESI 1, 2

and 3 (see Appendix A.6). Consequently, it appears that ESI 1 and ESI 2 would benefit from the time saved on ESI 3 by their common physicians. However, we observe from the table that extremely high time extensions (α=80% and 100%) could be slightly detrimental for ESI 1 and 2 while being slightly beneficial for ESI 3. From the latter, we deduce the necessity of a separated assignment of physicians between patients concerned by $\overline{SPSP}$ and those who are not, in the case of high service time extensions. Otherwise, if the observed service time extensions are high, and it is not possible to assign exclusive resources to the targeted patients, it would be better not to apply $\overline{SPSP}$.

6.6.2 Analyzing the impact of the system load

We assess the impact of the ED system load on the effectiveness of the collaborative strategy ($\overline{SPSP}$). Since the objective is to measure the impact of its application in Saint Camille, we do not include in our analysis the variation of the intrinsic features of the current system such as μ and p. Instead, we increase the patient arrivals. In Saint Camille, arrivals are time-dependent (vary with the hour of the day and the day of the week). This arrival pattern is quite typical for most EDs in the world. We assume arrivals to follow a non-homogenous Poisson process with a different average arrival rate for each hour of the week (see Figure 4.3). Note that resources are staffed in Saint Camille so as to match with the variation of patients arrivals. In order to vary the overall ED system load, we use the same arrival pattern which we increase by a percentage β (we multiply all the arrival rates by (1+β)). Arrivals are increased until the overall system stability limit. Figure 6.7, 6.8 and 6.9 summarize the results of the experiments for ESI 3, ESI 4 and ESI 5, respectively.

We observe from Figure 6.7 that the application of $\overline{SPSP}$ remains beneficial for ESI 3 in most scenarios, except for highest service time extensions α under the heaviest system loads (highest β), which is consistent with the results of Section 6.5. In contrast, for ESI 4 and ESI 5, the application of $\overline{SPSP}$ is always beneficial regardless of the overall system load. This robustness of ESI 4 and ESI 5 with respect to the system load could be explained by the low return probability of these patient types (probability to undergo examination tests) compared to more severe patients (ESI 1, 2 and 3). The threshold highlighted in Section 6.5 does not occur here because the system load considered in our experiments concerns the overall system, which means that patient categories with lower return probabilities (ESI 4 and ESI 5) could be less affected by the increase of the overall arrivals. Besides, the conducted experiments exclude a certain amount of patients, even among ESI 3, 4 and 5, like when result interpretation is conducted with juniors or jointly with a specialist. This reduces the occurrence of service

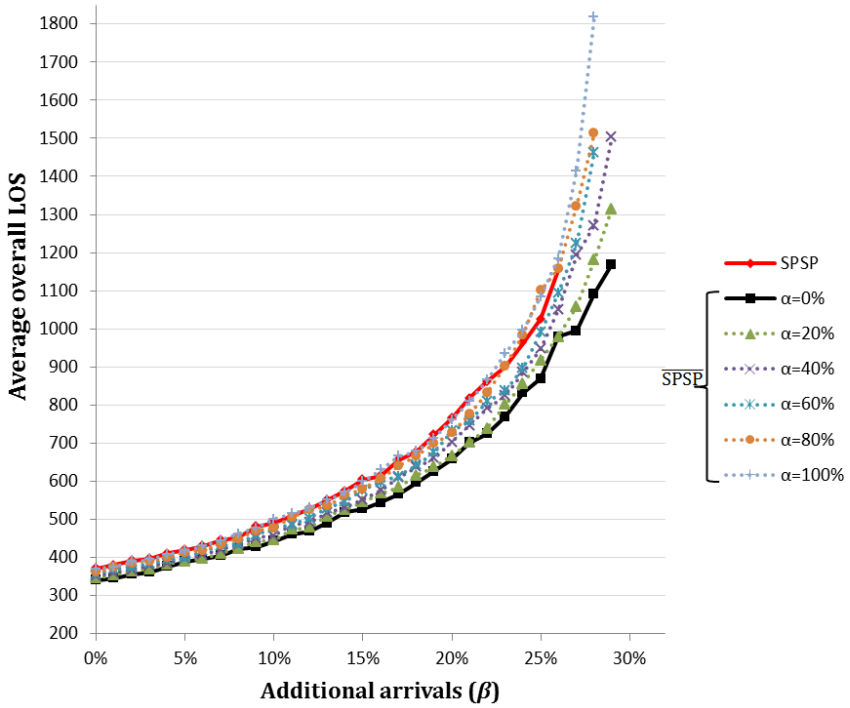

Figure 6.7: Impact of the system load on the effectiveness of $\overline{SPSP}$ for ESI 3 patients

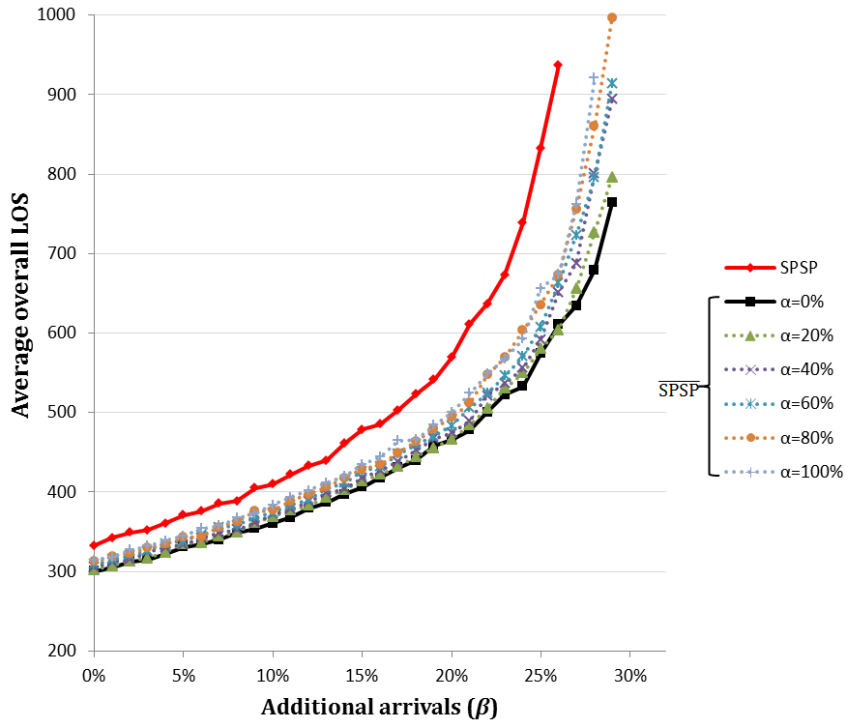

Figure 6.8: Impact of the system load on the effectiveness of $\overline{SPSP}$ for ESI 4 patients

time extensions and mitigate the negative effect of the application of the collaborative strategy. Finally, note that $\overline{SPSP}$ remains detrimental for ESI 1 and 2 for high service time extensions

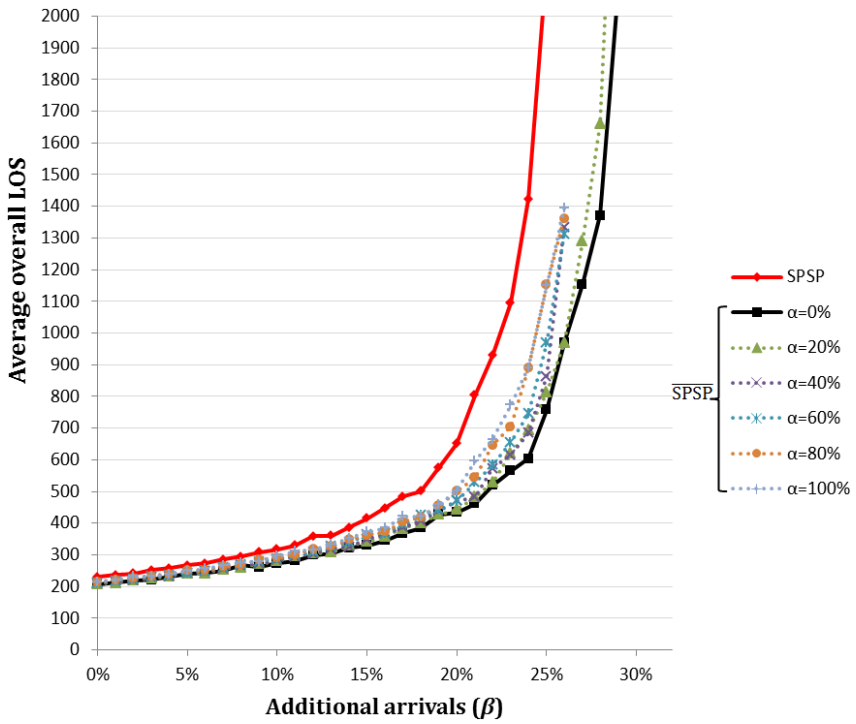

Figure 6.9: Impact of the system load on the effectiveness of $\overline{SPSP}$ for ESI 5 patients

(see Appendix A.7).

6.7 Going further: Analytical approximation of *SPSP* and $\overline{SPSP}$

An exact analytical comparison between *SPSP* and $\overline{SPSP}$ is too complex. The two architectures are queueing networks with complex routing mechanisms. It is however interesting to propose numerical algorithms that may substitute simulation. In what follows, we develop approximations for the performance analysis of *SPSP* and $\overline{SPSP}$.

6.7.1 Analytical approximation for $\overline{SPSP}$

New and returning patients wait in a common FCFS queue. Returning patients may be treated by any of the s servers. Hence, they may be handled either by the same physician who performed the initial consultation, or by a different one. We model the system as a Markovian birth-death process where a state i represents the total number of patients in the system (queue + service). The process is depicted in Figure 6.10. We first introduce a new parameter r that represents the probability that an occupied physician is working with a service rate μ. In other words, it is the probability that when a physician is treating a patient, the latter is either a new patient or a returning patient from her own slate of patients. In turn, $1 - r$ is the probability that a busy physician is operating with a service rate μ' (treating a returning patient from another

physician slate). The probability r is approximated as follows:

$r = \frac{\lambda}{\widehat{\lambda}} + \frac{(\frac{1}{s})(\lambda_{return})}{\widehat{\lambda}}$,

where $\frac{1}{s}$ represents the approximation of the probability that a returning patient will be handled by her first assigned physician (see Appendix A.4).

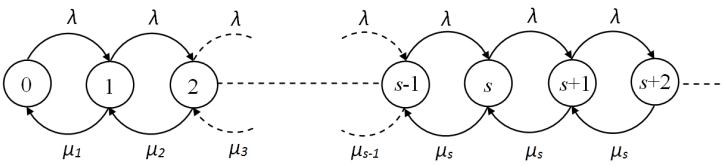

Figure 6.10: Markov chain associated to $\overline{SPSP}$ system

The quantity $\widehat{\lambda}$ is the "amplified arrival rate". It represents the total arrival rate of patients that includes the new and the returning flow. We have

$\widehat{\lambda} = \lambda + \underbrace{p\lambda + p(p\lambda) + p(p^2\lambda) + \cdots}_{\lambda_{return} = \lambda \frac{p}{(1-p)}}$

$\widehat{\lambda} = \lambda(1 + p + p^2 + p^3 + \cdots + p^\infty)$

$\widehat{\lambda} = \frac{\lambda}{(1-p)}$.

Therefore, $r = (1-p) + \frac{p}{s}$.

Let us focus on the formulation of the system load (ρ) for $\overline{SPSP}$. Because of the returning flow of patients, the arrival rate is amplified, $\widehat{\lambda} = \frac{\lambda}{(1-p)}$. In turn, the global service rate for $\overline{SPSP}$ is approximated as follows: $\mu^* = \frac{s}{r\frac{1}{\mu} + (1-r)\frac{1}{\mu'}}$. Hence, the system load is given as an expression of the key system parameters (α, p, μ and λ) as follows: $\rho = \frac{\frac{\lambda}{(1-p)}}{\frac{s}{r\frac{1}{\mu} + (1-r)(1+\alpha)\frac{1}{\mu}}}$.

The birth rate in any state i is the rate λ of new arrivals, for $i \geq 0$. The death rates are given below:

$$\mu_i = \begin{cases} i(1-p)(r\mu + (1-r)\mu'), \text{ for } i = 1 : (s-1) \\ s(1-p)(r\mu + (1-r)\mu'), \text{ for } i = s : n \end{cases}$$

Let us denote the stationary probabilities of the system states by π_i. Thus, $\pi_i = \frac{\lambda}{\mu_i}\pi_{i-1}$, for $i \geq 1$.

We may then write $\pi_i = \frac{\lambda^i}{\prod_{j=1}^{i} \mu_i}\ \pi_0$. Since all probabilities sum up to one, we obtain

$$\pi_0 = \left(1 + \sum_{i=1}^{\infty} \left(\frac{\lambda^i}{\prod_{j=1}^{i} \mu_i} \right) \right)^{-1}.$$

Having in hand the stationary probabilities, we deduce the average number of patients in the queue, $E(N) = \sum_{i=s+1}^{\infty} (i-s)\pi_i$. Using Little's law, the average waiting time is given by $E(W) = \frac{E(N)}{\widehat{\lambda}}$. Then, $E(W) = \frac{E(N)(1-p)}{\lambda}$.

6.7.2 Analytical approximation for *SPSP*

We model *SPSP* under a different routing policy. Each physician has her own queue containing new and returning patients. A patient upon her arrival is routed to the shortest queue and assigned to the corresponding physician. Within each physician queue, the discipline of service is FCFS. One can see that our model can be divided into s identical sub-systems. It suffices then to focus on the performance analysis of one of these sub-systems, i.e., a single physician and the associated queue. For each sub-system, let us define the birth-death process, as shown in Figure 6.11, where a state i represents the total number of patients in the sub-system (queue + service).

Let us denote the stationary probabilities of each sub-system states by π_i, for $i \geq 0$. We assume that the states of the sub-systems are independent, which is not true. For tractability, we also assume arrivals to each sub-system to follow a Poisson process. Because the routing rule consists in choosing the shortest queue, the arrival rate of patients to each sub-system depends on both the state of this sub-system, and those of all the other sub-systems. The arrival rate λ_i denotes the state-dependent arrival rate of patients to a sub-system at state $i-1$, for $i \geq 1$. For clarity of the exposition, let us consider the simplest case, $s = 2$. Three possibilities may happen:

$$\lambda_i = \begin{cases} 0, \text{if the considered sub-system has the longest queue} \\ \frac{\lambda}{2}, \text{if the two physicians are idle, or they are both busy and they have equal queue lengths} \\ \lambda, \text{if the considered sub-system has the shortest queue} \end{cases}$$

Let us denote the stationary probabilities of the system states by π_i, for $i \geq 0$. We may then write the state-dependent arrival rate as a function of the stationary probabilities of this sub-system: $\lambda_i = \lambda(\frac{\pi_{i-1}}{2} + \sum_{j=i}^{\infty} \pi_j)$, for $i \geq 1$.

From the Markov chain, we may write $\pi_i = \frac{\lambda_i}{(1-p)\mu} \frac{\lambda_{i-1}}{(1-p)\mu} \cdots \frac{\lambda_2}{(1-p)\mu} \frac{\lambda_1}{(1-p)\mu} \pi_0 = \frac{\prod_{j=1}^{i} \lambda_i}{\mu^i (1-p)^i} \pi_0$, for $i \geq 1$. Since all probabilities sum up to one, we obtain

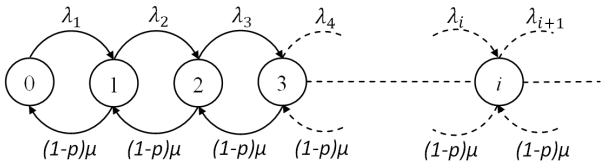

Figure 6.11: Markov chain associated to a sub-system of *SPSP*

$$\pi_0 = \left(1 + \sum_{i=1}^{\infty} \left(\frac{\prod_{j=1}^{i} \lambda_i}{\mu^i (1-p)^i} \right) \right)^{-1} .$$

Note that from the one hand, the state-dependent arrival rates λ_i are given as a function of the stationary probabilities π_i. From the other hand, π_i are given as a function of λ_i. As a consequence, we have a fixed point. We propose the following fixed point algorithm to compute it. In the first iteration, we choose an arbitrary value for λ_1. Then we compute π_i ($i \geq 0$). From these π_i, we next compute the new values of λ_i ($i \geq 1$). In the second iteration, we use the latter values of λ_i to compute π_i. From these new π_i, we compute the new values of λ_i. We do the same in the third iteration, and so on. We stop the algorithm when the values of π_i and λ_i converge to their limits with a given predefined precision.

Having in hand the stationary probabilities, we deduce the average number of patients in the queue: $E(N) = \sum_{i=2}^{\infty} (i-1)\pi_i$. Since arrival rates are time-dependent (λ_i is not the same for all i), the average arrival rate, denoted by $\overline{\lambda}$, has to be used in Little's law (Laguna and Marklund, 2013), i.e., $E(W) = \frac{E(N)}{\widehat{\overline{\lambda}}}$, where $\widehat{\overline{\lambda}}$ is the amplified average arrival rate: $\widehat{\overline{\lambda}} = \frac{\overline{\lambda}}{(1-p)}$ and $\overline{\lambda} = \sum_{i=1}^{\infty} \lambda_i \pi_{i-1}$. Therefore, $E(W) = \frac{(1-p)}{\overline{\lambda}} E(N)$.

Note that in steady state, the average arrival rate to each sub-system $\overline{\lambda}$ is $\frac{\lambda}{s}$ because the arrival rate λ is equally splitted over s identical sub-systems.

The most important approximation for the two models is that the external delay duration equal to zero (immediate return). Yom-Tov (2010) demonstrates that, in steady state conditions, standard quality measures of $Erlang - R$ model is independent of external delay duration ($1/\delta$). However, this independence does not hold for the studied models. Appendix A.2 is an illustration of the influence of external delay duration ($1/\delta$) on the queue performance (WT), for all scenarios of Figure 6.3. For instance, Figure 6.12 represents one of these scenarios (p=50%) and shows the impact of $1/\delta$ on the average WT. This dependency is due to the fact that in our models $1/\mu \neq 1/\mu'$, in contrast to $Erlang - R$ model. This figure confirms the independence between $1/\delta$ and WT for the scenario corresponding to $Erlang - R$ ($\overline{SPSP}$ with $\alpha = 0\%$). In contrast,

for scenarios with $\alpha > 0\%$, the curves always increase then tend to stabilize around a certain value, becoming independent of $1/\delta$.

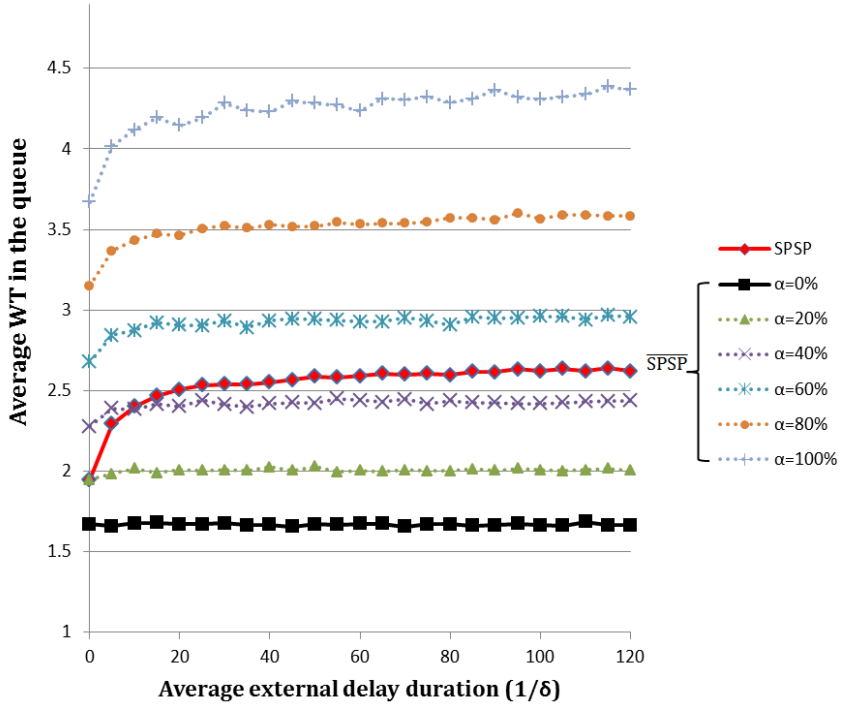

Figure 6.12: Impact of external delay duration on the average queue WT ($s = 2, 1/\lambda = 10, 1/\mu = 5, p = 50\%$)

6.8 Conclusions

We addressed the question of whether ED patients should be handled by the same physician all along the ED process (non-collaborative strategy or *SPSP*), or could be handled by a different one after the initial consultation (collaborative strategy or $\overline{SPSP}$). The conducted survey confirmed that *SPSP* stands as the standard practice in most EDs worldwide, and revealed that the majority of practitioners are reluctant about the application of a collaborative strategy (increased risk of error, time prolongation, human preference towards *SPSP* for both patient and practitioner, etc.). The survey also provided the practical conditions for a proper application of a collaborative strategy (straightforward cases requiring exams, proper handovers with reliable and exhaustive records, teams with homogeneous physician profiles, etc.). Mainly because of some task redundancies and a necessary time adaptation, the exam "interpretation and decision" step would suffer from a time extension when performed by a different physician. The latter would represent a percentage of the initial consultation. This time extension was justified and quantitatively estimated by experts.

We introduced the two system processes corresponding to *SPSP* and $\overline{SPSP}$ as complexity-augmented $Erlang - R$ queueing networks. We showed through simulation that the relevancy of the collaborative strategy depends on the system load. There is a certain threshold, related to the system load, under which the collaborative strategy ($\overline{SPSP}$) outperforms *SPSP*, and above which its application becomes detrimental. We have demonstrated that this threshold is a function of all the key system characteristics that compose the system load. Before reaching the threshold, $\overline{SPSP}$ dominates *SPSP* thanks to pooling effect. After the threshold, *SPSP* prevails because $\overline{SPSP}$ pooling is no longer an advantage (no or rare idleness situations) from the one hand, and because of service time extension from the other hand.

Numerical experiments under realistic conditions derived, for Saint Camille decision makers, useful insights that could stand as a strong argument against the reluctance of practitioners towards $\overline{SPSP}$. Experiments revealed that the collaborative strategy always improves the system performance in the current system. It is more beneficial for less severe cases and the amount of this improvement strongly depend on the amount of service time extension. $\overline{SPSP}$ remains beneficial for a wide range of overall system load. However, it could deteriorate the average LOS of critical patients for highest service time extensions. As a perspective for future experiments, it would be beneficial to perform a sharper analysis assessing the effect of $\overline{SPSP}$ within different periods of the day instead of considering a long term average LOS.

From this study one may summarize the recommendations and guidelines to ED managers as follows. The use of the collaborative strategy is recommended when the system load is low. For instance, the patients on which it should be applied are those requiring reasonable service times and with low probability to undergo examination tests (non-critical patients). Another key driver is the service time extension of the examination result interpretation. In order to minimize the latter, all the necessary actions must be taken in practice, primarily by using homogenous physicians (same team, experience and education), and improving the quality of physicians handovers. Following the initial consultation, physicians must provide reliable medical records filled properly.

Chapter 7

Process-related experiments: Modeling and analysis of triage nurse ordering

This chapter deals with a process-related intervention that takes place in the pre-diagnostic stage. We examine a modification in the current process called triage nurse ordering, which consists in allowing triage nurse to order tests before the patient is seen by the physician. We model the new patient path and assess its efficiency on the ED performance through simulation, while considering the length of stay as the key indicator. We examine the impact of triage nurse ability, system load and triage time extension on the benefits that might be derived from triage nurse ordering.

This work is published in the proceedings of the *IEEE International Conference on Industrial Engineering and Systems managements (IESM)* held in 2015, in Sevilla, Spain (Ghanes et al., 2015b).

7.1 Introduction

As already mentioned, the high emergency departments operating costs combined to budgetary limitations are urging the need for cost-effective solutions to address EDs inefficiencies and thus improve performance. Among these, triage nurse ordering (TNO) appears to be a promising approach that does not require any resource investment. It can be achieved using existing staff with little additional training (Rowe et al., 2011).

TNO is an *advanced triage intervention* that consists in allowing triage nurse to order tests and treatments before the patient is seen by the physician (Robinson, 2013; Seaberg and MacLeod, 1998; Pallin and Kittell, 1992). The common protocol in the ED is that triage nurse cannot order diagnosis tests. She is essentially responsible of making a first assessment of patients state and categorizing them into different acuity levels. The decision of requiring diagnosis tests or not is traditionally under the responsibility of the physician. However, the medical literature suggests that with an appropriate education and training, and adapted protocolled guidelines, triage nurses could be able to order some tests to a level comparable to that of physicians (Robinson, 2013; Free et al., 2009; Fry, 2001; Seaberg and MacLeod, 1998). Diagnostic imaging and laboratory tests are time-consuming processes in the ED that are associated with longest length of stay (LOS) (Robinson, 2013; Vegting et al., 2011; Yoon et al., 2003). If tests are early requested in the triage process, they could be undergone without waiting for the first examination by the ED physician, and test results could be reviewed by the latter as soon as she becomes available.

It is known that initial delays in EDs are associated with abandonment (see Chapter 2), and a relatively strong evidence suggests that *Triage interventions* in general (TNO, team triage and physician in triage) would reduce the number of patient LWBS (Oredsson et al., 2011). TNO has been related to enhanced patient satisfaction (Lindley-Jones and Finlayson, 2000; Parris et al., 1997; Lee et al., 1996). However, "little is known about the effectiveness of this intervention in improving ED time metrics" (Rowe et al., 2011). Only few medical papers reported that TNO could possibly reduce the ED LOS (Retezar et al., 2011; Cheung et al., 2002; Seaberg and MacLeod, 1998; Lee et al., 1996). As mentioned in Oredsson et al. (2011), there is only limited scientific evidence that having nurses to request certain tests results in shorter waiting time and LOS. Moreover, as highlighted by Rowe et al. (2011), the existing LOS improvements revealed in the literature may range widely (from 2.45 to 74 minutes). There is a real need to conduct studies that will legitimize the use of TNO in EDs in terms of LOS reduction (Robinson, 2013; Rowe et al., 2011).

In the present chapter, the objective is to analyze the effect of TNO on ED time metrics

taking into consideration the key parameters of such an intervention. We relied on an online survey that we performed with EDs from different countries in order to understand the current practices and obtain experts opinions. The survey questions focused on the relevancy of TNO, on which types of diagnosis tests exactly it could be applied and why. This survey helped us to understand the whys and wherefores of this problem in order to delimit the framework and include the most relevant parameters related to TNO in our model. Using simulation, we assess the effectiveness of TNO as a function of triage nurse ability level, and investigate other elements that could have an impact on this effectiveness (system load and service time extension). We derive useful insights that can assist decision makers when implementing a TNO intervention.

The rest of the chapter is organized as follows. A literature review on TNO is presented in Section 7.2. In Section 7.3, we report the results of the performed survey and describe in detail our TNO model. In Section 7.4, we conduct experiments. Finally, in Section 7.5, we summarize the main insights and highlight some future research.

7.2 Literature review

TNO, also called *advanced triage* (Rosmulder et al., 2009; Cheung et al., 2002), is a worldwide ED question that was addressed in North America (Retezar et al., 2011; Cheung et al., 2002), Europe (Rosmulder et al., 2009; Lindley-Jones and Finlayson, 2000), Asia (Than et al., 1999; Lee et al., 1996) and Australia (Parris et al., 1997). The prior literature examining TNO is almost exclusively addressed from a medical perspective. It consists in general on 2 types of empirical studies: examining the ability of triage nurses in initiating diagnosis tests properly (Seaberg and MacLeod, 1998) or assessing the effect of such an intervention on ED time metrics (Cheung et al., 2002; Lindley-Jones and Finlayson, 2000; Parris et al., 1997). For the first type, the method generally consists in using the attending physician as a standard to judge the accuracy of triage nurse orders, and for the second it consists of comparing statistically two samples of patients (one with the traditional ED process and another with TNO) in terms of time metrics, mainly LOS. However, to the best of our knowledge, no paper investigates how various nurses abilities could influence the system performance.

As highlighted in Rowe et al. (2011) and Oredsson et al. (2011), most of TNO interventions are limited to some radiographs, mostly joints and bones of distal limbs (Fry, 2001; Lindley-Jones and Finlayson, 2000; Lee et al., 1996). However, some include additional diagnostic test requesting such as blood tests, urinalysis, electrocardiogram (Cheung et al., 2002; Winn, 2001; Seaberg and MacLeod, 1998; Kirtland et al., 1995) and radiographs of other parts like the skull (Than et al., 1999). Even if there is some unanimity about distal limb radiographs, the choice

of the diagnosis type for TNO is rarely justified in the literature and still remains unclear.

TNO has been related in some papers to a decreased LOS. Rosmulder et al. (2009) report that LOS decreased by 27 minutes (18%) with foot/ankle X-rays initiated at triage. In Cheung et al. (2002), time savings is on average 46 min in the total LOS with TNO applied on some X-rays and blood tests. In Lee et al. (1996), the total LOS for patients with radiographs requested by the nurse is on average 18.59 minutes less than the overall average. Lindley-Jones and Finlayson (2000) report that a mean reduction of 37.2 min (36%) from time of triage to time of treatment decision was achieved in the group of patients with triage initiated X-rays compared to control group. However, according to Rowe et al. (2011) and Robinson (2013), there is a paucity of research examining the effect of TNO intervention on ED time metrics. Time reductions related to TNO may range widely (from 2.45 to 74 minutes according to Robinson (2013)) and some negative conclusions have also been reported. Parris et al. (1997) perform a comparison between a group of patients who had X-ray initiated in triage and a group with a regular pathway, and find that the difference in LOS between the two groups is not statistically significant. However, staff and patient satisfaction with this change is high, which justifies carrying on the practice in the ED.

Satisfaction is not formally measured but it is reported that physician satisfaction increases through the availability of diagnostic results since the first examination. Patients seem satisfied for using the waiting time more efficiently in addition to a greater sense of team working for all staff (Cheung et al., 2002; Fry, 2001; Lindley-Jones and Finlayson, 2000; Parris et al., 1997).

One of the arguments facing TNO is the fear of overrequesting diagnostic tests that would not have been ordered by the physician (Lee et al., 1996). The potential benefits of TNO in terms of time savings and satisfaction must be balanced with the disadvantages of such excessive requests: additional time, additional expense and increased resource utilization, unnecessary radiation exposure and potential morbidity (Seaberg and MacLeod, 1998; McArthur and Thomas, 1995). For instance, Lee et al. (1996) address the problem of triage nurse ability in ordering radiographs including 934 patients in their study. The triage nurse requests radiographs for 94.54% of patients (883), from which 5.44% (48 out of the 883) are considered unnecessary by the case physician.

Under-requesting is another type of possible error. Triage nurse could miss some necessary tests that will further be required by the physician. In Lee et al. (1996), among the same sample of 934 patients, triage nurse did not order any radio for 51 cases (5.5%), 65% (33/51) has an X-ray requested by the attending physician (3.5% of the total sample, 33/934). TNO presents also a risk of additional tests following the physician examination. Additional views of the same/adjacent or different regions can be ordered because the first view does not demonstrate

the problem or another injury is discovered during the consultation (Oredsson et al., 2011; Macleod and Freeland, 1992). "Additional trips to the radiology department become necessary, increasing both the time required for treatment and the inconvenience to the patient" (McArthur and Thomas, 1995). In Lee et al. (1996), 11% of ordered radios (97 out of 883) are followed by additional ones after physician assessment.

In addition to Lee et al. (1996), a few other papers examine the ability of triage nurses in initiating radiographs appropriately (using similar inclusion criteria). The reported statistics from most complete studies are provided in Table 7.1. Note that when additional tests are ordered by the physician, studies do not mention whether the ones ordered by the nurse were necessary or not. No statistics are formally reported about patients with both over-requesting and underrequesting (additional tests after unnecessary tests) except in rare papers like Seaberg and MacLeod (1998) where these patients represented 15%.

Table 7.1: TNO ability statistics reported in the literature

	Lee et al. (1996)	**Macleod and Freeland (1992)**	**Thurston and Field (1996)**	**Others**
Sample size on which TNO was applied	934	579	915	
Triage nurse Requesting rate	94.5%	72% (416/579)	78%	
Over-requesting rate (N+/P-)*	5.4% of ordered tests considered unnecessary by physician	6.5% of ordered tests considered unnecessary by physician	4% of ordered tests considered unnecessary by physician	4.5% and 8% in McArthur and Thomas (1995) and Rosmulder et al. (2009) respectively
Under-requesting rate (N-/P+)*	65% of situations of no tests ordered by nurse were followed by a physician order (3.5% of total sample)	47.2% of situations of no tests ordered by nurse were followed by a physician order (13.3% of total sample)	23.5% (66/281) of situations of no tests ordered by nurse were followed by a physician order (8% of total sample)	
Rate of additional tests requested by physician (N+/P++)*	11% of already x-rayed patients	5.3% (22/416) of already x-rayed patients	7.2% of already x-rayed patients	7.8% in Lindley-Jones and Finlayson (2000)

*The notations are explained later in Figure 7.2.

The success of TNO can likely be achieved using existing triage nurses with little additional training (Rowe et al., 2011). In most of the analyzed papers, triage nurses skills were extended before experiments with training programs on examination skills and inclusion/exclusion criteria for exams requisition (Cheung et al., 2002; Fry, 2001; Lee et al., 1996). As demonstrated by

Seaberg and MacLeod (1998), the ability of triage nurse in ordering tests can be improved with the use of test ordering guidelines. Lindley-Jones and Finlayson (2000) reports a reduced gap between triage nurse and physician ability in ordering radiographs after participating in a 1-day training program and by using carefully designed protocols. However, there are no standardized guidelines for TNO interventions (Rowe et al., 2011). Reported trainings are various in time and contents (Robinson, 2013; Rowe et al., 2011). In addition to an initial non-uniformity in nurse education between countries, and also within the same country (Free et al., 2009), different TNO trainings and protocols could have influenced results reported in studies (Robinson, 2013; Thurston and Field, 1996).

As explained above, TNO contributions consisted either in measuring the impact of such an intervention on time metrics or assessing triage nurse ability. However, it should be noted that no paper analyzed or quantified the impact of nurse ability on patient time in the ED so far. Service times extension is also an element that was not addressed. As mentioned in Lindley-Jones and Finlayson (2000) and Lee et al. (1996), the average time of triage and consultation could be lengthened under TNO which could also affect the results.

In the OR/OM domain, we identify very limited contributions. Kirtland et al. (1995) uses simulation to test several staffing and process alternatives in order to reduce the patients LOS in an ED (TNO, fast track, point of care testing, etc.). There was no significant time savings related to TNO (3.6 minutes on the total average LOS). However, the authors suggested that TNO would be more effective when the system is quite busy, but that was not demonstrated. With the exception of the considered types of exams (X-rays, lab tests and ECG), the paper does not provide any information about the parameters included in the TNO model (changes in the patient pathway, triage nurse ability, etc.).

7.3 Setting

In order to understand experts opinions about TNO and define the appropriate model framework, a survey is performed in many EDs in France, USA, the Netherlands, Germany, Belgium, Greece and Tunisia.

7.3.1 Survey results

The results are based on 36 practitioners (ED managers and ED physicians) from 24 different EDs. Experts from French EDs provided 75% of the answers. The majority of surveyed experts (86%) considered TNO as a potential relevant practice in general. However as shown in Table 7.2, the feasibility of TNO varies greatly from one test type to another. For each one, the experts

provided the practical reasons about the possibility to apply TNO or not.

Table 7.2: The possibility to apply TNO for the main types of tests

	For which types of tests could TNO be relevant?
X-ray	**97%**
CT scan	3%
MRI	0%
Echo	0%
Blood tests	47%
Urine analysis	83%
ECG	83%

Conventional radiology, also called standard radiology, radiographs or X-rays, were considered by the experts as the most appropriate diagnosis tests for a TNO intervention in their EDs. Only some particular types of radiographs are concerned which are those for simple extremity traumatology of stable patients. That includes bones and joints radiographs of distal limbs that are below the large joints like the hip and the shoulder (hand, wrist, elbow, foot, knee, ankle, etc.). They are routine tests, easy to perform with limited risk for patients (noninvasive). Multi trauma cases as well as radiographs of other parts such as spine, chest, abdomen and pelvis should be excluded.

For many reasons, CT scan, MRI and Echo were judged inappropriate for a TNO intervention. They are much more expensive and represent critical resources in the hospital. The application of TNO on these tests requires a medico-economic evaluation. Moreover, they are invasive and more specific tests. The decision of ordering such tests is complex and cannot be done without a complete (physical and clinical) examination by a physician.

Biological tests are also complex and costly and the opinion of experts about the application of TNO on them is mixed. The survey revealed 83% of favorable views for TNO of urinalysis. However, that was limited by experts to certain basic urine tests. The mainly mentioned candidate is a type of urine analysis called urine test strip. It is a basic and quick diagnosis tool that is used by practitioners in the ED without even resorting to the laboratory. Other kinds of basic urine tests were mentioned such as urine pregnancy tests. According to the survey, blood tests can be ordered by triage nurse only in certain cases (fever in a patient back from a tropical country or in a patient receiving chemo, diabetes, HIV testing, etc.). More specific and sophisticated blood tests require a clinical examination by a physician and must be discussed on a case-by-case basis. The disparity in the situations requiring biological testing makes them

difficult to generalize.

There is unanimity on allowing triage nurse to decide about an Electrocardiogram (ECG) in several cases like chest pain. This protocol is more common than other tests and is already applied in many surveyed EDs.

7.3.2 The TNO model description

Given the answers collected from the performed survey, the study will focus on low acuity level patients (ESI 4 and ESI 5) with distal limbs traumatology requiring conventional radiographs (X-rays). In France, Trauma injuries represent about half of ED visits (Potel et al., 2005; Baubeau and Carrasco, 2003). Among these, trauma to the extremity whether upper like wrist and hand injuries, or lower like ankle sprain represent the most common cases in EDs particularly among non-critical patients, with X-rays being the reference test (Ganansia, 2003).

As explained in Chapter 2, length of stay (LOS) is the key metric for this kind of patients. "Due to relatively minor nature of these injuries, those patients have often to wait a long time for treatment and investigation in EDs" (Parris et al., 1997). This group of patients rarely requires biological tests which make them free of any necessary sampling, and allow sending them to radiology right after triage.

The interviews with experts combined with the existing literature allowed to identify the most relevant parameters that could have an impact on the effectiveness of TNO, and that will further constitute the basis of our experiments:

- The accuracy of triage nurse in requesting tests (over, under and incomplete-requesting).
- The ED level of crowding.
- The impact of TNO on some service times.

As shown in Figure 7.1, compared to the traditional patient pathway, when triage nurse orders diagnosis tests, they are initiated right after triage. Since ordering tests is an additional task for triage nurse, the triage service time represented in our model by a random variable T could be increased by a certain amount of time ΔT. When tests are completed, the physician examines the patient for the first time and interprets her tests results during one single aggregated consultation. In our model, we make the assumption that this task has the same time distribution with a regular consultation.

The TNO path depicted in Figure 7.1 is a simplified representation that corresponds to a particular ideal scenario under TNO (see N+/P+ in Figure 7.2). According to the case and the triage nurse risk of error, we distinguish in total between six possible situations. The latter are depicted with the appropriate formalism in Figure 7.2.

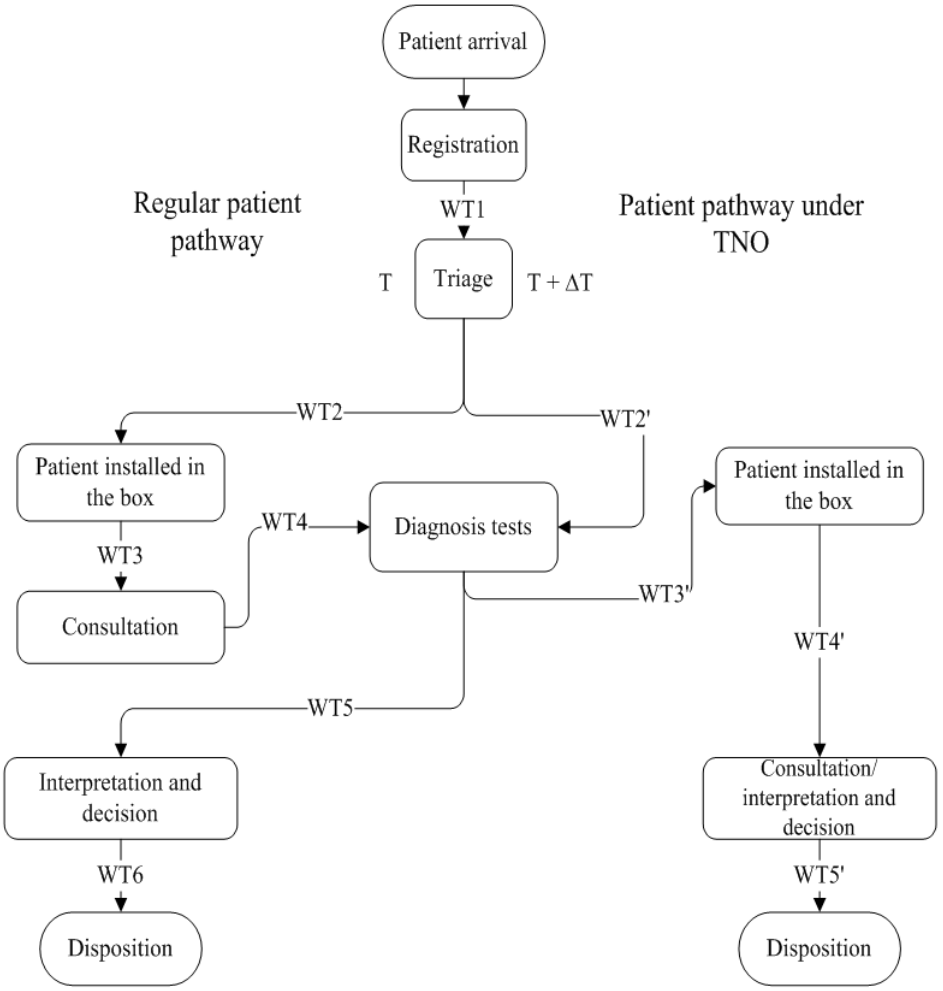

Figure 7.1: Regular and TNO patient pathway

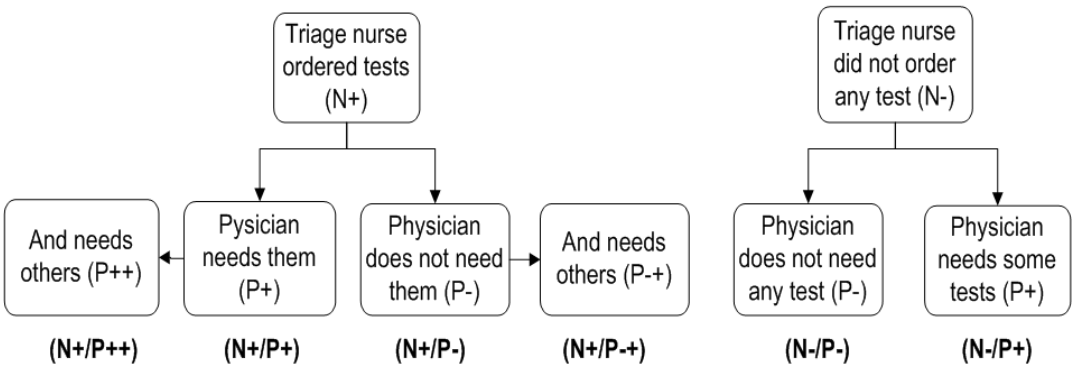

Figure 7.2: Possible situations under TNO

(N+/P+) is the ideal scenario that embodies the usefulness of TNO. Triage nurse orders the appropriate tests which would allow saving the time of a first consultation and its corresponding waiting times (WTs). In scenario (N-/P-), the nurse is right and this situation has neither advantages nor disadvantages. (N-/P+) is a harmless situation of underrequesting where the

TNO patient pathway is unintentionally turned into a regular one.

The rest of the scenarios are considered negative and would generate loss of time because of nurse errors made in ordering tests. (N+/P-) is a situation of over-requesting where time is lost for unnecessary tests; while without TNO, the patient would have been discharged right after the first consultation. (N+/P++) is a particular situation of under-requesting where triage nurse orders some necessary tests while missing others (the physician may order radiographs of other parts or different views of the same), which imposes additional trips to radiology. (N+/P-+) is a combination of both under and over requesting in which the patient is first sent unnecessarily to radiology by triage nurse, and then sent back later by the physician. In addition to potential loss of time, these situations have an impact on patient convenience and satisfaction.

As mentioned in the literature review, in empirical studies, reported statistics on the two situations (N+/P++) and (N+/P- +) are merged. Thus, they will also not be differentiated in our model and will all be considered as (N+/P++). In other words, we make the assumption that when the physician orders additional tests, the ones ordered by triage nurse are never completely useless. Consequently, we obtain five possible patient paths under TNO that are represented in Figure 7.3.

7.4 Experiments

Our experiments consist in a case study involving Saint Camille ED and are divided into 4 parts. In the first one, the objective is to understand the impact of nurse abilities and decisions on TNO effectiveness. We calculate the expected improvement in LOS as a function of a realistic range of TNO-related probabilities. In the second part, we extend the analysis by varying the different probabilities within a wider range of values in order to figure out what is the most harmful nurse error. In the third part, we assess the relationship between the ED load and TNO effectiveness. Finally, we assess the impact of triage processing time extension on the system performance. Collected data from Saint Camille ED indicate that eligible patients represent 17% of the total number of patients. We use the realistic discrete-event simulation model presented in Chapter 4 to conduct experiments on Saint Camille ED. For each simulation, we use a sufficiently long simulation period (semi-infinite).

7.4.1 The impact of realistic trained triage nurse ability on LOS

We identify 4 key probabilities related to a TNO intervention. The first one is triage nurse requesting rate. It represents the rate of patients sent to radiology by triage nurse among all eligible patients for TNO. The three other probabilities characterize triage nurse ability and

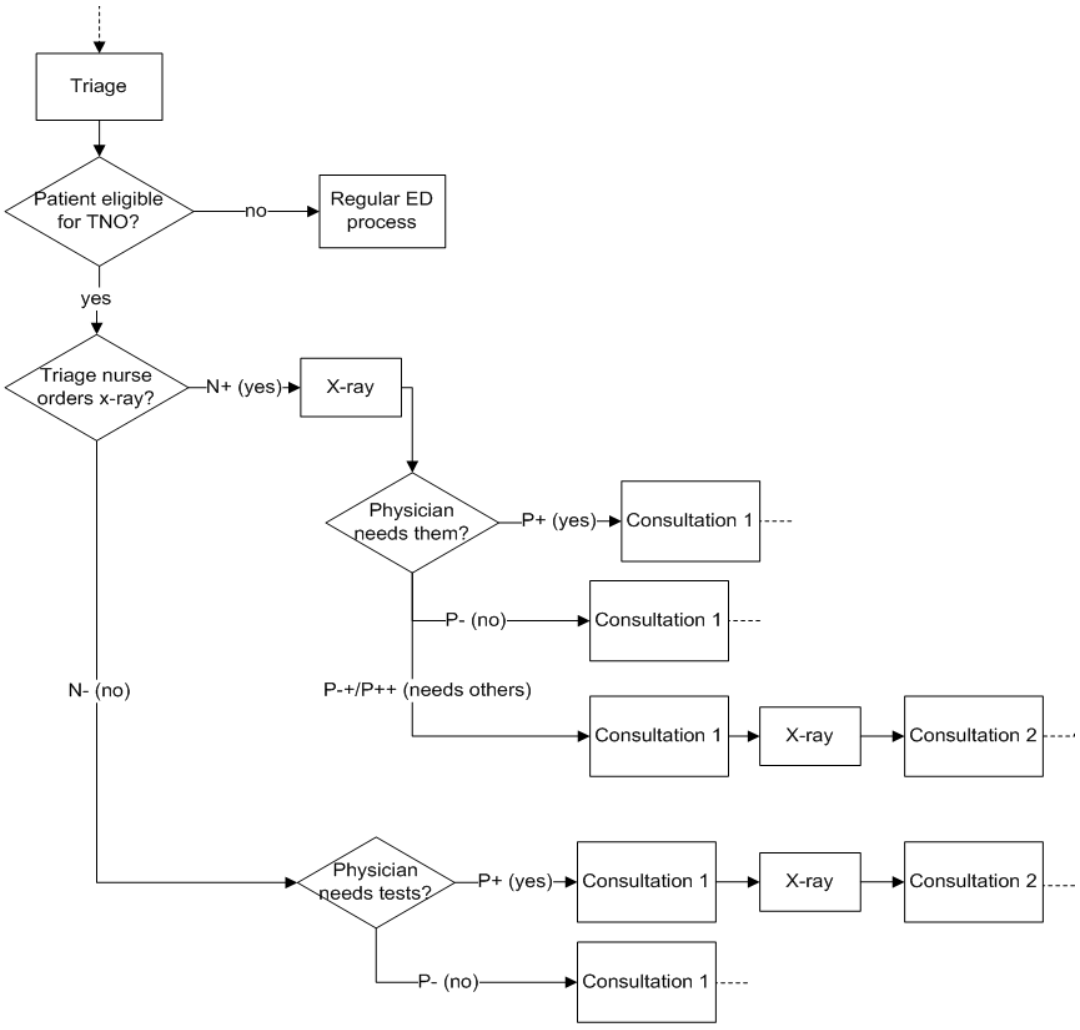

Figure 7.3: Possible patient pathways under TNO

her precision in requesting tests: over-requesting rate (N+/P-), underrequesting rate (N-/P+) and the rate of additional tests requested by physician (N+/P++). For each rate, we use the lowest and the highest value found in the literature (see Table 7.1) and generate all the possible combinations ($2^4 = 16$ in total). For instance, the scenario H-LLH refers to a TNO intervention where triage nurse has a high requesting rate, a low probability of over-requesting, a low probability of under-requesting and a high probability to require incomplete tests (that will be followed by a physician test order). The LOS improvement of each scenario for eligible patients is depicted in Figure 7.4 and Figure 7.5.

The derived insights can be summarized as follows:

- Within the used ranges of trained triage nurses ability reported by empirical studies, TNO is a beneficial intervention for all combinations.

- For any given trained nurse ability (for any set of over/under/incomplete requesting rate),

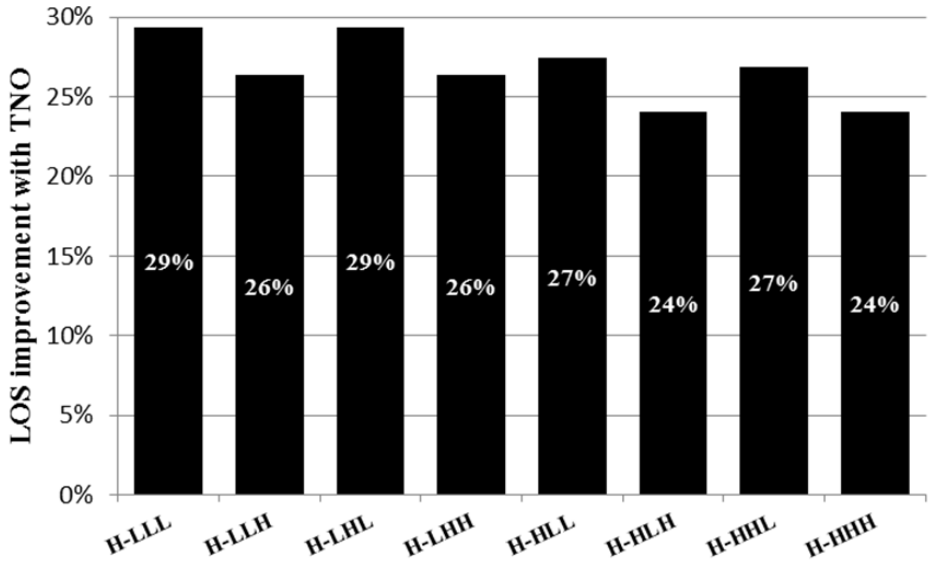

Figure 7.4: LOS improvement with TNO for high requesting rates scenarios

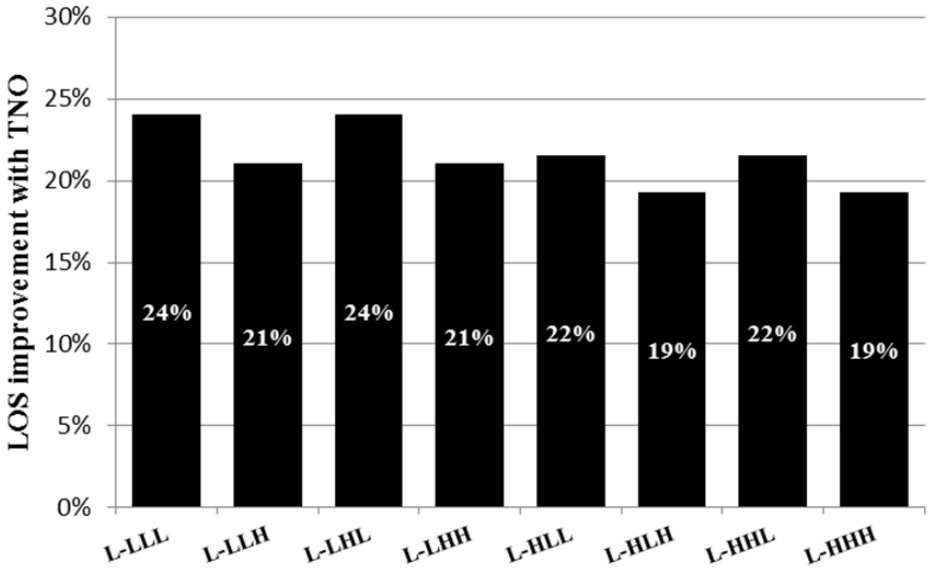

Figure 7.5: LOS improvement with TNO for low requesting rates scenarios

trained triage nurse should preferably order tests as much as possible while respecting the predetermined protocols.

- Under-requesting appears to be harmless. Over and incomplete requesting both reduce the benefit from TNO.

The worst scenarios are the situations L-HHH and L-HLH (19.31% of average LOS improvement among eligible patients and 4.49% of overall improvement) where triage nurse has the worst abilities while having a low requesting rate. The best scenario is the opposite situations

H-LLL and H-LHL (29.38% among eligible patients and 7.65% overall).

7.4.2 Extended analysis of TNO effectiveness as a function of the key probabilities

So far, no conclusion can be drawn about which triage nurse ordering error is the most harmful. This is because the used error probabilities coming from the literature are limited. They correspond to the ability of triage nurses that were preliminarily trained. In what follows we will experiment the best scenario H-LLL by varying one by one each probability rate from 0% to 100% (Figure 7.6).

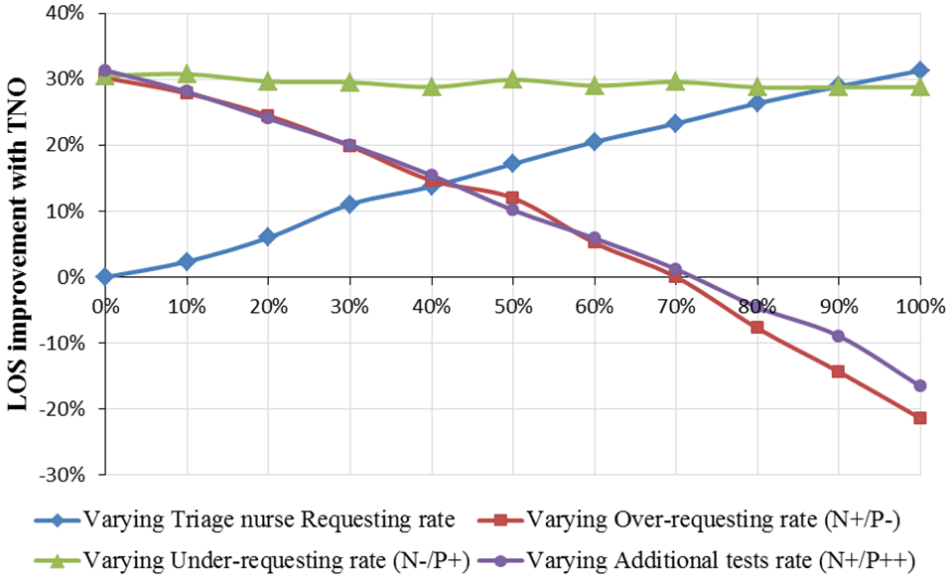

Figure 7.6: Sensitivity analysis on the 4 key probabilities

The following insights can be derived:

- The benefit from TNO is more apparent for higher requesting rates (with respect to protocols).
- The risk of under-requesting rate has no impact on TNO performance.
- The risk of over-requesting and the risk of incomplete requesting (additional tests further requested by physician) affect TNO performance and have similar impacts on it. This result is quite intuitive since both of them consist of an additional trip to radiology department. This result holds under the assumption that when the physician orders additional tests, the ones ordered by triage nurse were not completely useless. Otherwise, the rate of additional tests would be the most harmful.

- For over and incomplete requesting rates, there is a threshold under which TNO could be detrimental for the system performance.

7.4.3 TNO effectiveness as a function of the system load

The actual arrival pattern in Saint Camille ED depends on the day of the week and the hour of the day. Similarly to Yom-Tov and Mandelbaum (2014) and Ahmed and Alkhamis (2009), we assume that arrivals follow a non-homogenous Poisson process (7x24 arrival rates).

In order to assess the relationship between the system workload and the expected benefit from TNO, we perform a sensitivity analysis by varying arrival rates, in the best and the worst scenarios of Section 7.4.1 (see Figure 7.7). The following conclusion can be drawn: The benefit derived from TNO is more apparent for heavily loaded EDs.

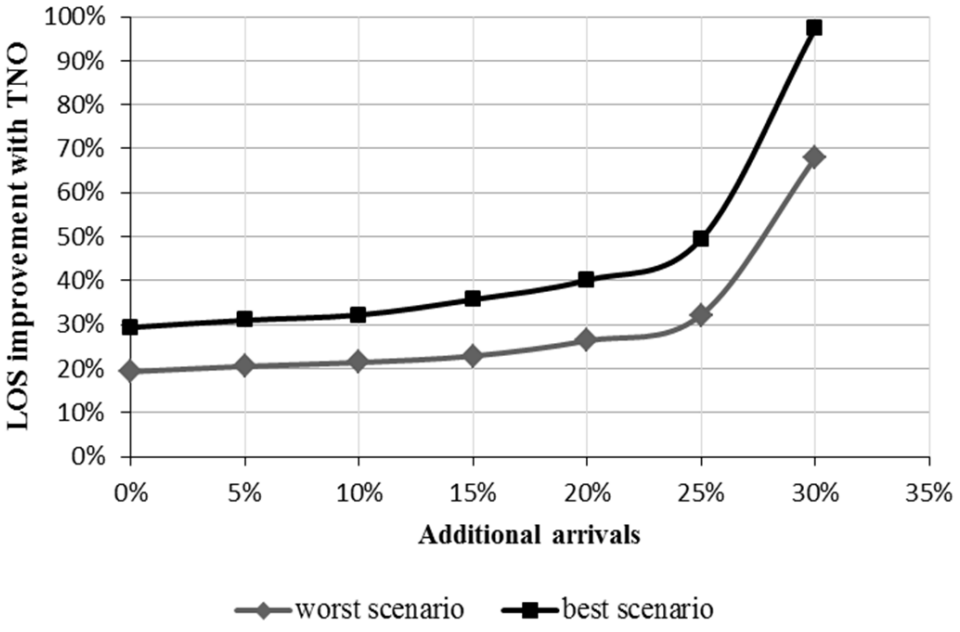

Figure 7.7: TNO effectiveness depending on arrivals

7.4.4 The impact of triage service time extension on TNO effectiveness

In what follows we address the question of triage service time extension because of TNO and assess its impact on TNO effectiveness. According to data collection (experts judgment in particular), the distribution of triage service time is assumed to be Normal (7, 1.5). Using the best and the worst scenarios from part 1 (H-LLL and L-HHH respectively), we perform a sensitivity analysis on triage time by extending it up to 200% (see Figure 7.8).

We derive the following insights:

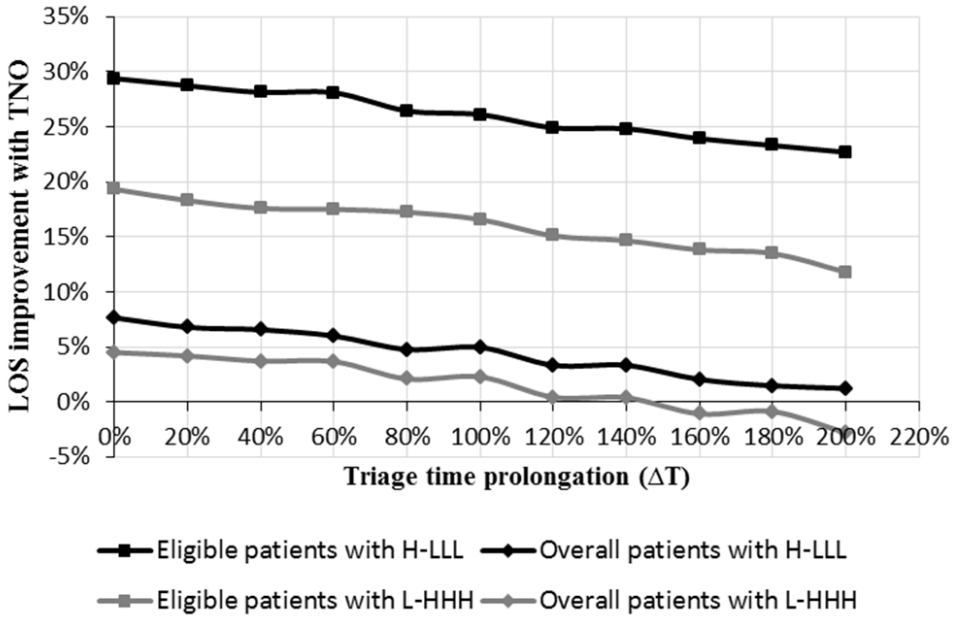

Figure 7.8: TNO effectiveness depending on triage time extension ΔT

- Triage time extension lowers TNO effectiveness, but should not be considered as a major concern for reasonable time extensions and nurse abilities.
- Triage time extension can make TNO detrimental for the overall system. For instance, for limited triage nurse abilities (L-HHH) and high triage extensions (from 150%), TNO remains beneficial for eligible patients but the overall system LOS is affected by longer waiting and processing times for triage.

7.5 Conclusions and perspectives

The present chapter is an ongoing work that represents the interface and the link between the two commonly addressed TNO issues, namely the assessment of triage nurse ability and the assessment of the effectiveness of this intervention in improving ED time metrics. We formalized with an OR approach the TNO model. We confirmed and quantified some intuitive elements regarding TNO and derived useful insights that will help decision makers for a successful implementation of TNO. For instance, TNO is always beneficial within a reasonable range of triage nurse ability level. However, there is a threshold on this ability under which TNO could be detrimental for the system performance. This confirms the importance of an adequate nurse training on inclusion criteria before implementation of TNO. The benefit derived from TNO is more apparent for heavily loaded EDs. We also demonstrated that triage time extension does not have a significant impact on eligible patients but can negatively affect the rest of the ED patients

and the overall ED performance. Although the modeling is based on a specific ED, qualitative conclusions hold for other ED frameworks. As a perspective, an analytical modeling for TNO would be helpful for the generalization of the aforementioned results. Similarly to initiating diagnostic tests earlier with TNO, it would be an interesting avenue for future research to assess the ability and the efficiency of triage nurse to initiate search for admission beds earlier (Potel et al., 2005; Kirtland et al., 1995) as a way to reduce transfer delays.

Chapter 8

Conclusion and perspectives

This chapter provides general conclusions and perspectives related to the research works presented throughout this manuscript. It summarizes the concluding remarks of the previous chapters through a holistic view, and gives plausible perspectives further to this research work.

An emergency department is the service within hospitals responsible for providing care, on a 24/7 basis, for patients with life-threatening cases and other severity levels. It is a key actor for public health and safety as a whole. However, EDs are facing a worldwide problem called overcrowding or congestion which leads to several negative effects on the quality of care, patients safety and satisfaction, working conditions as well as ED global revenue. Therefore, medical practitioners as well as OR/OM researchers are investigating solutions to alleviate overcrowding in EDs and to improve their performance. The purpose of this thesis is to provide ED managers with insights and cost-effective solutions so that to improve the performance of EDs. Several issues are addressed throughout the thesis. First, prospective studies are conducted in order to identify and understand the currently determinants of an ED performance, and how this latter should be measured. A special focus is made on internal operations management issues within EDs, i.e., resource-related and process-related issues. The experiments and the surveys that were conducted during this research work revealed both qualitative and quantitative results that might be very useful for practical management. All of the aforementioned is discussed in details in the first chapter.

In Chapter 2, a detailed literature review was given on the commonly used key performance indicators for emergency departments from an operations research and operations management perspective. Their respective features and their selection approach were identified and discussed. The study revealed that each KPI is used to measure specific ED aspects, and thus the choice of the appropriate KPI to be optimized is primordial. The advantages and drawbacks of each KPI were also highlighted. We underlined the value of combining complementary KPIs to provide relevant solutions in practice. Finally, we highlighted a number of missing literature on OR/OM such as the one related to fairness, universal measures of crowding, etc. In Chapter 3, statistical analyses were conducted to identify the factors that mainly contribute to longer stays in EDs. The study revealed that ED congestion is a multifactorial phenomenon with different factors simultaneously leading to ED longer LOS. The study concluded that improving ED performance requires a series of different remedial measures each focusing on a distinct influencing factor. For each factor of overcrowding, we provided an interpretation of how its influence is exerted in practice, and highlighted relevant corresponding remedial measures to alleviate this influence. The research avenues that have been derived in this chapter provided a basis to define (or confirm) the issues that were addressed in the next chapters.

In Chapter 4, a realistic ED discrete-event simulation model was presented. Thanks to a close collaboration with practitioners, essential structural and functional characteristics of EDs were identified and included in the model. Our experiments focused on ED internal human staffing

levels and provided useful insights to managers on how performance is affected by staffing budget, and how the latter could be rationally and efficiently increased. We also analyzed the impact of jointly considering the two main KPIs (overall LOS and $DTDT$ for critical patients) when optimizing performance. The results showed that the lower is the budget, the more apparent is the interdependency between these two KPIs. Some avenues for the improvement of the accurateness of the simulation model were highlighted, such as those concerning abandonment probabilities and the patient health status (which is supposed to be variable in practice). Moreover, an additional resource-related issue which consists in shift definition was investigated in Chapter 5. A method combining simulation-optimization and mathematical programming was proposed. This method optimizes the allocation of available resources, without increasing costs, while respecting the major constraints as encountered in practice. However, it is a time-consuming procedure that involves several software programs, which will require to develop an automated combination of these.

In Chapter 6, we assessed possible modifications in the ED process and their consequences on the performance. We focused on a tacit and widely used rule in EDs that we called *Same Patient Same Physician rule* (*SPSP*), and assessed the relevancy of ignoring such a rule in practice. A survey was therefore conducted, where the results confirmed that *SPSP* stands as the standard practice in most EDs worldwide. The survey also revealed the controversial nature of this issue and the reluctance of most practitioners towards the deletion of the *SPSP rule*. We introduced the two competing system processes as two complexity-augmented $Erlang - R$ queueing networks. We demonstrated how this additional complexity compromises mathematical tractability. Thus, we resorted to simulation in our experiments and showed that the relevancy of using or ignoring the *SPSP rule* depends on the system load. There is a certain threshold, related to the system load, under which ignoring *SPSP rule* is beneficial, and above which it becomes detrimental. Results were further validated under ED realistic conditions. As a perspective for future experiments, it would be beneficial to perform a sharper analysis assessing the effect of $\overline{SPSP}$ within different periods of the day instead of considering a long term average LOS. Another interesting extension would be to include ED nurses in the assessment of the *SPSP rule*.

In Chapter 7, a second process-related issue, namely *triage nurse ordering* (TNO), is addressed. This study represents the link between the two commonly addressed TNO issues in medical literature: the assessment of triage nurse ability on the one hand and the assessment of the effectiveness of TNO in improving ED time metrics on the other. Based on a field survey as well as on literature review, we were able to identify the key parameters needed for the study

and hence to formalize the TNO process. The simulation experiments that were conducted confirmed some of the intuitive elements with additional quantitative insights, and derived useful results that may help decision makers for a successful implementation of TNO. For instance, TNO is always beneficial within the range of trained triage nurse ability level reported in the literature. However, when these ranges are expanded, a threshold is thus to be defined on triage nurse ability under which TNO could be detrimental for the system performance. These conclusions confirm the reported importance of an adequate nurse training on inclusion criteria before implementation of TNO. Moreover, we assessed the effect of other TNO key parameters on the benefit derived from TNO, such as the system load and the expected triage time extension. As a perspective, an analytical modeling for TNO would be helpful for the generalization of the obtained results. Similarly to initiating diagnostic tests earlier with TNO, it would be an interesting avenue for future research to assess the ability and the efficiency of triage nurse to initiate search for admission beds earlier (Potel et al., 2005; Kirtland et al., 1995) as a way to reduce transfer delays.

List of Figures

List of Tables

Appendix A

Appendix of Chapter 6

This appendix deals with the analysis of Chapter 6. First, we provide the survey form used for $SPSP$ as well as the survey sample composition. Then, we provide additional insights related to the mathematical approximation: The impact of external delay durations on the average WT, the impact of using random routing in the analytical model of *SPSP*, and the assessment of the quality of the "1/s" approximation in the $\overline{SPSP}$ analytical model. Finally, we provide additional clarifications concerning the experiments under realistic conditions.

A.1 $SPSP$ survey

A.1.1 Survey form

"Same physician for a given patient during his stay in the ED?"

As demonstrated in the literature, emergency departments (EDs) problems can stem from the process itself and not from the staffing levels (Samaha et al., 2003). Researchers are nowadays developing methods that aim of modifying some protocols and organizational rules regarding the patient path in the ED.

ED physician are responsible of two major tasks which are: 1/ "the initial consultation" : the physician makes a first assessment of the patient's state and may decide, if necessary, to request diagnosis tests 2/ "the interpretation of diagnosis tests": the physician examines the results of the diagnosis tests and decides about the next steps (release, admission, transfer, etc.)

In current practices, the physician who conducts the initial consultation of a given patient will also be responsible of the interpretation of diagnosis results (in case when diagnosis are needed). This aforementioned process is known as the "Same Patient Same Physician ($SPSP$)" rule. However in some EDs, when the first physician is busy, the patient may be affected to another doctor in order to interpret and decide. Even though this strategy may reduce the

waiting time of the patient, it may however increase the duration of the interpretation step handed to the second physician. This is because the latter is not familiar with the patient's situation.

In our research work, we try to better understand the different practices and motivations related to the $SPSP$ rule. Therefore, we ask you to please fill out this form.

- Organization, E-mail address
- Which strategy do you apply in your ED ?
  - Always "Same Patient Same Physician" ($SPSP$)
  - Exams could sometimes be interpreted by another physician
- In which case do you think relevant to apply the second strategy? (Second strategy means allowing another physician to interpret results and take a decision). (open-ended question)
- For which reason wouldn't you apply the second strategy? (open-ended question)
- When the "initial consultation" and the "exam interpretation and decision" are performed by two different physicians, do you think that the duration of the second step would be prolongated?
  - Yes
  - No
- If yes, why? (open-ended question)
- If yes, by how much time this task is prolongated?
  - An additional duration equal to that of the "consultation step" performed by the first doctor
  - A percentage of the duration of the first consultation step
  - Others (open)
- If you answered "percentage" (second option of the previous question), by how much do you think the duration will be prolongated? (open)

A.1.2 Details concerning the sample of ED physicians used in the survey

Table A.1 summarizes the number of experts who participated in the survey classified by ED and country.

Table A.1: *SPSP* survey sample composition

Country	**Hospital**	**Answers**
France	Saint Camille	2
	SAU chartres	4
	Hôpital Tenon	1
	CHU Rouen	2
	Urgences Pontoise	1
	CHU Pitié-salpêtrière	1
	S.A.U. Hôpital Bichat	2
	CH Marne la Vallée	1
	Hôpital Ambroise-Paré	1
	Hôpital privé d'Antony	1
	CH Jossigny	1
	Hôpital privé de Marne la Vallée	1
	SAU Lariboisière	1
	CHU Besançon	1
	CHI Créteil	1
	Groupe hôspitalier Paris Saint Joseph	1
	Saint-Antoine	1
Belgium	Erasme hospital (ULB)	1
Greece	Aristotle University of Thessaloniki	1
Germany	Allgemeines krankenhaus celle	1
USA	Penn State Hershey Medical Center	5
the Netherlands	VU Medical Center	1
Tunisia	Charles Nicole	1
		33

A.2 The impact of external delay durations on the average WT

In this section, we assess the influence of the average external delay duration ($1/\delta$) on the average queue WT. We vary the values of $1/\delta$ in different set of parameters (with the same values of $s = 2$, $1/\lambda = 10$ and $1/\mu = 5$, and different returning rates p). In other words, we repeat the experiments illustrated in Figure 6.12 for other values of p (from 10% to 70%). Experiments confirm the independence between $1/\delta$ and WT for the scenario corresponding to $Erlang - R$ ($\overline{SPSP}$ with $\alpha = 0\%$). In contrast, for scenarios with $\alpha > 0\%$, the curves always increase then tend to stabilize around a certain value, becoming independent of $1/\delta$. The latter demonstrates that external delay duration does have an impact on the queue performance. It also highlights the inaccuracy of considering a null $1/\delta$ in the analytical modeling, and the deviation resulting from such an approximation.

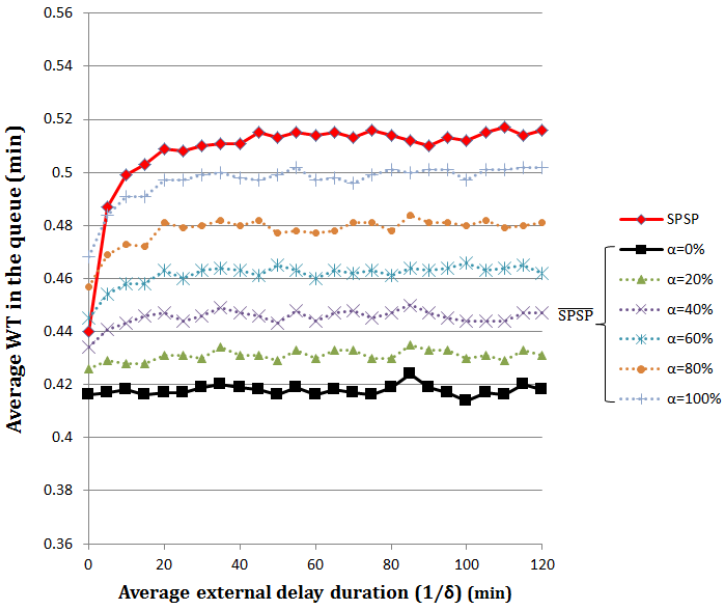

Figure A.1: The impact of external delay duration on the average queue WT ($s = 2, 1/\lambda = 10, 1/\mu = 5, p = 10\%$)

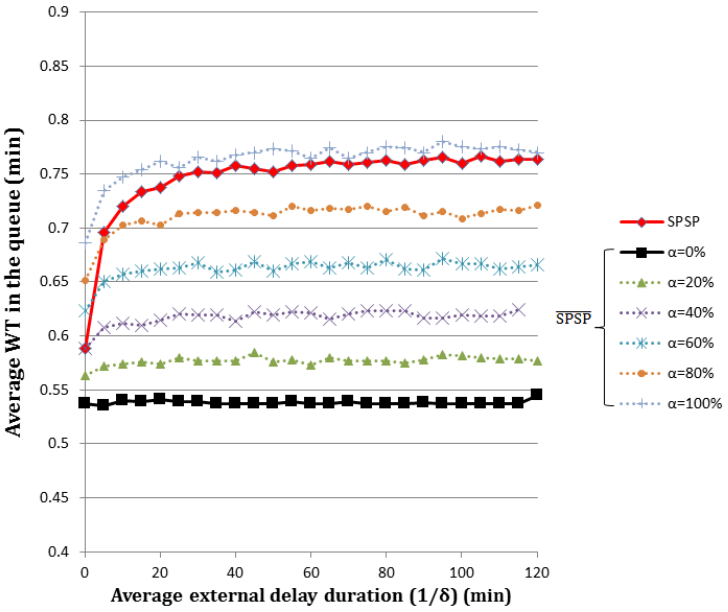

Figure A.2: The impact of external delay duration on the average queue WT ($s = 2, 1/\lambda = 10, 1/\mu = 5, p = 20\%$)

A.3 Using random routing in the analytical model of *SPSP*

Figure A.7 represents the performance comparison in terms of the average queue waiting time between several scenarios of $\overline{SPSP}$ and the numerical results obtained from the analytical approximation of *SPSP* under random routing.

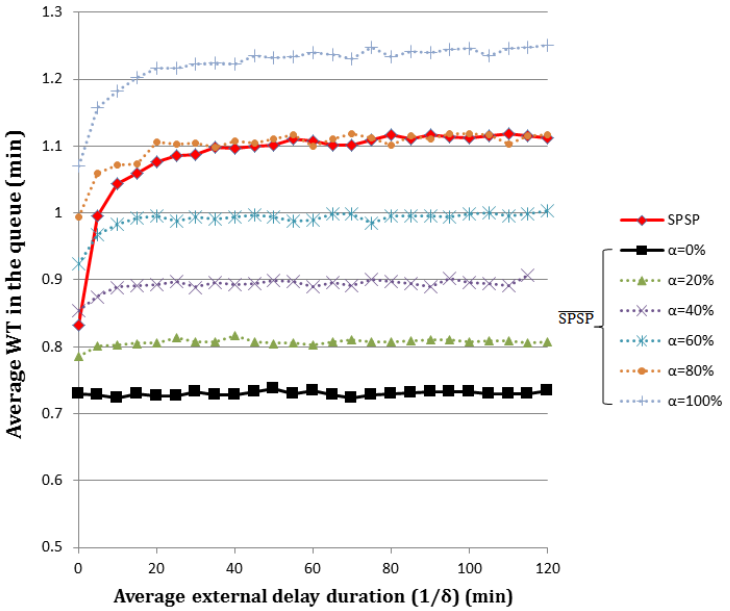

Figure A.3: The impact of external delay duration on the average queue WT ($s = 2, 1/\lambda = 10, 1/\mu = 5, p = 30\%$)

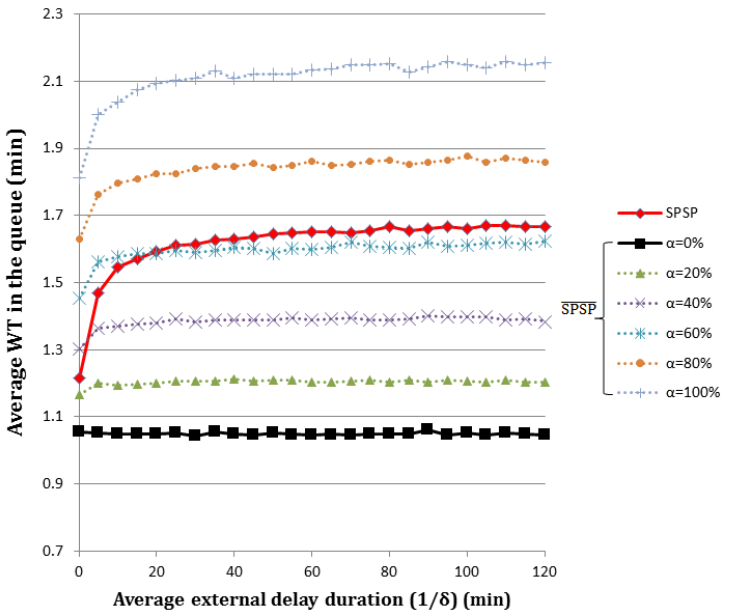

Figure A.4: The impact of external delay duration on the average queue WT ($s = 2, 1/\lambda = 10, 1/\mu = 5, p = 40\%$)

A.4 Assessing the quality of the approximation 1/s in the $\overline{SPSP}$ analytical model

In what follows, we assess the quality of the approximation ($\frac{1}{s}$) used in the analytical modeling of $\overline{SPSP}$, in Section 6.7.1. We use three different values of s: $s = 2$ which provides the

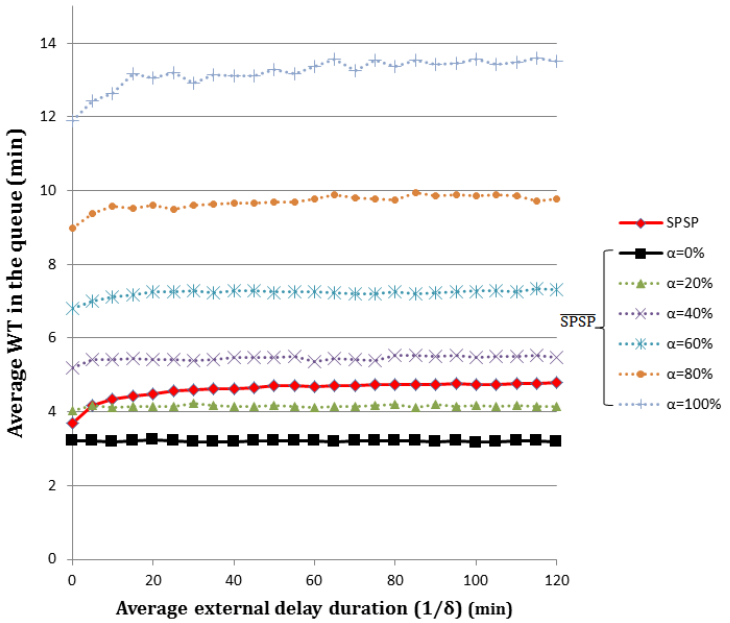

Figure A.5: The impact of external delay duration on the average queue WT ($s = 2, 1/\lambda = 10, 1/\mu = 5, p = 60\%$)

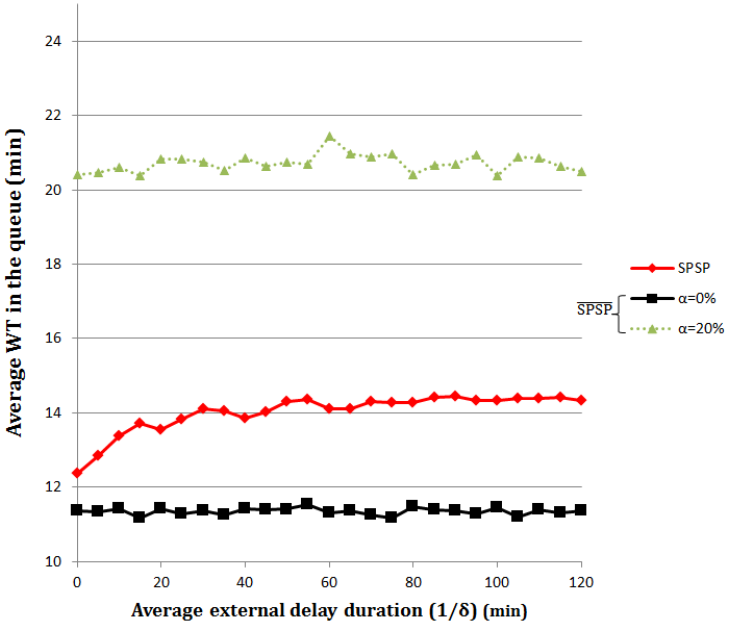

Figure A.6: The impact of external delay duration on the average queue WT ($s = 2, 1/\lambda = 10, 1/\mu = 5, p = 70\%$)

approximation 50%, $s = 5$ which provides the approximation 20% and $s = 10$ which provides the approximation 10%. We compare these analytical approximations with the real percentage given by simulation, for a given set of data. We also vary the values of p in order to observe the influence of the system load. Table A.2 summarizes simulation results and Table A.3 represents

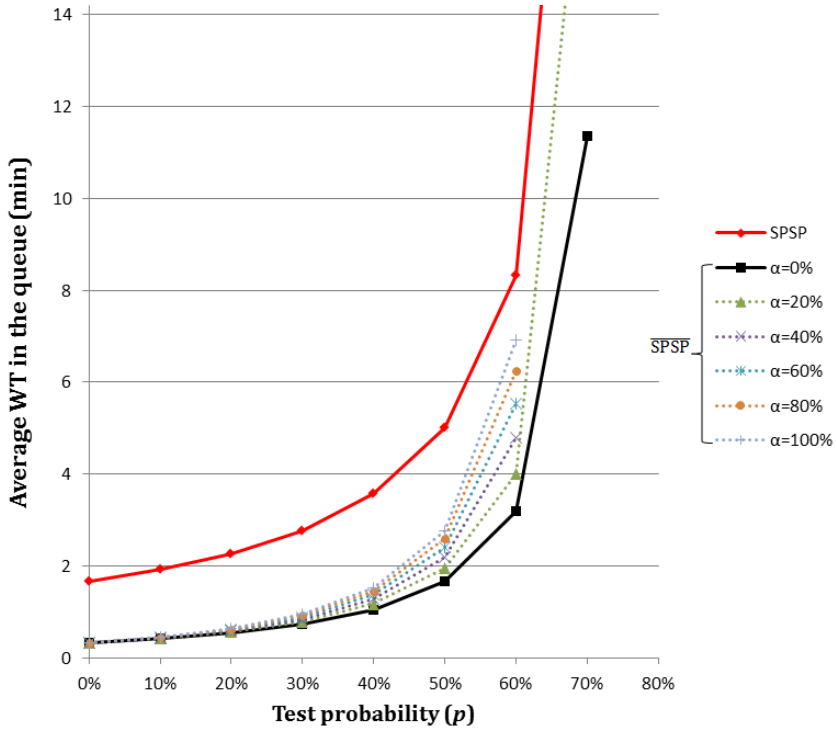

Figure A.7: Analytical approximation using random routing for *SPSP*: performance comparison as a function of p and α for the reference set of parameters ($s = 2, 1/\lambda = 10, 1/\mu = 5, 1/\delta = 60$)

the corresponding relative errors.

Table A.2: Simulated probability to see the same physician ($1/\lambda = 10, 1/\mu = 1, \alpha = 100\%, 1/\delta = 60$)

p\s	2	5	10
10%	0.500681	0.200125	0.100025
50%	0.500864	0.200138	0.100093
80%	0.501076	0.200302	0.100085

Table A.3: Relative errors of the approximation $\frac{1}{s}$ compared to simulation ($1/\lambda = 10, 1/\mu = 1, \alpha = 100\%, 1/\delta = 60$)

p\s	2	5	10
10%	0.136%	0.062%	0.025%
50%	0.173%	0.069%	0.093%
80%	0.215%	0.151%	0.085%

We observe that the higher is the system load (the smaller is s and the higher is p) the worse is the quality of the approximation. For instance, the worst approximation in Table A.3 is the lower left corner (the most loaded) with a relative error of 0.215%, while the best approximation

is the top right corner (the less loaded) with a relative error of 0.025%. As an explanation, we propose the following conjecture. The approximation $(\frac{1}{s})$ expresses a random routing of returning patients over the s servers. Things are more likely to proceed this way in situations where all servers are idle. The higher is the chance that servers are busy, the worse is the $(\frac{1}{s})$ approximation. Since the idleness of servers is obviously related to the system load, the latter has an influence on the quality of the approximation. Furthermore, we also test the influence of the average external delay duration $(1/\delta)$ on the quality of the approximation $(\frac{1}{s})$. To this end, we perform the same experiments of Table A.2 and Table A.3 by decreasing the value of $1/\delta$: 30, 5 and 0.

Table A.4: Relative errors of the approximation $\frac{1}{s}$ compared to simulation $(1/\lambda = 10, 1/\mu = 1, \alpha = 100\%, 1/\delta = 30)$

p\s	**2**	**5**	**10**
10%	**0.293%**	**0.156%**	**0.129%**
50%	**0.371%**	**0.187%**	**0.120%**
80%	**0.431%**	**0.280%**	**0.116%**

Table A.5: Relative errors of the approximation $\frac{1}{s}$ compared to simulation $(1/\lambda = 10, 1/\mu = 1, \alpha = 100\%, 1/\delta = 5)$

p\s	2	5	10
10%	1.486%	0.521%	0.295%
50%	1.936%	1.078%	0.587%
80%	2.123%	1.532%	0.839%

Table A.6: Relative errors of the approximation $\frac{1}{s}$ compared to simulation $(1/\lambda = 10, 1/\mu = 1, \alpha = 100\%, 1/\delta = 0)$

p\s	2	5	10
10%	9.685%	3.393%	0.129%
50%	14.271%	6.259%	0.120%
80%	20.696%	12.282%	0.116%

We observe that the value of the average external delay duration $(1/\delta)$ has an impact on the quality of the approximation $\frac{1}{s}$. The lower is $1/\delta$, the bigger is the deviation of the approximation $(1/\delta)$. The reason is that the lower is the delay before return, the more likely is the system to stay in the same state, which increases the probability that a patient will be processed by the

same server, and hence altering the probability of a random routing over the s different servers.

In what follows (Tables A.7, A.8, A.9 and A.10), we perform the same analysis using a more loaded set of parameters ($1/\lambda = 10, 1/\mu = 1, \alpha = 600\%$) which leads to the same observations done before.

Table A.7: Simulated probability to see the same physician ($1/\lambda = 10, 1/\mu = 1, \alpha = 600\%, 1/\delta = 60$)

p\s	2	5	10
10%	0.502220	0.200463	0.100047
50%	0.509291	0.202158	0.100619
80%	Unstable	0.202745	0.100916

Table A.8: Relative errors of the approximation $\frac{1}{s}$ compared to simulation ($1/\lambda = 10, 1/\mu = 1, \alpha = 600\%, 1/\delta = 60$)

p\s	2	5	10
10%	0.444%	0.231%	0.047%
50%	1.858%	1.079%	0.619%
80%	Unstable	1.372%	0.916%

Table A.9: Relative errors of the approximation $\frac{1}{s}$ compared to simulation ($1/\lambda = 10, 1/\mu = 1, \alpha = 600\%, 1/\delta = 30$

p\s	2	5	10
10%	0.773%	0.477%	0.268%
50%	3.442%	2.019%	1.086%
80%	Unstable	2.674%	1.710%

Table A.10: Relative errors of the approximation $\frac{1}{s}$ compared to simulation ($1/\lambda = 10, 1/\mu = 1, \alpha = 600\%, 1/\delta = 5$

p\s	2	5	10
10%	3.080%	1.569%	0.897%
50%	11.490%	7.542%	4.150%
80%	Unstable	12.414%	7.435%

Table A.11: Relative errors of the approximation $\frac{1}{s}$ compared to simulation $(1/\lambda = 10, 1/\mu = 1, \alpha = 600\%, 1/\delta = 0$

p\s	2	5	10
10%	11.633%	4.978%	0.268%
50%	27.551%	19.707%	1.086%
80%	Unstable	60.917%	1.710%

A.5 The average LOS values corresponding to Table 6.1

Table A.12 represents the values of the average LOS obtained from the ED-realistic simulations and corresponding to the percentages of Table 6.1.

A.6 List of the resources included in the model with appropriate assignments

Every single resource that can generate waiting times for patients in the ED are included in the model. The following table summarizes the way resources are split into the different subcategories of patients.

		Assigned to ESI:					Actual Min Staffing level	Actual Max Staffing level
		1	2	3	4	5		
Doctors	Senior LC	x	x	x			1	2
	Senior SC				x	x	0	1
	Senior LSC	x	x	x	x	x	0	1
	Junior 3			x			0	2
	Junior 4, 5				x	x	0	1
	Junior 3,4,5			x	x	x	0	1
Nurses	Triage Nurse	x	x	x	x	x	1	1
	Nurse LC	x	x	x			2	3
	Nurse SC				x	x	1	1
Stretcher bearer		x	x	x			1	2
Shock room places		x					3	3
Examination Rooms	Medium Boxes		x	x			9	9
	General Boxes				x		6	6
	Fast Track					x	1	1
Waiting Rooms	Int and Ext sit		x	x	x	x	Considered infinite	
	Intern lying		x				7	7
UHCD beds		x	x	x	x	x	12	12

Figure A.8: List of the resources included in the model with appropriate assignments

A.7 The impact of the system load on the effectiveness of $\overline{SPSP}$ for ESI 2 patients

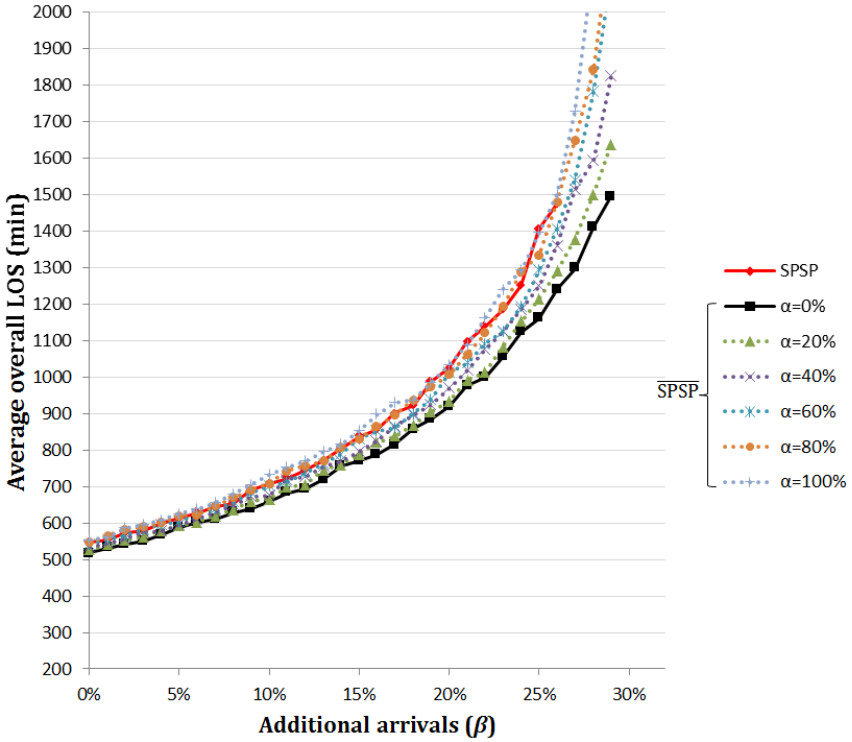

Figure A.9: Impact of the system load on the effectiveness of $\overline{SPSP}$ for ESI2 patients

Table A.12: The evolution of the average LOS when applying $\overline{SPSP}$ on ESI 3, 4 and 5

	Current system (SPSP)	$\overline{SPSP}$ applied on ESI 3, 4 and 5						
		α=0%	α=20%	α=30%	α=40%	α=60%	α=80%	α=100%
ESI 1	717.51	650.58	675.07	689.11	695.44	699.75	721.12	738.38
ESI 2	546.88	517.88	526.40	532.73	535.43	543.30	545.28	554.78
ESI 3	371.41	340.82	346.42	351.50	352.12	357.47	361.66	369.60
ESI 4	332.51	300.37	302.17	304.54	305.26	308.85	312.54	314.30
ESI 5	229.66	205.59	209.66	210.77	211.20	211.79	213.66	215.23
Overall	359.71	330.06	333.43	337.64	338.16	342.72	346.04	350.79

Appendix B

Appendix of Chapter 7

This appendix deals with the analysis of Chapter 7. We provide the survey form used in the TNO issue as well as the survey sample composition.

B.1 TNO survey

B.1.1 Survey form

"Should triage nurse be allowed to order diagnosis tests?"

As demonstrated in the literature, emergency departments (EDs) problems can stem from the process itself and not from the staffing levels (Samaha et al., 2003). Researchers are nowadays developing methods that aim of modifying some protocols and organizational rules regarding the patient path in the ED. It has been revealed that giving the Triage nurse the possibility to initiate diagnostic testing, without waiting for the consultation of the physician, may improve patients' satisfaction and possibly decrease their length of stay. This aforementioned practice is known as "Triage Nurse Ordering" (TNO). This last point is being investigated in our research work which is a collaboration between Ecole Centrale Paris (ECP) and Agence régionale de santé (ARS). Consequently, we kindly ask you to help us by filling the present form.

- Organization, E-mail address
- In your opinion, is "Triage Nurse Ordering" (TNO) a relevant practice in EDs?
  - Yes
  - No
  - Other (open)

- In your opinion, on which diagnosis tests can TNO be applied and why? (X-ray, Scanner, MRI, Echo, Blood test, urine test, ECG) (open-ended question)
- In your opinion, on which diagnosis tests TNO must not be applied and why? (X-ray, Scanner, MRI, Echo, Blood test, urine test, ECG).
- If you have any comments regarding the TNO issue, please write them down here. (open)

B.1.2 Details concerning the sample of ED physicians used in the survey

Table B.1 summarizes the number of experts who participated in the survey classified by ED and country.

Table B.1: TNO survey sample composition

Country	**Hospital**	**Answers**
France	Saint Camille	4
	SAU chartres	3
	Hôpital Tenon	1
	CHU Rouen	2
	Urgences Pontoise	1
	CHU Pitié-salpêtrière	1
	S.A.U. Hôpital Bichat	2
	CH Marne la Vallée	1
	Hôpital Ambroise-Paré	1
	Hôpital privé d'Antony	1
	CH Jossigny	2
	Hôpital privé de Marne la Vallée	1
	SAU Lariboisière	1
	CHU Besançon	1
	CHI Créteil	1
	Groupe hôspitalier Paris Saint Joseph	1
	Saint-Antoine	1
	SAU-Hôpital Louis Pasteur	2
Belgium	Erasme hospital (ULB)	1
Greece	Aristotle University of Thessaloniki	1
Germany	Allgemeines krankenhaus celle	1
USA	Penn State Hershey Medical Center	4
the Netherlands	VU Medical Center	1
Tunisia	Charles Nicole	1
		36

Bibliography

Abo-Hamad, W. and Arisha, A. (2013). Simulation-based framework to improve patient experience in an emergency department. *European Journal of Operational Research*, 224:154–166.

Adenso-Diaz, B. and Laguna, M. (2006). Fine-tuning of algorithms using fractional experimental designs and local search. *Operations Research*, 54(1):99–114.

Ahmed, M. and Alkhamis, T. M. (2009). Simulation optimization for an emergency department healthcare unit in kuwait. *European Journal of Operational Research*, 198(3):936–942.

Ajami, S., Ketabi, S., Yarmohammadian, M. H., and Bagherian, H. (2011). Waiting time in emergency department by simulation. *Studies in Health Technology and Informatics*, 164:196–200.

Al-Najjar, S. and Husain Ali, S. (2011). Staff and scheduling emergency rooms in two public hospitals: A case study. *International Journal of Business Administration*, 2(2):137–148.

Alavi-Moghaddam, M., Forouzanfar, R., Alamdari, S., Shahrami, A., Kariman, H., Amini, A., Pourbaba, S., and Shirvan, A. (2012). Application of queuing analytic theory to decrease waiting times in emergency department: Does it make sense? *Archives of Trauma Research*, 1:101–107.

Allon, G., Deo, S., and Lin, W. (2013). The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research*, 61(3):544–562.

American Hospital Association (2010). 2010 rapid response survey: Telling the hospital story. *American Hospital Association*.

Arendt, K., Sadosty, A., Weaver, A., Brent, C., and E.T., B. (2003). The left-without-being-seen patients: What would keep them from leaving? *Annals of Emergency Medicine*, 42(3):317–323.

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., and Yom-Tov, G. B. (2011). Patient flow in hospitals: A data-based queueing-science perspective. *Working Paper, Technion-Israel Institute of Technology, Technion City, Haifa.*

Ashour, O. M. and Kremer, G. E. O. (2013). A simulation analysis of the impact of fahp-maut triage algorithm on the emergency department performance measures. *Expert Systems with Applications*, 40(1):177–187.

Audit commission (2001). Review of national findings: Accident and emergency in london. *Audit Commission.*

Avramidis, A., Chan, W., Gendreau, M., L'Ecuyer, P., and Pisacane, O. (2010). Optimizing daily agent scheduling in a multiskill call center. *European Journal of Operational Research*, 41(6):483–497.

Avramidis, A., Chan, W., and L'Ecuyer, P. (2009). Staffing multi-skill call centers via search methods and a performance approximation. *IIE Transactions*, 200:822–832.

Baker, D., Stevens, C., and Brook, R. (1991). Patients who leave a public hospital emergency department without being seen by a physician. causes and consequences. *The Journal of the American Medical Association*, 266(8):1085–1090.

Batt, R. J. and Terwiesch, C. (2014). Doctors under load: An empirical study of state-dependent service times in emergency care. Working paper, University of Wisconsin-Madison, Madison.

Batt, R. J. and Terwiesch, C. (2015). Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science*, 61(1):39–59.

Baubeau, D. and Carrasco, V. (2003). Les usagers des urgences, premiers résultats d'une enquête nationale. *Paris: Direction de la recherche, des études de l'évaluation et des statistiques, Études et Résultats*, 212.

Beaulieu, H., Ferland, J., Gendron, B., and Michelon, P. (2000a). A mathematical programming approach for scheduling physicians in the emergency room. *Helath Care Management Science*, 3:193–200.

Beaulieu, H., Ferland, J. A., Gendron, B., and Michelon, P. (2000b). A mathematical programming approach for scheduling physicians in the emergency room. *Health care management science*, 3(3):193–200.

Bernstein, S., Verghese, V., and Leung, W. e. a. (2003). Development and validation of a new index to measure emergency department crowding. *Academic Emergency Medicine*, 10:938–942.

Beveridge, R., Clarke, B., and Janes, L. e. A. (1998). Emplementation guidelines for the canadian emergency department triage and acuity scale (ctas). *Canadian Association of Emergency Physicians*.

Bhattacharjee, P. and Ray, P. K. (2014). Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: A review and reflections. *Computers & Industrial Engineering*, 78:299–312.

Bolandifar, E., DeHoratius, N., Olsen, T., and Wiler, J. L. (2014). Modeling the behavior of patients who leave the emergency department without being seen by a physician. Chicago Booth Research Paper 12-14, University of Chicago, Chicago.

Bonastre, J., Journeau, F., Nestrigue, C., and Or, Z. (2013). Activité, productivité et qualité des soins des hôpitaux avant et après la t2a. *Questions d'économie de la santé*, 186:1–8.

Brandeau, M. L., Sainfort, F., and Pierskalla, W. P. (2004). *Operations research and health care: a handbook of methods and applications*, volume 70. Springer Science & Business Media.

Brick, C., Lowes, J., Lovstrom, L., Kokotilo, A., Villa-Roel, C., Lee, P., Lang, E., and Rowe, B. H. (2014). The impact of consultation on length of stay in tertiary care emergency departments. *Emergency Medicine Journal*, 31(2):134–138.

Broyles, J. R. and Cochran, J. K. (2007). Estimating business loss to a hospital emergency department from patient reneging by queuing-based regression. pages 613–618.

Broyles, J. R. and Cochran, J. K. (2011). A queuing based statistical approximation of hospital emergency department boarding. *Proceedings of the International Conference on Computers and Industrial Engineering*.

Burke, E. K., De Causmaecker, P., vanden Berghe, G., and Van Landeghem, H. (2004). The state of the art of nurse rostering. *Journal of scheduling*, 7:441–499.

Burström, L., Nordberg, M., Örnung, G., Castrén, M., Wiklind, T., Engström, M., and Enlund, M. (2012). Physician-led team triage based on lean principles may be superior for efficiency and quality? a comparison of three emergency departments with different triage models. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 20.

Burt, C. W., McCaig, L. F., and Valverde, R. H. (2006). Analysis of ambulance transports and diversions among us emergency departments. *Annals of emergency medicine*, 47(4):317–326.

Campello, F., Ingolfsson, A., and Shumsky, R. A. (2013). Queueing models of case managers. Technical report, Working paper.

Carmen, R. and Van Nieuwenhuyse, I. (2014). Improving patient flow in emergency departments with or techniques: A literature overview. Working Paper.

Centeno, M., Giachetti, R., Linn, R., and Ismail, A. (2003). Emergency departments ii: a simulation-ilp based tool for scheduling er staff. *Proceedings of the 2003 Winter Simulation Conference, New Orleans*, pages 1930–1938.

Chan, C., Yom-Tov, G. B., and Escobar, G. (2014). When to use speedup: An examination of service systems with returns. *Operations Research*, 62(2):462–482.

Chan, T., Killeen, J., Kelly, D., and Gussm, D. (2005). Impact of rapid entry and accelerated care at triage on reducing emergency department patient wait times, lengths of stay, and rate of left without being seen. *Annals of Emergency Medicine*, 46(6):491–497.

Chase, V. J., Cohn, A. E., Peterson, T. A., and Lavieri, M. S. (2012). Predicting emergency department volume using forecasting methods to create a “surge response" for noncrisis events. *Emergency Medicine Journal*, 19(5):569–576.

Cheung, W. W. H., Heeney, L., and Pound, J. L. (2002). An advance triage system. *Accident and emergency nursing*, 10(1):10–16.

Chonde, S., Parra, C., and Chang, C. (2013). Minimizing flow-time and time-to-first-treatment in an emergency department through simulation. pages 2374–2385.

Clark, T. D. J. and Waring, C. (1987). A simulation approach to analysis of emergency services and trauma center management. *Proceedings of the 19th Winter Simulation Conference*, pages 925–934.

Cochran, J. and Roche, K. (2009). A multiclass queuing network analysis methodology for improving hospital emergency department performance. *Computers Operations Research*, 36:1497–1512.

Connelly, L. and Bair, A. (2004). Discrete event simulation of emergency department activity: A platform for system-level operations research. *Academic Emergency Medicine*, 11:1177–1185.

Cooke, M., Wilson, S., and Pearson, S. (2012). The effect of a separate stream for minor injuries on accident and emergency department waiting times. *Journal of Emergency Medicine*, 19:28–30.

Deo, S. and Gurvich, I. (2011). Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science*, 57:1300–1319.

Derlet, R. and Richards, J. (2000). Overcrowding in the nation's emergency departments: Complex causes and disturbing effects. *Annals of Emergency Medicine*, 35:63–68.

Ding, R., McCarthy, M., J.S., D., Lee, J., Aronsky, D., and Zeger, S. (2010). Characterizing waiting room time, treatment time, and boarding time in the emergency department using quantile regression. *Academic Emergency Medicine*, 17(8):813–823.

Do, H. and Shunko, M. (2013). Pareto improving coordination policies in queueing systems: Application to flow control in emergency medical services. *Available at SSRN 2351965.*

Dobson, G., Tezcan, T., and Tilson, V. (2013). Optimal workflow decisions for investigators in systems with interruptions. *Management Science*, 59(5):1125–1141.

Duguay, C. and Chetouane, F. (2007). Modeling and improving emergency department systems using discrete event simulation. *Simulation*, 83(4):311–320.

Emergency Nurses Association (2011). Emergency department violence surveillance study. https://www.ena.org/practice-research/research/Documents/ENAEDVSReportNovember2011.pdf.

Ernst, A. T., Jiang, H., Krishnamoorthy, M., Owens, B., and Sier, D. (2004). Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research*, 153(3):3–27.

Evans, G., Gor, T., and Unger, E. (1996). A simulation model for evaluating personnel schedules in a hospital emergency department. *Proceedings of the 1996 Winter Simulation Conference, Coronado*, pages 1205–1209.

Fayyaz, J., Khurshed, M., Mir, M., and Mehmood, A. (2013). Missing the boat: odds for the patients who leave ed without being seen. *BMC Emergency Medicine*, 13(1).

Fermann, G. J. and Suyama, J. (2002). Point of care testing in the emergency department. *The Journal of emergency medicine*, 22(4):393–404.

Fernandes, C., Daya, M., Barry, S., and Palmer, N. (1994). Emergency department patients who leave without seeing a physician: the toronto hospital experience. *Annals of Emergency Medicine*, 24(6):1092–1096.

Fernandes, C. M. B., Price, A., and Christenson, J. M. (1997). Does reduced length of stay decrease the number of emergency department patients who leave without seeing a physician? *The Journal of emergency medicine*, 15(3):397–399.

Ferrin, D. M., Miller, M. J., and McBroom, D. L. (2007). Maximizing hospital finanacial impact and emergency department throughput with simulation. In *Proceedings of the 39th conference on Winter simulation: 40 years! The best is yet to come*, pages 1566–1573. IEEE Press.

Forster, A., Stiell, I., and Wells, G. e. a. (2003). The effect of hospital occupancy on emergency department length of stay and patient disposition. *Annals of Emergency Medicine*, 10(2):127–133.

Free, B., Lee, G. A., and Bystrzycki, A. (2009). Literature review of studies on the effectiveness of nurses ability to order and interpret x-rays. *Australasian Emergency Nursing Journal*, 12(1):8–15.

Freitas, A., Silva-Costa, T., Lopes, F., Garcia-Lema, I., and Texeira-Pinto, A. (2012). Factors influencing hospital high length of stay outliers. *BMC Health Services Research*, 12(265).

Fry, M. (2001). Triage nurses order x-rays for patients with isolated distal limb injuries: A 12-month ed study. *Journal of Emergency Nursing*, 27(1):17–22.

Ganansia, O. (2003). Quand ne pas prescrire de radio en traumatologie des extrémités chez l'adulte. *Société Française de Médecine d'Urgence*.

Garcia, M. L., Centeno, M., Rivera, C., DeCario, N., et al. (1995). Reducing time in an emergency room via a fast-track. In *Proceedings of the 1995 Winter Simulation Conference*, pages 1048–1053. IEEE.

Gentile, S., Durand, A. C., Vignally, P., Sambuc, R., and Gerbeaux, P. (2009). Do nonurgent patients presenting to an emergency department agree with a reorientation towards an alternative care department? *Revue d'Épidémiologie et de Santé Publique*, 57(1):3–9.

George, G., Jell, C., and Todd, B. (2006). Effect of population ageing on emergency department speed and efficiency: a historical perspective from a district general hospital in the uk. *Emergency Medicine Journal*, 23(5):379–383.

Ghanes, K., Diakogiannis, A., Wargon, M., Jouini, O., and Jemai, Z. (2015a). A heuristic for definition of shifts in an emergency department. In *Proceedings of the 45th International Conference on Computers and Industrial Engineering, Metz*.

Ghanes, K., Jouini, O., Jemai, Z. ., Diakogiannis, A., and Wargon, M. (2014a). Key performance indicators for emergency departments: A survey from an operations management perspective. Under revision in IIE Transactions on Healthcare Systems Engineering.

Ghanes, K., Jouini, O., Jemai, Z., and Wargon, M. (2015b). Modeling and analysis of triage nurse ordering in emergency departments. In *Proceedings of the IEEE International Conference on Industrial Engineering and Systems Management IESM'15, Sevilla*, pages 228–235.

Ghanes, K., Jouini, O., Jemai, Z., Wargon, M., Hellmann, R., Thomas, V., and Koole, G. (2014b). A comprehensive simulation modeling of an emergency department: A case study for simulation optimization of staffing levels. In *Proceedings of the 2014 Winter Simulation Conference, Savannah*, pages 1421–1432.

Ghanes, K., Wargon, M., Jouini, O., Jemai, Z., Diakogiannis, A., Hellmann, R., Thomas, V., and Koole, G. (2015c). Simulation-based optimization of staffing levels in an emergency department. *Simulation*, 90(10):942–953.

Goodacre, S. and Webster, A. (2005). Who waits longest in the emergency department and who leaves without being seen? *Annals of Emergency Medicine*, 22:93–96.

Gorelick, M., Yen, K., and Yun, H. (2005). The effect of in-room registration on emergency department length of stay. *Emergency Medicine Journal*, 45(2):128–133.

Green, L. (2006). Using operations research to reduce delays for healthcare. *Tutorials in Operations Research*.

Green, L., Savin, S., and Savva, N. (2013). ′Nursevendor Problem′: Personnel staffing in the presence of endogenous absenteeism. *Management Science*, 59(10):2237–2256.

Green, L., Soares, J., Giulio, J., and Green, R. (2006). Using queuing theory to increase the effectiveness of physician staffing in the emergency department. *Academic Emergency Medicine*, 13(1):61–68.

Green, L. V., Kolesar, P. J., and Whitt, W. (2007). Coping with time varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39.

Grosgurin, O., Cramer, B., Schaller, M., Sarasin, F., and Rutschmann, O. (2013). Patients leaving the emergency departmentwithout being seen by a physician: a retrospective database analysis. *Swiss Medical Weekly*, 143.

Günal, M. and Pidd, M. (2010). Discrete event simulation for performance modelling in health care: a review of the literature. *Journal of Simulation*, 4(1):42–51.

Guttmann, A., Schull, M., Vermeulen, M., and Stukel, T. (2011). Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from ontario, canada. *British Medical Journal*, 342.

Hagtvedt, R., Griffin, P., Keskinocak, P., Ferguson, M., and Jones, F. (2009). Cooperative strategies to reduce ambulance diversion. *Proceedings of the 2009 Winter Simulation Conference, Austin*, pages 1861–1874.

Han, J., France, D., Levin, S., Jones, I., Storrow, A., and Aronsky, D. (2010). The effect of physician triage on emergency department length of stay. *Journal of Emergency Medicine*, 39(2):227–233.

Hanna, E., Giglio, R., and Sadowski, R. (1974). A simulation analysis of a hospital emergency department. *Proceedings of the 1974 Winter Simulation Conference, Washington*, pages 379–388.

Harrison, J. and Ferguson, E. (2011). The crisis in united states hospital emergency services. *International journal of health care quality assurance*, 24(6):471–483.

Helm, J., AhmadBeygi, S., and van Oyen, M. (2011). Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management*, 20(3):359–374.

Hong, K. J., Shin, S. D., Song, K. J., Cha, W. C., and Cho, J. S. (2013). Association between ed crowding and delay in resuscitation effort. *The American journal of emergency medicine*, 31(3):509–515.

Hoot, N. and Aronsky, A. (2008). Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine*, 52(2):126–136.

Hoot, N. R., LeBlanc, L. J., Jones, I., Levin, S. R., Zhou, C., Gadd, C. S., and Aronsky, D. (2008). Forecasting emergency department crowding: a discrete event simulation. *Annals of Emergency Medicine*, 52(2):116–125.

Hopp, W. and Lovejoy, W. (2012). Hospital operations, principles of high efficiency health care. *FT Press*.

Hsia, R. Y., Kellermann, A. L., and Shen, Y. (2011). Factors associated with closures of emergency departments in the united states. *Jama*, 305(19):1978–1985.

Huang, J., Carmeli, B., and Mandelbaum, A. (2012). Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. Working Paper.

Hui, M. K. and Tse, D. K. (1996). What to tell consumers in waits of different lengths: An integrative model of service evaluation. *Journal of Marketing*, 60(2):81–90.

Hwang, U., McCarthy, M. L., Aronsky, D., Asplin, B., Crane, P. W., Craven, C. K., Epstein, S. K., Fee, C., Handel, D. A., Pines, J. M., Rathlev, N. K., Schafermeyer, R. W., Zwemer, J. F. L., and Bernstein, S. L. (2011). Measures of crowding in the emergency department: A systematic review. *Academic Emergency Medicine*, 18:527–538.

Hwang, U., Richardson, L. D., Sonuyi, T. O., and Morrison, R. S. (2006). The effect of emergency department crowding on the management of pain in older adults with hip fracture. *Journal of the American Geriatrics Society*, 54(2):270–275.

IRDES (2015). L'hôpital en france: Eléments de bibliographie. *Institut de Recherche et Documentation en Economie de la Santé*.

Izady, N. and Worthington, D. (2012). Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments. *European Journal of Operation Research*, 219:531–540.

Jahangirian, M., Taylor, S. J., Eatock, J., Stergioulas, L. K., and Taylor, P. M. (2015). Causal study of low stakeholder engagement in healthcare simulation projects. *Journal of the Operational Research Society*, 66:369–379.

Johnson, M., Myers, S., Wineholt, J., Pollack, M., and Kusmiesz, A. (2009). Patients who leave the emergency department without being seen. *Journal of Emergency Nursing*, 35:105–108.

Jones, S. S. and Evans, R. S. (2008). An agent based simulation tool for scheduling emergency department physicians. *AMIA Annual Symposium Proceedings*, pages 338–342.

Kelen, G., Scheulen, J., and Hill, P. (2001). Effect of an emergency department (ed) managed acute care unit (acu) on ed overcrowding and emergency medical services diversion. *Academic Emergency Medicine*, 8(11):1095–1100.

Khare, R., Powell, E., Reihardt, G., and Lucenti, M. (2009). Adding more beds to the emergency department or reducing admitted patient boarding times: which has a more significant influence on emergency department congestion? *Annals of Emergency Medicine*, 53(5):575–585.

Kirtland, A., Lockwood, J., Poisker, K., Stamp, L., and Wolfe, P. (1995). Simulating an emergency department. In *wsc*, pages 1039–1042. IEEE.

Kleijnen, J. and Wan, J. (2007). Optimization of simulated systems: Optquest and alternatives. *Simulation Modelling Practice and Theory*, 15(3):354–362.

Kocher, K. E., Meurer, W. J., Desmond, J. S., and Nallamothu, B. K. (2012). Effect of testing and treatment on emergency department length of stay using a national database. *Academic Emergency Medicine*, 19(5):525–534.

Kolker, A. (2008). Process modeling of emergency department patient flow: Effect of patient length of stay on ed diversion. *Journal of Medical Systems*, 32:389–401.

Komashie, A. and Mousavi, A. (2005). Modeling emergency departments using discrete event simulation techniques. *Proceedings of the 2005 Winter Simulation Conference, Orlando*, pages 2681–2685.

Kuo, Y., Leung, J., and Graham, C. (2012). Simulation with data scarcity: Developing a simulation model of a hospital emergency department. *Proceedings of the 2012 Winter Simulation Conference, Berlin*, pages 1–12.

LaCalle, E. and Rabin, E. (2010). Frequent users of emergency departments: The myths, the data, and the policy implications. *Annals of Emergency Medicine*, 56:42–48.

Laguna, M. and Marklund, J. (2013). *Business process modeling, simulation and design*. CRC Press.

Laskowski, M., McLeod, R., Friesen, M., Podaima, B., and Alfa, A. (2009). Models of emergency departments for reducing patient waiting times. *Public Library of Science One*, 4(7).

Lau, F. and Leung, K. (1997). Waiting time in an urban accident and emergency department–a way to improve it. *Journal of Accident & Emergency Medicine*, 14(5):299–303.

Law, A. and McComas, M. (2001). How to build valid and credible simulation models. *Proceedings of the 2001 Winter Simulation Conference, Arlington, VA*, pages 24–33.

Le Spegaque, D., Cauterman, M., and Kletz, F. (2006). Réduire les temps de passage aux urgences-recueil de bonnes pratiques organisationnelles. tome 1.

Lee, K. M., Wong, T. W., Chan, R., Lau, C. C., Fu, Y. K., and Fung, K. H. (1996). Accuracy and efficiency of x-ray requests initiated by triage nurses in an accident and emergency department. *Accident and Emergency Nursing*, 4(4):179–181.

Lee-Lewandrowski, E., Corboy, D., Lewandrowski, K., Sinclair, J., et al. (2003). Implementation of a point-of-care satellite laboratory in the emergency department of an academic medical center: impact on test turnaround time and patient emergency department length of stay. *Archives of pathology & laboratory medicine*, 127(4):456.

Liao, H., Liaw, S., Hu, P., Lee, K., Chen, C., and Wang, F. (2002). Emergency department patients who leave without being seen by a doctor: the experience of a medical center in northern taiwan. *Chang Gung Medical Journal*, 25(6):367–373.

Liew, D. and Kennedy, M. P. (2003). Emergency department length of stay independently predicts excess inpatient length of stay. *The Medical Journal of Australia*, 179(10):524–526.

Lin, D., Patrick, J., and Labeau, F. (2013). Estimating the waiting time of multi-priority emergency patients with downstream blocking. *Health Care Management Science*, 17(1):88–99.

Lindley-Jones, M. and Finlayson, B. J. (2000). Triage nurse requested x rays: are they worthwhile? *Journal of accident & emergency medicine*, 17(2):103–107.

Locker, T. E., Mason, S. M., et al. (2005). Analysis of the distribution of time that patients spend in emergency departments. *BMJ*, 330(7501):1188–1189.

Lowe, R. A., Bindman, A. B., Ulrich, S. K., Norman, G., Scaletta, T. A., Keane, D., Washington, D., and Grumbach, K. (1994). Refusing care to emergency department patients: evaluation of published triage guidelines. *Annals of emergency medicine*, 23(2):286–293.

Macleod, A. J. and Freeland, P. (1992). Should nurses be allowed to request x-rays in an accident & emergency department? *Archives of Emergency Medicine*, 9(1):19–22.

Mandelbaum, A., Momcilovc, P., and Tsetlyn, Y. (2012). On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science*, 58(7):1273–1291.

Mayhew, L. and Smith, D. (2008). Using queuing theory to analyze the government′s 4-h completion time target in accident and emergency departments. *Health Care Management Science*, 11(1):11–21.

McArthur, C. L. and Thomas, M. (1995). Comparison of triage nurse versus emergency physician ordering of extremity radiographs. *The American journal of emergency medicine*, 13(2):248–250.

McCaig, L. and Burt, C. (2004). National hospital ambulatory medical care survey: 2002 emergency department summary. *National Center for Health Statistics*, 340.

McGuire, F. (1994). Using simulation to reduce length of stay in emergency departments. *Journal of the Society for Health Systems*, 5(3):81–90.

McMullan, J. and Vesser, F. (2004). Emergency department volume and acuity as factors in patients leaving without treatment. *Southern Medical Journal*, 97(8):729–733.

Medeiros, D. J., Swenson, E., and DeFlitch, C. (2008). Improving patient flow in a hospital emergency department. *Proceedings of the 2008 Winter Simulation Conference, Miami*, pages 1526–1531.

Miller, H. E., Pierskalla, W. P., and Rath, G. J. (1976). Nurse scheduling using mathematical programming. *Operations Research*, 24(5):857–870.

Miró, O., Sánchez, M., Espinosa, G., Coll-Vinent, B., Bragulat, E., and Millá, J. (2003). Analysis of patient flow in the emergency department and the effect of an extensive reorganisation. *Emergency Medicine Journal*, 20:143–148.

Mohsin, M., Forero, R., Ieraci, S., Bauman, A., Young, L., and Santiano, N. (2007). A population follow-up study of patients who left an emergency department without being seen by a medical officer. *Emergency Medicine Journal*, 24(3):175–179.

Montimore, A. and Cooper, S. (2007). The "4-hour target": emergency nurses' views. *Emergency Medicine Journal*, 24(6):402–404.

Monzon, J., Friedman, S., Clarke, C., and Arenovich, T. (2005). Patients who leave the emergency department without being seen by a physician: a control-matched study. *Canadian Journal of Emergency Medicine*, 7(2):107–113.

Murray, R. P., Leroux, M., Sabga, E., Palatnick, W., and Ludwig, L. (1999). Effect of point of care testing on length of stay in an adult emergency department. *The Journal of emergency medicine*, 17(5):811–814.

National Center for Health Statistics (2012). Health, united states, 2012: With special feature on emergency care. hyattsville, md. 2013. *Library of Congress Catalog Number 76-641496*

For sale by Superintendent of Documents U.S. Government Printing Office. Washington, DC 20402.

Niska, R., Bhuiya, F., and Xu, J. (2010). National hospital ambulatory medical care survey: 2007 emergency department summary. *National Center for Health Statistics*, 26.

Olshaker, J. S. and Rathlev, N. K. (2006). Emergency department overcrowding and ambulance diversion: The impact and potential solutions of extended boarding of admitted patients in the emergency department. *Journal of Emergency Medicine*, 30:351–356.

Oredsson, S., Jonsson, H., Rognes, J., Lind, L., Göransson, K., Ehrenberg, A., Asplund, K., and Castrén, M. (2011). A systematic review of triage-related interventions to improve patient flow in emergency departments. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 19(43).

Orr, J. (2008). The good, the bad, and the four hour target. *British Medical Journal*, 337.

Ospina, M., Bond, K., Schull, M., Innes, G., Blitz, S., Friesen, C., and Rowe, B. (2006). Measuring overcrowding in emergency departments: a call for standardization. *Ottawa: Canadian Agency for Drugs and Technologies in Health.*

Pallin, A. and Kittell, R. (1992). Mercy hospital: simulation techniques for er processes. *Industrial Engineering*, 24(2):35–37.

Parekh, K., Russ, S., Amsalem, D., Rambaran, N., and Wright, S. (2013). Who leaves the emergency department without being seen? *BMC Emergency Medicine*, 13(10).

Parris, W., McCarthy, S., Kelly, A. M., and Richardson, S. (1997). Do triage nurse-initiated x-rays for limb injuries reduce patient transit time? *Accident and Emergency Nursing*, 5(1):14–15.

Pateron, D. (2012). Une organisation des flux au sein des urgences. *Les premi ères Assises de l'Urgence*, page 25.

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

Pham, J., Ho, G., Hill, P., McCarty, M., and Pronovost, P. (2009). National study of patient, visit, and hospital characteristics associated with leaving an emergency department without being seen: predicting lwbs. *Academic Emergency Medicine*, 16(10):949–955.

Pham, J., Patel, R., Millin, M., Kirsch, T., and Chanmugam, A. (2006). The effects of ambulance diversion: a comprehensive review. *Academic Emergency Medicine*, 13(11):1220–1227.

Pines, J., Decker, S., and Hu, T. (2012). Exogenous predictors of national performance measures for emergency department crowding. *Academic Emergency Medicine*, 60(3):293–298.

Pitts, S., Niska, R., and Xu, J. a. (2008). National hospital ambulatory care surve: 2006 emergency department summary. *National Health Statistics Reports*, 7:1–39.

Plambeck, E., Ang, E., Bayati, M., and Kwasnick, S. (2014). Improving the prediction of emergency department waiting times. Working Paper. Stanford University.

Pot, A., Bhulai, S., and Koole, G. (2008). A simple staffing method for multiskill call centers. *Manufacturing and Service Operations Management*, 10(3):421–428.

Potel, G., Lauque, D., Bouget, J., et al. (2005). L'organisation de l'aval des urgences: état des lieux et propositions.

Powell, E. S., Khare, R., and Reinhardt, G. (2007). Using computer simulation to evaluate the effect of point-of-care testing on emergency department patient flow. *Annals of Emergency Medicine*, 50(3):S70.

Public Health and Injury Prevention Committee (2011). mergency department violence: An overview and compilation of resources. http://www.acep.org/workarea/DownloadAsset.aspx?id=81782.

Ramirez-Nafarrate, A., Hafizoglu, A. B., Gel, E. S., and Fowler, J. W. (2014). Optimal control policies for ambulance diversion. *European Journal of Operational Research*, 236(1):298–312.

Retezar, R., Bessman, E., Ding, R., Zeger, S. L., and McCarthy, M. L. (2011). The effect of triage diagnostic standing orders on emergency department treatment time. *Annals of emergency medicine*, 57(2):89–99.

Robbins, T. and Harrison, T. (2010). A stochastic programming model for scheduling call centers with global service level agreements. *European Journal of Operational Research*, 207:1608–1619.

Robinson, D. (2013). An integrative review: Triage protocols and the effect on ed length of stay. *Journal of Emergency Nursing*, 39(4):398–408.

Robinson, S. (1994). Simulation projects: building the right conceptual model. *Industrial Engineering*, 26(9):34–36.

Roche, K. T. and Cochran, J. K. (2007). Improving patient safety by maximizing fast-track benefits in the emergency department-a queuing network approach. In *Proceedings of the 2007 Industrial Engineering Research Conference*, page 619. Institute of Industrial Engineers-Publisher.

Rosmulder, R. W., Krabbendam, J. J., Kerkhoff, A. H., Schinkel, E. R., Beenen, L. F., and Luitse, J. S. (2009). Advanced triage improves patient flow in the emergency department without affecting the quality of care. *Nederlands tijdschrift voor geneeskunde*, 154:1109–1109.

Rossetti, M., Trzcinski, G., and Syverud, S. (1999). Emergency department simulation and determination of physician staffing schedules. *Proceedings of the 1999 Winter Simulation Conference, Phoenix*, 2:1532–1540.

Rowe, B., Channan, P., Bullard, M., Blitz, S., Saunders, L., Rosychuk, R., Lari, H., Craig, W., and Holroyd, B. (2006). Characteristics of patients who leave emergency departments without being seen. *Academic Emergency Medicine*, 13(8):848–852.

Rowe, B. H., Villa-Roel, C., Guo, X., Bullard, M. J., Ospina, M., Vandermeer, B., Innes, G., Schull, M. J., and Holroyd, B. R. (2011). The role of triage nurse ordering on mitigating overcrowding in emergency departments: a systematic review. *Academic Emergency Medicine*, 18(12):1349–1357.

Ruohonen, T., Neittaanmaki, P., and Teittinen, J. (2006). Simulation model for improving the operation of the emergency department of special health care. *Proceedings of the 2006 Winter Simulation Conference, Monterey*, pages 453–458.

Saghafian, S., Austin, G., and Traub, S. J. (2015). Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering*, to appear.

Saghafian, S., Hopp, W., van Oyen, M., Desmond, J., and Kronick, S. (2012). Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research*, 60(5):1080–1097.

Saghafian, S., Hopp, W. J., van Oyen, M., Desmond, J. S., and Kronick, S. L. (2014). Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing and Service Operations Management*, 16(3):329–345.

Samaha, S., Armel, W., and Starks, D. (2003). The use of simulation to reduce the length of

stay in an emergency department. *Proceedings of the 2003 Winter Simulation Conference, New Orleans*, pages 1907–1911.

Saunders, C., Makens, P., and Leblanc, L. (1989). Modeling emergency department operations using advanced computer simulation systems. *Annals of Emergency Medicine*, 18(2):134–140.

Schafermeyer, R. and Asplin, B. (2003). Hospital and emergency department crowding in the united states. *Emergency Medicine Australasia*, 15(1):22–27.

Seaberg, D. C. and MacLeod, B. A. (1998). Correlation between triage nurse and physician ordering of ed tests. *The American journal of emergency medicine*, 16(1):8–11.

Shi, P., Chou, M., Dai, J. G., Ding, D., and Sim, J. (2014). Models and derlet inpatient operations: Time-dependent ed boarding time. *Management Science*.

Shuman, L., Spears, R. J., and Young, J. (1975). Operations research in health care: A critical analysis. *Baltimore, MD: Johns Hopkins University Press*.

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

Sinreich, D. and Marmor, Y. (2005). Emergency department operations: The basis for developing a simulation tool. *IIE Transactions*, 37:233–245.

Skaikh, S., Jerrard, D., Witting, M., Winters, M., and Brodeur, M. (2012). How long are patients willing to wait in the emergency department before leaving without being seen? *Western Journal of Emergency Medicine*, 13(6):463–467.

Smith, M. and Feied, C. (1999). Available on: http://necsi.org/projects/yaneer/emergencydeptcx.pdf.

Solberg, L., Asplin, B., Weinick, R., and Magid, D. (2003). Emergency department crowding: consensus development of potential measures. *Annals of emergency medicine*, pages 824–834.

Song, L., Tucker, A., and Murell, K. (2013). The impact of pooling on throughput time in discretionary work settings: An empirical investigation of emergency department length of stay. Working Paper.

SoRelle, R. (2002). Fairness in practice: Views from the front. *Emergency Medicine News*, 24(2):30–31.

Sorup, C. M., Jacobsen, P., and Forberg, J. L. (2013). Evaluation of emergency department performance - a systematic review on recommended performance and quality-in-care measures. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 21.

Spirivulis, P., Da Silva, J., Jacobs, I., Frazer, A., and Jelinek, G. (2006). The association between hospital overcrowding and mortality among patients admitted via western australian emergency departments. *Medical Journal of Australia*, 184(5):208–212.

Stock, L., Bradley, G., Lewis, R., Baker, D., Sipsey, J., and Stevens, C. (1994). Patients who leave emergency departments without being seen by a physician: magnitude of the problem in los angeles county. *Annals of Emergency Medicine*, 23(2):294–298.

Subash, F., Dunn, F., McNicholl, B., and Marlow, J. (2004). Team triage improves emergency department efficiency. *Emergency Medicine Journal*, 21(5):542–544.

Tan, B. A., Gubaras, A., and Phojanamongkolkij, N. (2002). Schedule evaluation: simulation study of dreyer urgent care facility. In *Proceedings of the 34th Conference on Winter Simulation*, pages 1922–1927. IEEE Press.

Tanabe, P., Myers, R., Zosel, A., Bricem, J., Ansari, A., Evans, J., Martinovich, Z., Todd, K., and Paice, J. (2007). Emergency department management of acute pain episodes in sickle cell disease. *Academic Emergency Medicine*, 14.

Than, K. C., Leong, Y. L., and Ngiam, B. S. (1999). Initiation of x-rays by the triage nurse: competency and its effect on patients' total time spent in the accident and emergency department. *Annals of Emergency Medicine*, 34(4):S60.

Thurston, J. and Field, S. (1996). Should accident and emergency nurses request radiographs? results of a multicentre evaluation. *Journal of accident & emergency medicine*, 13(2):86–89.

Tropea, J., Sundararajan, V., Gorelik, A., Kennedy, M., Cameron, P., and Brand, C. (2012). Patients who leave without being seen in emergency departments: An analysis of predictive factors and outcomes. *Academic Emergency Medicine*, 19(4):439–447.

Trzeciak, S. and Rivers, E. (2003). Emergency department overcrowding in the united states: An emerging threat to patient safety and public health. *Emergency Medicine Journal*, 20(5):402–405.

Tseytlin, Y. (2009). *Queueing systems with heterogeneous servers: On fair routing of patients in emergency departments*. PhD thesis, Technion-Israel Institute of Technology.

Van den Bergh, J., Beliën, J., De Bruecker, P., Demeulemeester, E., and De Boeck, L. (2013). Personnel scheduling: A literature riview. *European Journal of Operational Research*, 226:367–385.

van Dyke, K. J., McHugh, M., Yonek, J., and Moss, D. (2011). Facilitators and barriers to the implementation of patient flow improvement strategies. *Quality Management in Healthcare*, 20(3):223–233.

Vegting, I. L., Alam, N., Ghanes, K., Jouini, O., Mulder, F., Vreeburg, M., Biesheuvel, T., van Bokhorst, J., Go, P., Kramer, M. H. H., et al. (2015). What are we waiting for? factors influencing completion times in an academic and peripheral emergency department. *The Netherlands journal of medicine*, 73(7):331–340.

Vegting, I. L., Nanayakkara, P. W., van Dongen, A. E., Vandew alle, E., van Galen, J., and Kramer, M. H. e. a. (2011). Analysing completion times in an academic emergency department: coordination of care is the weakest link. *The Netherlands journal of medicine*, 69(9):392–398.

Vertesi, L. (2004). Does the canadian emergency department triage and acuity scale identify non-urgent patients who can be triaged away from the emergency department? *Cjem*, 6(05):337–342.

Vieth, T. and Rhodes, K. (2006). The effect of crowding on access and quality in an academic ed. *American Journal of Emergency Medicine*, 24(7):787–794.

Vilke, G., Brown, L., Skogland, P., Simmons, C., and Guss, D. (2004). Approach to decreasing emergency department ambulance diversion hours. *Journal of Emergency Medicine*, 26(2):189–192.

Wallace, R. and Whitt, W. (2005). A staffing algorithm for call centers with skill-based routing. *Manufacturing and Service Operations Management*, 7:276–294.

Wang, J., Li, J., and Howard, P. K. (2013). A system model of work flow in the patient room of hospital emergency department. *Health care management science*, 16(4):341–351.

Wang, J., Li, J., Tussey, K., and Ross, K. (2012). Reducing length of stay in emergency department: A simulation study at a community hospital. *IEEE Transactions on Systems, Man, and Cybernetics*, 42(6):1314–1322.

Wang, Q. (2004). Modeling and analysis of high risk patient queues. *European Journal of Operational Research*, 155(2):502–515.

Wang, X. (2013). Emergency department staffing: A separated continuous linear programming approach. *Mathematical Problems in Engineering*.

Wargon, M., Guidet, B., Hoang, T. D., and Heublum, G. (2009). A systematic review of models for forecasting the number of emergency department visits. *Emergency Medicine Journal*, 26(6):395–399.

Warner, M. (2013). *Personnel staffing and scheduling, in Patient Flow: Reducing Delay in Healthcare Delivery*, volume 206. Springer Science & Business Media.

Weiss, S., Derlet, R., Arndahl, J., Ernst, A., Richards, J., Fernández-Frankelton, M., Schwab, R., Stair, T. O., Vicellio, P., Levy, D., et al. (2004). Estimating the degree of emergency department overcrowding in academic medical centers: results of the national ed overcrowding study (nedocs). *Academic Emergency Medicine*, 11(1):38–50.

Welch, S. (2009). Quality matters: solutions for a safe and efficient emergency department. *Academic Emergency Medicine*. Joint Commission Resources. Oakbrook Terrace (IL).

Welch, S., Asplin, B., Stone-Griffith, S., Davidson, S., Augustine, J., and Schuur, J. (2011). Emergency department operational metrics, measures and definitions: Results of the second performance measures and benchmarking summit. *Annals of Emergency Medicine*, 58(1):33–40.

Welch, S., Augustine, J., Camargo Jr., C., and Reese, C. (2006). Emergency department performance measures and benchmarking summit. *Academic Emergency Medicine*, 13(10):1074–1080.

Weng, S., Cheng, B., Kwong, S., Wang, L., and Chang, C. (2011). Simulation optimization for emergency department resources allocation. *Proceedings of the 2011 Winter Simulation Conference, Phoenix*, pages 1231–1238.

White, C. (2007). Health care spending growth: how different is the united states from the rest of the oecd? *Health Affairs*, 26(1):154–161.

Wiler, J., Gentle, C., and Halfpenny, J. e. a. (2010). Optimizing emergency department front-end operations. *Annals of Emergency Medicine*, 55:142–160.

Wiler, J. L., Bolandifar, E., Griffey, R. T., Poirier, R. F., and Olsen, T. (2013). An emergency department patient flow model based on queueing theory principles. *Academic Emergency Medicine*, 20(9):939–946.

Wiler, J. L., Griffey, R. T., and Olsen, T. (2011). Review of modeling approaches for emergency department patient flow and crowding research. *Academic Emergency Medicine*, 18(12):1371–1379.

Winn, K. (2001). *Emergency department efficiency through utilization of triage nurse protocols*. PhD thesis, Master thesis, Texas Tech University Health Science Center, Texas, USA.

World Health Organization (2014). *World Health Statistics*. World Health Organization.

Xiao, J., Osterweil, L., and Wang, Q. (2010). Dynamic scheduling of emergency department resources. *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 590–599.

Xu, K. and Chan, C. (2013). Using future information to reduce waiting times in the emergency department. Working Paper.

Yankovic, N. and Green, L. (2011). Good nursing levels: A queuing approach. *Operations Research*, 59(4):942–955.

Yom-Tov, G. (2010). Queues in hospitals: Queueing networks with reentrant customers in the qed regime.

Yom-Tov, G. and Mandelbaum, A. (2014). Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299.

Yoon, P., Steiner, I., and Reinhardt, G. (2003). Analysis of factors influencing length of stay in the emergency department. *Cjem*, 5(03):155–161.

Zayas-Caban, G., Xie, J., Green, L., and Lewis, M. (2013). Optimal control of an emergency room triage and treatment process. Working Paper.

Zeltyn, S., Marmor, Y. N., Mandelbaum, A., Carmeli, B., Greenshpan, O., Mesika, Y., Wasserkrug, S., Vortman, P., Shtub, A., Lauterman, T., et al. (2011). Simulation-based models of emergency departments:: Operational, tactical, and strategic staffing. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 21(4):24.

Zetlyn, S., Marmor, Y., Mandelbaum, A., Carmeli, B., Greenshpan, O., Mesika, Y., Vortman, P., Shtub, A., Lauterman, T., Schwartz, D., Moskovitch, K., Tzafrir, S., and Basis, F. (2011). Simulation-based models of emergency departments: Operational, tactical and strategic staffing. *ACM Transactions on Modelling and Computer Simulation*, 21(4). Article No. 24.